UNIVERSITÉS FRANCOPHONES

André Clas
Université de Montréal

Philippe Thoiron
Université Lumière Lyon-2

Henri Béjoint
Université Lumière Lyon-2

LEXICOMATIQUE ET DICTIONNAIRIQUES

Actes du Colloque de Lyon
1995

Ouvrage publié
avec l'aide
de
Rank Xerox
Research Centre
Meylan - France

actualité scientifique

FMA

LEXICOMATIQUE
ET
DICTIONNAIRIQUES

ISBN 2-92 00 21 70 2

Bibliothèque nationale du Québec
Bibliothèque nationale du Canada
Bibliothèque nationale de France
Imprimé au Liban

LEXICOMATIQUE ET DICTIONNAIRIQUES

IV[es] Journées scientifiques du réseau thématique
«Lexicologie, Terminologie, Traduction»
Lyon, France, 28, 29, 30 septembre 1995

Sous la direction de :

André CLAS, Université de Montréal, Canada
Philippe THOIRON, Université Lumière Lyon-2
Henri BÉJOINT, Université Lumière Lyon-2

1996

FMA
Beyrouth

AUPELF • UREF
B.P 400, succ. Côte-des-Neiges
Montréal (Québec) Canada
H3S 2S5

Avant-propos

La diffusion de l'information scientifique et technique est un facteur essentiel du développement. Aussi dès 1988, l'Agence francophone pour l'enseignement supérieur et la recherche (AUPELF-UREF), mandatée par les Sommets francophones pour produire et diffuser revues et livres scientifiques, a créé la collection Universités francophones.

Lieu d'expression de la communauté scientifique de langue française, Universités francophones vise à instaurer une collaboration entre enseignants et chercheurs francophones en publiant des ouvrages, coédités avec des éditeurs francophones, et largement diffusés dans les pays du Sud, grâce à une politique tarifaire préférentielle.

Quatre séries composent la collection :

– Les manuels : cette série didactique est le cœur de la collection. Elle s'adresse à un public de deuxième et troisième cycles universitaires et vise à constituer une bibliothèque de référence couvrant les principales disciplines enseignées à l'université.

– Actualité scientifique : dans cette série, à laquelle appartient le présent ouvrage, sont publiés les actes des journées scientifiques organisées par les réseaux thématiques de recherche de l'UREF. A ce jour, 16 réseaux thématiques, rassemblant plusieurs milliers de chercheurs de toute la francophonie, sont constitués au sein de l'UREF.

– Prospectives francophones : s'inscrivent dans cette série des ouvrages de réflexion donnant l'éclairage de la francophonie sur les grandes questions contemporaines.

– Savoir plus Université : cette nouvelle série se compose d'ouvrages de synthèse qui font un point précis sur des sujets scientifiques d'actualité.

Notre collection, en proposant une approche plurielle et singulière de la science, adaptée aux réalités multiples de la francophonie, contribue efficacement à promouvoir la recherche dans l'espace francophone et le plurilinguisme dans la recherche internationale.

Professeur Michel Guillou
Directeur général de l'AUPELF
Recteur de l'UREF

Sommaire

Liste des auteurs

Alonso Ramos, Margarita, Département de linguistique générale, Université de La Corunha, Campus da Zapateira s/n, 15071 La Corunha, Espagne

Apresjan, Jurij D., Russie 101 447, Moscou YSP 4, rue Ermolova 19, IPPI RAN, Laboratoire 15

Barbaud, Philippe, Département de linguistique, UQAM C.P. 8888, succ. Centre-ville, Montréal (Québec), H3C 3P8, Canada

Ben Ahmed, Mohamed, Laboratoire de recherche en informatique arabisée et documentique intégrée (RIADI), Tunis, Tunisie

Blache, Philippe, 2LC CNRS, 1361 route des Lucioles, Sophia Antipolis, 06560 Valbonne, France

Blampain, Daniel, Département de linguistique française, Groupe de recherche TERMISTI, Institut supérieur de traducteurs et interprètes, 34 rue Joseph Hazard, 1180 Bruxelles, Belgique

Blanc, Étienne, GETA CLIPS, Institut IMAG, Université Joseph Fourier Grenoble I, B.P. 53, 150 rue de la Chimie, 38041 Grenoble Cedex 9, France

Bouchard, Lorne, Département d'informatique, Université du Québec à Montréal, C.P. 8888, succ. Centre-ville, Montréal (Québec), H3C 3P8, Canada

Bouillon, Pierrette, ISSCO (Institut pour les études sémantiques et cognitives), Université de Genève, 54 route des Acacias, 1227 Genève, Suisse

Bourigault, Didier, Centre d'analyse et de mathématiques sociales, Unité mixte EHESS CNRS Paris Sorbonne, EDF Direction des études et recherches,1 avenue du Général de Gaulle, 92141 Clamart Cedex, France

Bouveret, Myriam, Praxiling, Université de Montpellier, France

Breidt, Élisabeth, Seminar für Sprachwissenschaft, Universität Tübingen, Wilhelmstr. 113, D 72074 Tübingen, Allemagne

Brunet, Étienne, Institut National de la langue française, UPR 6861, UFR Lettres, Arts et Sciences Humaines, 98 bd Herriot, B.P. 209, 06204 Nice Cedex 3, France

Campenhoudt, Marc Van, Département de linguistique française, Centre de recherche TERMISTI, Institut Supérieur de traducteurs et interprètes, 34 rue Joseph Hazard, 1180 Bruxelles, Belgique

Chibout, Karim, Groupe Langage et Cognition, LIMSI CNRS, B.P. 133, 91403 Orsay Cedex, France

Daille, Béatrice, IRIN, Université de Nantes, 2 rue de la Houssinière, 44072 Nantes Cedex 03, France

Delisle, Sylvain, Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières, 3351 bd des Forges, Trois Rivières (Québec), G9A 5H7, Canada

Delpui, Mireille, 2LC CNRS, 1361 route des Lucioles, Sophia Antipolis, 06560 Valbonne, France

Deville, Guy, École de Langues vivantes, Facultés universitaires de Namur, 61 rue de Bruxelles, B 5000 Namur, Belgique

Emirkanian, Louisette, Département de linguistique, Université du Québec à Montréal, C.P. 8888, succ. Centre-ville, Montréal (Québec), H3C 3P8, Canada

Fontenelle, Thierry, Département d'anglais, Université de Liège, 3 Place Cockerill, B 4000 Liège, Belgique

Gaudin, François, Université de Rouen, CNRS URA 1164 IRED, 7 rue Thomas Becket, 76130 Mont Saint Aignan, France

Habaili, Hussein, 2 rue du 3 septembre 1934, Tunis 1007, Tunisie

Herbigniaux, Emmanuel, École de Langues vivantes, Facultés universitaires de Namur, 61 rue de Bruxelles, B 5000 Namur, Belgique

Jouis, Christophe, UFR DIST/CREDO, Université Charles de Gaulle Lille III, B.P. 149, 59653 Ville d'Ascq, France

Knowles, Frank, Language Studies Unit, Aston University, Birmingham, B4 7ET, Royaume-Uni

Lelubre, Xavier, 5 rue Nicolaï, 69007 Lyon, France

Leroy-Turcan, Isabelle, Université Jean Moulin Lyon 3, 18 rue Chevreul, 69007 Lyon, France

Mantha, Suzanne, Université de Montréal, Département de linguistique, C.P. 6128 succ. Centre-ville, Montréal (Québec), H3C 3J7, Canada

Masson, Nicolas, Groupe Langage et Cognition, LIMSI CNRS, B.P. 133, 91403 Orsay Cedex, France

Mathieu-Colas, Michel, Laboratoire de linguistique informatique, Université Paris XIII CNRS INaLF, Villetaneuse, France

Mustafa-Elhadi, Widad, UFR DIST/CREDO, Université Charles de Gaulle Lille III, B.P. 149, 59653 Ville d'Ascq, France

Nirenburg, Sergei, Computing Research Laboratory, New Mexico State University, Box 30001, Las Cruces, New Mexico 88003 8001, États-Unis

Prince, Violaine, LIMSI CNRS, B.P. 133, 91403 Orsay Cedex, France

Radanyi, Dominique, CIEH, Université de Paris III, France

Segond, Frédérique, Rank Xerox Research Centre, 6 chemin de Maupertuis, 38240 Meylan, France

Sérasset, Gilles, GETA CLIPS IMAG, Université Joseph Fourier Grenoble I, B.P. 53, 150 rue de la Chimie, 38041 Grenoble Cedex 9, France

Szende, Thomas, Centre interuniversitaire d'études hongroises, Université Sorbonne Nouvelle Paris III, 1 rue Censier, 75005 Paris, France

Taifi, Miloud, 20 rue Abdallahben Maskoud, Quartier Prince Héritier, Fès, Maroc

Tanaka, Kumiko, Takeichi Laboratory, Information Engineering Course, Graduate School of Engineering, The University of Tokyo, 1-3-7 Hongo Bunkyouku Tokyo, 113, Japon

Thomas, Patricia, Unit 54, Smithbrook Kilns, Cranleigh, GU6 8JJ, Royaume-Uni

Tutin, Agnès, URA SILEX, Université Lille III, B.P. 149, 59653 Villeneuve d'Ascq Cedex, France

Viegas, Évelyne, Computing Research Laboratory, New Mexico State University, Box 30001, Las Cruces, New Mexico 88003 8001, États-Unis

Walther, Catherine, Laboratoire d'analyse et de technologie du langage, Département de linguistique, Université de Genève, 1211 Genève 4, Suisse

Wehrli, Éric, Laboratoire d'analyse et de technologie du langage, Département de linguistique, Université de Genève, 1211 Genève 4, Suisse

Wooldridge, Russon, University of Toronto, Toronto (Ontario), Canada

Membres du Comité de réseau « LTT »

Chad, Mohammed, Professeur, doyen de la Faculté des lettres, Université Sidi Mohamed Ben Adballah, Fès, Maroc

Clas, André, Coordonnateur du réseau, professeur, directeur du GRESLET, Université de Montréal, Montréal, Canada

Ouoba, Benoît Bendi, Professeur, Université de Ouagadougou, Ouagadougou, Burkina Faso

Thoiron, Philippe, Professeur, directeur du CRTT, Université Lumière Lyon-2, Lyon, France

Goffin, Roger, Professeur, Université Libre de Bruxelles, Bruxelles, Belgique

Préface

André CLAS

Coordonnateur du réseau LTT, Université de Montréal, Canada

Lexicomatique et dictionnairiques

Le réseau lexicologie, terminologie, traduction (LTT), créé en 1988, a tenu à l'Université Lumière-Lyon 2, du 28 au 30 septembre 1995, ses Quatrièmes Journées scientifiques sur le thème général **LEXICOMATIQUE et DICTIONNAIRIQUES**.

Après Fès en 1989 (*Visages du français. Variétés lexicales de l'espace francophone*), il y a eu, en 1991, Mons (*L'environnement traductionnel. La station de travail du traducteur de l'an 2001*), puis en 1993, Montréal (*TA-TAO : recherches de pointe et applications immédiates*). Ainsi après avoir exploré des questions de variations lexicales en Francophonie (signalons que l'on trouve dans les *Actes de Fès*, en plus des contributions variées de divers spécialistes de la lexicographie une excellente *bibliographie scientifique concernant la langue française en Afrique noire,* et une présentation d'une proposition de *Trésor informatisé des vocabulaires francophones*), le réseau a abordé les autres domaines de sa compétence et notamment le rôle et l'importance de la traduction, le thème général se situait dans le secteur des postes de travail de traducteurs et l'automatisation de certaines tâches permettait d'aborder dans des perspectives résolument futuristes la question des stations de travail. On examine dans les *Actes de Mons* toute la problématique de l'utilisation de l'ordinateur et de son importance grandissante dans tout le secteur de ce que l'on peut regrouper sous l'étiquette de traitement des langues. Il était normal que les Journées suivantes se consacrent aux questions de la traduction assistée et automatique : un bilan des recherches en cours et des applications en vue s'imposait. On comprend donc ainsi mieux le thème retenu pour les Journées de Lyon. Toutes ces rencontres, on nous permettra de le dire, car nous avons des preuves éclatantes et des témoignages indubitables, ont été extrêmement fructueuses à de très nombreux égards. En plus de permettre de faire connaissance, de tisser de nouveaux liens, d'opérer certains rapprochements, de conclure des ententes pour des travaux en commun ou de mieux cibler certains travaux, elles ont été l'occasion pour tous de mieux savoir ce qui se faisait, de comprendre les progrès faits, d'examiner les nouvelles voies dans la recherche, bref d'aller de l'avant, de mieux savoir ce qui se faisait et de mieux connaître ceux qui le

faisaient. Ainsi de nouvelles recherches ont été lancées, de nouvelles orientations ont été trouvées et indubitablement des idées neuves sont nées. Il me semble donc pouvoir affirmer que le bilan a été, à notre humble avis, nettement positif. Il convenait aussi d'explorer de nouvelles avenues dans la recherche en génie linguistique et les transformations technologiques créant ce qu'on appelle les industries de la langue, il fallait examiner les domaines de la lexicologie, de la lexicographie, de la terminologie et de la terminographie. En un mot la préparation de nouveaux outils, d'outils linguistiques électroniques, d'outils qui puissent donner accès à des bases documentaires, à des réseaux des autoroutes de l'information, à des prolégomènes du cyberespace. Le thème de ces Journées devait permettre de mettre en perspective les différents outils destinés aux « langagiers », la variété des outils imposant en plus une réflexion sur les méthodes et les méthodologies de la recherche lexicographique. La *lexicomatique* se définissant comme l'ensemble des méthodes, des techniques et des pratiques qui utilisent l'informatique pour les explorations sémantiques, grammatico-lexicales du lexique d'une langue. On inclut ainsi facilement les débordements vers l'étude de corpus informatisés, de codages de corpus, de gestion de corpus bilingues et même vers la préparation de nouveaux outils pour la traduction assistée. La *dictionnairique* concerne toutes les questions techniques qui servent à élaborer de nouveaux dictionnaires et plus particulièrement des dictionnaires électroniques. On a ainsi pu montrer que la lexicologie et la lexicographie venaient d'entrer dans une nouvelle ère.

Si l'on examine quelque peu l'histoire de la lexicographie, on peut découvrir diverses étapes. Celle de la pragmatique d'abord : les lexicographes sont avant tout des praticiens, même si des relents théoriques percent ici et là, ce sont des dépouilleurs d'œuvres, des collectionneurs d'attestations, surtout littéraires. Peu à peu les sciences interviennent dans les dictionnaires, le dictionnaire a de plus en plus une vocation sociale, ce sont les Larousse et les Littré qui donnent la mesure. Aujourd'hui, c'est l'informatique qui apporte sa contribution à l'élaboration de dictionnaires. L'ordinateur a bouleversé la lexicographie, la lexicologie, et le dictionnaire même. Comme on le sait, le dictionnaire s'appuie de plus en plus largement sur des bases de données, donc des textes aux milliers d'occurrences de lexies, bien plus le dictionnaire est devenu lui-même base de données, où l'on peut retrouver les attestations les plus diverses. Les nouveaux supports informatiques permettent un emmagasinage gigantesque et un questionnement multiple et varié. La répartition entre dictionnaire de langue générale et dictionnaire spécialisé semble s'amenuiser, le nouveau type de dictionnaire étant en mesure de répondre en fonction des intérêts de l'usager. C'est l'ère du dictionnaire à la carte qui s'ouvre.

Les textes dans ces *Actes* abordent toutes ces questions, présentent les diverses possibilités et explorent avec clarté les voies suivies ou à suivre. Ils fournissent dans maints cas des cheminements à entreprendre, des orientations à développer. Les méthodologies sont plurielles, les propositions sont multiples, et comme dans toute recherche scientifique on trouve des constructions de schémas ou des créations de modèles qui méritent d'être exploitées et réalisées.

C'est ainsi que l'on pourra mesurer tout l'intérêt que présente la proposition de **Jurij D. Apresjan** pour un *Enseignement du lexique assisté par ordinateur*, où l'on voit percer toute une nouvelle pédagogie stimulant le désir de savoir par des exercices programmés mais aussi par des jeux. **Étienne Brunet** avec sa communication sur son logiciel *Hyperbase* montre tout ce qu'on peut tirer d'un logiciel convivial et remet la

recherche philologique et littéraire à l'ordre du jour en intégrant les différentes versions d'un même texte sur ordinateur et surtout donne accès sur Internet à celui qui a illustré le mieux l'invention verbale, la saveur et la richesse linguistique : Rabelais. **Thierry Fontenelle** explore les *Réseaux sémantiques et les dictionanires bilingues électroniques* en s'appuyant sur la théorie Sens ⇔ Texte. L'utilisateur peut ainsi moduler à son gré des requêtes sur le réseau lexico-sémantique des fonctions lexicales mel'čukiennes. Avec **Étienne Blanc**, on aborde la présentation d'*Une maquette de base lexicale multilingue à pivot lexical : PARAX* qui est structurée pour l'expérimentation d'acceptions ou sens de mots et de leurs correspondants dans les autres langues et démontre son utilité notamment dans les questions qui appartiennent à la traduction automatique. **Lorne H. Bouchard** et **Louisette Ermikanian** présentent l'*Élaboration d'un dictionnaire informatisé pour le traitement automatique de la langue* dont les informations sont de nature morphologique et qui doit permettre le développement d'une grammaire computationnelle du français. Leur exposé est illustré par l'extraction de connaissances liées aux verbes de mouvement. **Guy Deville** et **Emmanuel Herbigniaux** décrivent leur projet *ANTHEM* dont l'objectif est le développement d'un prototype portable d'interface visant, par la création d'un modèle de représentation sémantique, la traduction et l'encodage automatiques de diagnostics médicaux. L'article de **Frédérique Segond** et **Élisabeth Breidt** présente *IDAREX* une *description formelle des expressions à mots multiples en français et en allemand dans le cadre de la technologie des états finis*, c'est-à-dire une expérimentation de compréhension des quasi-phrasèmes et des phrasèmes ainsi que des collocations. On voit tout l'intérêt que présente une telle étude pour un traitement de la langue tant en traduction automatique qu'en indexation automatique, par exemple. **Thomas Szende** et **Dominique Radanyi** traitent d'*aspects informatiques et pratiques dans la cadre de la lexicographie bilingue*. Leur expérience s'appuie sur la préparation du dictionnaire bilingue hongrois-français et montre, si besoin en était, que le rédacteur doit rester maître de son entreprise. **Patricia Thomas** et **Frank Knowles** s'attachent à déterminer l'orientation, ou mieux l'identification des bases et des collocataires dans les collocations en langues de spécialité. **Béatrice Daille** décrit *ACABIT : une maquette d'aide à la construction automatique de banques terminologiques* qui facilite la tâche des chercheurs en proposant un format et des structures morphosyntaxiques, grâce à des données linguistiques et statistiques, pour déterminer le recensement de termes spécialisés. **Didier Bourigault** décrit les problèmes théoriques et méthodologiques que présente la *Conception et l'exploitation d'un logiciel d'extraction de termes*. **Kumiko Tanaka** et **Violaine Prince** proposent un algorithme pour l'*Amélioration automatique incrémentale de dictionnaires bilingues utilisant un corpus monolingue* en s'appuyant sur des heuristiques telles que des informations morphologiques, la connaissance des synonymes et des valeurs de cooccurrence. **Xavier Lelubre** décrit sa *Conception d'un dictionnaire terminologique et phraséologique trilingue anglais/français-arabe dans le domaine de l'optique*. **Hussein Habaili** et **Mohamed Ben Ahmed** dans leur étude *Génération automatique de néologismes arabes* spécifient les règles lexicales qui permettent de dériver automatiquement des mots à partir des mots de base. **Miloud Taifi** dans son article *Construction des formes de mot et classification des entrées lexicales* présente de façon conventionnelle les problèmes particuliers de la lexicographie berbère. Les études de **Gilles Sérasset**, *Informatisation du Dictionnaire explicatif et combinatoire : le projet Nadia-DEC,* d'**Agnès Tutin**, *La formalisation des collocations pour le traitement automatique du langage naturel : le modèle des fonctions lexicales syntagmatiques*, et de **Margarita Alonso Ramos** et **Suzanne Mantha**, *Description lexicographique des collocatifs dans un dictionnaire*

explicatif et combinatoire : articles de dictionnaire autonome ?, appartiennent toutes trois, sous des aspects différents, informatisation et théorisation, à la même théorie linguistique, celle du Sens ⇔ Texte d'Igor Mel'čuk. **Christophe Jouis** et **Widad Mustafa-Elhadi** proposent un *nouvel outil interactif d'aide à la conception de dictionnaires électroniques spécialisés*. **Sylvain Delisle** dans son article *La construction de dictionnaires à partir de l'analyse informatisée de corpus bruts* vise à faciliter la rédaction d'un dictionnaire spécialisé, plus spécifiquement à préciser les propriétés syntaxiques et sémantiques des verbes. **Marc Van Campenhoudt** démontre dans son article *Réseau notionnel, intelligence artificielle et équivalence en terminologie multilingue : essai de modélisation* l'usage du réseau notionnel pour gérer divers problèmes d'équivalence dans une base de données multilingue. **Russon Wooldridge** et **Isabelle Leroy-Turcan** analysent dans *Les mots-clés métalinguistiques comme outil d'interrogation structurante des dictionnaires anciens* les difficultés qu'il faut envisager pour une exploitation optimale informatisée des dictionnaires anciens. **Michel Mathieu-Colas** dans son article *Représentation de la polysémie dans un dictionnaire électronique* propose le dégroupement maximal comme préalable à la description des liens qui lient les mots polysèmes. **Catherine Walther** et **Éric Wehrli** dans leur communication *Une base de données lexicale multilingue interactive* montrent tout l'intérêt qu'il y a à explorer les formes morphologiques associées à un lexème lorsqu'elles sont exprimées sous forme d'un ensemble de relations entre entrées indépendantes. **Évelyne Viegas** et **Sergei Nirenburg** démontrent dans leur article *Acquisition semi-automatique du lexique* le processus d'acquisition lexicale (Spanlex) dans un système de traduction automatique (Mikrokosmos). **François Gaudin** et **Myriam Bouveret** décrivent les *Pistes de description sémantique* telles que présentées dans *Biolex, dictionnaire des bio-industries*. **Pierrette Bouillon** dans son article *Le lexique génératif : une alternative au traitement de la polysémie* décrit le traitement de la polysémie des adjectifs qui dénotent un état mental dans le cadre de la théorie du lexique génératif. **Daniel Blampain** tantôt linguiste, tantôt usager de la langue décortique toute la problématique qu'il faut envisager *Quand l'informatique tutoie le dictionnaire des difficultés de la langue française*. **Philippe Barbaud** dans *Choix de grammaire et organisation du lexique* remet à jour la question du choix d'une grammaire et propose un réaménagement, plus spécifiquement une syntaxe dérivationnelle, pour le traitement informatisé des mots composés. **Philippe Blache** et **Mireille Delpui** décrivent un *Outil d'intégration de bases de connaissances lexicales aux analyseurs syntaxiques* et plaident pour une construction de lexiques adaptés à un formalisme donné. **Karim Chibout** et **Nicolas Masson** présentent comme thème d'étude *Un réseau lexico-sémantique des verbes construit à partir du dictionnaire pour le traitement automatique du français.*

On trouvera dans *Lexicomatique et dictionnairiques* une richesse expérimentale lexicologique et lexicographique tout à fait remarquable. Il y a là les modèles les plus variés, les applications les plus profondes, les réflexions les plus proligères en lexicologie, en lexicographie, en sémantique pour un développement fantastique des industries de la langue. On voit que le travail artisanal de la lexicographie a donné naissance à un savoir technique imprégné de connaissance scientifique. C'est en effet la science qui exige une réduction à des schémas qui permettent des substitutions. On voit facilement la complémentarité qui existe entre les données linguistiques et l'informatique. Il y a là des rapports étroits, un appui mutuel, si l'on peut dire, qui décrit bien la diversité des méthodes mais l'unicité de l'esprit scientifique qui s'appuie sur la réalité en vue de décrire, d'expliquer et de valider. Les schémas, les modèles, s'ils sont nécessairement abstraits doivent cependant subir la procédure de vérification, en

linguistique comme dans les autres sciences. La langue étant une des plus extraordinaires créations de l'Homme.

Pour de simples raisons financières, il n'a pas été possible d'inclure dans ce volume toutes les excellentes communications présentées aux IV^{es} Journées scientifiques de Lyon. Nous le regrettons sincèrement et nous nous en excusons auprès des collègues. Ces communications seront cependant publiées dans des revues spécialisées reconnues internationalement.

Enseignement du lexique assisté par ordinateur

Jurij D. Apresjan

Institut pour les problèmes de transmission d'information, Académie des Sciences de Russie, Moscou, Russie

Introduction

L'objectif final de ce projet est de créer un manuel multilingue sur PC destiné à enseigner le lexique des langues étrangères à travers le lexique de la langue maternelle et à apprendre le bon usage (correct et idiomatique) de ce dernier. Ce manuel s'adresse aux étudiants des écoles secondaires, des collèges, des universités, ainsi qu'aux personnes qui veulent individuellement apprendre une langue étrangère. Il leur permettra de se familiariser avec le lexique d'une langue étrangère ou de leur langue maternelle grâce aux techniques linguistiques originales basées sur les théories linguistiques modernes.

Dans le prototype que nous présentons aujourd'hui et qui fonctionne sur un IBM PC 486, seules deux langues sont accessibles : le russe et l'anglais. La taille des dictionnaires est également assez réduite : ils ne comportent qu'un millier de lexèmes dans chacune des langues. Comme on fournit systématiquement, pour chaque lexème, au moins un équivalent dans l'autre langue, nous avons pu garder un assez bon isomorphisme entre les deux dictionnaires. Actuellement, dans le cadre d'un projet INTAS, nous travaillons à la création d'un dictionnaire allemand. Dans les prochaines années, il est prévu d'y rajouter la version française et de porter la taille des dictionnaires à 3 000, voire 4 000 unités lexicales.

Sur le plan théorique notre manuel est essentiellement basé sur la théorie « Sens-Texte » de Mel'čuk (STM), dont il reprend les principaux concepts et notamment les fonctions lexicales (FL), le système de paraphrasage et le dictionnaire explicatif et combinatoire (DEC). Il s'appuie également sur le métalangage sémantique conçu par Apresjan pour décrire la sémantique des langues naturelles et sur sa théorie de la décomposition lexicale (définitions).

Le présent article se décompose en plusieurs sections : dans un premier temps, nous allons passer en revue quelques concepts généraux de la théorie STM (section

1), ensuite nous nous concentrerons sur la notion des fonctions lexicales de Mel'čuk (section 2), dans la 3e section nous montrerons comment les principes de la décomposition sémantique ont été mis en œuvre lors de la rédaction de notre manuel, enfin la 4e section décrit les « jeux linguistiques » sur PC issus des théories indiquées ci-dessus.

1. Concepts généraux de la théorie STM

Le modèle STM est un instrument logique destiné à établir pour une langue donnée des correspondances multiples entre les représentations sémantiques et les textes dans cette langue. Ce modèle s'oriente beaucoup plus vers l'expression que vers la compréhension, c'est-à-dire qu'il est plus adapté à la production des textes qu'à leur interprétation. Plus précisément, il permet de simuler trois aptitudes principales de la personne qui parle une langue quelconque, ces aptitudes caractérisant en fait la maîtrise de cette langue, à savoir :

a) l'aptitude à choisir les mots, les formes grammaticales et les constructions syntaxiques qui en principe pourraient être utilisés pour exprimer une pensée concrète. Par exemple, si l'on devait exprimer une pensée comme : « Le fait que la température de l'air ambiant est devenue soudain beaucoup plus basse a eu pour effet que les oiseaux nouveau-nés ont cessé de vivre ». Une personne parlant couramment l'anglais produirait sans difficulté un texte comme : *The nestlings died as a result of the cold wave* (Les couvées périrent à cause de l'arrivée d'un front froid).

b) l'aptitude à combiner correctement les mots, les formes grammaticales et les constructions syntaxiques retenus. On sait bien que les restrictions combinatoires ne s'expliquent pas toujours par des raisons sémantiques. Même les synonymes peuvent avoir un potentiel combinatoire différent. Ainsi, le mot anglais *word* (parole) dans une de ses acceptions est synonymique au mot *promise* (promesse) ; cf. *to give a promise <one's word> to do something* (donner la promesse <sa parole> de faire quelque chose) ; *to keep one's promise <one's word>* (tenir sa promesse <sa parole>) ; *to break one's promise <one's word>* (manquer à sa promesse <sa parole>). Cependant, seul le mot *promise* (promesse) peut s'employer avec les verbes *to make* et *to fulfil* (faire et remplir), cf. *to make a promise, to fulfil a promise* (faire une promesse, remplir la promesse), avec un adjectif numéral ordinal, cf. *three promises* (trois promesses) et avec un article défini ou indéfini, cf. *a promise, the promise* (une promesse, la promesse). Alors que *to make a word, to fulfil a word, three words, a word, the word* (faire une parole, remplir la parole, trois paroles, une parole, la parole) sont strictement inadmissibles pour cette acception du mot *word* (parole).

c) l'aptitude à exprimer la même pensée de plusieurs façons différentes si le message n'a pas été interprété correctement ou tout simplement suivant les circonstances, le caractère stylistique et le contexte général dans lequel se déroule un acte de communication : *The nestlings died as a result of the cold wave, The nestlings perished as a result of the cold wave, The nestlings died due to the cold wave, The cold wave killed the nestlings, The cold wave caused the death of the nestlings, The cold wave led to the death of the nestlings, The death of the nestlings was a consequence of the cold wave, The death of the nestlings was due to the cold wave, The death of the nestlings was a result of the cold wave, The nestlings died as a result of a sudden drastic*

drop in temperature, The nestlings perished as a result of a sudden drastic drop in temperature, The nestlings died due to a sudden drastic drop in temperature, etc.

Toutes ces aptitudes peuvent être décrites dans le cadre de la théorie « Sens-Texte » en termes de décomposition sémantique et de fonctions lexicales. Ces dernières permettent à la fois de formuler les règles de la cooccurrence des lexèmes et celles du paraphrasage.

2. Fonctions lexicales

Une FL, selon Mel'čuk, est un **sens** suffisamment abstrait et général qui peut être exprimé par un grand nombre de lexèmes différents ; le choix d'un lexème concret est chaque fois déterminé par le mot-clé (**argument**) auquel ce **sens** abstrait est associé. En d'autres mots, les FL sont des **sens**, dont l'expression est soumise à toutes sortes de restrictions lexicales. Prenons quelques exemples avec la FL bien connue MAGN (intensificateur) qui signifie « le haut degré de ce qui est exprimé par le lexème-argument ».

Fonction	**Argument**	**Valeur**
MAGN	*disease*	*grave*
MAGN	*contrast*	*sharp*
MAGN	*control*	*strict*
MAGN	*sleep (verb)*	*soundly*
MAGN	*know*	*firmly*

Un autre exemple des FL sont les verbes-supports appartenant à la famille OPER-FUNC, à savoir : $OPER_1$, $OPER_2$, $FUNC_1$, $FUNC_2$, $LABOR_{1-2}$, $LABOR_{1-2}$. Ce sont des verbes sémantiquement « vides » qui en fait reprennent une partie du sens exprimé par le mot-clé (argument). Leur apport sémantique à la signification de la phrase est quasiment nul. Notons que du point de vue syntaxique ces verbes sont des conversifs les uns par rapport aux autres.

$OPER_1$ prend le premier actant de la situation exprimée par le mot-clé en tant que sujet grammatical et il prend le mot-clé en tant que complément d'objet direct :

Fonction	**Argument**	**Valeur**
$OPER_1$	*resistance*	*put up (resistance)*
$OPER_1$	*control*	*exercise (control)*
$OPER_1$	*operation (surgical)*	*make (an operation)*
$OPER_1$	*fear*	*feel (fear)*
$OPER_1$	*proposal*	*make (a proposal)*
$OPER_1$	*sound*	*utter (a sound)*
$OPER_1$	*care*	*take (care) of (smb)*

$OPER_2$ prend le deuxième actant de la situation exprimée par le mot-clé en tant que sujet grammatical et il prend le mot-clé en tant que complément d'objet direct :

$OPER_2$	*resistance*	*meet (resistance)*
$OPER_2$	*control*	*be under (control)*
$OPER_2$	*operation*	*undergo (an operation)*

$FUNC_1$ prend le mot-clé en tant que sujet grammatical et il prend le premier actant de la situation exprimée par le mot-clé en tant que complément d'objet direct :

$FUNC_1$	*fear*	*possesses (smb)*
$FUNC_1$	*proposal*	*comes from (smb)*
$FUNC_1$	*sound*	*escapes (smb)*

$LABOR_{1-2}$ prend le premier actant de la situation exprimée par le mot-clé en tant que sujet grammatical, il prend le deuxième actant en tant que complément d'objet direct et le mot-clé en tant que complément d'objet indirect :

$LABOR_{1-2}$	*control*	*keep (smth) under (control)*
$LABOR_{1-2}$	*criticism*	*subject (smb) to (criticism)*
$LABOR_{1-2}$	*care*	*have (smb) in (one's care)*

Les quatre traits spécifiques des FL indiqués ci-dessous les rendent particulièrement attrayantes pour l'étude des langues : a) universalité ; b) adaptation au paraphrasage ; c) spécificité intralinguistique (idiomatisme) ; d) spécificité interlinguistique (idiomatisme).

a) Dans le cadre de la théorie « Sens-Texte » on définit environ 50 FL élémentaires, que l'on suppose être universelles pour toutes les langues et qui sont, par conséquent, décrites au moyen d'un métalangage formel standard. Ce métalangage qui s'appuie sur les assises théoriques très solides permet d'établir des équivalences entre les expressions de n'importe quelles langues sans jamais recourir à la traduction « mot à mot » – péché commun d'un grand nombre de dictionnaires. Nous avons déjà cité plus haut quelques exemples des FL MAGN en anglais, il serait intéressant de voir maintenant leurs équivalents russes :

Fonction	**Argument**	**Valeur**
MAGN	*bolezn' 'disease'*	*tjazhelaja*, lit. *'heavy'*
MAGN	*kontrast 'contrast'*	*rezkij*, lit. *'cutting'*
MAGN	*kontrol' 'control'*	*strogij*, lit. *'strict'*
MAGN	*spat' 'to sleep'*	*krepko*, lit. *'firmly'*
MAGN	*znat' 'to know'*	*tverdo*, lit. *'hard'*

b) La bonne adaptation des FL au système de paraphrasage peut être illustrée sur l'exemple de la famille OPER-FUNC. Comme nous l'avons déjà signalé plus haut, l'apport sémantique des FL, faisant partie de cette famille, à la signification du mot-clé et à toute la phrase est minime. Aussi constituent-elles un kit linguistique idéal permettant de formuler quelques règles de paraphrasage très générales et d'établir les relations de synonymie entre les énonciations. Notamment, chaque verbe « significatif » X peut être paraphrasé moyennant une expression constituée par un verbe support, vide de sens, appartenant à la famille OPER-FUNC et un substantif – nom d'action, d'état ou de processus – dérivé du verbe X.

$X = OPER_1 + S_0(X)$
$X = OPER_2 + S_0(X)$
$X = LABOR_{1-2} + S_0(X)$
$X = FUNC_1 + S_0(X)$ etc.

Les équations ci-dessus permettent de construire les équivalences suivantes : $OPER_1 + S_0(X) = OPER_2 + S_0(X) = LABOR_{1-2} + S_0(X) = FUNC_1 + S_0(X)$ etc. Cf. :

The president controls (X) *the situation* =
The president has ($OPER_1$) *control* ($S_0(X)$) *of the situation* =
The president keeps ($LABOR_{1-2}$) *the situation under control* ;
The surgeon operated (X) *on the wounded* =
The surgeon made ($OPER_1$) *an operation* ($S_0(X)$) *on the wounded* =
The wounded underwent ($OPER_2$) *an operation (by the surgeon)* ;
He was afraid (X) =
He felt ($OPER_1$) *fear* ($S_0(X)$) =
Fear possessed ($FUNC_1$) *him* ;
He didn't utter ($OPER_1$) *a sound* =
No sound escaped ($FUNC_1$) *him.*

L'aptitude à trouver ces paraphrases fait partie des facultés linguistiques qui assurent la maîtrise parfaite d'une langue, qu'il s'agisse d'une langue étrangère ou maternelle.

c) La notion de spécificité intralinguistique est assez transparente et ne nécessite pas de longues explications, il suffit juste de regarder quelques exemples des FL cités ci-dessus pour se rendre à l'évidence qu'il n'existe aucune corrélation sémantique directe entre la valeur d'une FL et le sens du mot-clé. En effet, il serait très difficile d'expliquer pourquoi *orders are issued*, alors que *demands are made* ; ou encore pourquoi nous disons *sleep soundly*, mais *know firmly*, et pourquoi le contraire est impossible : **sleep firmly*, **know soundly* – malgré le fait que du point de vue sémantique les deux adverbes ont une signification très proche. Ce phénomène devient très spectaculaire lorsqu'on compare deux synonymes, pour lesquels les restrictions de cooccurrence lexicale sont différentes. Ainsi, il n'existe aucune raison sémantique pour que le substantif *word* lorsqu'il signifie 'promesse' ne puisse être employé avec les verbes *to make* et *to fulfil* (*to *make a word*, **to fulfil a word*), – qui par ailleurs s'emploient facilement avec le substantif *promise*. Ce qui plus est, deux substantifs, dont les acceptions principales sont synonymiques, par exemple, *desire* et *wish*, sont parfois soumis à des restrictions différentes ; cf. *strong/keen/intense/fervent/ardent/ overwhelming desire*, mais seulement *strong/fervent/ardent wish*. Par conséquent, une substitution synonymique libre est impossible pour ces deux mots. Ainsi, toute personne qui prétend parler couramment une langue, doit pour chaque mot-clé apprendre tout bêtement par cœur le lexème qui signifie le haut degré de ce qui est exprimé par ce mot-clé, puisqu'il n'existe aucun critère sémantique qui lui permette de choisir le mot approprié dans chaque cas concret.

d) Compte tenu de la très grande variété d'expressions d'une fonction lexicale au sein d'une langue donnée, on peut présumer que la spécificité interlinguistique des FL est encore plus prononcée. Pour ne pas citer de nouveaux exemples, nous renvoyons à ceux de la FL MAGN que nous avons considérés plus haut dans le paragraphe (a) en comparant les substantifs russes à leurs équivalents anglais : *bolezn'/disease, kontrast/ contrast, kontrol'/control, spat'/to sleep, znat'/to know*. Il existe cependant un autre aspect – beaucoup plus fondamental – de la spécificité interlinguistique des FL : les langues sont différentes non seulement parce qu'elles expriment la même idée par des mots différents (comme nous l'avons vu dans les exemples précédents), mais aussi et surtout

parce que la fréquence de l'expression de cette idée varie d'une langue à l'autre. Cette dernière différence est profondément ancrée dans la typologie d'une langue donnée et dans sa base conceptuelle inhérente (la soi-disant « vision naïve de l'univers »). Ainsi, la langue russe qui tend à masquer le vrai agent d'une action et à personnifier les états physiques et mentaux en leur prêtant une volonté autonome et indépendante (d'où la profusion des constructions impersonnelles et indéfinies en russe), utilise beaucoup plus que l'anglais les verbes supports qui prennent le mot-clé (argument d'une FL) en tant que leur sujet grammatical. Alors que l'anglais ou le français, avec leur rationalisme et leur tendance à mettre en relief l'agent d'une action, évitent les constructions impersonnelles et essayent de les remplacer par les constructions agentives. Cf. :

U nego ['*he*', complément d'objet] *byli opasenija*
['*apprehensions*', sujet grammatical] *po aetomu povodu* -
He [sujet grammatical] *had certain apprehensions*
[complément d'objet] *on this account*.

où les mots russes *on, opasenija* et leurs équivalents anglais *he* et *apprehensions* jouent les rôles syntaxiques opposés. Cf. aussi :

Toska glozhet <snedaet, tochit> ego [complément d'objet] -
He [sujet grammatical] *is consumed with melancholy,*
Radost' vladeet im [complément d'objet] -
He [sujet grammatical] *is full of joy,*
Ljubopytstvo razbiraet ego [complément d'objet] -
He [sujet grammatical] *cannot suppress his curiosity,*
Revnost' muchit ego [complément d'objet] -
He [sujet grammatical] *is a martyr to jealousy,*
U nego [complément d'objet] *na ume chto-to drugoe* -
He [sujet grammatical] *has something else on his mind etc.*

Cette construction syntaxique s'applique en russe non seulement aux substantifs qui désignent les émotions et les états mentaux, mais aussi aux noms d'états physiques, comme par exemple *kashel'* 'toux', *golod* 'faim', *zhazhda* 'soif', etc. :

Ego [complément d'objet] *bil <donimal> kashel'* -
He [sujet grammatical] *had a bad fit of coughing,*
Ego [complément d'objet] *muchit golod <zhazhda>* -
He's [sujet grammatical] *going through the tortures of hunger <thirst>.*

Cf. aussi :
Èta melodija do six por zvuchit u menja [complément d'objet] *v ushax* -
I [sujet grammatical] *can still hear this melody,*
Protivnyj vkus ryby do six por u menja [complément d'objet] *vo rtu* -
I [sujet grammatical] *can still feel the foul taste of fish in my mouth,* etc.

Ce type de divergences entre les langues peut être explicité facilement en termes de FL. Ainsi, en décrivant la façon d'exprimer les états physiques, mentaux et émotionnels dans les différentes langues, on pourra dire que l'anglais préfère les constructions à $OPER_1$ là où le russe emploie les constructions à $FUNC_1$.

Dans le cadre de la théorie « Sens-Texte », on distingue deux grandes catégories de FL : les FL (dites **paradigmatiques**) qui s'emploient *à la place* du mot-clé et les FL **syntagmatiques** qui s'emploient *avec* le mot-clé (par exemple, la FL MAGN).

Parmi les FL paradigmatiques on retrouve : SYN (synonymes), ANTI (antonymes), CONV (conversifs, comme *to buy* et *to sell*), GENER (mot générique ou hyperonyme), ainsi que les dérivés syntaxiques d'un mot-clé : cf. S_0(*to teach*) = *teaching*, S_1(*to teach*) = *teacher* ('celui qui enseigne'), S_2(*to teach*) = *subject* ('ce qu'on enseigne'), S_3(*to teach*) = *pupil* ('celui à qui on enseigne') ; S_0(*to sell*) = *sale*, S_1(*to sell*) = *seller*, S_2(*to sell*) = *article*, S_3(*to sell*) = *buyer*, S_4(*to sell*) = *cost*, etc. Ces FL sont largement utilisées dans certains types de paraphrases, cf. *He teaches me* = *He is my teacher* = *I am his pupil* (basées sur les règles universelles ci-dessous. X = copula + S_1(X) = copula + S_2(X), etc.).

Les LF syntagmatiques comprennent avant tout les verbes supports $OPER_1$, $OPER_2$, $LABOR_{1-2}$ etc. ; elles participent également à un certain nombre de paraphrases (voir les exemples ci-dessus).

Par ailleurs, les FL se subdivisent en FL **simples ou élémentaires** (voir les exemples ci-dessus) et en FL **complexes**, ces dernières étant généralement construites par superposition de deux ou plusieurs FL simples. Voici quelques exemples de FL complexes (dont les noms sont suffisamment parlants et ne demandent pas d'explications supplémentaires) : $INCEPOPER_1$, $FINOPER_1$, $INCEPFUNC_1$, $FINFUNC_1$, etc.

Le nombre total des principales FL simples et complexes, dont on aura besoin pour formuler les règles du choix lexical collocationnel et du paraphrasage, s'élève à environ 80. La liste des FL a été revue et adaptée aux besoins des jeux linguistiques. Pour chacune des FL faisant partie de cette liste nous avons fourni une définition en anglais et en russe et un ensemble de mots (plusieurs dizaines pour certaines fonctions) qui servent à illustrer cette fonction et qui constituent la « matière linguistique » pour le jeu. Par rapport à la version initiale de la théorie STM, les définitions des FL ont été systématisées et standardisées.

3. Décomposition sémantique (définition)

Une entrée du DEC russe ou anglais contient les informations suivantes : 1) une vedette ou nom du lexème (par lexème on entend un mot dans une de ses acceptions lexicales) ; 2) une définition du lexème (d'une acception lexicale) ; 3) une partie du discours ; 4) un équivalent du lexème dans la langue cible ; 5) les principales FL paradigmatiques pertinentes pour ce lexème ; 6) les principales FL syntagmatiques pertinentes pour ce lexème.

En dehors des FL, l'information la plus importante d'une entrée du DEC est la définition. Les définitions, quoique un peu simplifiées, sont formulées dans un métalangage sémantique et s'appuient sur quelques principes généraux posés par l'auteur du manuel dans ses publications récentes.

Le métalangage sémantique n'est en fait qu'un sous-langage d'une langue naturelle donnée. Ceci est vrai tant pour son lexique que pour sa grammaire. Ce mé-

talangage comprend des mots et des constructions syntaxiques relativement simples. Chaque élément du métalangage sémantique doit répondre aux critères de la non-synonymie (idéalement, à chaque sens doit correspondre un seul mot) et de la non-homonymie (idéalement, à chaque mot nous devons attribuer un seul sens). La synonymie est autorisée dans deux cas bien précis : lorsqu'il s'agit d'une transposition syntaxique : *to make/making* et dans le cas des unités sémantiques élémentaires, cf. *to want* et *good*, ou *to know* et *true*, bien qu'il y ait un très grand recoupement sémantique entre les deux paires de mots ci-dessus.

Même ce descriptif très sommaire montre que notre métalangage sémantique n'est pas universel, mais qu'il est spécifique pour une langue donnée. Ceci n'exclut pas la possibilité de l'adapter à une autre langue à condition qu'il s'agisse de langues assez proches du point de vue culturel, par exemple le russe, l'anglais, le français et l'allemand. Dans la plupart des cas une simple traduction mot à mot du lexique du métalangage suffit.

Le noyau du métalexique est constitué par les soi-disant **unités sémantiques élémentaires** – mots qui ne peuvent pas être décomposés en éléments plus simples dans une langue donnée sans qu'il y ait un cercle vicieux. Ceci dit, une unité sémantique élémentaire ne doit pas nécessairement correspondre au plus simple sens possible. Le seul critère auquel elle doit répondre est le suivant : chaque fois que l'on veut discerner un sens encore plus simple dans une unité sémantique élémentaire, il faut trouver un mot approprié dans la langue en question pour exprimer ce sens. Il est évident, par exemple, que les verbes *to want, to wish* et *to desire* du point de vue sémantique ont une partie commune importante. Mais cette partie commune n'est en fait qu'une pure abstraction, que l'on n'arrive pas à verbaliser en anglais. En réalité, le verbe *to want* ne peut pas prétendre à exprimer un sens plus élémentaire, puisqu'en plus du noyau sémantique commun pour tous ces verbes il contient aussi l'idée de la 'nécessité', qui est totalement étrangère aux verbes *wish* et *desire* ; à son tour le verbe *wish* peut exprimer l'idée de la 'futilité' des désirs du sujet, alors que le verbe *desire* rajoute l'idée de 'volonté' et de 'résolution'.

La liste des unités sémantiques élémentaires en anglais contient les éléments, tels que : *time, space, world, object, person, property, part, state, process, eye, ear,* ainsi que *good, true, all, one, first, to want, to think, to know, to feel, to say, to do, to happen, to be, to perceive, to cause, can, not, or, and that, what, which* etc. En dehors des unités sémantiques élémentaires, le métalangage contient aussi quelques éléments sémantiques intermédiaires, comme : *obligatory = such that it is impossible not to do it ; impossible = not possible ; possible = such that can happen or be done.*

Passons maintenant aux définitions à proprement parler. Elles doivent répondre aux critères suivants :

1) absence de cercles vicieux ;
2) exhaustivité et non-redondance : la définition d'un lexème doit contenir tous ceux et seulement ceux des éléments sémantiques qui constituent son sens ;
3) réductibilité : il faut que toute définition puisse être décomposée en unités sémantiques élémentaires, directement, ou en passant par des stades intermédiaires ;
4) systématicité : l'ensemble des définitions doit être construit de sorte que l'on puisse identifier facilement les liens sémantiques systématiques qui existent entre les lexèmes d'une langue.

Pour illustrer les concepts généraux exposés ci-dessus, nous allons proposer les définitions d'un groupe de lexèmes (dits « temporaux ») qui présentent des traits sémantiques communs : *autumn, dark, day$_1$, day$_2$, evening, hour, January, light, minute, Monday, month$_1$, month$_2$, morning, night, period, season, second, to see, spring, summer, sun, sunrise, sunset, today, tomorrow, week, winter, year, yesterday.*

Autumn = the season of the year between summer and winter when it starts to be getting colder.
Dark = such as has little or no light.
Day$_1$ = the part of time between two consecutive sunrises.
Day$_2$ = the lightest part of day$_1$ which follows morning, precedes evening and ends between 16 and 17 o'clock.
Evening = the part of day$_1$ when it starts to be getting dark, which includes sunset and ends between 23 and 0 o'clock.
Hour = the 24-th part of day$_1$.
January = the first month of the year.
Light = that property of the world which makes it possible to see objects.
Minute = the sixtieth part of an hour.
Monday = the first working day of the week.
Month$_1$ = one of the twelve parts into which a year is divided.
Month$_2$ = any period of approximately 30 days$_1$.
Morning = the part of day$_1$ when it starts to be getting light which includes sunrise and ends between 11 and 12 o'clock.
Night = the darkest part of day$_1$ which follows evening, precedes morning and ends between 4 and 5 o'clock.
Period = any part of time.
Season = one of the four large parts into which a year is divided on the basis of the natural cycle considerations.
Second = the sixtieth part of a minute.
See = to perceive with eyes.
Spring = the season of the year between winter and summer when it starts to be getting warmer.
Summer = the warmest season of the year with the longest days$_2$ and shortest nights.
Sun = the brightest object in the sky which gives light and warmth.
Sunrise = the process of the sun appearing on the horizon and the initial stages of its going up in the sky, or the time during which this process takes place.
Sunset = the process of the sun disappearing beyond the horizon and the final stages of its going down in the sky, or the time during which this process takes place.
Today = that day$_1$ through which the speaker is living at the time of speech.
Tomorrow = the day$_1$ immediately after today.
Week = a period of seven days$_1$ singled out on the basis of human activity cycle considerations.
Winter = the coldest season of the year with the shortest days$_2$ and longest nights.
Year = the part of time equal to 365 days$_1$, in which the earth makes one full circle around the sun.
Yesterday = the day$_1$ immediately before today.

4. Jeux linguistiques

Le manuel permet de jouer à quatre jeux linguistiques différents basés sur les définitions des lexèmes et sur les FL et qui consistent à : 1) trouver une traduction pour le lexème proposé par l'ordinateur ; 2) indiquer les valeurs de toutes les FL affichées par l'ordinateur pour un lexème choisi par le joueur ; 3) indiquer les valeurs d'une FL choisie par le joueur pour tous les lexèmes affichés par l'ordinateur ; 4) trouver le nom du lexème qui correspond à la définition affichée par l'ordinateur.

Dans sa version initiale le manuel supposait que tous les jeux seront joués dans le cadre d'une même langue. Mais maintenant chacun des jeux énumérés ci-dessus peut aussi être joué en mode bilingue, où la connaissance de sa langue maternelle (quelle qu'elle soit) aide le joueur à trouver une réponse correcte. Ceci est possible dans la mesure où l'on arrive à garder l'isomorphisme des DEC.

Le manuel permet également de vérifier la qualité des connaissances linguistiques d'un joueur. À chaque étape du jeu on affiche le nombre de points que le joueur a gagnés par sa dernière réponse et le score total du jeu.

Nous avons prévu de doter le manuel d'un système de dialogue moderne (actuellement notre programmeur travaille à la mise au point de ce menu) basé sur une présentation idéographique qui permettra aux joueurs de choisir le domaine lexico-sémantique qui les intéresse, le type et le mode du jeu.

Nous travaillons également à la création d'un logiciel qui doit permettre l'extension et la mise à jour des DEC et des autres ressources linguistiques du manuel.

Par la suite nous envisageons également d'introduire de nouveaux jeux linguistiques, dont un jeu de paraphrasage qui consiste à trouver le maximum de paraphrases possibles pour une proposition donnée construite à partir des unités lexicales présentes dans les DEC.

L'hypertexte *Hyperbase*

Étienne Brunet

Institut national de la langue française, UPR 6861 (CNRS), Nice, France

On a quelque scrupule à présenter un produit qui n'est plus tout jeune ni tout à fait inconnu. Mais sa notoriété n'est pas suffisante pour autoriser la prétérition et son âge – six ans déjà – mesure en réalité le destin d'une famille où plusieurs générations sont enveloppées. Tout logiciel qu'on a l'imprudence de livrer au public est en effet un boulet qu'on traîne pendant des années, avec la nécessité de le renouveler sans cesse, pour le corriger, l'améliorer, le compléter et surtout l'adapter aux changements rapides des machines et des systèmes. Quelle différence avec le livre, bon ou mauvais, qui s'envole au moment de la publication et dont l'auteur se libère d'un coup, en songeant au suivant. Quand on met un logiciel sur le marché on ne peut jamais dire *alea jacta est*. Le lancement d'un tel produit est au mieux celui d'un cerf-volant captif dont il faut gouverner les sautes en surveillant le vent, en sorte que l'auteur est plus captif encore.

Curieusement les auteurs de logiciels ont pris l'habitude, chaque fois qu'ils publient une version nouvelle, de faire l'historique de leur produit, en dressant le catalogue de leurs fautes, de leurs corrections, de leurs variations. Quel soin scrupuleux pour noter les erreurs et les repentirs, alors que les écrivains partagent généralement le souci inverse, qui est d'effacer les ratures et de laisser croire au premier jet lumineux de l'inspiration. Et ce blanchissement est facilité de nos jours par les traitements de texte qui anéantissent les brouillons et toutes les traces de la transpiration. Cette discrétion orgueilleuse étant réservée à l'écrivain, on s'en tiendra donc à la pratique des auteurs de logiciels dont la modestie excuse l'impudeur.

1. La première version d'*Hyperbase*

La tradition des concordances est ancienne et l'idée d'en confier la réalisation à l'ordinateur est venue très tôt, dès les années 60. Les premières réalisations, à Besançon (avec Quemada), à Liège (avec Delatte), à Gallarate (avec Busa), datent de cette époque héroïque. Dix ans plus tard, les gros systèmes offraient à peu près partout dans le monde de tels services documentaires : à Nancy, à Paris, à Pise, à Oxford, à Göte-

borg, à Montréal et dans beaucoup d'universités américaines. Et dès 1970 nous proposions une chaîne de programmes pour le traitement documentaire et statistique des données textuelles[1]. À cette époque déjà le *TLF* avait mené à son terme l'indexation du corpus des XIX^e^ et XX^e^ siècles, soit plus de 70 millions de mots. Et disposant des données dérivées de ce traitement (cela s'appelait les fichiers-répertoires), nous avons créé une panoplie d'outils spécialisés pour l'étude quantitative de cette matière textuelle à demi traitée, au besoin en faisant le chemin inverse et en reconstituant le texte à partir des index[2]. Dix ans plus tard, au cours des années 80, le paysage technologique change radicalement avec l'avènement de la micro-informatique et l'extension des réseaux. À l'Institut national de la langue française se constitue la base de données *Frantext*, qu'on peut interroger par le réseau *Transpac* et qui offre à la communauté scientifique un immense champ de recherche (près de 3 000 textes complets et quelque 160 millions de mots), en même temps qu'un produit dérivé, *Discotext 1*, présente une large part de cette base (500 titres) sur CD-ROM, au standard PC.

De notre côté nous nous étions lancé dans un projet, complémentaire des précédents, et destiné aux utilisateurs (nombreux chez les littéraires) du standard Apple. Et un contrat de développement, signé avec la compagnie Apple-France, prévoyait une gamme de produits diversifiés. Le premier prototype d'*Hyperbase* était destiné, comme *Discotext*, aux grands corpus. Conçue en 1989 pour une manifestation du Bicentenaire à Beaubourg, cette version première du logiciel a mis à la disposition du public du Centre Pompidou un ensemble de textes de la Révolution, issu du *TLF* et représentant 30 millions de caractères. Il n'est guère utile d'en décrire le détail, puisque deux publications sont explicites là-dessus[3]. Il suffit de représenter, dans la figure 1, le menu qui accueille l'utilisateur et qui lui est offert systématiquement dès qu'une action est accomplie. Deux voies principales s'ouvrent à la sélection : à droite on s'engage dans la recherche documentaire ; à gauche on s'oriente vers les traitements statistiques.

1. « Programmes linguistiques, série 1 », *CUMFID*, n° 2, 1970, 175 p. Cette chaîne de programmes intégrés a servi en particulier à l'élaboration des *Index de J. J. Rousseau* en 26 volumes, publiés aux Éditions Slatkine, Genève.

2. Plusieurs monographies sont nées de cette exploitation intensive du gisement de Nancy, sur Proust, Zola, Hugo, ouvrages publiés aux Éditions Slatkine.

3. « Hyperbase : logiciel documentaire et statistique pour l'exploitation des grands corpus », *Tools for Humanists*, Toronto, 1989, pp. 33-36.

« Computer processing and quantitative text analysis : *Hyperbase*, an interactive software for large corpora », *Data Analysis, Learning Symbolic and Numeric Knowledge*, INRIA, Nova Science Publishers, New York, Budapest, 1989, pp. 207-214.

« What do the tables say ? What do the figures say ? », Colloque ALLC de Toronto *The Dynamic Text*, juin 1989, in *Literary and Linguistic Computing*, Oxford, 1989, pp. 70-82.

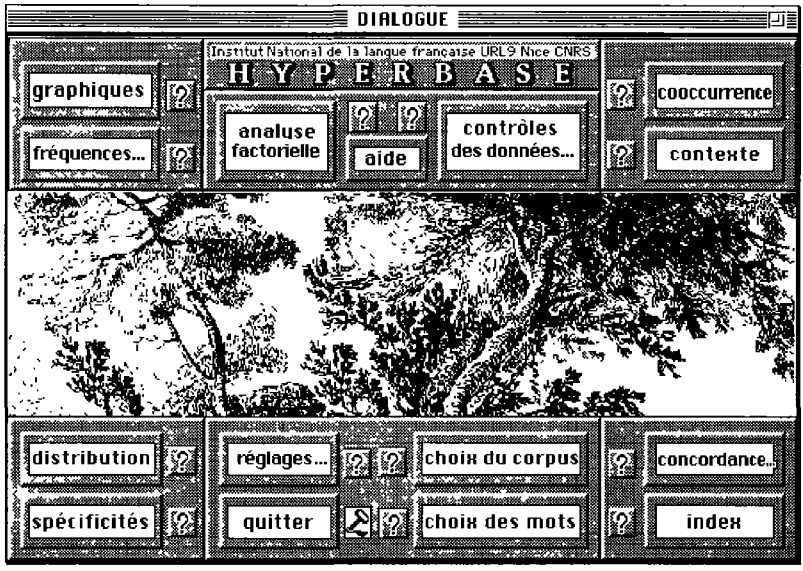


FIGURE 1 : *Hyperbase* première version. Page d'accueil.

Le choix du traitement s'accompagne d'autres sélections qui s'exercent parmi les textes et parmi les mots et que proposent les deux écrans ci-dessous (figures 2 et 3). Ces fonctions de sélection et d'exclusion se justifient dans le cas des très grands corpus qui réunissent une multitude de textes, de genres, d'époques et d'auteurs. Elles sont naturellement disponibles dans *Frantext* et *Discotext 1*. Elles permettent au chercheur de se constituer un lot de sous-corpus spécifiques et d'écarter provisoirement le reste. De même on peut à sa guise créer des listes de mots (au besoin à l'aide de filtres comme la finale, l'initiale ou la fréquence), les appliquer à des corpus différents et les retrouver d'une séance à l'autre.

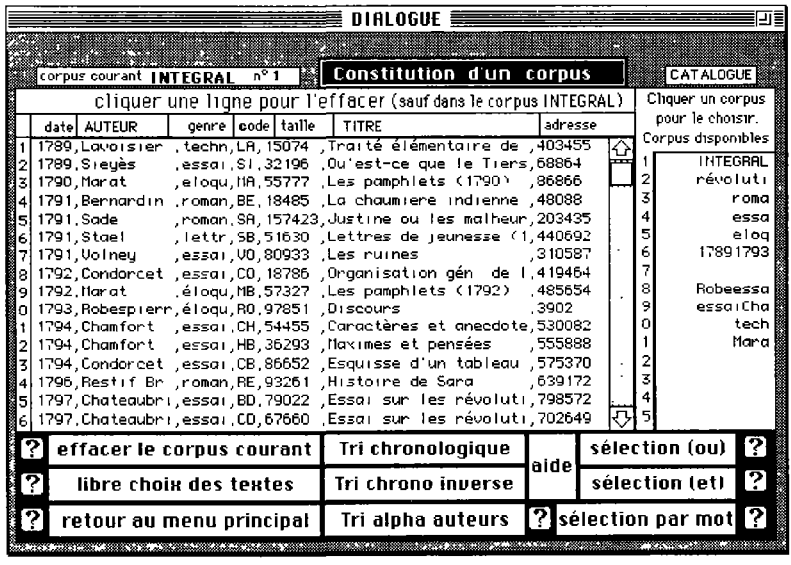


FIGURE 2 Choix du corpus.

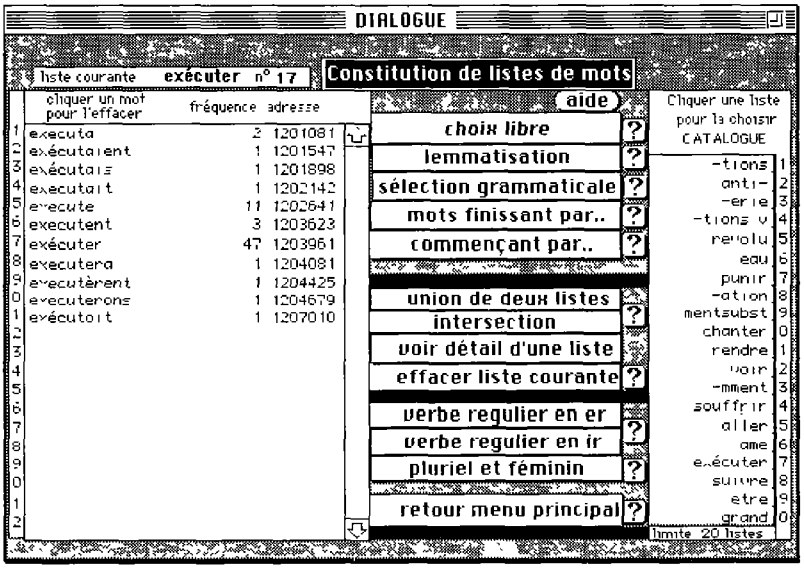


FIGURE 3 : Choix des mots

Nous ne cacherons pas notre préférence pour ce premier-né, qui nous a servi fidèlement et que nous gardons pour nos besoins personnels ou pour quelque grande entreprise, comme le CD-ROM Balzac envisagé à l'occasion du centenaire de 1999. Quand un corpus comporte une centaine de textes, il est déraisonnable de penser que les utilisateurs exploreront toujours la totalité indivisible, et peu rentable de poursuivre les recherches sur l'ensemble quand seule une fraction est jugée digne d'intérêt.

En dehors de cette liberté dans le choix de la portée, restreinte ou étendue, des traitements, la version d'origine offrait des fonctionnalités qui ont été abandonnées par la suite, à notre grand regret. Les options les plus précieuses et les plus difficiles à mettre en œuvre sont relatives à la lemmatisation et au codage grammatical. Comme on disposait d'un fichier de correspondance entre les vocables et les formes, établi à Nancy pour les besoins du *TLF*, on avait là le moyen de procéder à un regroupement des formes sous le même lemme. De plus le code grammatical emprunté à la même source permettait d'ajouter un filtre à la sélection des mots et par exemple de réunir les mots en *-tions* qui ne soient pas des formes verbales (à la 1re personne du pluriel). Enfin des données quantitatives puisées dans le fichier du *TLF* permettaient d'établir pour chaque vocable une comparaison entre les fréquences observées dans le corpus de la Révolution et dans un corpus plus vaste englobant la première moitié du XIXe siècle.

2. La version commercialisée d'*Hyperbase*

Comme le prévoyait le contrat de développement signé avec *Apple*, une nouvelle version du logiciel a été créée, qui se fonde sur des principes différents et dont la logique est celle des logiciels commerciaux, qui sont vides de données (comme les traite-

ments de texte, les tableurs, etc.), mais riches d'outils variés, conçus pour aider l'utilisateur à traiter son bien propre. Le programme a certes un jeu de données provisoires, ce qui facilite l'apprentissage. Mais la base peut être vidée de son contenu et recevoir d'autres données (sous forme de texte *ASCII*). Programmes de préparation et d'exploitation sont fournis conjointement dans le même produit.

Hyperbase dans cette version à tout faire vise à la généralité et à la simplicité. Le produit est conçu pour s'adapter immédiatement aux données de l'utilisateur et réaliser l'indexation dans un temps acceptable et sans manipulation excessive. Ces contraintes ont empêché de fournir les outils spécialisés de la lemmatisation, laquelle est nécessairement propre à chaque langue et ne peut jamais être complètement automatique. Comme cette version indifférenciée du logiciel devait s'appliquer à tous les textes qui utilisent un alphabet latin et donc à la plupart des langues occidentales[4], il était difficile de fournir pour toutes ces langues un dictionnaire-machine doté non seulement des codes grammaticaux indispensables mais aussi d'indications de fréquence[5]. Si donc la lemmatisation a été sacrifiée, faute de pouvoir être universelle, les deux objectifs antérieurs ont été maintenus qui orientent l'exploitation vers la recherche documentaire et la statistique.

a - Le programme d'exploitation répond, par les méthodes de l'hypertexte, aux besoins classiques du traitement automatique des textes : concordances de type *kwic* (avec tri des expansions droite ou gauche du contexte), index sélectifs ou systématiques, dictionnaires des fréquences, sélection de contextes larges, cooccurrences, filtrage et masquage des mots et constitution de listes, recherche des parties de mots (début, fin ou chaîne quelconque) ou des groupes de mots, limitation ou extension du corpus de travail, etc.

Nous renvoyons le lecteur à des publications antérieures où l'on trouvera une description plus détaillée de ces fonctions documentaires[6]. Un mode d'emploi d'une centaine de pages, plus explicite encore, accompagne le logiciel et guide l'utilisateur. Celui-ci a encore à sa disposition une aide en ligne plus concentrée, et, qui plus est, une page d'explication, immédiatement disponible, pour chaque fonction. Si le symbolisme d'un bouton n'est pas compris tout de suite et si son nom laisse perplexe, on peut par précaution faire apparaître sur l'écran les instructions précises avant de déclencher l'action correspondante.

On se bornera à titre d'exemple à la fonction *contexte* dont on montrera les options (figure 4) et le résultat (figure 5). Le corpus exploré est ici celui de Julien Gracq, soit l'ensemble de son œuvre (17 titres).

4 Une version particulière a été adaptée au grec moderne La même opération pourrait être envisagée pour le cyrillique

5. Un seul dictionnaire est fourni qui est limité au français et aux 10 000 formes les plus fréquentes.

6 « Un hypertexte statistique · HYPERBASE », *JADT 1993*, TELECOM, Paris, 1994, pp 1-16
« Hyperbase, synopsis », *Traitements informatisés de corpus textuels*, Didier Érudition, 1994, pp 169-184

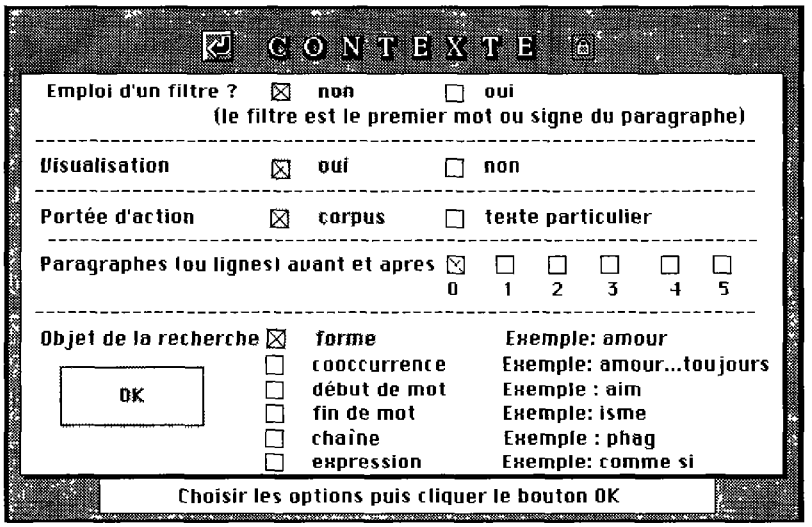


FIGURE 4 Dialogue proposé par la fonction *contexte*.

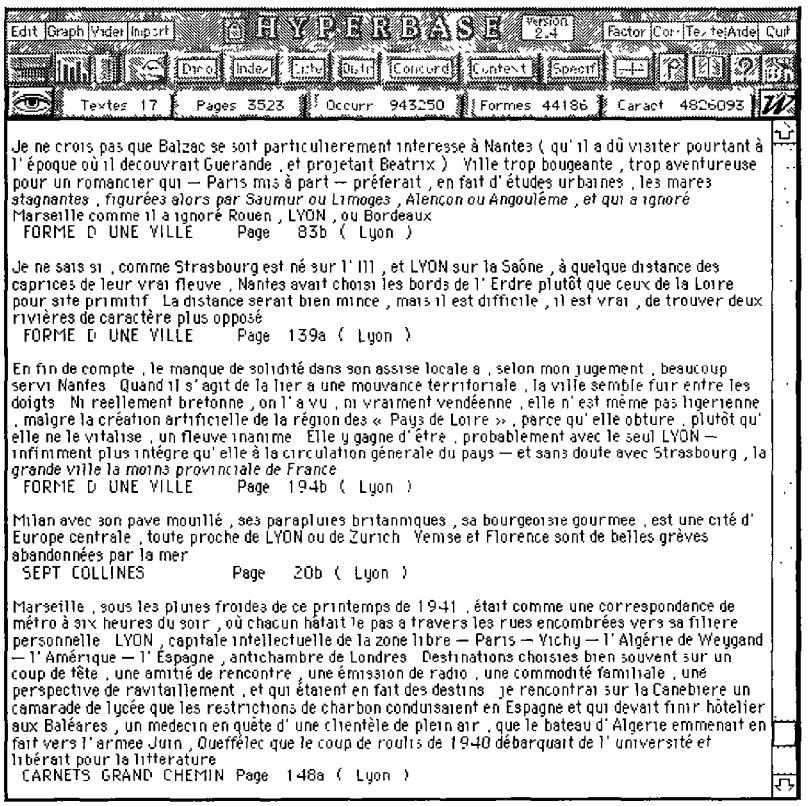


FIGURE 5 : Résultat de la fonction *contexte* dans l'œuvre de Gracq
(8 occurrences de la ville de *Lyon*, extrait).

b - *Hyperbase* se distingue des hypertextes similaires par l'orientation statistique donnée au produit. D'une part, s'il s'agit d'un texte français, une comparaison est faite[7], sous forme d'écart réduit, avec le corpus du *Trésor de la langue française* (XIX-XX^{es}, soit 70 millions de mots). D'autre part, le corpus peut être partitionné pour permettre des comparaisons internes. *Hyperbase* restitue ainsi les mots-clés propres à chaque texte, de même qu'il dresse le profil caractéristique du corpus dans son ensemble, se détachant sur la toile de fond de l'usage littéraire de la langue depuis 1789.

De même le profil d'un mot (ou de plusieurs que l'on superpose) est dessiné par *Hyperbase*, les sous-fréquences observées, judicieusement pondérées, se transformant à volonté en histogramme (figure 6). Des tableaux peuvent être constitués qui, à partir de critères, automatiques ou non, procèdent aux regroupements de mots ou de textes. Et ainsi peut-on pallier l'absence de lemmatisation. Par exemple *Hyperbase* permet de circonscrire une catégorie grammaticale, un champ thématique, voire même le système de la ponctuation. Une fois constituées, ces listes – ce sont en réalité des tableaux à deux dimensions – peuvent être soumises aux méthodes multidimensionnelles (un programme d'analyse factorielle a été intégré à *Hyperbase*).

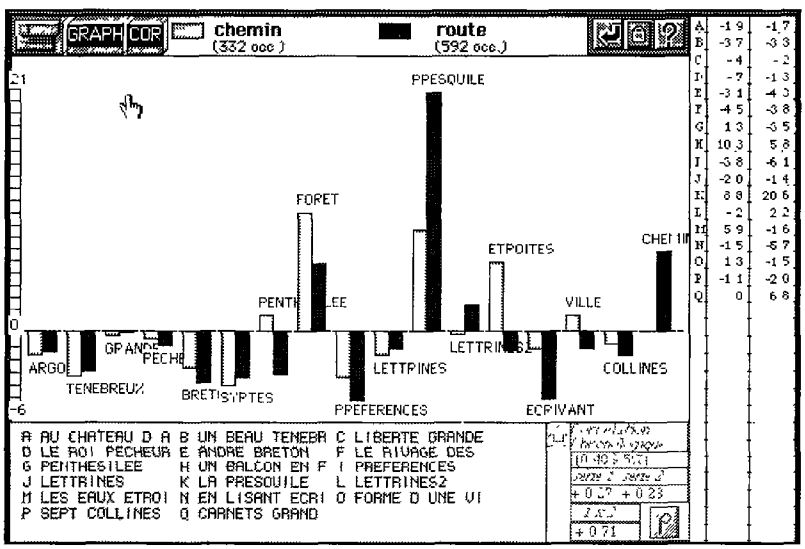


FIGURE 6 Représentation graphique de *route* et *chemin*.

D'autres calculs lexicométriques sont assurés qui permettent d'apprécier la richesse relative du vocabulaire, la distribution des classes de fréquences, l'abondance, si l'on peut dire, des mots rares (ou hapax), l'accroissement et l'évolution du vocabulaire, etc. En particulier une fonctionnalité nouvelle est apparue dernièrement (version 2.5, juillet 1995), qui mesure la distance que chaque texte établit avec tous les autres du même corpus, et qui est le rapport entre les vocables communs aux deux textes que l'on confronte et les vocables exclusifs que chacun des deux se réserve. Cela, qui peut s'appeler aussi la connexion lexicale, permet d'établir une typologie des textes à partir des similarités lexicales et principalement de leur composante sémantique[8].

7 Là où le calcul se justifie, c'est-à-dire quand la fréquence est suffisante dans le modèle (soit f = 500 dans le *TLF*).
8 La fréquence ne joue ici aucun rôle. Les effectifs sont constitués sur le seul critère présence/absence.

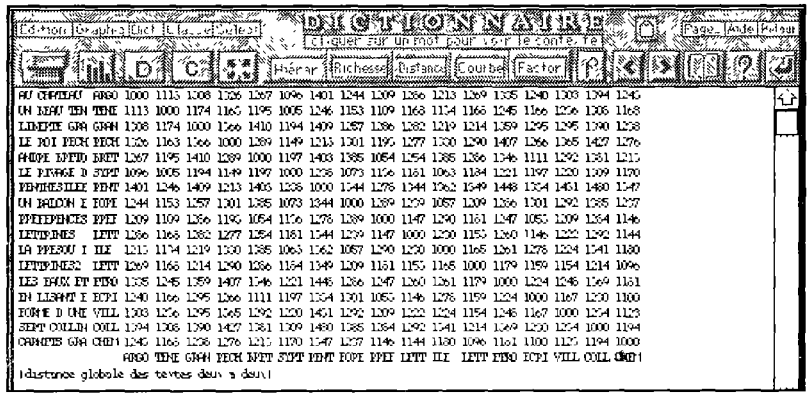


FIGURE 7 · Le tableau des distances lexicales dans l'œuvre de Gracq.

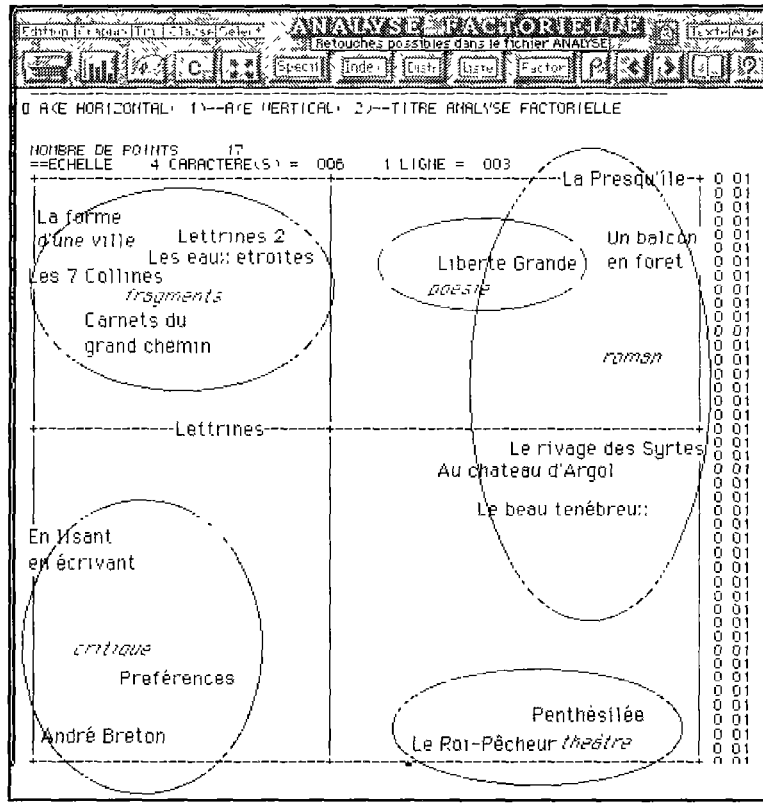


FIGURE 8 : Analyse factorielle des distances lexicales

c - Le traitement d'un nouveau texte. À la différence de beaucoup de logiciels où les fonctions sont absolument séparées des données, les unes et les autres sont mêlées dans une pile Hypercard, surtout lorsqu'il s'agit du type « standalone ». La pile originale doit donc être recopiée dès que l'on veut traiter des données nouvelles. Sous son nouveau nom elle garde ses programmes – qu'il faut conserver – et ses anciennes

données – qu'il faut évacuer. L'élimination de celles-ci se fait en sollicitant le bouton *Vider* (voir ci-dessous). Le résultat est une pile vierge, qui peut servir de modèle pour toutes les applications ultérieures. L'incorporation d'un texte est assurée par le bouton *Importer* (voir ci-dessous), qui montre la première page du fichier des données et s'enquiert des options souhaitables en affichant le dialogue de la figure 9.

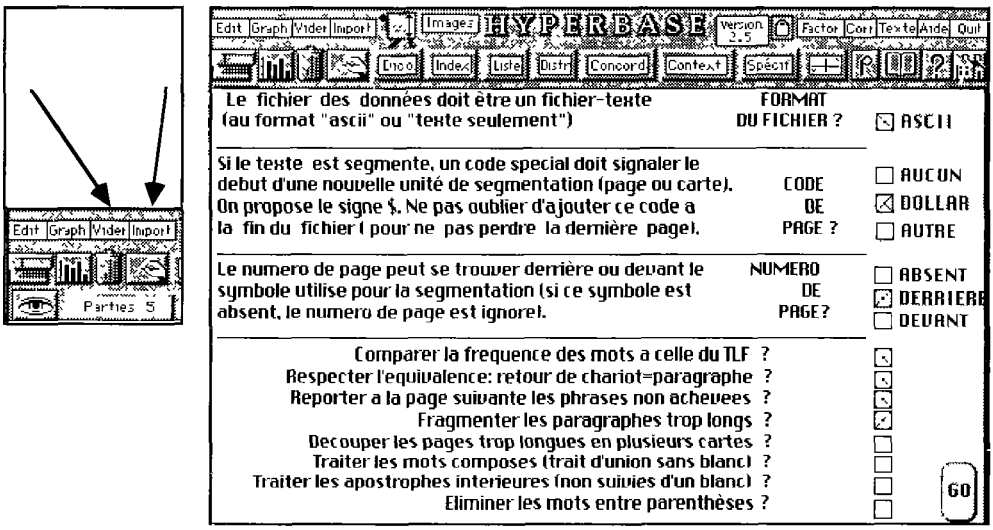


FIGURE 9 Entrée des données.

Les données textuelles doivent se trouver dans un fichier ASCII (ou « texte seulement »). On a pris en compte la plupart des alphabets européens. Aucun formatage particulier n'est obligatoire, le logiciel se chargeant de la pagination et de la partition, si elles sont absentes du fichier. En ce cas, les cartes (ou pages) ont environ 200 mots et l'ensemble du texte est découpé en dix parties de longueur voisine.

Mais il vaut mieux suivre le découpage naturel des données, s'il existe. Deux conventions doivent alors être respectées :

– les parties doivent être précédées d'une ligne où l'on indiquera le titre (en 20 caractères maximum, sans virgules ni apostrophes) en utilisant devant et derrière le symbole composite &&& (sans blanc). Veiller à bien choisir le dernier mot du titre qui sert d'abréviation lorsque la place manque, par exemple dans les graphiques, et qui doit être unique et distinctif.

– les pages sont indiquées en ajoutant une ligne (au début) et en y portant le numéro, immédiatement précédé ou suivi d'un code spécial (par exemple le symbole $; mais on peut choisir un autre code, si le symbole $ apparaît dans le texte même). Exemple :

```
&&&La vie en rose&&&
$1
texte de la page 1
$2
texte de la page 2, etc.
&&&Le travail au noir&&&
$62
texte de la page 62
$63
texte de la page 63, etc.
```

Le traitement d'un texte nouveau s'opère en trois phases : la première libère l'espace requis et transfère le texte dans la base, à raison d'une carte par page. C'est l'occasion de transcoder le texte afin d'uniformiser la présentation et en particulier de standardiser la ponctuation. En même temps, est constitué un fichier formaté qui va servir d'entrée à la phase 2. Celle-ci suit la première étape de façon automatique ou manuelle. On choisira ce dernier mode, afin de libérer le maximum de mémoire, si le fichier à traiter est de grande taille et si la machine est de faible puissance. Dans ce cas la pile est abandonnée à l'issue de la phase 1 et un double clic sur l'icône *Tripart1.6* (ci-dessous) mettra en œuvre le programme d'indexation, en lui réservant toutes les ressources de l'ordinateur.

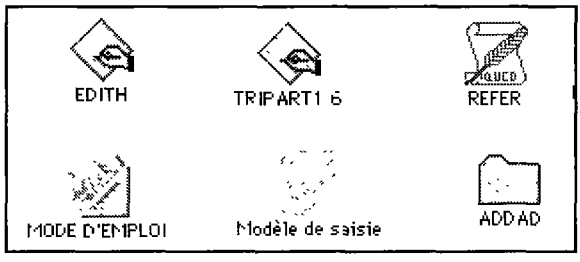


FIGURE 10 Le programme d'indexation.

Quand le tri est achevé et que les traitements subsidiaires ont pris fin, il suffit de revenir dans la pile par un double clic pour que les multiples résultats obtenus dans la phase 2 soient communiqués à la pile et définitivement enregistrés dans la phase 3. Les fichiers intermédiaires peuvent alors être détruits, la pile disposant de toutes les informations dont elle a besoin, et particulièrement du dictionnaire inverse.

Prévoir 15 minutes pour le dépouillement d'un texte de 500 pages (avec un Mac équipé d'un microprocesseur *8030*) et 20 minutes pour le tri. Après ce temps de préparation (qui comporte un transcodage, un découpage en cartes, un tri des formes, un dictionnaire des fréquences et divers tests statistiques), la pile est exploitable. Si l'on dispose d'un microprocesseur *8040* ou *PowerPC*, le temps de préparation est fortement raccourci.

d - Spécifications techniques. On ne s'appesantira pas sur les spécifications techniques. Il suffit de préciser qu'*Hyperbase* comprend un programme de préparation

(écrit en *Pascal*), un éditeur de texte (écrit en langage *C*), un programme d'analyse factorielle (écrit en *Fortran* et emprunté à *ADDAD*) et un programme d'exploitation (écrit en *Hypertalk* et complété par de nombreuses commandes externes). La configuration requise est peu exigeante et se contente d'une mémoire vive de 2 000 Ko pour son usage propre (il faut ajouter la mémoire requise par le système et celle que réclament épisodiquement les applications externes auxquelles *Hyperbase* fait appel, traitement de texte ou analyse factorielle[9]). Le disque dur est indispensable et l'écran couleur recommandé. *Hyperbase* fonctionne indifféremment sur système 6 ou 7 et sur toute la gamme *Apple*, du *Mac Plus* au *PowerMac*. Précisons que la dernière version, totalement refondue (2.5), s'est affranchie de l'environnement Hypercard. Étant du type « standalone », l'application est devenue parfaitement autonome et ne dépend plus de l'installation de l'utilisateur. Mais elle échappe aussi à toute intervention intempestive dans le code même des programmes. Les scripts sont hors d'atteinte et ni la fenêtre de commande, ni la barre de menus ne livrent accès aux intrus.

Au bout de trois ans de commercialisation, les clients sont en majorité des chercheurs du secteur public, littéraires, linguistes, historiens et sociologues mais aussi des entreprises privées, spécialisées dans la veille technologique ou les systèmes experts. À l'étranger le logiciel est plus connu au Japon, aux États-Unis, au Canada et au Brésil. Le succès est venu là où on ne l'attendait pas ; dans les instituts de sociologie ou de sondage. *Infométrie*, l'agence *Harris* et la *Sofres* se servent d'*Hyperbase*, comme le prouve l'étude du langage des principaux candidats publiée par la *Sofres*, à la veille du scrutin présidentiel de 1995.

3. La version CD-ROM d'*Hyperbase*

À partir des mêmes données relatives à l'œuvre intégrale de Julien Gracq, nous avons réalisé un premier CD-ROM, en orientant différemment le traitement. Ce CD-ROM a été présenté à Julien Gracq lui-même et communiqué à plusieurs chercheurs spécialisés dont certains œuvraient à l'édition de Gracq dans la Pléiade. C'est précisément parce que le second tome de cette édition de référence n'était pas encore paru qu'on a préféré différer la commercialisation. Il valait mieux attendre un an ou deux pour profiter du texte corrigé et pouvoir disposer de la pagination nouvelle. Au reste, Julien Gracq qui avait d'abord donné son accord, a jugé prudent d'attendre un peu afin de protéger les droits de son éditeur tant que la jurisprudence n'était pas fermement établie, en matière de *copyright*, relativement au nouveau support du CD-ROM. Actuellement des pourparlers sont en cours avec plusieurs maisons d'édition pour la diffusion du produit.

Mais entre-temps une proposition nous avait été faite qui ne soulevait pas de problèmes de cette sorte. S'agissant de Rabelais, les héritiers ne pouvaient guère avoir de prétentions non plus que les éditeurs, puisque le texte était puisé à la source. L'entreprise a démarré en juin 1994, à l'initiative de M. L. Demonet et du laboratoire *EQUIL XVI* (de l'Université de Clermont-Ferrand), qui avaient établi le texte, assuré son

9 Le seul moment où l'on a avantage à disposer d'une mémoire abondante et d'une machine rapide est celui de l'indexation, qui transforme un texte ASCII en base de données. Ce traitement exige précautions et patience mais, comme il n'a lieu qu'une fois pour chaque corpus, l'effort consenti se justifie sans peine.

transfert sur support informatique, choisi les documents et rédigé les commentaires. Restait à réunir ces matériaux sur l'étroite surface d'un CD-ROM et à en faire une base de données. En réalité, si léger que soit ce support, sa contenance est très large et dépasse de très loin le volume de la base (grosse de 120 Mo), et, en y accumulant 400 documents iconographiques, des visites guidées, une concordance complète, un index, un dictionnaire et un manuel, c'est à peine si l'on utilise la moitié (soit 314 Mo) de la surface disponible, comme le montre l'image de son contenu (figure 11). Il est vrai que la moitié restante a été consacrée au standard PC, où tous ces éléments sont accessibles, à l'exception de la base elle-même, qui est trop étroitement liée à la *Toolbox Apple* pour être portable sur un autre équipement.

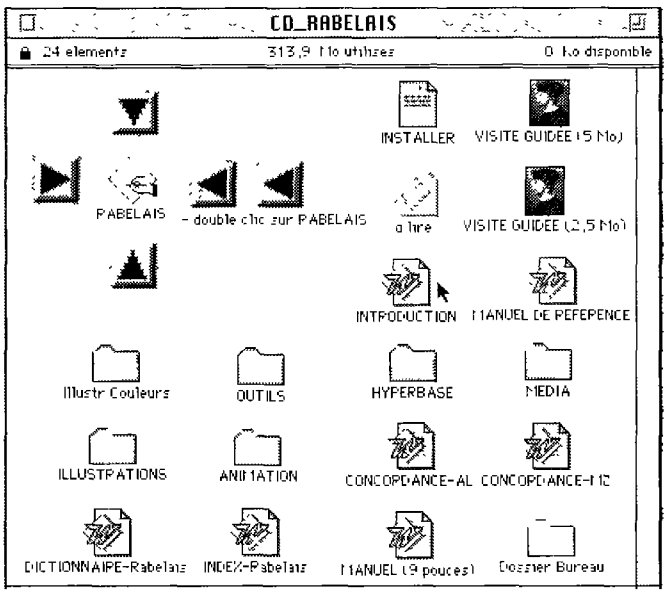


FIGURE 11 . Le contenu du CD-ROM Rabelais

Il faut souligner le caractère collectif de la réalisation, qui a bénéficié de collaborations multiples mettant en cause des organismes publics et des entreprises privées. Outre le laboratoire EQUIL XVI, à qui revient l'initiative et la conduite du projet, et le centre niçois (UPR 63861, INaLF, CNRS) qui en a assuré la réalisation technique, l'opération a reçu le concours de la Bibliothèque municipale de Lyon, de la Bibliothèque nationale de France, du Centre national du Livre et de la société APPLE-France, à quoi s'est ajoutée en dernier ressort l'industrie de l'éditeur[10]. On trouvera ci-dessous la liste des contributeurs, telle qu'elle apparaît lors de la mise en route (figure 12).

Les particularités de cette version tiennent aux spécifications du support, qui est réputé lent, tant pour l'accès que pour le débit. On a donc renversé l'ordre des priorités, en privilégiant le paramètre temps, quitte à perdre de l'espace et à redoubler l'in-

10 Éditions *Les Temps qui courent*, 118-130 bd J Jaurès, 75019 Paris

formation pour la rendre accessible à l'endroit où on la réclame, en évitant les voyages inutiles. Au demeurant le corpus est d'une taille modeste, même s'il contient dix textes qui enveloppent non seulement l'œuvre de Rabelais, de *Gargantua* au *Cinquième Livre*, mais aussi celle des devanciers et des imitateurs, soit cinq textes pararabelaisiens.

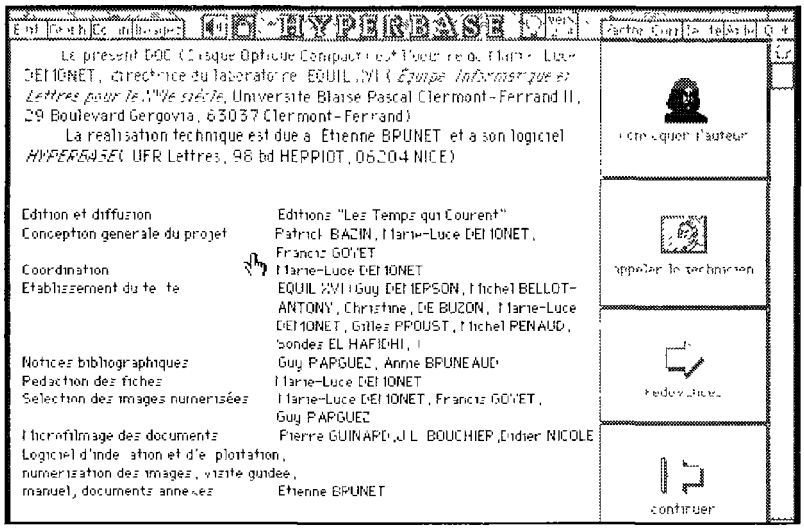


FIGURE 12 : Les contributions.

Mais la base a pris de l'embonpoint, parce qu'elle voulait être agile, si paradoxal que cela puisse paraître. On peut vérifier la vitesse du traitement en proposant la bonne ville de Lyon à la fonction *Contexte*. Les 19 passages où le corpus de Rabelais en fait mention sont restitués en 4 secondes (figure 13). Il en faut 13 pour quérir un mot plus fréquent, qui est familier à Rabelais sans être inconnu à Lyon, le mot *vin* (200 occurrences, 95 000 caractères). Pour une concordance, quelle que soit la fréquence, il suffit d'une ou deux secondes.

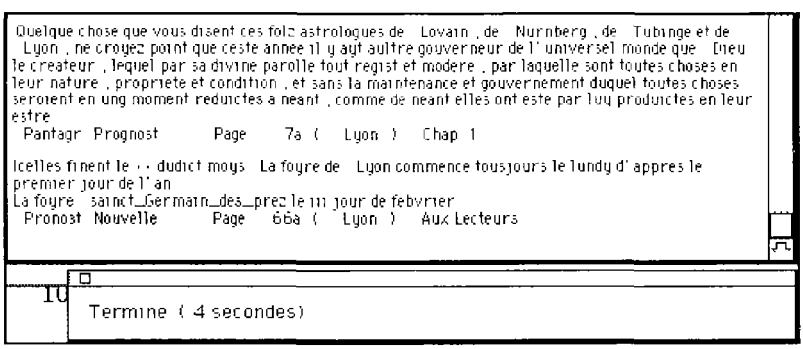


FIGURE 13 : La fonction *Contexte* dans le CD-ROM Rabelais
(Extrait des 19 emplois du mot Lyon).

L'originalité du produit ne tient pas seulement aux méthodes techniques, que l'utilisateur a le droit d'ignorer, mais aux fonctions disponibles dont certaines sont nouvelles. Outre les outils documentaires et statistiques habituels, le CD-ROM Rabelais offre une comparaison sous forme synoptique de plusieurs éditions du même texte, du moins là où les variantes sont les plus intéressantes, c'est-à-dire dans le cas du *Pantagruel* et du *Quart Livre*. Un clic sur n'importe quel mot de l'une des versions renvoie au mot correspondant de l'autre version. Voir figure 14.

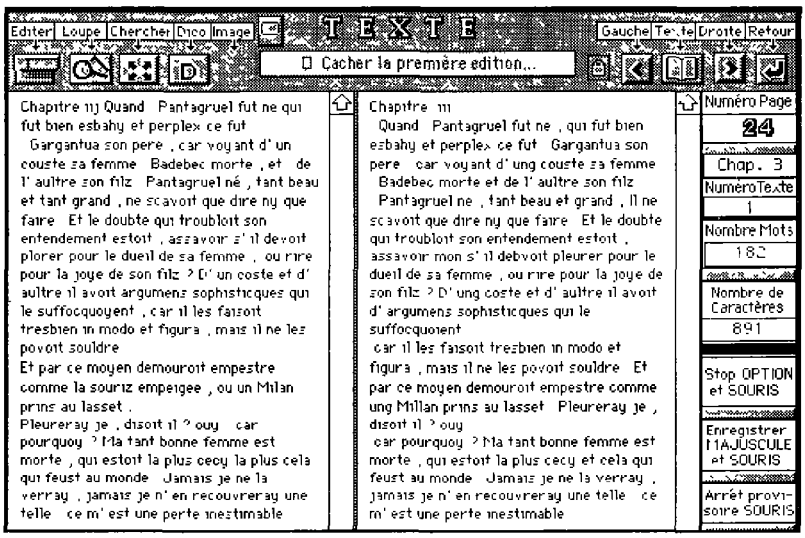


FIGURE 14 Comparaison de deux éditions du *Pantagruel*

En outre certains mots du texte de Rabelais ont été mis en relation avec les dictionnaires ou glossaires de l'époque. Ils apparaissent en caractères gras et réagissent au clic de la souris, montrant la définition du *Thresor* de Jean Nicot ou le commentaire de la *Brève déclaration*. L'italique désigne par ailleurs d'autres mots ou expressions qui bénéficient d'une explication et d'une illustration. Et de la même façon le clic sur le passage en italique fait apparaître successivement l'une et l'autre. Quatre cents documents iconographiques qui ont un rapport avec le texte du *Gargantua* ont été fournis par la Bibliothèque municipale de Lyon. Ils datent tous de l'époque de Rabelais et éclairent certains aspects méconnus du texte. Ces illustrations accompagnent les pages du *Gargantua* dans leur défilement mais elles constituent aussi une base de données autonome qu'on peut consulter librement. L'exemple de la figure 15 est emprunté à un très bel ouvrage de botanique qui avait cours au temps de Rabelais et que présente la figure 16.

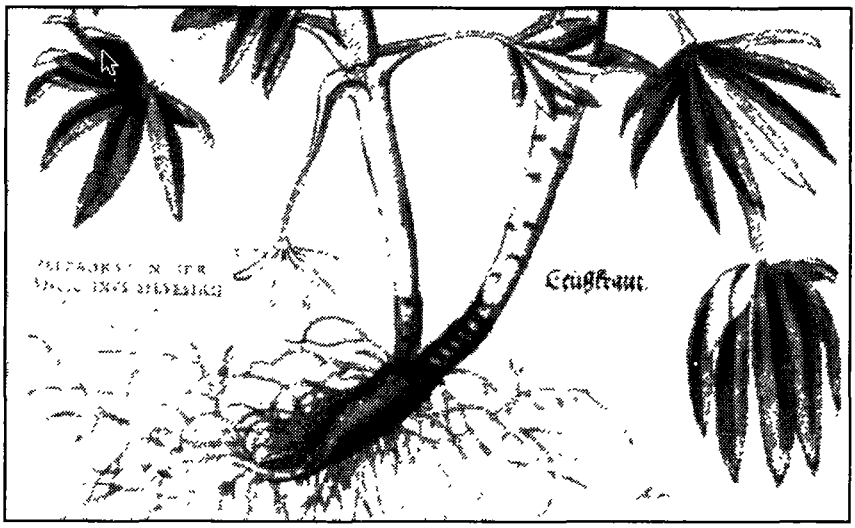


FIGURE 15 : Un document iconographique (extrait).

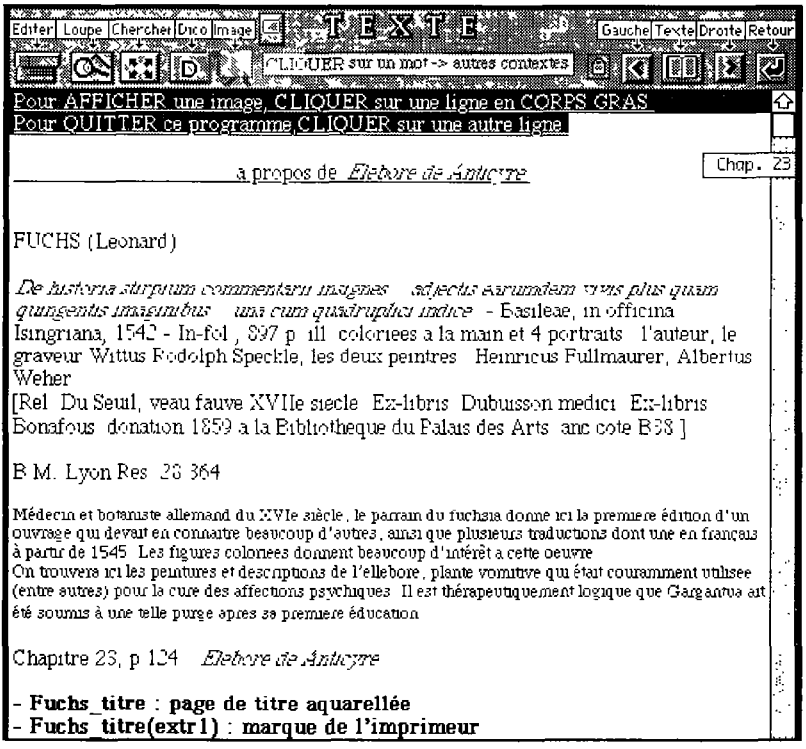


FIGURE 16 : Le commentaire de l'image.

Les facilités multimédia du CD-ROM ont été mises à profit, non seulement pour le traitement de la couleur (pourquoi s'en priver sur écran puisque le noir et blanc n'est pas plus économique), mais aussi pour l'incrustation de séquences animées et sonores qui expliquent et illustrent le fonctionnement de la base. Quand un bouton fait mystère, l'utilisateur a le moyen d'en exiger l'explication, brève ou détaillée. Un écran lui est d'abord montré avec un commentaire approprié et si cela ne suffit pas une séquence *Quicktime* lui est proposée, qui décompose les phases de l'opération, selon l'invitation reproduite dans la figure 17.

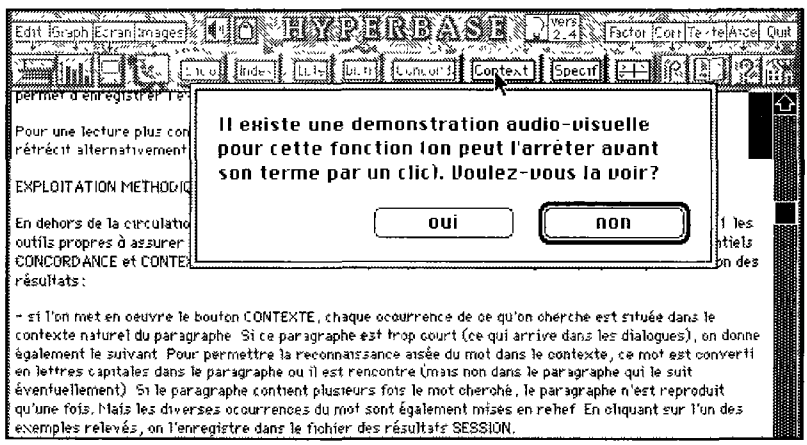


FIGURE 17 : Les aides multimédia (la fonction *Contexte* expliquée).

4. La version Internet d'*Hyperbase*

La version *Internet* d'*Hyperbase* pallie certaines des lacunes du CD-ROM. Même si le *CD-ROM Rabelais* est bi-standard, il n'offre pas les mêmes fonctionnalités sous *Windows* que sur matériel *Apple*. Et si les fichiers dérivés de la base (dictionnaires, index et concordances) ou associés à son emploi (documents iconographiques et leurs commentaires) sont accessibles sur *PC*, l'usage interactif de la base elle-même est réservé aux possesseurs de machines *Apple*. Quand au contraire on interroge une base sur le réseau, les résultats sont strictement les mêmes quel que soit le poste d'interrogation, qui peut être aussi bien une station *Unix*, un compatible *Dos* ou *Windows* et bien entendu une machine *Apple*. Pour atteindre ce haut degré de compatibilité, il faut passer par un langage précis et contraignant qui impose un codage particulier des messages à transmettre, dont chaque élément est doté de balises univoques. Ce langage *HTML*, instanciation de la norme plus générale *SGML*, gouverne les communications asymétriques entre le serveur et le client. Le client envoie des ordres brefs : clic sur un bouton, sur l'item d'un menu déroulant ou sur une zone soulignée (cela s'appelle une ancre ou un lien) et dans certains cas écrit un mot ou une expression dans une zone précise du dialogue. Sur un ordre exprès (bouton *submit* ou *OK*) ce message est transmis au serveur sous forme de chaîne de caractères, avec une ponctuation spécifique qui permet d'isoler et d'interpréter les paramètres. La réponse du serveur peut être immédiate s'il s'agit de transmettre un fichier déjà existant, qui peut être une page d'information, une image ou un fichier composite. Elle peut être légèrement différée si un traitement préalable est nécessaire, comme c'est le cas avec notre base Rabelais. Une

concordance, une recherche de contexte, un graphique, une liste de mots ou une analyse factorielle sont nécessairement des travaux sur mesure qu'il faut exécuter sur commande, au lieu que les documents iconographiques, les spécificités et certains résultats relatifs à la structure lexicale sont catalogués dans des fichiers tout prêts à l'emploi et immédiatement transmissibles. Dans tous les cas, à l'aller comme au retour, les données transmises doivent respecter les conventions du langage et en particulier coder les caractères accentués selon une norme commune à tous les standards.

Il serait oiseux d'entrer plus profondément dans les détails techniques. Peu importe quelles transformations ont été apportées au logiciel d'interrogation. Les plus importantes sont dues à une plus grande lourdeur des échanges quand s'interpose la distance. En traitement local, le dialogue peut être aussi vif ou laconique qu'on le souhaite. On peut répondre par oui ou par non et progresser rapidement dans la suite des répliques. À distance les répliques se transforment en tirades et le rythme des échanges se ralentit. Il faut quelques secondes pour obtenir une réponse et, dans l'attente du résultat, l'utilisateur est plongé dans un trou noir où il n'a d'autre alternative que d'attendre ou de suspendre. Afin de réduire ce silence, nous avons volontairement limité à 60 secondes le temps d'un échange, ce qui interdit les demandes impudentes ou imprudentes dont le résultat se ferait attendre trop longtemps (par exemple la concordance de tous les mots qui finissent par la lettre *s*). De toutes façons on a fixé à 1 000 le nombre maximum de lignes dans une concordance et à 100 000 caractères la taille de tout fichier transmis. Ces bornes raisonnables ont été établies pour éviter qu'un client gêne les autres ou se gêne lui-même par maladresse. On trouvera ci-dessous (figure 18) l'écran qui accueille l'utilisateur quand il prend contact avec la base à l'adresse :

(http ://ancilla.unice.fr/rabelais.html) ou (http ://134.59.31.3/rabelais.html)

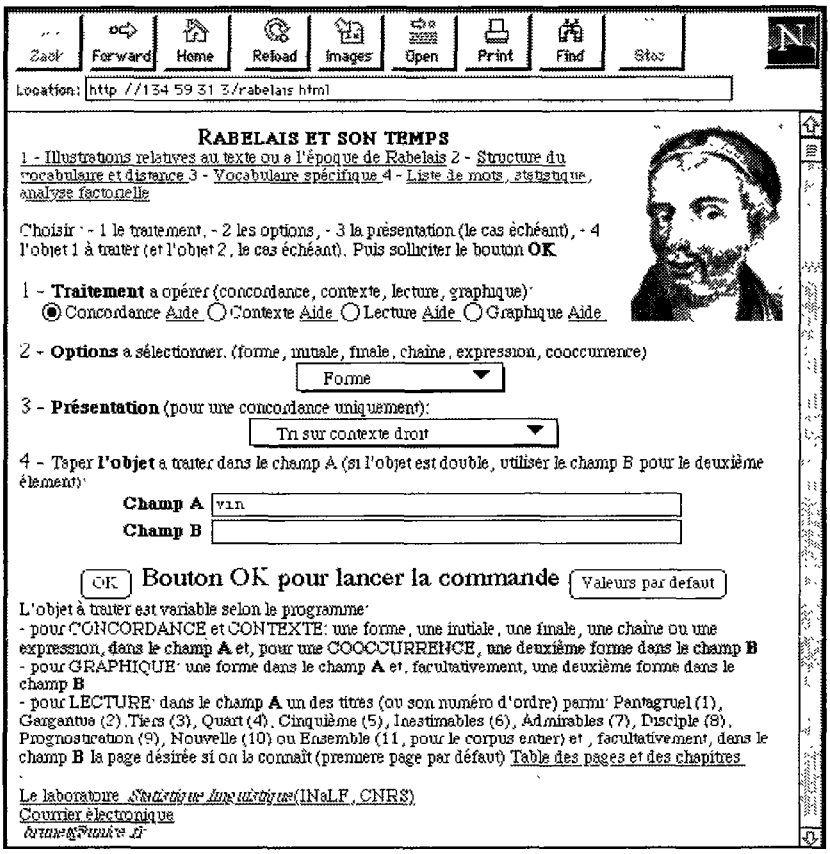


FIGURE 18 : La page d'accueil de la base *Rabelais* sur *Web*

Dans cet exemple, les paramètres sont réglés pour obtenir la concordance complète du mot *vin* dans le corpus avec une présentation ordonnée sur l'environnement à droite. Il ne faut guère que deux secondes d'exécution pour la collecte des 200 contextes de ce mot, à quoi s'ajoute le temps du transcodage et de la transmission, variable selon le type des liaisons.

```
CI 153a|                c' est à dire , en vin verité . Les deux parties estoient
PA 108a|      d' un grand hanat plein de vin vermeil , disant . Compere tout
DI  33f|    le monde . ? Le tiers est de vin vermeil qui passe en bonte tous les
CI 146a|signification evidente .  que le vin vos est en mespris , et par vos
CI  62d| . ¶Le cordieu vous aurez vostre vin à ceste heure :  je le vous promets
CI 188b|      estoit bouteille Pleine de vin a un aureille  ¶De vin , je dis ,
CI 180a|. C' est , dist frere  Jean , du vin à une aureille . Puis le vestit d'
CI 184a|   , et naturel flascon plein de vin  Phalerne ' lequel elle fist tout


Paramètres:
Concordance Forme vin( tridroit)
Temps d'exécution : 2 seconde(s)
Nombre d'appels de cette base: 912
```

FIGURE 19 : Extrait de la concordance triée du mot *vin*.

Outre la fonction *Concordance*, la page d'accueil présente les fonctions *Contexte*, *Lecture* et *Graphique* qui sont en tous points semblables à celles du CD-ROM et qui s'appliquent pareillement non seulement aux formes isolées mais aussi à une expression et à tout paradigme fondé sur une initiale, une finale ou une chaîne quelconque.

La visualisation des documents iconographiques est aussi simple et presque aussi rapide que dans la version CD-ROM. Un index (figure 20) est proposé qui en donne la liste en rattachant chaque illustration à son commentaire et à la page du *Gargantua* qu'elle complète. Ainsi le document précédemment reproduit dans la figure 15 – il s'agit de l'ellébore noir – est accessible à la ligne 230 de l'index, et de la même façon la page 124 du *Gargantua* à laquelle est liée cette illustration (figure 21) et la description bibliographique de l'ouvrage de botanique qui traite de l'ellébore et qui fait l'objet de la figure 16.

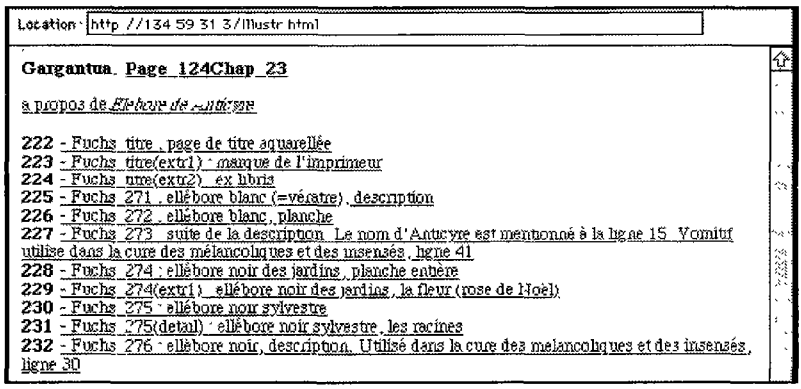


FIGURE 20 : L'index des illustrations.

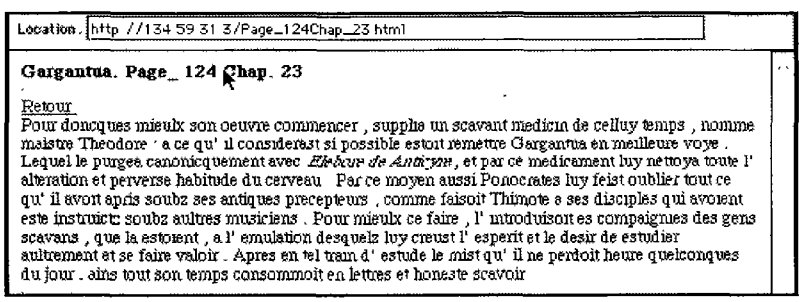


FIGURE 21 · La page 124 du *Gargantua* (chapitre 23).

Quant aux traitements statistiques, ceux qui ont trait au vocabulaire spécifique ou à la structure lexicale ont été préétablis et leur résultat peut être immédiatement transmis en sollicitant les ancres du haut de l'écran. Par contre, les traitements libres, soumis à la discrétion de l'utilisateur, font l'objet d'un écran particulier dont les options sont assez complexes. Elles permettent de constituer des tableaux à deux dimensions, avec les mots en ligne et les textes en colonne, et de les soumettre à l'analyse factorielle, avec différents types de pondération. La fonction *Graphique* a ici des possibilités étendues et s'applique aussi bien aux colonnes (profil d'un texte) qu'aux

lignes (distribution d'un mot). Elle s'applique également aux totaux marginaux comme aux individus. On prendra garde toutefois à ne pas allonger démesurément les listes, car la limite des 60 secondes veille là aussi à prévenir les excès.

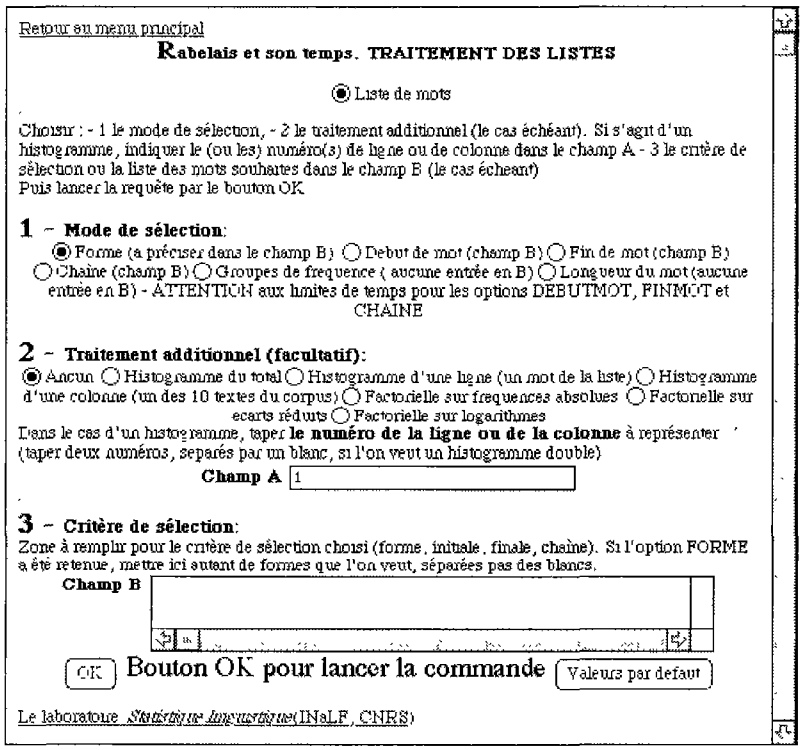


FIGURE 22 · Les traitements statistiques.

Conclusion

Il n'est pas certain que ces quatre variations d'un même thème aient épuisé les virtualités. Tout d'abord mille autres approches sont possibles que d'autres logiciels ont mises en œuvre avec succès. Mais aussi d'autres modes de diffusion peuvent être envisagés, qui appartiennent déjà aux opérations de routine. Rien n'est plus commun de nos jours que le téléchargement des fichiers et des logiciels. *Hyperbase* dans sa version à tout faire pourrait comme beaucoup d'autres être proposé en « shareware ». Mais, même sans donnée aucune, son poids reste dissuasif, avec deux millions d'octets à transmettre en format compressé. Et nos deux serveurs, qui ne disposent que d'un débit modeste de 65 kilobits, pourraient s'étouffer dans ce rude transfert. Une autre solution serait d'offrir un travail à façon, délocalisé, les données venant de l'extérieur sous forme de fichier *ASCII* à traiter, et les résultats retournant à l'envoyeur après traitement. Mais de tels services exigent des serveurs puissants et serviables et cela dépasse provisoirement les moyens du serviteur dévoué qui signe ces lignes.

Réseaux sémantiques et dictionnaires bilingues électroniques

Thierry FONTENELLE

Université de Liège, Belgique

1. Introduction

Dans cet article, je me propose d'aborder le problème de la réutilisation de ressources lexicales bilingues disponibles sur support informatique dans la perspective de la création et de l'amélioration de la composante lexicale de systèmes de traitement automatique du langage naturel. Le problème de la réutilisation des ressources lexicales existantes est en effet un problème épineux, tout particulièrement pour les ressources en langue française. Depuis le début des années 1980, les dictionnaires électroniques attirent l'attention des chercheurs en linguistique computationnelle. Ces derniers voient en effet dans les dictionnaires commerciaux électroniques une source inespérée de données lexicales d'ordre morphologique, syntaxique et sémantique cruciales pour des applications aussi diverses que la traduction automatique, la recherche documentaire, les interfaces de bases de données en langage naturel ou encore l'enseignement assisté par ordinateur. Depuis une dizaine d'années, les congrès et autres ateliers se sont multipliés et ont rassemblé des chercheurs du monde entier avides de partager leurs expériences, leurs succès, leurs frustrations aussi, dans l'extraction et la formalisation des informations lexicales contenues dans ces dictionnaires électroniques (voir entre autres, Boguraev & Briscoe, 1989 ; Zernik, 1991 ; Byrd, 1989 ; Atkins & Zampolli, 1994 ; Zampolli *et al.*, 1994). Force est de constater, cependant, que la grande majorité des efforts fournis dans le domaine concernent directement l'anglais. Les raisons en sont multiples. Il est intéressant de les analyser brièvement.

Il est incontestable que la tradition britannique des dictionnaires d'apprenants a joué un rôle considérable dans ce contexte. La richesse des dictionnaires monolingues pour apprenants étrangers comme le *Longman Dictionary of Contemporary English* (LDOCE, Procter, 1978), le *Collins Cobuild English Language Dictionary* (Sinclair, 1987), l'*Oxford Advanced Learner's Dictionary of English* (OALD, Cowie, 1989) ou, tout récemment, le *Cambridge International Dictionary of English* (CIDE, Procter, 1995) ne pouvait laisser indifférents tous ceux qui cherchent à automatiser la création,

la gestion et la mise à jour des lexiques informatisés utilisés dans des applications toujours plus sophistiquées. Les concepteurs sont en effet face à un cruel dilemme : soit ils font appel à des équipes de lexicographes spécialisés pour élaborer les lexiques dont ils ont besoin, ce qui, eu égard à la taille de ces lexiques pour des applications opérationnelles, s'avère extrêmement coûteux, soit ils essaient d'automatiser, ou du moins de semi-automatiser, le processus d'acquisition lexicale en examinant les possibilités d'extraction de ces informations de dictionnaires existants[1]. Or, il se fait que tous les dictionnaires mentionnés ci-dessus se distinguent des dictionnaires dits de langue par le recours systématique à un codage extrêmement élaboré des informations morphologiques, syntaxiques et parfois même sémantiques. Ainsi, des codes grammaticaux permettent de rendre compte de façon très détaillée de l'environnement syntaxique des items lexicaux, indiquant par exemple que tel verbe se construit avec une proposition infinitive en anglais, qu'il admet également les subordonnées introduites par *that* ou que tel nom, dans son sens indénombrable, gouverne la préposition *on*. Dans le cas de certains dictionnaires, on note même la présence de descriptions sémantiques s'apparentant aux règles de sélection. Ainsi, le LDOCE possède, dans sa version informatisée, une hiérarchie de codes sémantiques permettant de spécifier le trait sémantique inhérent d'un nom (par exemple, [+humain]) ou les contraintes sémantiques imposées par un verbe sur ses arguments (par exemple, le verbe X requiert un objet direct [+abstrait]).

La tradition lexicographique britannique n'a malheureusement pas eu d'émules dans le monde francophone. On ne connaît à ce jour aucun dictionnaire français disponible dans le commerce et possédant une richesse grammaticale comparable. Certains projets se sont penchés sur l'exploitation du dictionnaires Zyzomys de Hachette (Bouchard & Emirkanian, 1994 ; Bouchard *et al.*, 1991 ; Ide *et al.*, 1994), un dictionnaire de langue disponible sur CD-ROM. Il faut néanmoins constater que les quelques fragments de taxonomies ainsi que les descriptions morphosyntaxiques extraits de ce dictionnaire souffrent difficilement la comparaison avec les résultats obtenus sur les dictionnaires anglais. Si l'on considère que certains mettent en doute l'utilité même des recherches effectuées ces 15 dernières années sur les dictionnaires anglais (Veronis & Ide, 1994), on peut raisonnablement affirmer que les tentatives de réutiliser les dictionnaires monolingues français n'ont pas révolutionné la lexicographie computationnelle, principalement par manque de ressources adéquates.

Les considérations qui précèdent ont toutes trait aux tentatives de réutilisation de dictionnaires monolingues. Les dictionnaires bilingues, quant à eux, ont été encore plus négligés. Une des premières raisons est que les bandes magnétiques des dictionnaires bilingues comportant le français comme langue cible ou source n'ont été mises à la disposition que de quelques centres de recherche. Le *Robert & Collins dictionnaire anglais-français, français-anglais* (Atkins & Duval, 1978), par exemple,

1 Je passe sous silence le rôle non négligeable joué par les corpus de textes dans la problématique de l'acquisition lexicale. Il est évident que les dictionnaires informatisés ne peuvent fournir qu'une fraction de l'information lexicale nécessaire à un système élaboré du TALN. Les chercheurs se sont donc tout naturellement également tournés vers les ressources textuelles dont le traitement statistique permet de dégager des descriptions lexicales reflétant mieux l'usage, la fréquence d'emploi, la distribution, etc. À priori, la situation du français devrait être meilleure puisque la création de corpus ne dépend pas d'une quelconque tradition lexicographique anglo-saxonne, comme c'est le cas pour les dictionnaires. Dans les faits, cependant, on remarque que la plupart des logiciels permettant de traiter « en profondeur » un corpus (analyseurs syntaxiques) ne sont valables que pour les corpus anglais

n'est disponible, dans sa version non abrégée, que dans notre propre laboratoire à Liège ainsi qu'au Lexical Systems Group d'IBM à Yorktown Heights (Byrd, 1989 ; Byrd *et al.*, 1987 ; Boguraev, 1991). D'autres groupes ont eu accès à des versions de poche des dictionnaires bilingues publiés par Collins (le groupe de l'ISSCO par exemple, cf. Petitpierre *et al.*, 1994 et Robert, 1995 pour la version anglais-français ou le groupe de Pise pour l'anglais-italien – cf Picchi *et al.*, 1992), mais la petite taille de ces dictionnaires ne les rend pas susceptibles d'un traitement intéressant dans le cadre des recherches qui nous occupent ici, à savoir le dépistage et la formalisation des collocations.

Les dictionnaires informatisés bilingues ont été quelque peu négligés pour une autre raison, plus fondamentale. En règle générale, en effet, le format de ces dictionnaires est beaucoup moins structuré que celui des dictionnaires monolingues anglais dont il est question plus haut. Alors que ces derniers se rapprochent des bases de données lexicales dont les linguistes ont besoin pour leurs travaux (leur organisation logique permettant plus facilement l'identification de chaque information – partie du discours, définition, exemple, codes grammaticaux...), les dictionnaires bilingues comme le Robert & Collins ne sont le plus souvent que des dictionnaires lisibles par machine, en ce sens que les fichiers qui sont mis à la disposition des chercheurs ne sont rien d'autre que les bandes magnétiques ayant servi à la photocomposition de l'ouvrage. La structure logique des fichiers est dès lors limitée à la spécification de codes permettant le changement de typographie (passage à l'italique, etc.). La transformation de ces fichiers en véritables bases de données est loin d'être chose aisée et réclame une analyse approfondie de la microstructure des entrées. La nouvelle génération de dictionnaires bilingues, basée sur la norme SGML, devrait faciliter l'exploitation de ces ouvrages de référence en permettant l'identification immédiate des éléments composant les entrées lexicales (à ce sujet, on lira avec intérêt les résultats obtenus dans le cadre du projet COMPASS coordonné par le Centre de Recherche de Rank Xerox à Grenoble : le but de ce projet est d'exploiter les versions informatisées de dictionnaires bilingues pour construire la composante lexicale d'un système interactif d'aide à la compréhension de textes ; outre le dictionnaire Collins-Klett anglais-allemand dont il est question ailleurs dans ce volume – cf. Segond & Breidt – le projet COMPASS utilise la version informatisée du récent dictionnaire Oxford-Hachette (Corréard & Grundy, 1994), le tout premier dictionnaire bilingue construit à partir d'un corpus et balisé à l'aide du langage SGML – cf. Bauer *et al.*, 1995 ; Segond & Zaenen, 1994).

2. Construction d'une base de données lexico-sémantique à partir du dictionnaire Robert & Collins

Les travaux présentés dans cet article ont été réalisés dans le cadre d'un doctorat en linguistique anglaise (Fontenelle, 1995). L'idée de base était de réutiliser la version lisible par machine de la partie anglais-français du dictionnaire Robert & Collins afin d'exploiter l'information collocationnelle qu'il contient. Ces recherches ont été entreprises suite à la constatation que l'appareil métalinguistique du Robert & Collins s'avère être une véritable mine de renseignements sur les propriétés combinatoires des items lexicaux. Le traitement explicite des restrictions quant aux sujets et objets des verbes, par exemple, rend ce dictionnaire bilingue extrêmement utile comme source

d'information pour la construction semi-automatique d'une base de données bilingue de collocations.

L'appareil métalinguistique utilisé par les lexicographes du Robert & Collins couvre toute une série d'informations cruciales pour la désambiguïsation. Ces indications, données en italiques, vont de la spécification de la partie du discours (*n, adj, vt, vi...*) aux codes matières (*Bio, Comput, St Ex* [Stock Exchange], *Mus...*), en passant par les niveaux de langues (*fml, infml, fig, lit*) et les restrictions de sélections et autres collocations. C'est à cette dernière catégorie que je me suis plus particulièrement intéressé, dans le but de rendre ces contraintes collocationnelles accessibles à l'utilisateur humain ou à la machine. Le système appliqué par les lexicographes mérite d'être présenté et illustré afin de donner par la suite une idée plus claire des potentialités de la base de données.

- Les noms sujets typiques d'un verbe apparaissent entre crochets.
- Les noms typiquement utilisés comme compléments d'un autre nom apparaissent également entre crochets.
- Les objets directs d'un verbe et les noms typiquement modifiés par un adjectif apparaissent à côté de la partie du discours (pas de crochets ni de parenthèses).
- Les adjectifs, verbes et adverbes modifiés par un adverbe apparaissent sans crochets ou parenthèses.

Les exemples suivants illustrent cette pratique :

Objets typiques
arouse *vt* **(b)** (*cause*) *suspicion, curiosity etc* éveiller, susciter; *anger* exciter, provoquer; *contempt* susciter, provoquer.
dissipate *vt fog, clouds, fears, suspicions* dissiper; *hopes* anéantir; *energy, efforts* disperser, gaspiller; *fortune* dissiper, dilapider.
harbour 3 *vt* **(b)** *suspicions* entretenir, nourrir; *fear, hope* entretenir.

Sujets typiques
billow 2 *vi [sail]* se gonfler; *[cloth]* onduler; *[smoke]* s'élever en tourbillons *or* en volutes, tournoyer.
flap 3 *vi* **(a)** *[wings]* battre; *[shutters]* battre, claquer; *[sails]* claquer.
puff up 1 *vi [sails etc]* se gonfler; *[eye, face]* enfler.

Combinaisons N+N
flap 1 *n* **(a)** *[wings]* battement, coup; *[sails]* claquement...
beard *n [fish, oyster]* barbe; *[goat]* barbiche; *[grain]* barbe, arête...

Combinaisons Adj+N
baseless *adj accusation etc* sans fondement; *suspicion* sans fondement, injustifié.
well-founded *suspicion* bien fondé, légitime.
well-grounded *suspicion* bien fondé, légitime.

Les entrées ci-dessus appellent plusieurs commentaires. Tout d'abord, il est clair que l'accès à l'information dépend du public visé par le dictionnaire. Dans le cas présent, nous avons affaire à un dictionnaire qui permet à un utilisateur francophone de déterminer le sens exact, et par conséquent, la traduction d'un mot anglais en fonction

de son contexte. Ce contexte est décrit par le lexicographe à l'aide du vocabulaire métalinguistique en italiques dont la richesse et l'abondance contribuent à la qualité de l'ouvrage. Il s'agit donc, dans la perspective de l'utilisateur francophone, d'un dictionnaire de décodage. Le même utilisateur qui souhaiterait se servir de cette partie du dictionnaire pour encoder, c'est-à-dire générer, du texte en anglais se verrait confronté au problème crucial de l'accès aux collocations. Si l'on prend comme hypothèse de base que, comme le souligne Hausmann (1979), les collocations sont des combinaisons polaires comportant une base et un collocatif, la première étant responsable de la sélection du second dont le sens dépend de sa mise en contexte avec la base, on s'aperçoit aisément que, dans les entrées ci-dessus, la base est l'élément en italiques alors que l'entrée proprement dite constitue le collocatif. L'espace me manque ici pour reprendre tous les arguments en faveur de la création de dictionnaires combinatoires où les informations collocationnelles seraient classées sous la base, et non, comme c'est le cas ici, sous le collocatif (voir à ce sujet Hausmann, 1985, 1989 ; Cowie, 1986 ; Heid, 1994). Les informations présentées ici ne permettent pas de répondre facilement à des questions lexicographiquement aussi intéressantes que :

- Que peut-on faire à un 'soupçon' ? (≈ Quels verbes peuvent prendre le mot *suspicion* comme objet direct ?)
- Quelles peuvent être les caractéristiques d'un 'soupçon' ? (≈ Quels adjectifs peuvent qualifier le mot *suspicion* ?)
- Que peut faire une voile ? (≈ Quels verbes peuvent prendre le mot *sail* comme sujet ?)

Ces questions sont en fait celles auxquelles le lexicographe est confronté dans son travail quotidien. On notera également que ces problèmes sont au cœur de nombreuses recherches actuellement menées en linguistique informatique où l'acquisition de collocations par des méthodes statistiques dans des corpus de textes est un sujet brûlant (on lira avec intérêt les travaux de Smadja, 1991, 1993 ; Grefenstette, 1994a et b ; Church & Hanks, 1990 ; Church *et al.*, 1994 ; Zernik, 1991). Les applications en sont aussi bien la construction de nouvelles ressources lexicographiques que l'élaboration de lexiques pour la génération automatique du langage naturel (Smadja, 1993).

Dans le cas qui nous occupe, la recherche des collocations se fait non pas dans des corpus de textes, mais dans la version informatisée du dictionnaire Robert & Collins qui est lui-même considéré comme un corpus pré-digéré d'informations lexicales. Comme on l'a vu plus haut, les collocations pertinentes que nous recherchons sont présentes dans le dictionnaire, mais l'organisation même de celui-ci, par le classement alphabétique des collocatifs, ne permet pas à l'utilisateur de découvrir le lien étroit qui unit des entrées comme *dissipate, arouse, harbour, well-founded* ou *baseless*. La présence de l'item *suspicion* dans la micro-structure de chacune de ces entrées permet cependant à une machine de reconstruire l'environnement collocationnel d'une base donnée en extrayant les occurrences de cette base et les entrées sous lesquelles elle apparaît. Pour autant que le dictionnaire soit organisé en base de données, il est alors possible de repérer toutes les occurrences du mot *suspicion* en italiques dans la totalité du dictionnaire. Un utilisateur qui n'aurait que la version papier du dictionnaire à sa disposition devrait alors parcourir les 800 pages de la partie anglais-français, alors que cette requête ne prend que quelques fractions de seconde dans notre base de données. La liste résultant de cette interrogation pour le nom *suspicion* comprend les entrées suivantes :

arouse, avert, awake, baseless, confirm, dissipate, drive away, eliminate, entertain, harbour, just, quieten, remove, rest, rouse, suspicious (2x), *suspiciously* (2x), *suspiciousness* (2x), *unsuspicious* (2x), *verify, well-founded, well-grounded.*

Il ne m'est pas possible, pour des raisons d'espace, de décrire les problèmes liés à la transformation de la bande magnétique du Robert & Collins en une base de données permettant des accès multiples à l'information lexicale. Il importe néanmoins de noter que cette transformation fut loin d'être triviale et qu'elle a occupé Jacques Jansen, informaticien dans notre département, pendant de nombreux mois. Le modèle relationnel a été choisi pour diverses raisons, entre autres parce que d'autres dictionnaires, par exemple le LDOCE (Procter, 1978), étaient déjà disponibles dans le même format, ce qui facilite les fusions et applications se servant de ressources multiples. Le Robert & Collins est actuellement disponible sur PC et sur station UNIX (Sun Sparc Station) et des programmes d'application, écrit en C par Luc Alexandre et en Clipper par moi-même, permettent une interrogation souple de la base de données. Certaines critiques ont été émises à l'encontre du modèle relationnel (Boguraev *et al.*, 1992), mais, comme la base de données comporte toute une série de tables distinctes (pour les entrées, les parties du discours, les traductions, les indicateurs métalinguistiques, les exemples, etc.), le fait qu'une entrée comporte trois traductions alors qu'une autre n'en comporte qu'une n'est en fait pas un problème. Comme les différentes tables sont liées par des champs communs, la redondance est évitée et les programmes d'application sont chargés de reconstruire, à l'intention de l'utilisateur, une vue non fragmentaire des données lexicales. Des arguments plus détaillés en faveur de l'utilisation du modèle relationnel sont proposés par Michiels (1995). La structure proprement dite de la base de données du Robert & Collins est, quant à elle, décrite en détail dans Fontenelle (1995) et dans certains rapports internes du projet DECIDE dans lequel s'inscrit ce travail (Jansen & Fontenelle, 1994).

3. Enrichissement de la base de données

Dès le départ, il est apparu qu'il serait souhaitable de pouvoir structurer les informations collocationnelles de notre base de données. En effet, si l'utilité des renseignements repris plus haut sur les possibilités combinatoires du mot *suspicion* est incontestable, il n'en reste pas moins que les relations lexico-sémantiques unissant ce mot aux entrées sous lesquelles on le trouve sont de nature hétérogène. Ainsi, les verbes *harbour* (entretenir, nourrir) et *dissipate* (dissiper) peuvent tous deux prendre *suspicion* comme objet direct mais, du point de vue sémantique, renvoient à des sens diamétralement opposés. De la même façon, *baseless* (sans fondement) et *well-founded/ well-grounded* (légitime) sont, dans le contexte de *suspicion*, liés par une relation d'antonymie. Comme le principe même d'une base de données est de permettre à l'utilisateur non seulement d'extraire de l'information par des chemins d'accès divers, mais aussi de mettre à jour les données et de les enrichir, il fut décidé d'ajouter systématiquement, pour l'ensemble du vocabulaire métalinguistique (collocations et restrictions de sélection), un code permettant de formaliser la relation lexico-sémantique existant entre la base et le collocatif. Pour ce faire, je me suis basé sur la Théorie Sens⇔Texte d'Igor Mel'čuk et plus particulièrement sur le mécanisme des fonctions lexicales. Ce mécanisme est, je crois, suffisamment connu pour m'épargner une nouvelle introduction au domaine (on consultera les trois volumes du *Dictionnaire explicatif et combinatoire du français contemporain* – Mel'čuk *et al.*, 1984, 1988, 1992

– ainsi que les contributions d'Apresjan, Tutin, Sérasset et Alonso-Ramos & Mantha dans ce volume). Je me contenterai de rappeler ici que le terme 'fonction lexicale' désigne une relation de sens assez abstraite telle que l'expression linguistique de cette fonction dépend du lexème auquel elle vient se joindre. La notation traditionnelle, empruntée au modèle mathématique, est $f(x) = y$, où f est la fonction lexicale (FL), x est le mot-clé et y la valeur de la fonction. L'exemple typique est la fonction Magn, qui dénote l'intensification ou le degré élevé de quelque chose, comme dans Magn (*célibataire*) = *endurci* ou Magn (*menteur*) = *fieffé*. On a également des fonctions censées codifier les relations de verbes supports, c'est-à-dire les verbes pratiquement vides de sens qui se combinent avec le mot-clé, comme $Oper_1$ (*attention*) = *faire*, $Oper_1$ (*crime*) = *commettre* ou $Oper_1$ (*question*) = *poser*. On notera que les fonctions lexicales présentent l'ensemble de la cooccurrence lexicale restreinte d'un lexème donné. Dans le cas des collocations, le mot-clé est la base au sens où l'entend Hausmann (1979) et la valeur de la fonction est le collocatif. Le modèle de Mel'čuk, qui comprend une soixantaine de fonctions lexicales de base, permet en outre de formaliser des relations sémantiques d'ordre paradigmatique, et non pas seulement syntagmatique (par exemple, des relations de synonymie (FL=Syn), d'antonymie (FL=Anti), de dérivation morphologique (A_0(*loi*)=*légal*), etc.).

La majeure partie de la tâche à laquelle je me suis attelé pendant environ deux ans a été de déterminer la nature de la relation lexico-sémantique unissant les 70 000 indicateurs métalinguistiques en italiques et l'entrée sous laquelle on trouve ces indicateurs. Pour ce faire, j'ai écrit un programme d'application me permettant de créer une nouvelle table pour accueillir cette relation sémantique. Le processus d'encodage, guidé par une série de menus déroulants destinés à réduire les risques d'incohérence, a permis de créer, pour chaque item lexical apparaissant dans le vocabulaire métalinguistique du dictionnaire, une sorte de réseau sémantique où la base est liée aux autres mots par une relation empruntée au modèle mel'čukien. Dans le cas du mot *suspicion* illustré plus haut, cela signifie que j'ai attribué, pour chacune des occurrences de ce mot dans une entrée du dictionnaire (*arouse, awake, baseless, harbour*, etc.), une étiquette correspondant à une fonction lexicale. De plus, comme le dictionnaire est un ouvrage bilingue, j'ai également traduit l'item en italiques afin de créer une véritable base de données lexico-sémantique permettant d'accéder à l'information collocationnelle par le biais de l'anglais ou du français (la base – l'indicateur à l'origine en italiques, sa traduction française, ajoutée manuellement, le collocatif – l'entrée anglaise ou la traduction française donnée par le dictionnaire). Le réseau sémantique pour les 26 termes gravitant autour de *suspicion* se présente comme suit, en respectant la notation proposée par Mel'čuk, à savoir f(x) = y :

$causfunc_0$ (suspicion / soupçon) = arouse (éveiller) [S]
liqu (suspicion/soupçon) = avert (écarter) [S]
$causfunc_0$ (suspicion/soupçon) = awake (éveiller) [S]
antiver (suspicion/soupçon) = baseless (sans fondement) [S]
$real_1$ (suspicion/soupçon) = confirm (confirmer) [S]
liqu (suspicion/soupçon) = dissipate (dissiper) [S]
liqu (suspicion/soupçon) = drive away (chasser) [S]
liqu (suspicion/soupçon) = eliminate (éliminer) [S]
$oper_1$ (suspicion/soupçon) = entertain (nourrir) [S]
$oper_1$ (suspicion/soupçon) = harbour (entretenir) [S]
ver (suspicion/soupçon) = just (justifié) [S]
liqu (suspicion/soupçon) = quieten (calmer) [S]
liqu (suspicion/soupçon) = remove (dissiper) [S]

(suspicion/soupçon) = rest (fonder) [S]
$causfunc_0$ (suspicion/soupçon) = rouse (éveiller) [S]
a_1 (suspicion/soupçon) = suspicious (soupçonneux) [P]
a_2 (suspicion/soupçon) = suspicious (suspect) [P]
adv_1 (suspicion/soupçon) = suspiciously (avec méfiance) [P]
adv_2 (suspicion/soupçon) = suspiciously (d'une manière suspecte) [P]
s_0qual_1 (suspicion/soupçon) = suspiciousness (caractère soupçonneux) [P]
s_0qual_2 (suspicion/soupçon) = suspiciousness (caractère suspect) [P]
$antia_1$ (suspicion/soupçon) = unsuspicious (peu soupçonneux) [P]
$antia_2$ (suspicion/soupçon) = unsuspicious (qui n'a rien de suspect) [P]
$real_1$ (suspicion/soupçon) = verify (vérifier) [S]
ver (suspicion/soupçon) = well-founded (bien fondé) [S]
ver (suspicion/soupçon) = well-grounded (bien fondé) [S]

Sans entrer dans les détails, on remarque que ce réseau sémantique contient des informations collocationnelles cruciales pour l'encodage. Ainsi, la fonction **Liqu** (<'liquider', 'détruire'>) permet de répondre à la question : Quels verbes expriment la 'suppression' d'un soupçon ? (*avert, dissipate, drive away, eliminate, quieten, remove* pour l'anglais ; *écarter, dissiper, chasser, éliminer, calmer* pour le français). La fonction **$Oper_1$**, qui correspond à la notion de verbe 'support' pratiquement vide de sens, ou du moins sémantiquement appauvri, se retrouve dans les combinaisons *entertain_suspicion* ou *harbour_suspicion* (on *entretient* ou on *nourrit* des soupçons à l'égard de quelqu'un ≈ on soupçonne quelqu'un). La fonction **Ver**, qui permet de générer des adjectifs signifiant que le mot-clé est 'correct' ou 'tel qu'il doit être', permet de codifier le lien unissant *suspicion* et *just, well-founded* ou *well-grounded.* Inversement, **Ver** se combine avec la FL **Anti** pour former la fonction complexe **Anti-Ver** et ainsi étiqueter la relation unissant *suspicion* et *baseless.*

L'élément apparaissant entre crochets à la fin de chaque ligne correspond au codage de la nature typographique de l'élément en italiques. Ainsi, [S] signifie que le mot suspicion apparaît à ce que nous avons appelé le niveau de 'surface' (d'où le 's') dans les entrées *arouse, avert, awake,* c'est-à-dire sans parenthèses ni crochets. Comme je l'ai indiqué plus haut, cette information typographique est cruciale puisqu'elle permet de déterminer la nature du lien syntaxique entre l'indicateur métalinguistique et son entrée. Ici, le niveau de surface [S] correspond au statut d'objet direct pour un verbe transitif (something *arouses* somebody's suspicions – quelque chose *éveille* les soupçons de quelqu'un). Ce même niveau [S] code le lien entre un adjectif et le nom qu'il modifie (*just* suspicions – des soupçons *justifiés*). Dans le cas présent, nous n'avons pas de verbes intransitifs dont *suspicion* pourrait être le sujet. Si tel était le cas, l'indicateur apparaissant alors entre crochets, l'information typographique serait représentée par [C]. La formalisation de cette information dans la base de données permet de poser des questions comme, par exemple, « Quels sont les verbes transitifs qui peuvent avoir l'item X comme sujet/objet possible ? ».

On remarque en outre que, dans certains cas, l'information typographique indique que le terme *suspicion* apparaît au niveau [P], c'est-à-dire entre parenthèses dans la version imprimée du dictionnaire. Les parenthèses sont en fait utilisées pour ajouter des micro-définitions. On a par exemple l'entrée suivante :

unsuspicious *adj* *(feeling no suspicion)* peu soupçonneux, peu méfiant
(causing no suspicion) qui n'a rien de suspect, qui n'éveille aucun soupçon

Si l'on considère que le mot *suspicion* implique la présence de deux actants (la personne qui soupçonne et ce qui est soupçonné), on arrive au cas de figure suivant où A_1 et A_2 sont les fonctions lexicales utilisées par Mel'čuk pour générer l'adjectif correspondant au premier et au second actant respectivement :

$Anti A_1$ (suspicion) = unsuspicious (peu soupçonneux)
$Anti A_2$ (suspicion) = unsuspicious (qui n'a rien de suspect)

Il ne s'agit bien sûr plus ici d'information collocationnelle puisque la relation est de nature paradigmatique. Comme elle est présente dans le dictionnaire, cependant, il aurait été dommage de s'en passer et, dans le cadre d'un réseau sémantique et d'une application possible dans la recherche documentaire, ce genre de relation se doit d'être formalisé. On notera également qu'il a été possible de semi-automatiser l'attribution de ces fonctions lexicales grâce à une analyse des structures récurrentes dans les micro-définitions. Ainsi, les structures « causing + N » ou « feeling + N » sont utilisées comme des indices permettant de déterminer la nature de la relation lexico-sémantique en question. Je me suis inspiré ici des travaux d'Amsler (1980), Michiels & Noël (1984) ou Ahlswede & Evens (1988) pour identifier les formules définitoires récurrentes dans les micro-définitions du Robert & Collins.

4. Collocations, terminologie et fonctions lexicales

Le Robert & Collins n'est pas un dictionnaire permettant d'étudier les langues de spécialité. Il ne contient donc normalement que des informations sur la langue générale, même si le vocabulaire plus technique est loin d'en être absent. La base de données que nous avons construite permet cependant de transformer le dictionnaire en une sorte de thésaurus où l'encyclopédique se mêle au lexical. Comme le note Frawley (1988), les lexiques spécialisés ne fournissent généralement pas d'information sur l'environnement collocationnel des termes (même si cette situation est en train de changer ; cf. Cohen, 1986 ; Verlinde *et al.*, 1992). C'est pourtant de cette information que les traducteurs, rédacteurs et apprenants ont besoin. Pour formaliser le discours spécialisé utilisé pour parler d'un terme donné, les théories de Mel'čuk ne sont probablement pas les plus appropriées parce qu'elles ne permettent de coder que des relations standard de la langue générale et les langues de spécialité ont le plus souvent recours à des relations très spécifiques (cf. Blampain *et al.*, 1992 et Blampain, 1993 qui identifient un certain nombre de relations notionnelles spécifiques à un domaine donné et pour lesquelles le modèle de Mel'čuk n'offre aucun équivalent). On constate néanmoins que la base de données collocationnelle du Robert & Collins contient des informations précieuses sur la combinatoire de certains termes et, d'un point de vue phraséologique, permet de répondre à certaines questions fréquemment posées par les terminologues travaillant sur des corpus spécialisés. Si l'on considère que notre base de données, grâce aux nombreux index sur tous les types d'information, permet d'accéder aux données via n'importe quelle clé, séparément ou en combinaison, il est possible de formuler des questions combinant syntaxe et sémantique dont les réponses nécessiteraient une conversation approfondie avec un spécialiste du domaine, comme le souligne Frawley (1988).

Prenons l'exemple du mot anglais *sail* (fr. voile) qui, nous l'avons vu plus haut, peut se combiner avec des verbes tels que *flap, billow, puff up* et bien d'autres encore.

L'analyse des occurrences de *sail* dans les rubriques métalinguistiques des entrées du Robert & Collins nous permet d'obtenir une image assez détaillée du discours nautique utilisé pour parler des voiles en anglais et en français. La formalisation des collocations à l'aide des fonctions lexicales permet de poser les questions suivantes à notre base de données :

Q : Y a-t-il un terme désignant un ensemble régulier de voiles ?
A : Mult (sail) = set (Fr. jeu)

Q : Y a-t-il un verbe désignant le son typique des voiles ?
A : Son (sail) = flap (Fr. claquer)

Q : Y a-t-il un nom désignant le son typique des voiles ?
A : S_0Son (sail)= flap (Fr. claquement)

Q : Y a-t-il des verbes transitifs prenant le mot *sail* comme objet direct et signifiant que le but typique associé aux voiles est réalisé ?
A : $Real_1$ (sail) = get up, haul up, hoist, put up, shake up, spread (Fr. hisser, déployer)

Q : Y a-t-il des verbes transitifs prenant le mot *sail* comme objet direct et signifiant que le but typique associé aux voiles n'est plus réalisé ?
A : $AntiReal_1$ (sail) = haul down, (Fr affaler), lower (Fr. abaisser), strike (Fr. amener)

Q : Quelles sont les parties typiques d'une voile ?
A : Part (sail) = belly (Fr. creux), gore (Fr. pointe)

Q : La voile fait-elle partie d'une entité plus importante ?
A : Whole (sail) = barge (Fr. barge), yacht (Fr. yacht, voilier)

On notera que les deux dernières relations ne sont pas couvertes par le modèle mel'čukien parce qu'elles ne correspondent pas à des fonctions à proprement parler, mais plutôt à des relations de 1 → n. Dans la base de données, j'ai introduit les relations **Part** et **Whole** pour rendre compte de liens méronomiques présents dans le dictionnaire. Même si ces relations sont de nature sémantique, et non lexicale, elles n'en sont pas moins cruciales pour la recherche documentaire ou pour l'analyse automatique du langage[2].

2. L'analyse des constituants d'une phrase et la levée d'ambiguïtés potentielles peuvent être grandement facilitées si les entrées lexicales comportent une indication des relations « a_comme_partie » (correspondant à notre **Part**), « fait_partie_de » (**Whole** dans notre base de données) ou « a_comme_instrument_typique » (correspondant à la fonction lexicale $\mathbf{S_{instr}}$ chez Mel'čuk) Considérons les deux phrases suivantes
(a) *John painted the jeep with the new wheels.* (John a peint la jeep aux roues neuves)
(b) *John painted the jeep with the new brush* (John a peint la jeep avec le nouveau pinceau)
Il est clair que [the jeep with the new wheels] ne forme qu'un seul et même constituant alors qu'on a deux constituants distincts, [the jeep] et [with the new brush] (attaché au prédicat *paint*), dans (b) Ce type d'analyse ne peut reposer que sur l'indication explicite des liens suivants dans le lexique (Gener est également une fonction lexicale qui renvoie à l'hyperonyme du mot-cle)
S_{instr} (paint) = brush ..
Part (vehicle) = wheel...
Gener (jeep) = vehicle

Il faut admettre que l'enrichissement de la base de données à l'aide des fonctions lexicales a ses limites. Je n'ai pas essayé à tout prix d'imposer une fonction lexicale pour modéliser une relation trop spécifique. Plutôt que d'attribuer une FL qui ne correspondrait que de loin à la relation unissant une base et son collocatif, j'ai préféré m'abstenir dans un certain nombre de cas. J'ai également évité de multiplier les FL pour ne pas créer des relations qui ne seraient utilisées que deux ou trois fois dans une base de données contenant, rappelons-le, plus de 70 000 enregistrements. L'absence de fonction lexicale standard pour certaines combinaisons n'en empêche pas moins de compléter le réseau sémantique évoqué ci-dessus. Ainsi, on peut poursuivre le dialogue imaginaire relatif au nom *sail* et poser les questions suivantes à la base de données :

Q : Y a-t-il d'autres verbes transitifs pouvant prendre *sail* comme objet direct et indiquant ce que l'on peut faire à une voile ?
A : *bend* (Fr. enverguer), *furl* (Fr. ferler), *gore* (Fr. mettre une pointe à), *puff* (Fr. gonfler), *reef* (Fr. prendre un ris dans), *swell* (Fr. gonfler), *trim* (Fr. gréer).

Q : Y a-t-il des verbes prenant *sail* comme sujet et indiquant ce qu'une voile peut faire ?
A : Les voiles peuvent *billow* ou *billow out* (Fr. se gonfler), *fill out* (Fr. gonfler), *puff out* ou *puff up, swell* ou *swell out* (Fr. se gonfler), elles peuvent aussi *gibe* (passer d'un bord à l'autre du mât).

Q : Y a-t-il des combinaisons Adj+N ou N+N servant à décrire les voiles ? (par ex. quels adjectifs peuvent qualifier le nom *sail* ?)
A : Les voiles peuvent être *billowy* (Fr. gonflé par le vent), *full* (Fr. plein) ou *swelling* (Fr. gonflé). Le syntagme « the *billow* of a sail » (Fr. gonflement) est également attesté, à côté d'autres collocations N+N correspondant à des fonctions lexicales standard comme *flap of sails* ci-dessus (S_0 Son).

5. Conclusions

Comme on a pu le constater, l'appareil métalinguistique du Robert & Collins offre un point de départ intéressant pour la construction de réseaux sémantiques et les combinaisons reprises plus haut attestent de la richesse de ce dictionnaire. Ces combinaisons représentent ce que l'on serait tenté d'appeler « du matériau linguistique d'aide à la décision » puisqu'elles permettent de rafraîchir la mémoire parfois défaillante d'un traducteur ou d'activer une série de termes parmi lesquels l'utilisateur du dictionnaire électronique pourra opérer un choix selon le sens qu'il souhaite rendre. Les différentes clés d'index de la base de données permettent d'interroger celle-ci via plusieurs chemins d'accès, qu'il s'agisse de la base anglaise de la collocation (l'item en italiques), de sa traduction (ajoutée manuellement), du collocatif (l'entrée anglaise du dictionnaire), du collocatif français ou de la fonction lexicale. Comme les clés d'accès peuvent être combinées, il est possible de poser des questions sémantiquement très complexes (par exemple, quels sont les verbes supports de type inchoatif pouvant prendre comme objet direct le nom anglais *habit* ? ⇔ IncepOper_1 (habit) = *acquire, contract, develop, form, take to* – fr. *prendre, contracter).* On se rapproche dès lors du dictionnaire des collocations tel que l'envisage Hausmann (1979, 1985)

puisque l'utilisateur module à son gré des requêtes portant sur un réseau lexico-sémantique où le nœud est la base de la collocation et les flèches unissant ce nœud aux collocatifs et termes associés correspondent aux fonctions lexicales mel'čukiennes (enrichies d'autres relations non prévues par le modèle standard).

La souplesse des programmes d'application et l'ajout de l'information lexico-sémantique a radicalement modifié la conception du dictionnaire bilingue. Alors que l'utilisateur était au départ soumis aux contraintes inhérentes de l'ordre alphabétique, il peut à présent utiliser la base de données comme un thésaurus informatisé couplé à un dictionnaire combinatoire. Le linguiste dispose alors d'un outil souple lui permettant de tester, sur un vaste échantillon des vocabulaires anglais et français, des hypothèses relatives à la structure des lexiques de ces deux langues. Le traducteur et l'apprenant disposent quant à eux d'une ressource lexicale bilingue leur permettant de poser des questions qu'ils n'avaient peut-être pas encore imaginées et dont les réponses devraient faciliter l'encodage. Dans une perspective purement computationnelle, enfin, la base de données lexicale du Robert & Collins devrait pouvoir être utilisée pour désambiguïser les textes et donc, dans une certaine mesure, résoudre l'épineux problème de la détermination du sens exact d'un mot dans son contexte (et de sa traduction si l'on se place dans un contexte multilingue)[3].

Remerciements

Les recherches décrites dans cet article s'inscrivent dans le cadre d'une thèse de doctorat soutenue par l'auteur en octobre 1995. Elles s'inscrivent également dans le cadre des activités du projet de recherche DECIDE (*Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora* – MLAP 93/19) financé en partie par la Commission des Communautés européennes[4]. Elles n'auraient pas abouti sans la contribution essentielle de Jacques Jansen qui a analysé avec minutie les bandes magnétiques du dictionnaire Robert & Collins et conçu la base de données relationnelle. Qu'il en soit vivement remercié, ainsi que Luc Alexandre, qui a écrit certains des programmes d'application pour PC et pour stations UNIX. Il faut en outre souligner que ce travail n'aurait pu être réalisé sans la confiance des éditeurs (Harper Collins Publishers et les Dictionnaires Le Robert) qui ont accepté de mettre à notre disposition la version électronique de leurs dictionnaires à des fins de recherche fondamentale. Ils méritent également notre gratitude.

3 Pour rester dans le domaine nautique que nous avons évoqué plus haut, songeons simplement à la polysémie du verbe *amener* dans « Le matelot amène la voile » et « Le matelot amène ses enfants à la voile »

4 Outre le département d'anglais de l'Université de Liège (coordinateur du projet), le consortium du projet DECIDE comprend l'Institut fur maschinelle Sprachverarbeitung de l'Université de Stuttgart et le Centre de Recherche Rank Xerox de Grenoble B T S Atkins (anciennement éditeur en chef du Robert & Collins) participe également au projet en tant que consultante

Une maquette de base lexicale multilingue à pivot lexical (« acceptions multilingues ») : PARAX

Étienne BLANC

GETA-CLIPS, Institut IMAG (UJF & CNRS), Grenoble, France

• *Abstract* •

An Acception Based Multilingual Lexical Database has been designed under Hypercard. The aim of this mock-up, which contains about 200 words in five languages (English, French, German, Russian and Chinese), is to contribute to the validation of the concept of Multilingual Acceptions in the building of Multilingual Lexical Databases.

1. Introduction

On sait l'importance majeure qu'ont pris les dictionnaires dans les systèmes de Traduction Automatique ou les systèmes d'Aide à la Traduction, et le développement consécutif récent des travaux dans le domaine des ressources lexicales réutilisables et notamment des bases lexicales multilingues.

Le GETA qui est engagé dans un projet multilingue de Traduction Automatique Fondée sur le Dialogue, le projet LIDIA (Boitet & Blanchon, 1993 ; Blanchon, 1994a et 1994b), a naturellement été amené lui aussi à s'intéresser aux bases lexicales multilingues (Sérasset, 1994a et 1994b).

Notre réflexion a surtout porté sur un type particulier de structure de base multilingue, la structure de type pivot lexical ou pivot « par acceptions ». Malgré l'intérêt de ce type de structure, il n'a été encore que peu utilisé, une exception notable étant le système de TAO « ULTRA » de l'Université du Nouveau Mexique (Farwell, Guthrie & Wilks, 1993).

Nous présentons, dans cet article, une maquette de base lexicale de structure pivot « par acceptions » vue dans la perspective de l'utilisateur. Le but de cette maquette étant de permettre une expérimentation de ce type de structure du point de vue

linguistique, nous nous sommes efforcés de la concevoir simple d'emploi et facilement extensible à un grand nombre de langues (elle en comporte cinq actuellement : allemand, anglais, chinois, français et russe). Pour le moment, nous n'avons pas attaché trop d'importance ni à la structure linguistique, relativement rudimentaire, ni à la couverture lexicale (environ 500 acceptions actuellement).

Le choix d'Hypercard pour la réalisation est, entre autres, motivé par la grande souplesse de réalisation que permet ce logiciel. Nous pouvons ainsi faire évoluer notre base pour l'adapter à des représentations linguistiques variées, ce qui nous a en particulier permis d'intégrer des descriptions lexicales et sémantiques utilisées dans Mel'čuk (1988) dont l'extension multilingue nous paraît particulièrement intéressante.

2. La structure pivot « par acceptions »

2.1. Bases multilingues de type transfert et de type pivot

On peut schématiquement envisager deux types de structure pour une base multilingue traitant n langues :

- ou bien n dictionnaires monolingues et $n(n-1)$ liaisons unidirectionnelles (ou $n(n-1)/2$ bidirectionnelles) entre ces dictionnaires deux à deux ; c'est la structure multibilingue, encore appelée « de transfert » par analogie avec la technique de TAO de même nom ;

- ou bien n dictionnaires monolingues et un dictionnaire d'un langage pivot (ou interlingue), et $2n$ liaisons unidirectionnelles (ou n bidirectionnelles) entre ces dictionnaires et le dictionnaire pivot ; c'est la structure dite « pivot ».

La seconde approche semble la plus appropriée à des applications typiquement multilingues, ne serait-ce qu'en raison du nombre moindre de liaisons interdictionnaires à réaliser et c'est effectivement celle qui a été utilisée pour des projets comme ULTRA (Farwell, Guthrie & Wilks, 1993) et CICC (Yaoliang & Zhendong, 1991) qui font intervenir un grand nombre de langues (anglais, allemand, chinois, japonais et espagnol pour ULTRA, japonais, chinois, malais, indonésien et thai pour CICC).

Mais si la structure pivot semble la mieux adaptée à des applications multilingues, il reste qu'elle est de réalisation et de maintenance plus complexe que la structure multibilingue.

2.2. Interlangue ontologique et interlangue lexicale : les acceptions multilingues

Une des principales difficultés de l'approche pivot réside dans la réalisation de l'interlingue.

L'interlingue la plus séduisante, mais la plus complexe à réaliser, est l'interlingue ontologique (dictionnaire de concepts). C'est le choix adopté pour le plus grand projet mondial de base lexicale bilingue, le projet EDR (EDR, 1993), qui comporte un dictionnaire de 640 000 concepts. Notons cependant que, pour tenir compte

des difficultés d'une approche purement ontologique, EDR a adopté une structure mixte pivot-transfert : les dictionnaires monolingues anglais et japonais sont chacun mis en correspondance avec le dictionnaire de concepts (structure pivot), mais sont également mis en correspondance entre eux (structure de transfert).

À la complexité de l'interlingue ontologique s'oppose la simplicité de l'interlingue lexicale « par acceptions ». Dans cette approche, on ne cherche, en effet, pas à définir intrinsèquement les éléments de l'interlingue. Ces éléments, que nous nommerons ici « acceptions multilingues », ou par abréviation « acceptions », peuvent être définis de la façon suivante :

Un lemme donné d'une langue L a en général plusieurs sens, ou « acceptions monolingues ». Si l'on peut faire correspondre à une acception monolingue de ce lemme des acceptions monolingues sémantiquement identiques de lemmes dans les langues L', L'', etc. on dira que l'ensemble de ces acceptions monolingues représente l'acception multilingue associée à ces diverses acceptions monolingues. Le lexique « par acceptions » est constitué de l'ensemble de ces acceptions multilingues.

Une acception multilingue n'est donc pas définie intrinsèquement, comme l'est un concept dans une base ontologique. On peut cependant lui attribuer l'ensemble des propriétés sémantiques partagées par les acceptions monolingues qu'elle représente, on retrouve ainsi au moins en partie l'intérêt d'une base ontologique, tout en évitant la difficulté d'une description complète de concepts.

Cette première définition élémentaire de l'acception multilingue suppose que l'on puisse associer à une acception donnée dans une langue donnée des équivalents sémantiques dans toutes les langues de la base que l'on veut construire. Il est clair que la situation n'est pas toujours aussi simple, et que, si pour certaines acceptions monolingues (les termes techniques par exemple) la définition d'une acception multilingue ne présente pas de difficultés, pour d'autres une correspondance directe ne sera pas possible et on aura à considérer des correspondances plus complexes (analogues, par exemple, aux relations de sous-relation, super-relation et synonymie utilisées dans EDR). Et il n'est pas évident qu'une solution satisfaisante puisse toujours être trouvée.

C'est pourquoi il nous a paru important de tester expérimentalement la possibilité de construction d'une base de ce type sur un nombre suffisamment grand de langues appartenant à des familles différentes. C'est la raison d'être de la maquette que nous décrivons ici.

3. Structure générale de PARAX

Le dictionnaire pivot et les dictionnaires monolingues de PARAX sont chacun constitués d'une pile HYPERCARD (figure 1).

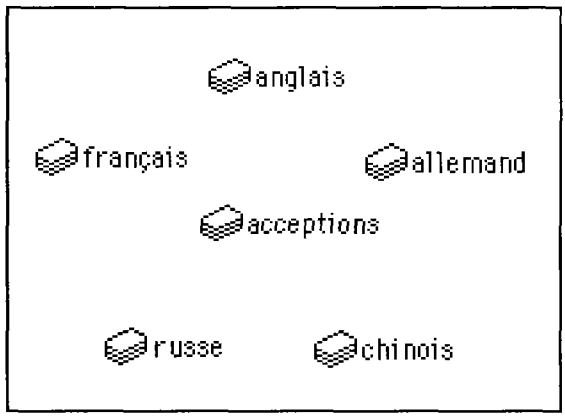


FIGURE 1 Les dictionnaires de PARAX.

Les dictionnaires monolingues sont actuellement au nombre de cinq (français, anglais, allemand, chinois et russe) et contiennent environ 200 mots chacun. Le dictionnaire pivot (dictionnaire d'acceptions) contient environ 500 acceptions.

Notons que l'ajout de nouvelles langues est chose aisée. Il suffit, pour créer un nouveau dictionnaire, de dupliquer une pile « patron » et de compléter un fichier descripteur. Les liens avec le dictionnaire d'acceptions sont créés automatiquement lors de l'entrée des termes.

Nous allons commencer la description de PARAX en y simulant une navigation, et pour cela, cliquons sur l'icône d'un dictionnaire monolingue, le dictionnaire français, par exemple.

4. Les dictionnaires monolingues de PARAX

On accède ainsi à la première page de ce dictionnaire (figure 2) qui présente la liste des lemmes indexés. En cliquant sur un élément de cette liste, le lemme *argent* par exemple, on accède à la page courante correspondante du dictionnaire (figure 3).

On voit que deux sens ont été retenus pour le lemme *argent*. On voit également que chacun de ces sens est décrit de façon autonome, morphosyntaxiquement dans la colonne de gauche, sémantiquement dans la colonne centrale. La séparation des descriptions morphosyntaxique et sémantique reflète le fait que cette dernière description, commune à tous les équivalents dans les autres langues, n'est qu'une copie provenant de la pile pivot.

D'autres informations peuvent apparaître sur demande : en cliquant sur le mot « FLEXICALES » qui suit la description morphosyntaxique d'un sens, on obtient la liste des fonctions lexicales de Mel'čuk. En cliquant sur le mot « EXEMPLES », on obtient des exemples d'utilisation du sens correspondant du lemme. Ces informations étant également accessibles à partir du dictionnaire d'acceptions, nous en parlerons plus en détail dans le paragraphe consacré à ce dictionnaire.

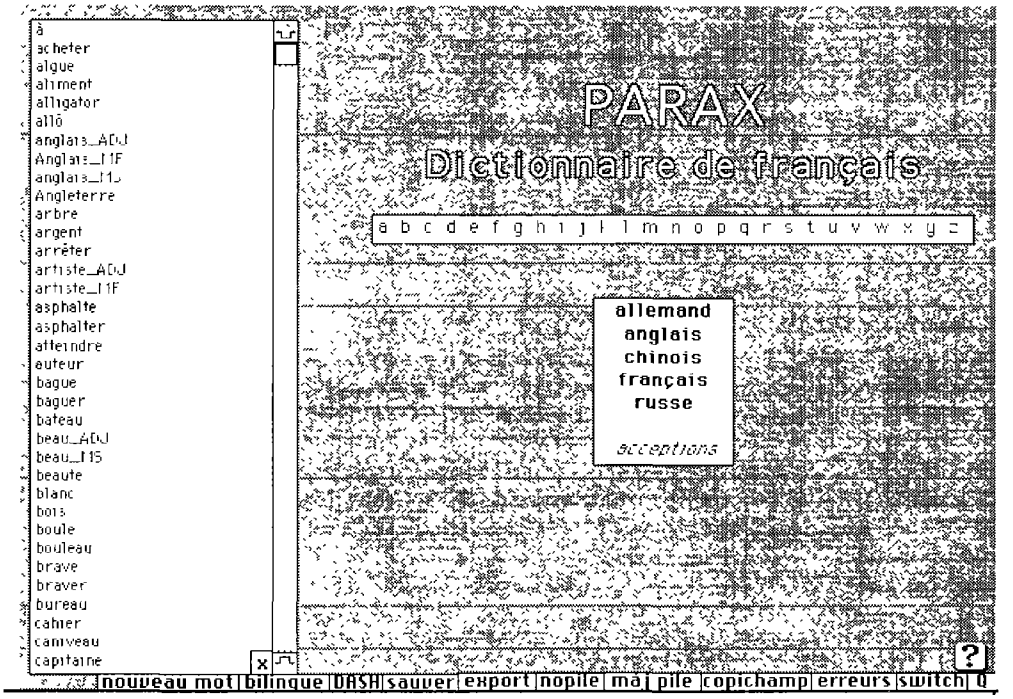


FIGURE 2 : La première page du dictionnaire français.

Dict français	argent	
argent #money	**#money**	
SENS 2 CAT nc GNR mas	{toute sorte de monnaie}	
[1IONOPIVOT/FLEXICALES/EXEMPLES]	SEMENT argent	
argent #silver_metal	**#silver_metal**	
SENS 1 CAT nc GNR mas	{metal noble de numéro atomique 47}	
[1IONOPIVOT/FLEXICALES/EXEMPLES]	SEMENT matiere	

FIGURE 3 · Le lemme *argent* dans le dictionnaire français.

Une carte donnée est consacrée à un lemme et non à une famille dérivationnelle ou unité lexicale. Cependant, les dérivations sont indiquées et le passage du lemme dérivé au lemme origine, ou inversement du lemme origine aux lemmes dérivés possibles, est immédiat (le passage a plus précisément lieu non pas entre lemmes origines et lemmes dérivés, mais entre acceptions de lemmes origines et acceptions de lemmes dérivés).

La figure 4a représente ainsi un fragment de la carte du lemme « beauté » décrivant l'une des trois acceptions retenues pour ce lemme : la beauté en tant que qualité esthétique. Le fait que cette acception (notée #beauté_esthet) dérive du lemme « beau » dans l'acception #beau_esthet, suivant le schéma adjectif → nom abstrait, est indiqué par la présence de la variable morphologique « DRVA » (dérivation de l'adjectif), par la valeur « nabst » (nom abstrait) de cette variable et par l'indication du lemme origine et de son acception dans la notation *£beau @beau_esthet*. Cette nota-

tion reflète le fait qu'en cliquant sur *@beau_esthet* on accède à la description de l'acception origine #beau_esthet sur la carte du lemme origine « beau » (figure 4b). Les mots composés et les locutions sont traités de façon analogue.

Dict. français	**beauté**
beauté #beaute_esthet	**#beaute_esthet**
SENS:1 CAT: vbimp DRVA: nabst CAT:	{qualité de ce qui fait éprouver une émotion
nc GNR fem. DRVA: nabst *£beau*	esthétique }
@beau_esthet	SEMENT: état.
[MONOPIVOT/FLEXICALES/EXEMPLES]	

FIGURE 4a : Dérivation adjectif → nom abstrait : en cliquant sur « @beau_esthet », on accède au lemme origine de la figure 4b

Dict. français	**beau**
beau #beau_esthet	**#beau_esthet**
SENS:1. CAT: adj. DRVAP: nabst. *£beauté*	{qui fait éprouver une émotion esthétique }
@beauté_esthet	SEMPROPR: qualif.
[MONOPIVOT/FLEXICALES/EXEMPLES]	

FIGURE 4b : Dérivation potentielle adjectif → nom abstrait.

Revenons maintenant à la carte relative au lemme « argent » de la figure 3 et cliquons sur le mot « MONOPIVOT » qui termine l'article relatif au sens 2, auquel correspond le code d'acception.multilingue #silver_metal. On accède alors à la carte du dictionnaire d'acceptions décrivant cette acception multilingue (figure 5).

5. Le dictionnaire d'acceptions

La figure 5 montre la carte du dictionnaire d'acceptions décrivant l'acception #silver_metal.

Source Français	FR	**#silver_metal**	Target
argent #silver_metal		**#silver_metal °AL °AN °CH °FR °RU**	
SENS 1 CAT nc GNR mas		{métal noble de numéro atomique 47}	
[MONOPIVOT/FLEXICALES/EXEMPLES]		SEMENT matière	

FIGURE 5 : Une carte du dictionnaire d'acceptions.

La colonne de gauche comporte la recopie de la description morphosyntaxique du représentant de cette acception dans le dictionnaire « source », c'est-à-dire le dictionnaire monolingue à partir duquel on vient d'accéder à la pile d'acceptions.

L'acception elle-même est décrite dans le champ central avec indication des langues dans lesquelles elle est représentée à l'aide des symboles AL (allemand), AN (anglais), CH (chinois), FR (français) et RU (russe). En cliquant sur l'un de ces symboles, on fait apparaître dans le champ de droite la description morphosyntaxique du représentant de l'acception dans la langue correspondante (figure 6).

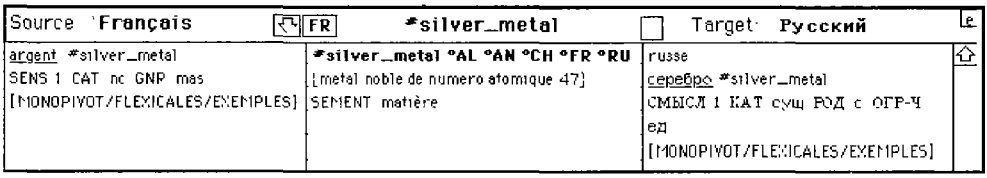


FIGURE 6 . La carte du dictionnaire pivot après choix du russe comme langue cible.

La figure 7 correspond au cas où la langue source est le chinois, les langues cibles sont multiples, et où l'on a fait apparaître les exemples dans les différentes langues, accessibles, comme nous l'avons dit, à partir des piles monolingues comme à partir de la pile d'acceptions. On peut de même faire apparaître les fonctions lexicales de Mel'čuk dans la ou les langues où elles ont été entrées.

On voit également, sur cette figure, que la définition sémantique est maintenant rédigée en chinois. Un bouton dans le bandeau supérieur de la carte permet, en effet, de choisir la langue de définition sémantique parmi toutes les langues de la base. Cette possibilité n'est pas uniquement une commodité d'utilisation pour des locuteurs de différentes langues, mais a un intérêt plus fondamental. En effet, si l'on est parvenu à donner d'une même acception des définitions correspondantes dans les diverses langues de la base et que ces définitions sont illustrées par des exemples identiques, ou au moins proches dans les diverses langues, on peut penser que l'on est parvenu à cerner correctement cette acception multilingue.

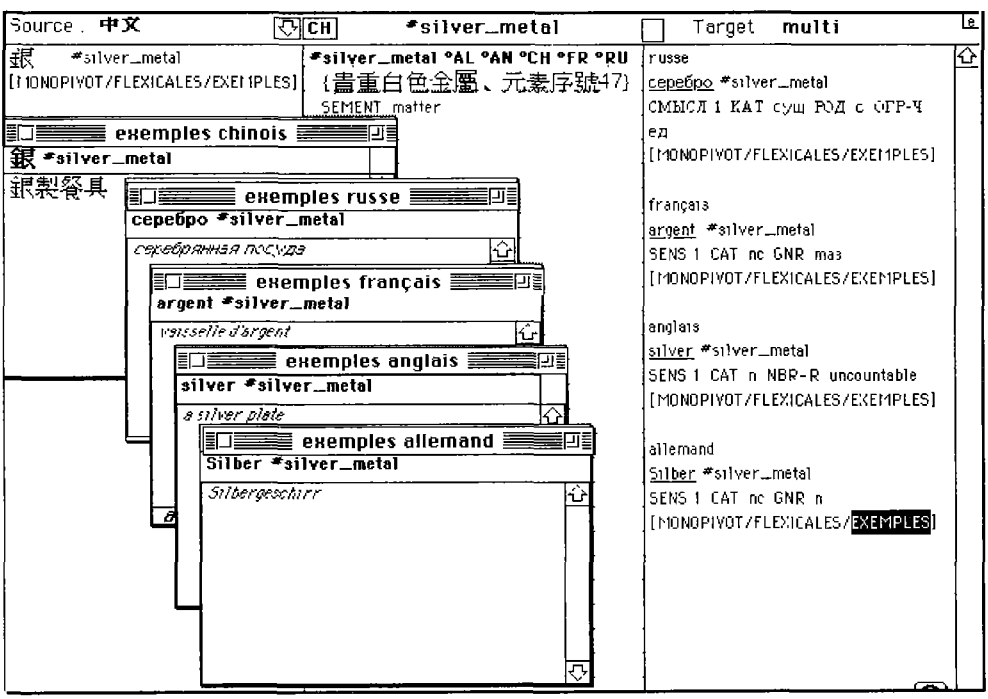


FIGURE 7 : Le chinois est choisi comme langue source et comme langue de description sémantique ; on a par ailleurs fait apparaître les exemples dans les différentes langues

Pour résumer cette présentation de la structure de la pile d'acceptions, on peut dire que cette pile stocke les informations sémantiques et propose, pour faciliter la consultation, des copies des informations lexicales dans les diverses langues (alors que les piles monolingues stockent les informations lexicales et proposent des copies des informations sémantiques).

5. Sur-acceptions et sous-acceptions

Il arrive bien entendu que l'on ait du mal à faire une correspondance suffisamment exacte entre des acceptions monolingues pour aboutir à une acception multilingue satisfaisante.

Un exemple élémentaire que nous avons traité correspond à la notion de sous-acception ou de sur-acception. Considérons, par exemple, le sens 2 du lemme français *capitaine* : « officier qui commande une compagnie d'infanterie, un escadron de cavalerie, une batterie d'artillerie » (figure 8a). Lorsqu'on passe au dictionnaire pivot (figure 8b), on voit que, si l'on choisit l'allemand comme langue cible, il existe un niveau de raffinement supérieur : avant de cliquer sur le symbole « AL », il faut choisir une des sous-acceptions de l'acception #captain_officer ; l'équivalent allemand viendra alors se placer dans le champ de droite en face de cette sous-acception.

Dict français	capitaine	
capitaine #captain_navy SENS 1 CAT nc GNP mas [MONOPIVOT/FLEXICALES/EXEMPLES]	**#captain_navy** {officier qui commande un navire de commerce ou de pêche}	
capitaine #captain_officer$ SENS 2 CAT nc GNP mas [MONOPIVOT/FLEXICALES/EXEMPLES]	**#captain_officer$** {officier qui commande une compagnie d'infanterie, un escadron de cavalerie, une batterie d'artillerie *artillerie/air/cavalerie/infanterie*}	
capitaine #captain_sportteam SENS 3 CAT nc GNP mas [MONOPIVOT/FLEXICALES/EXEMPLES]	**#captain_sportteam** {chef d'une équipe sportive}	

FIGURE 8a · Le lemme français *capitaine*.

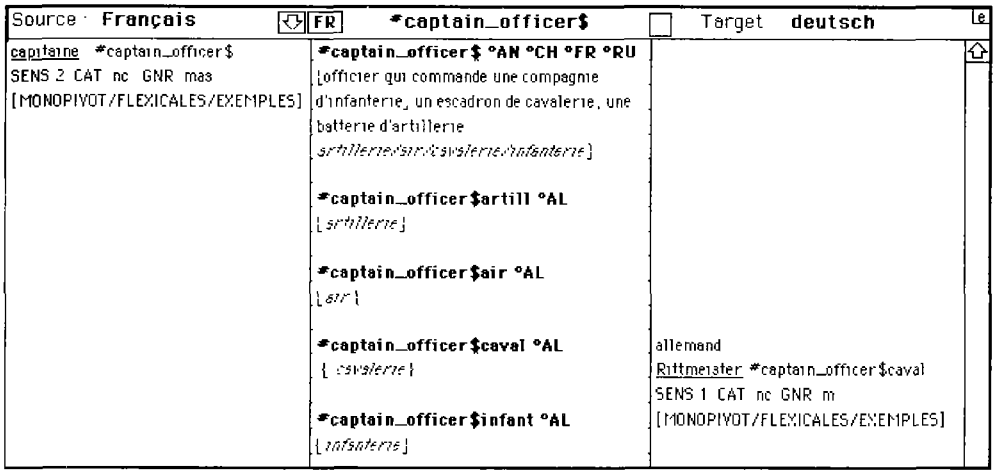


FIGURE 8b : Le passage d'une acception à une sous-acception.

6. Traitement des collocations

Considérons l'expression allemande *die Faust ballen*. Elle est décrite dans la carte *Faust* du dictionnaire monolingue allemand (figure 9a) avec une décomposition la rattachant à l'acception multilingue #formerboule du verbe *ballen*, (à laquelle, rappelons-le, on accède directement en cliquant sur *£ballen @formerboule* dans le champ de gauche).

L'expression anglaise correspondante *to clench the fist* est décrite dans la carte du dictionnaire monolingue anglais avec rattachement à l'acception multilingue #empoigner_serrer du verbe *to clench*.

Il est clair qu'il est préférable d'effectuer le passage de l'expression allemande à l'expression anglaise par l'intermédiaire d'une acception multilingue #serrer le poing (figure 9b), plutôt que de créer artificiellement une acception multilingue associant le verbe allemand *ballen* au verbe anglais *to clench*.

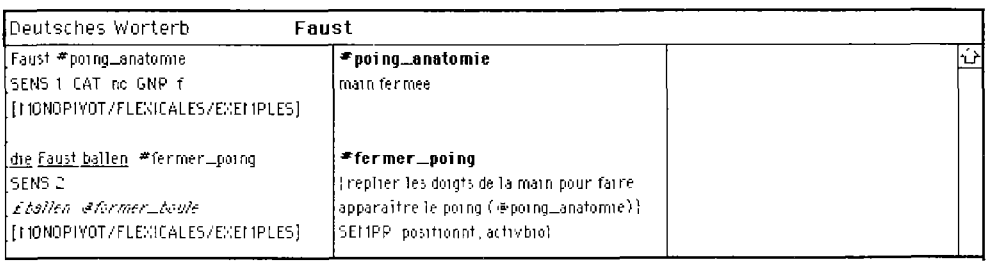


FIGURE 9a · L'expression idiomatique allemande *die Faust ballen*.

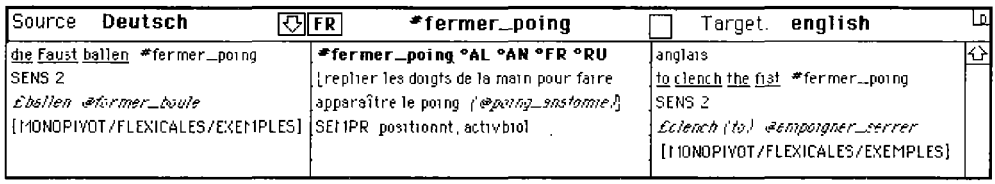


FIGURE 9b : Passage à l'expression idiomatique anglaise *to clench the fist.*

7. Exploitation de la base PARAX

Le principal objectif d'une base lexicale n'est en général pas de proposer une consultation directe de l'information qu'elle contient, mais de permettre d'utiliser cette information pour la réalisation d'outils lexicographiques, notamment de dictionnaires, dictionnaires destinés à une utilisation humaine ou à une application machine.

7.1. Outil d'aide à l'écriture de dictionnaires machine

Dans le cadre du projet de LIDIA de TAO multilingue fondée sur le dialogue, nous avons d'abord utilisé PARAX comme outil d'aide à l'écriture et à la maintenance des dictionnaires machine en l'associant à une interface hypertextuelle de documentation et de commande du générateur de systèmes de TAO ARIANE (l'interface DASH, pour Documentation Ariane sous Hypercard).

La figure 10 montre ainsi une copie d'écran comportant en arrière-plan une page du dictionnaire pivot de PARAX et, en premier plan, la page correspondante du fichier source d'un dictionnaire de transfert ARIANE tel qu'il apparaît dans l'interface DASH. Ce dictionnaire, qui a été envoyé à DASH par messagerie électronique depuis le serveur ARIANE, peut ensuite être modifié (ou même créé) en tenant compte des informations contenues dans PARAX, puis être réexpédié au serveur ARIANE pour compilation et exploitation.

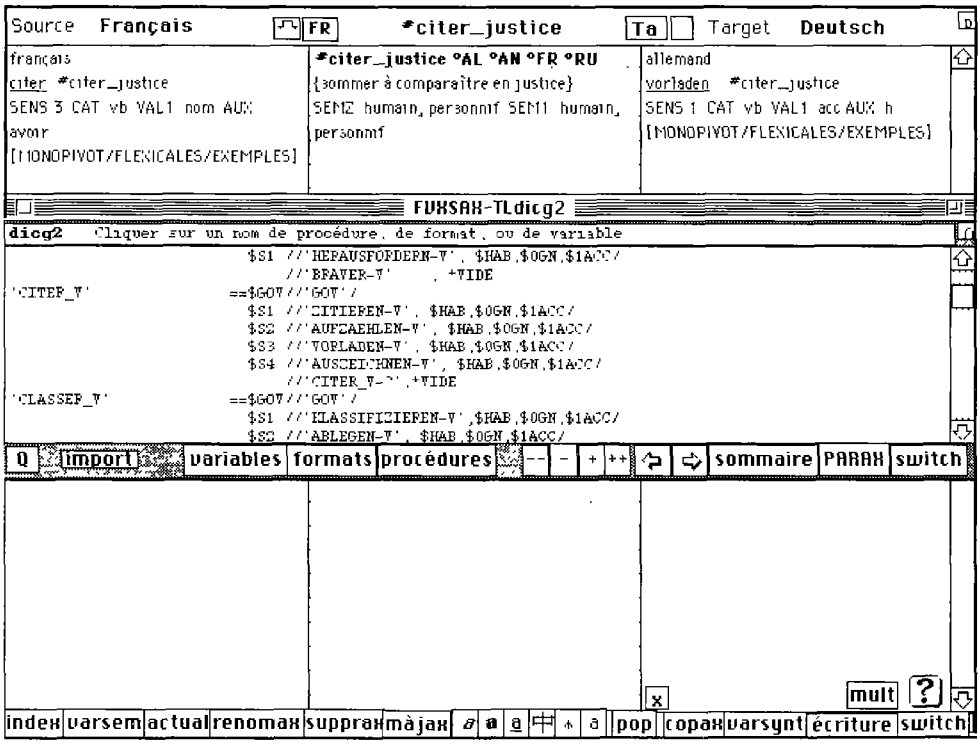


FIGURE 10 Utilisation de PARAX à la maintenance de dictionnaires machine.

7.2. Extraction de dictionnaires bilingues

On peut par ailleurs extraire de PARAX des dictionnaires bilingues sous forme de piles Hypercard avec comme langue source et comme langue cible un couple quelconque des langues représentées dans PARAX.

Ces dictionnaires s'obtiennent simplement à partir de la première carte du dictionnaire monolingue de la langue source (cf. figure 2), en cliquant sur le bouton « bilingue » et en indiquant la langue cible désirée.

Un tel dictionnaire bilingue Hypercard comporte une carte index (analogue à la carte index d'une pile monolingue) permettant d'accéder aux cartes courantes relatives aux lemmes de la langue source. Le format de ces dictionnaires permet de les obtenir sous forme de dictionnaires papier par la simple commande d'impression de la pile.

La figure 11 montre la carte d'un dictionnaire russe-allemand relative au mot allemand *Rittmeister* relevant du passage de sous-acception à sur-acception décrit plus haut. La langue choisie pour la description sémantique peut être arbitrairement sélectionnée parmi les langues représentées dans PARAX, c'est ainsi que l'on a ici un dictionnaire russe-allemand décrivant les sens en français.

russe - allemand	**капитан**	
капитан #captain_officer$_ СМЫСЛ 1 КАТ сущ РОД мр	**#captain_officer$** {officier qui commande une compagnie d'infanterie, un escadron de cavalerie, une batterie d'artillerie *artillerie/air/cavalerie/infanterie*} **#captain_officer$artill** *artillerie* **#captain_officer$air** *air* **#captain_officer$infant** *infanterie* **#captain_officer$caval** { *cavalerie* }	#captain_officer$ Batteriechef #captain_officer$artill SENS 1 CAT nc GNR m Hauptmann #captain_officer$air SENS 1 CAT nc GNR m Hauptmann #captain_officer$infant SENS 2 CAT nc GNR m Rittmeister #captain_officer$caval SENS 1 CAT nc GNR m
капитан #captain_navy СМЫСЛ 2 КАТ сущ РОД мр	**#captain_navy** {officier qui commande un navire de commerce ou de pêche}	Kapitan #captain_navy SENS 1 CAT nc GNR m
капитан #captain_sportteam СМЫСЛ 3 КАТ сущ РОД мр	**#captain_sportteam** {chef d'une equipe sportive}	Mannschaftskapitan #captain_sportteam SENS 2 CAT nc GNR m

FIGURE 11 · Une carte d'un dictionnaire bilingue russe-allemand extrait de PARAX

8. Enrichissement et maintenance de la base

L'enrichissement et la maintenance sont des problèmes délicats dans une base de structure pivot, même de dimensions réduites, du fait de l'interaction entre tous les dictionnaires monolingues par l'intermédiaire du dictionnaire pivot.

Nous n'allons pas décrire ici en détail la méthodologie utilisée pour l'enrichissement et la maintenance de PARAX mais seulement indiquer quelques-uns des outils que nous avons développés.

8.1. Entrée d'une nouvelle acception ; les cartes de description sémantique et le filtrage des acceptions

L'entrée d'une nouvelle acception s'effectue à partir d'un dictionnaire monolingue. En effet, une acception n'est pas définie d'une manière intrinsèque mais a besoin d'au moins un représentant dans une langue pour être définie.

Un problème préliminaire est de connaître avec suffisamment de précision l'état de la base : est-on sûr que cette acception n'a pas déjà été entrée, et si non, comment les éventuelles acceptions voisines ont-elles été définies ?

À cette fin, on peut faire apparaître, dans le champ supérieur droit des cartes courantes des dictionnaires monolingues, une liste des acceptions existantes filtrées suivant la catégorie sémantique. La figure 12 montre, sur une carte en cours de création, la liste des acceptions filtrées par la catégorie sémantique « animal ». Le choix de la

catégorie sémantique de filtrage s'effectue à l'aide de cartes de description sémantique comme celles de la figure 13 : en cliquant sur une catégorie sémantique figurant dans les cases du tableau, on obtient la liste des acceptions correspondant à cette catégorie sémantique et aux catégories hiérarchiquement inférieures.

Enfin, en cliquant sur le nom d'une acception, dans la liste filtrée des acceptions, on obtient sa description, dans le champ inférieur droit, ainsi que la liste des langues où elle est représentée.

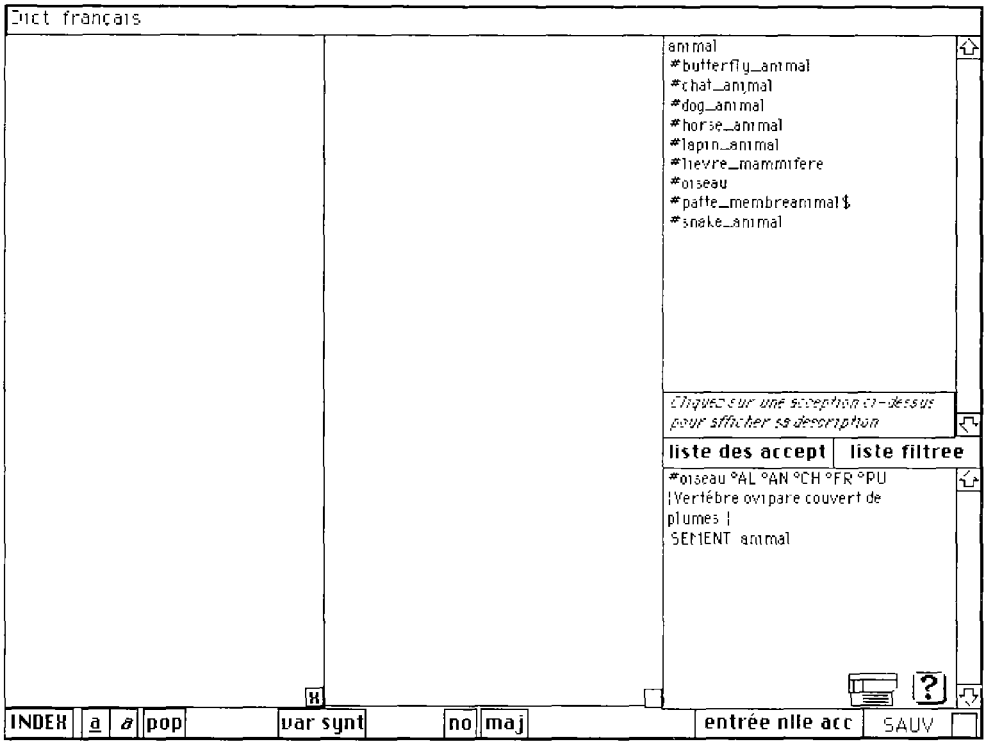


FIGURE 12 : Une carte en cours de création de dictionnaire monolingue comportant une liste filtrée d'acceptions.

8.2. Entrée d'un nouveau lemme

Alors qu'une nouvelle acception est entrée à partir du dictionnaire monolingue de son premier représentant, un nouveau lemme associé à une acception existante est entré à partir du dictionnaire d'acceptions, ce qui permet d'avoir sous les yeux les équivalents déjà entrés dans d'autres langues.

8.3. Écriture contrôlée des variables et valeurs sémantiques et morphosyntaxiques

Les cartes de descriptions sémantiques comme celle de la figure 13, et les cartes de descriptions morphosyntaxiques comme celle représentée pour le français à la figure

15, permettent par ailleurs d'écrire de façon rapide et contrôlée la partie variables-valeurs des descriptions sémantiques et morphosyntaxiques : en cliquant sur le nom des variables et des valeurs désirées, celles-ci s'inscrivent avec une ponctuation normalisée dans le champ inférieur droit de la carte, d'où elles sont ensuite copiées vers les emplacements qui leur reviennent dans le dictionnaire monolingue (variables morphosyntaxiques) ou le dictionnaire d'acceptions (variables sémantiques).

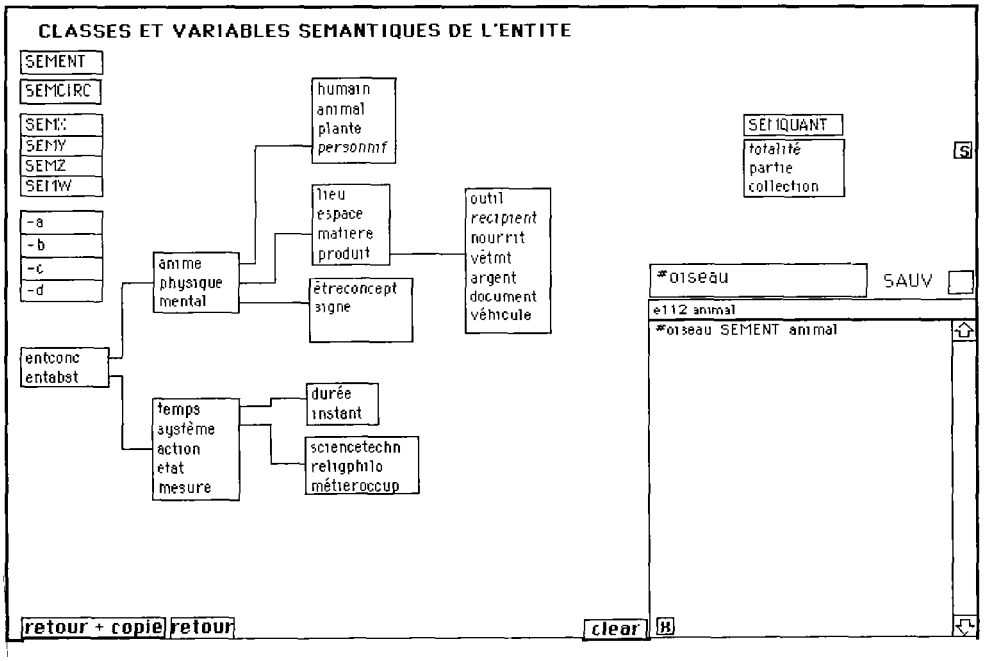


FIGURE 13 . La carte de description sémantique de l'entité.

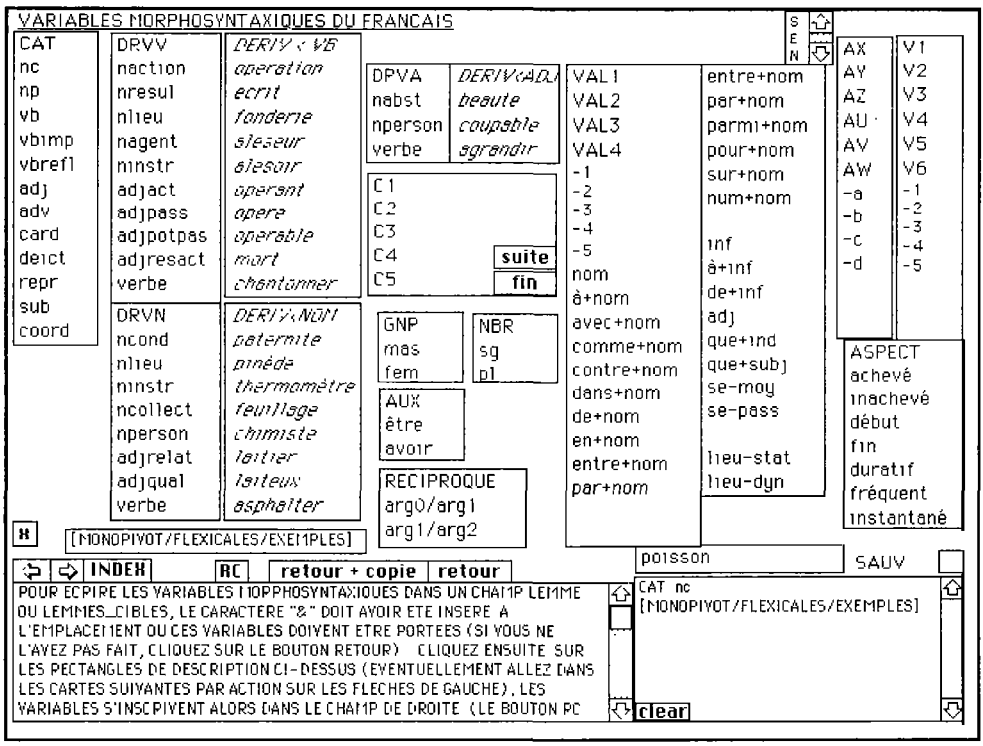


FIGURE 14 : La carte de description morphosyntaxique du français.

8.4. Actualisations diverses

8.4.1. Actualisation des dictionnaires monolingues

Rappelons qu'une carte de pile monolingue contient la description morphosyntaxique « originale » des différents sens d'un lemme donné, mais seulement des copies des descriptions sémantiques des acceptions multilingues correspondantes (cf. 4.2.). Il peut donc être nécessaire de réactualiser ces copies lorsqu'on a été amené à modifier des descriptions sémantiques.

Deux modes de réactualisation sont disponibles :

- dès qu'une modification d'une description d'acception est effectuée dans le dictionnaire d'acceptions, un bouton d'actualisation répercute automatiquement cette modification dans tous les dictionnaires monolingues ;

- par précaution, une procédure de réactualisation, en principe redondante, est disponible dans les piles monolingues : un bouton présent sur chaque carte de lemme permet de remplacer automatiquement les descriptions sémantiques éventuellement périmées par les descriptions présentes dans la pile d'acceptions. Cette réactualisation peut se faire au niveau de la pile entière.

8.4.2. Actualisation du dictionnaire d'acceptions

De même, une carte du dictionnaire d'acceptions contient la description « originale » d'une acception, mais seulement la copie des descriptions morphologiques des représentants de cette acception dans les langues de la base.

Une actualisation de ces informations morphologiques dans le dictionnaire d'acceptions n'est cependant pas vraiment nécessaire, puisqu'elle se réalise automatiquement lors de chaque consultation manuelle (cf. § 4), et que l'extraction automatique de dictionnaires bilingues (cf. § 5) n'utilise que les informations originales des dictionnaires.

Par contre, d'autres fonctionnalités, que nous ne décrirons pas ici, permettent de maintenir la cohérence de la base vis-à-vis de modifications sur les acceptions telles qu'un changement de dénomination de l'acception, une structuration en sous-acceptions, une suppression pure et simple...

Conclusion

La structure de PARAX s'est avérée représenter assez bien le compromis que nous nous étions assignés : assez simple et évolutive pour permettre une expérimentation aisée sur la notion de base multilingue par acceptions, mais disposant de fonctionnalités suffisantes pour que cette expérimentation, qu'il nous reste maintenant à poursuivre, soit réaliste.

Remerciements

Que soient ici remerciés J. P. Guilbaud, N. Nédobejkine et G. Sérasset pour de fructueuses discussions et M. Axtmeyer, D. Levenbach, F. Tchéou et à nouveau N. Nédobejkine pour leur contribution prépondérante à l'entrée des termes anglais, allemands, chinois et russes.

Élaboration d'un dictionnaire informatisé pour le traitement automatique de la langue[1]

Lorne H. Bouchard et Louisette Emirkanian

Département d'informatique et Département de linguistique, Université du Québec à Montréal, Canada

Introduction

Cette recherche s'inscrit dans un programme de recherche dont un des buts est le développement d'une grammaire computationnelle du français (GCF) (Emirkanian & Bouchard, 1992). Cette grammaire, élaborée dans le cadre de la théorie de la Grammaire Syntagmatique Généralisée (GSG) (Gazdar *et al.*, 1985), a été implantée à l'aide de l'outil *Grammar Development Environment* (GDE) (Boguraev, 1988). Du point de vue pratique, une grammaire computationnelle doit avoir une large couverture, si elle veut être utile. D'un point de vue théorique, une grammaire computationnelle doit avoir une couverture large, si elle veut être un modèle crédible de la performance (Bouchard, Emirkanian & Morin, 1992). L'importance du lexique en tant que lieu central de stockage des informations phonologique, morphologique, syntaxique et sémantique est mise de l'avant par la plupart des théories linguistiques contemporaines. Ainsi, une grammaire computationnelle à large couverture doit s'appuyer sur une base de données lexicales importante dans laquelle ces diverses informations sont représentées efficacement. N'ayant trouvé aucune base de données lexicales facilement disponible, nous avons entrepris d'en construire une. La construction d'une telle base de données est une tâche d'envergure qui doit être automatisée le plus possible, ou tout au moins assistée par ordinateur. Nous nous sommes donc concentrés sur le développement de stratégies et d'outils permettant d'extraire les informations nécessaires des dictionnaires sur support informatique et des corpus.

L'ambiguïté, une des caractéristiques des langues naturelles, se manifeste à divers niveaux : nous nous intéresserons ici à l'ambiguïté catégorielle (*le* peut être déterminant ou pronom, par ex.) et à l'ambiguïté structurelle (dans l'analyse de *Paul ex-*

1 Recherche subventionnée par le CRSHC (410-93-0607) et le FCAR (95 ER 1198)

pédie des porcelaines de Chine le syntagme *de Chine* peut être rattaché au syntagme nominal *des porcelaines*, au syntagme verbal *expédie des porcelaines* ou à la phrase comme adjoint de lieu). En faisant intervenir des connaissances – de la langue, du monde et du contexte d'élocution – on peut lever quelquefois cette ambiguïté. L'identification des connaissances requises, leur extraction et leur exploitation est une tâche d'envergure. Néanmoins malgré tout, certaines ambiguïtés subsisteront : leur résolution devra faire intervenir l'humain.

Cependant, puisque certaines informations requises par l'analyseur, la sous-catégorisation et les restrictions de sélection en particulier, ne sont pas notées explicitement dans les dictionnaires disponibles sur support informatique, elles doivent être extraites de l'analyse des exemples utilisés dans les dictionnaires ou de celle de corpus. L'extraction des informations provenant du lexique est un processus itératif qui se développe en plusieurs phases. À chaque étape, une information initiale, déjà disponible ou facile à extraire, est utilisée comme amorce d'un processus qui permet d'extraire de nouvelles informations qui, à leur tour, permettent d'alimenter l'étape suivante.

Interface entre la GCF et le lexique

La GSG est une théorie linguistique qui s'inscrit dans le courant des grammaires à base de contraintes (Shieber, 1986). Dans cette théorie toutes les informations sont véhiculées à l'aide d'ensembles de traits-valeurs, la valeur d'un trait pouvant être un ensemble. Le module d'accès lexical doit produire les ensembles de traits-valeurs correspondant aux mots de la phrase à analyser. Cette tâche comprend essentiellement l'analyse des flexions, la catégorisation et la désambiguïsation catégorielle en contexte s'il y a lieu.

Analyse morphologique

L'analyse morphologique est la première étape de l'analyse de la phrase. Étant donné un mot fléchi, elle produit le radical ou le lemme du mot (c'est-à-dire le mot d'entrée dans le dictionnaire), la catégorie du mot et les informations morphologiques (genre, nombre, personne, temps, mode, par ex.) dérivées de l'analyse de la flexion du mot.

Diverses stratégies peuvent être utilisées pour réaliser des analyseurs morphologiques.

Dictionnaire des formes fléchies

La stratégie la plus simple consiste à utiliser une liste exhaustive des formes fléchies les plus fréquentes. Cette technique a été utilisée dans le module de catégorisation *gramr* de SATO (Daoust, 1992). La liste des formes fléchies a été compilée manuellement en utilisant quelques procédures d'extraction automatique. La liste contient environ 500K mots. La maintenance d'une liste de cette taille pose de sérieux problèmes, d'autant plus que seules les formes fléchies les plus fréquentes sont stockées.

Analyse des suffixes utilisant un dictionnaire des exceptions

Une stratégie originale exploite les régularités de la morphologie du français. La technique est décrite par Meunier (1970). Par exemple, on observe en français que la plupart des mots se terminant en *-able* sont des adjectifs, sauf un petit nombre d'exceptions. Nous avons effectué une étude systématique de la nomenclature du Grand Robert (Rey, 1988). Des 99 134 mots dans la nomenclature, 1 185 se terminent en *-able* et seulement 25 de ces mots correspondent à un nom commun simple ou composé. On trouve également 15 verbes qui ont *-abler* pour suffixe. On peut construire un analyseur morphologique basé sur cette technique, en consultant d'abord un fichier contenant la liste des exceptions à une règle donnée : les mots qui se terminent par *-able* sont des adjectifs. Tous les mots qui n'ont pas été catégorisés lors de la consultation du fichier sont donc catégorisés *adjectif.*

Automate à suffixes et dictionnaire des radicaux

Lors de la construction d'un prototype d'analyseur du français capable de détecter et de corriger certaines erreurs d'orthographe d'usage et de syntaxe (Emirkanian & Bouchard, 1989), nous avions mis au point un analyseur morphologique.

L'algorithme utilisé par cet analyseur morphologique était basé sur un automate à suffixes (Aho & Corasick, 1975). L'analyse de la forme fléchie d'un mot est effectuée de droite à gauche ; lors de ce balayage, diverses hypothèses sont émises, hypothèses qui sont confirmées ou non par une consultation du dictionnaire des radicaux[2]. Dans ce dictionnaire, on associe un radical à la liste des suffixes qui peuvent le compléter. Cette représentation a l'avantage de mettre en facteur la liste des suffixes.

Construction et minimisation d'automates à états finis

Un lexique informatisé a été compilé par le groupe intelligence artificielle et langage naturel du CNET à Lannion (CNET, 1990). Dans ce lexique, un mot est représenté à l'aide de trois informations : le lemme, une liste de suffixes et une liste de paramètres qui définissent la jonction entre le lemme et les suffixes. Fouqueré (1994) a construit un prototype d'analyseur morphologique basé sur ce lexique. Les lemmes sont d'abord développés pour obtenir toutes les formes fléchies, puis un automate à états finis est construit et minimisé. L'analyseur résultant est petit (400K) et très rapide.

Morphologie à deux niveaux : compilateur de règles et de lexique

L'analyse morphologique à deux niveaux (Koskenniemi, 1983) s'inspire de la morphologie concrète : elle postule deux niveaux de représentation, à savoir le niveau lexical et le niveau de surface. La correspondance entre ces deux niveaux est main-

2. Par *radical* on entend dans cette implantation la plus longue partie invariante d'un mot fléchi.

tenue par des transducteurs à états finis bidirectionnels. La morphologie à deux niveaux a été grandement popularisée par un logiciel du domaine public pour l'IBM PC et le Macintosh (Antworth, 1990). L'élégance et la clarté de ce modèle font qu'il a été adopté comme formalisme de choix pour la définition du module d'analyse morphologique de certains systèmes de traitement automatique de la langue naturelle (Ritchie, Russell, Black & Pulman, 1992). En utilisant les techniques de manipulation des automates à états finis, les règles de ce système peuvent être compilées efficacement. Plusieurs travaux portent sur l'analyse morphologique à deux niveaux du français (Karttunen, 1993 ; Karttunen & Beesley, 1992 ; Lun, 1983).

L'analyseur morphologique du français développé par Xerox (Xerox, 1993) utilise cette technologie. Nous précisons que nous avons utilisé cet analyseur comme une boîte noire, puisqu'il nous a été fourni sous la forme d'un programme compilé.

Comparaison des résultats obtenus

On peut comparer sommairement ces divers analyseurs morphologiques[3] en utilisant un fragment d'un corpus d'articles de presse que l'on retrouve à la figure 1. La figure 2 résume les résultats obtenus en utilisant la procédure *gramr* de SATO, l'analyseur XLT de Xerox et le « prototype » d'analyseur construit sur le lexique du CNET. À première vue, il semble y avoir un consensus entre les divers analyseurs, ce qui est encourageant. Cependant une étude plus minutieuse montre que certains mots n'ont pas été analysés, certaines catégories sont absentes et certains mots ont une catégorisation erronée. Par exemple, *n'et aujourd'hui* n'ont pas été analysés par le « prototype » et la catégorie est notée « XXX » dans la figure 2. Les mots *contre* et *grand* ne sont pas notés « Noun » par XLT. La catégorisation de *s'* comme « conjonction » par *gramr* est sûrement une erreur, de même que le grand nombre de « CONJ » produit par le prototype.

Les codes lexicaux utilisés par les analyseurs morphologiques doivent être traduits dans le système de traits-valeurs utilisé par la GCF. Il a été relativement facile de traduire les codes lexicaux produits par les analyseurs morphologiques que nous avons utilisés. En conclusion donc, les analyseurs morphologiques existants peuvent être utilisés avantageusement comme module de pré-traitement de la grammaire GCF.

Désambiguïsation des codes lexicaux en contexte

Quoiqu'on puisse désambiguïser lors de l'analyse syntaxique (Bouchard, Emirkanian & Ratté, 1989), il est préférable de le faire avant, afin de réduire le temps d'analyse. Une solution très répandue consiste à utiliser un algorithme probabiliste qui se fonde sur un modèle de Markov caché (Kupiec, 1992). Cependant, une solution basée sur une exploitation directe des contraintes des langues naturelles (Brill, 1992 ; Chanod & Tapanainen, 1995) nous semble préférable, parce que plus sûre et plus transparente.

3. Il s'agit des analyseurs qui étaient disponibles comme outils autonomes au moment où nous avons effectué notre étude comparative

Nous avons préféré utiliser, comme nous le verrons plus loin, un modèle discret basé sur les n-grammes pour découvrir dans un corpus donné les contraintes fortes qui peuvent servir à désambiguïser les codes lexicaux. Les résultats obtenus semblent meilleurs que ceux obtenus en utilisant un modèle stochastique et surtout plus faciles à justifier, du moins rationnellement.

« Il est difficile de vivre avec des rêves rétrécis Le grand froid qui a figé la fin de l'année 1993 exacerbe l'adversité qui semble aujourd'hui s'acharner contre tous les projets généreux Si nous n'y prenons garde, les temps qui viennent pourraient être ceux de la résignation, ou pire, de la dureté »

Lise Bissonnette, *Le Devoir*,
Vendredi 31 décembre 1993, p A6

FIGURE 1 Fragment d'un article de journal.

mot	gramr	XLT	prototype
a	(aux,v_conj)	(Noun,Verb)	(N,V)
acharner	v_inf	Verb	V
adversité	nomc	Noun	N
année	nomc	Noun	N
aujourd'hui	adv	Adv	XXX
avec	prép	(Adv,Prep)	PREP
ceux	p_indéf	Pro	PRON
contre	(v_conj,nomc,prép)	(Adv,Prep,Verb)	(ADVE,CONJ,N,PREP,V)
de	prép	(Det,Prep)	PREP
des	(artind,artpart,prép)	(Det,Prep+Det)	(CONJ,DET)
difficile	adj	Adj	A
dureté	nomc	Noun	N
est	(aux,v_conj,nomc)	(Noun,Verb)	(A,N,V)
être	(aux,v_inf,nomc)	(Noun,Verb)	(N,V)
exacerbe	v_conj	Verb	V
figé	(adj,ppassé)	(Adj,Verb)	(A,V)
fin	(adj,nomc)	(Adj,Adv,Noun)	(A,ADVE,N)
froid	(adj,nomc)	(Adj,Noun)	(A,N)
garde	(v_conj,nomc)	(Noun,Verb)	(N,V)
généreux	adj	Adj	A
grand	(adj,nomc)	(Adj,Adv)	(A,ADVE,N)
il	p_pers	PC	(CONJ,PRON)
l'	(artdéf,p_pers)	(Det,PC)	(DET,PRON)
la	(artdéf,nomc,p_pers)	(Det,Noun,PC)	(DET,N,PRON)
le	(artdéf,p_pers)	(Det,PC)	(DET,PRON)
les	(artdéf,p_pers)	(Det,PC)	(DET,PRON)
n'	adv	Neg	XXX
nous	p_pers	(PC,Pro)	PRON
ou	conjonction	Coord	CONJ
pire	adj	(Adj,Noun)	(A,N)
pourraient	v_conj	Verb	V
prenons	v_conj	Verb	V
projets	nomc	Noun	N

qui	p_relatif	Pro	PRON
résignation	nomc	Noun	N
rétrécis	(adj,v_conj,ppassé)	Verb	(A,V)
rêves	(v_conj,nomc)	(Noun,Verb)	(N,V)
s'	conjonction	(PC,SConj)	PRON
semble	v_conj	Verb	V
si	(adv,conjonction,nomc)	(APrMod,Noun,SConj)	ADVE,CONJ,N,PREP)
temps	nomc	Noun	(N,V)
tous	(détindéf,p_indéf)	(Det,Pro)	(A,DET,PRON)
viennent	v_conj	Verb	V
vivre	(v_inf,nomc)	(Noun,Verb)	(N,V)
y	(nomc,p_pers)	(Noun,PC)	(N,PRON)
1993	détnum	Card	XXX

FIGURE 2 · Liste de codes lexicaux affectés.

Le problème du rattachement des syntagmes prépositionnels

L'ambiguïté du rattachement des syntagmes prépositionnels donne lieu à une explosion combinatoire dans le nombre des analyses produites par un analyseur (Church & Patil, 1982).

Par exemple, lorsqu'il apparaît dans la position immédiatement à droite du syntagme nominal objet direct du verbe, le syntagme prépositionnel peut être rattaché à trois endroits différents : comme adjoint du syntagme nominal qui le précède immédiatement, comme argument du verbe (le plus souvent facultatif) ou comme adjoint du syntagme verbal. À défaut d'autres informations, l'analyseur doit produire trois analyses dans ce cas.

Trois types d'information permettent de modéliser les préférences au niveau du rattachement (Wilks, Huang & Fass, 1985) : la sous-catégorisation, les restrictions de sélection et certaines connaissances du domaine. La sous-catégorisation, une information de nature syntaxique, informe sur la structure des compléments, obligatoires ou facultatifs, d'une tête lexicale. Les restrictions de sélection, une information de nature sémantique, informent sur les types sémantiques des compléments de la tête. Les connaissances du domaine sont des connaissances de nature encyclopédique qui établissent des liens entre les éléments du domaine.

Sous-catégorisation dans GSG et la GCF

Les règles lexicales de dominance immédiate sont la composante de la GSG qui établit le pont entre la grammaire et le lexique. Dans ces règles, la valeur du trait SUB de la tête lexicale représente de façon succincte les divers compléments obligatoires ou facultatifs de la tête. Nous ne traiterons, dans les exemples qui suivent, que de la sous-catégorisation des verbes.

La sous-catégorisation des verbes a été traitée abondamment dans la GCF (GIREIL, 1994). Le verbe *penser* (figure 3) porte les traits de sous-catégorisation SUB20+, SUB33+, SUB50+ et SUB51+.

Traits	Schémas associés	Exemples
SUB20	P3[à,CPR N3]	Gilles pense à Mireille
SUB33	P3[à,CPR V3]	Gilles pense à visiter Boston
SUB50	Q3[que]	Mireille pense que cette thèse est très bonne
SUB51	V3[VFORM er]	Mireille pense venir

FIGURE 3 . Schéma de sous-catégorisation pour *penser*.

Dans le cas d'une sous-catégorisation stricte, c'est-à-dire lorsqu'un complément est obligatoirement requis par la tête, le trait de sous-catégorisation permet d'obtenir une seule analyse syntaxique. Le plus souvent cependant, l'analyseur est confronté à des cas de sous-catégorisation où le complément est spécifié comme facultatif.

Détermination des cas de sous-catégorisation

S'il est fréquent que les dictionnaires courants notent systématiquement la distinction transitif / intransitif pour les verbes, la sous-catégorisation apparaît peu ou le plus souvent elle n'apparaît pas du tout. Les cas de sous-catégorisation doivent être extraits de l'analyse des exemples contenus dans le dictionnaire ou d'une analyse de corpus.

Analyse des exemples d'un dictionnaire

Le ZYZOMYS (ZYZOMYS, 1989) est la version CD-ROM du *Dictionnaire de notre temps* (Guerard, 1989). Sur le CD-ROM, un fichier contient la base textuelle du dictionnaire.

On distingue quatre sources d'information pour les verbes dans le ZYZOMYS : la catégorie lexicale et le modèle de conjugaison spécifiés dans l'entrée du verbe, les informations entre parenthèses, et les exemples. Pour extraire l'information, nous avons adopté une représentation commune permettant d'unifier en un tout cohérent les informations provenant de diverses sources.

Dans la plupart des cas (90 %), un verbe est spécifié comme *transitif, intransitif, impersonnel* ou *pronominal*. Dans seulement 10 % des cas il est noté *transitif direct, transitif indirect, copule* ou *auxiliaire*. Le ZYZOMYS utilise 83 modèles de conjugaison. En consultant le modèle du verbe on peut retrouver divers renseignements : l'auxiliaire requis, *avoir* ou *être* ou les deux, et si le verbe est *pronominal* ou *défectif*.

Le ZYZOMYS utilise les parenthèses dans les articles pour signaler des restrictions de sélection sur les arguments des verbes.

Par exemple, dans l'article de abandonner, on trouve les informations suivantes :

```
renoncer à (qqch)
laisser (qqch) à (qqn)
mettre (qqch) à la disposition de (qqn)
délaisser (qqch)
```

Cependant, une analyse systématique des informations signalées entre parenthèses dans les entrées révèle que cette information est très éparse, comme l'indique l'affichage en ordre inverse des fréquences à la figure 4. Nous n'avons retenu de cette information que la distinction *qqn/qqch* et nous avons encodé toutes les autres distinctions en termes de celle-là.

```
342   (qqn)
164   (qqch)
 87   (S. comp.)
 50   (choses)
 37   (personnes)
 34   (e)
 27   (+ inf.)
 17   (une chose)
 15   (I)
 13   (sens 2)
 12   (un navire)
 12   (qqn, qqch)
 12   (Re'cipr.)
 12   (E)
 11   (un corps)
 11   (un animal)
 11   (Personnes.)
 11   (Choses.)
 10   (un objet)
 10   (un liquide)
 10   (un lieu)
...
```

FIGURE 4 · Informations notées entre parenthèses dans les articles du ZYZOMYS.

Comme dernière source d'information sur les verbes dans le ZYZOMYS, nous avons les exemples utilisés dans les articles. Ces exemples sont simples et l'information qu'on peut en extraire, les cas de sous-catégorisation, permet d'enrichir les informations extraites par ailleurs.

À titre indicatif, voici quelques exemples utilisés pour illustrer l'article d'*abandonner* :

```
Abandonner un projet.
Abandonner un emploi.
De nombreux coureurs ont abandonné au cours de cette étape.
J'abandonne la capitale pour m'établir dans une petite ville.
Abandonner sa famille.
```

Pour analyser ces exemples nous avons développé une grammaire hors-contexte empirique permettant d'isoler les syntagmes qui gravitent autour du verbe. Un « chart parser » piloté par cette grammaire permet d'analyser les exemples.

Les articles des verbes sont d'abord extraits du dictionnaire et traduits en format LISP à l'aide d'un programme **lex**. Un lexique des formes fléchies uniques est construit à partir de ces données et catégorisé à la main. Les exemples sont ensuite analysés à l'aide du « chart parser » et les données en format LISP sont enrichies du résultat de l'analyse syntaxique. La figure 5 montre un fragment du résultat correspondant à l'analyse du verbe *abandonner*.

```
'(ABANDONNER (CAT V)
        ((SENS 1
                ((SCAT TRANSITIF (TAKES SN))
                 (AUX AVOIR)
                 (SUBDEF 1
                        ((SPECIFIEUR (ANIME -))
                         (EXEMPLE (TAKES SN)
                                (ABANDONNER UN
                                        PROJET))
                         (EXEMPLE (TAKES SN)
                                (ABANDONNER SON
                                        EMPLOI))
                         (DOMAINE SPORT
                                ((TAKES 0)
                                 (EXEMPLE AMBIGUE
                                        (DE NOMBREUX
                                           COUREURS
                                           ONT
                                           ABANDONNE03
                                           AU
                                           COURS
                                           DE
                                           CETTE
                                           E03TAPE))))))
                 (SUBDEF 2
                        ((TAKES
                          (SN (ANIME -)
                             SP
                             ((HUMAIN +) (PREP A))))
                         (SPECIFIEUR (ANIME -))
                         (SPECIFIEUR (HUMAIN +))))
                 (SUBDEF 3
                        ((SPECIFIEUR (ANIME -))))
                 (SUBDEF 4
                        ((EXEMPLE AMBIGUE
                                (J ABANDONNE
                                   LA
                                   CAPITALE
                                   POUR
                                   M
                                   E03TABLIR
                                   DANS
                                   UNE
                                   PETITE
                                   VILLE))))
                 (SUBDEF 5
                        ((EXEMPLE (TAKES SN)
                                (ABANDONNER SA
                                        FAMILLE))))))
...
```

FIGURE 5 Fragment de l'information extraite du ZYZOMYS pour le verbe *abandonner*.

La figure 6 permet d'apprécier le taux de réussite d'une première expérience que nous avons effectuée. On peut observer que la transcription des entrées du dictionnaire en formes LISP ne s'effectue pas toujours correctement à cause d'imperfections dans le programme de transcription. On peut voir également que les échecs de l'analyse des exemples sont attribués, à part égale, à deux causes principales : les ambiguïtés et les erreurs d'analyse. L'ambiguïté provient des analyses mutiples, dues notamment au problème du rattachement. Les erreurs d'analyse sont principalement dues aux limitations de la grammaire empirique.

entrées	nombre	« lispifiées »	exemples	corrects	ambiguës	erronés
a-	411	387 (94%)	764 (100%)	309 (40%)	222 (30%)	233 (30%)
b-	236	125 (52%)	87 (100%)	44 (50%)	19 (22%)	24 (28%)
c-	403	299 (74%)	409 (100%)	172 (42%)	116 (28%)	121 (30%)
d-	776	687 (89%)	876 (100%)	406 (46%)	243 (28%)	229 (26%)
e-	641	599 (93%)	884 (100%)	407 (46%)	272 (31%)	205 (23%)
Total	2467	2097 (85%)	3020 (100%)	1338 (44%)	872 (29%)	812 (27%)

FIGURE 6 Statistiques sur le taux de réussite de l'extraction des exemples du ZYZOMYS

L'analyse du ZYZOMYS a été effectuée en utilisant un « chart parser » piloté par une grammaire hors-contexte façonnée à la main. Dans des travaux subséquents, nous avons cherché à éliminer cette étape dans la mesure où elle est très laborieuse et dépendante du corpus. Une première direction de recherche a exploré l'induction automatique de grammaires régulières, tandis que la seconde a étudié l'exploitation directe d'un modèle discret des n-grammes, ce qui élimine complètement la grammaire.

Analyse d'un corpus

Le corpus d'Alain Guillet (1990) est un ensemble de phrases simples permettant d'exemplifier la classification des verbes dans le lexique-grammaire (Leclère, 1990). Ce corpus contient 20 603 phrases et 175 021 mots, dont 17 503 formes fléchies uniques.

Considérons la première direction.

Induction de grammaire régulière

Dans une première expérience nous avons cherché à induire automatiquement une grammaire régulière (Ejerhed, 1988) à partir d'un échantillon du texte à analyser. Cela revient à effectuer une induction à partir de données positives seulement et certaines précautions sont nécessaires afin d'éviter la sur-généralisation (Angluin, 1980). Pour cela, il suffit d'ordonner les données présentées pour que le processus d'induction converge sur le langage le plus restreint qui soit compatible avec les données, stratégie connue sous le nom de principe des sous-ensembles (Berwick, 1986). L'ordonnancement requis consiste à toujours présenter les phrases et les syntagmes plus

simples avant les plus complexes. Nous avons utilisé comme procédure d'induction la méthode de fusion des queues (Miclet, 1980) qui est particulièrement simple à implanter.

Manuellement, un échantillon du corpus a été catégorisé, découpé en constituants et ordonné selon le principe des sous-ensembles. Puisque l'algorithme semblait sur-généraliser, du moins par rapport au type de grammaire recherché, nous avons dû noter par un & les constituants qui étaient vides dans les phrases utilisées pour l'apprentissage et modifier l'algorithme d'induction en conséquence.

La figure 7 illustre la séquence d'apprentissage utilisée pour induire une grammaire de phrases simples qui découpe les constituants qui gravitent autour du verbe, par exemple le sujet, le complément direct, le complément indirect. Les | servent à délimiter les syntagmes superficiels de la phrase et les &, qui marquent des positions vides, servent à guider l'induction dans une position donnée.

```
| &       | V | &       | &       |.
| NP      | V | &       | &       |.
| DET N   | V | &       | &       |.
| &       | V | NP      | &       |.
| &       | V | DET N   | &       |.
| &       | V | &       | P NP    |.
| &       | V | &       | P DET N |.
...
```

FIGURE 7 : Une séquence d'apprentissage bien ordonnée.

L'automate à états-finis résultant est converti manuellement en un transducteur qui permet de découper le reste du corpus en constituants. Ce transducteur est donc un analyseur syntaxique primitif qui permet de mettre en relief les arguments du verbe.

Afin de cerner le positionnement de la négation et des clitiques autour du verbe conjugué simple, on doit avoir recours à une séquence d'apprentissage comme le montre la figure 8. Dans cette séquence et la suivante, le signe « * » identifie un mot absent et le signe « & » , un constituant vide.

```
&  | * * * V *                  | &  | &  | &.
&  | * * * V ADV *              | &  | &  | &.
&  | * * CLI_I V *              | &  | &  | &.
&  | NEG1 * * V NEG2            | &  | &  | &.
&  | NEG1 * * V ADV NEG2        | &  | &  | &.
&  | NEG1 CLI_D * V NEG2        | &  | &  | &.
&  | NEG1 * CLI_I V NEG2        | &  | &  | &.
...
```

FIGURE 8 : Séquence d'apprentissage pour le verbe avec positionnement des clitiques et de la négation.

Afin de cerner le positionnement de la négation et des clitiques autour du verbe conjugué à un temps composé, on doit avoir recours à une séquence d'apprentisage comme celle de la figure 9.

```
&  | NEG1 CLI_D * AUX NEG2 P_PAS  | & | & | &.
&  | NEG1 * CLI_I AUX NEG2 P_PAS  | & | & | &.
&  | NEG1 * * AUX ADV NEG2 P_PAS  | & | & | &.
&  | * CLI_D * AUX * P_PAS        | & | & | &.
&  | * * CLI_I AUX * P_PAS        | & | & | &.
&  | * * * AUX ADV * P_PAS        | & | & | &.
...
```

FIGURE 9 Séquence d'apprentissage pour l'auxiliaire et le participe passé avec positionnement des clitiques et de la négation

Cette méthode présente deux désavantages majeurs. En effet, le premier est que les données doivent être prétraitées manuellement, ce qui demande un travail fastidieux. Si on cherche à induire une grammaire particulière, le travail de préparation des séquences d'apprentissage devient rapidement comparable à celui qui est requis pour synthétiser manuellement la grammaire désirée. Le second désavantage est qu'il est difficile de prévoir la séquence d'entraînement nécessaire pour obtenir une grammaire qui couvre un corpus donné.

De plus, malgré les précautions prises, l'algorithme sur-généralise et il en résulte que de nombreuses phrases non-grammaticales sont acceptées par la grammaire. Cette limitation a peu d'impact dans notre application. Cependant, nous avons observé que les grammaires obtenues en désactivant l'étape d'induction demeuraient de taille acceptable et sur-généralisaient moins. L'induction présente donc peu d'intérêt en pratique pour ce genre d'application, ce qui nous amène à exposer la deuxième direction de recherche que nous avons explorée.

Désambiguïsation catégorielle en contexte

Dans cette seconde approche nous avons expérimenté avec un modèle discret des n-grammes, c'est-à-dire ne faisant aucunement intervenir les notions de probabilité ou de statistique. Plutôt que de réitérer l'expérience précédente et de catégoriser manuellement, nous avons préféré utiliser un programme pour le faire. Cela nous permettra de discuter à nouveau du problème de la désambiguïsation et de proposer une nouvelle solution.

Un lexique des formes fléchies uniques extraites du corpus a été catégorisé en utilisant le tagger XLT de Xerox (Xerox, 1993) afin d'obtenir pour chaque forme fléchie sa classe d'ambiguïté, c'est-à-dire l'ensemble des étiquettes de catégorie qui peuvent la caractériser. On peut voir affichées à la figure 10 les classes d'ambiguïté correspondant aux premières entrées dans le lexique du corpus. Puis nous avons étiqueté le corpus directement avec une version expérimentale du tagger stochastique de Xerox (Kupiec, 1992 ; Xerox, 1995).

```
A NOUN_PL NOUN_SG PREP_A
a NOUN_PL NOUN_SG VAUX_P3SG
à PREP_A
a NOUN_PL NOUN_SG VAUX_P3SG
à PREP_A
a-t-elle PC+VAUX_P3SG
a-t-il PC+VAUX_P3SG
abaissé PAP_SG
abaissement NOUN_SG
abandon NOUN_SG
abandonne VERB_P1P2 VERB_P3SG
abandonné PAP_SG
abandonnée PAP_SG
abandonnent VERB_P3PL
abasourdi PAP_SG
abasourdissement NOUN_SG
abat NOUN_SG VERB_P3SG
abat-jour NOUN_PL NOUN_SG
...
```

FIGURE 10 Fragment du lexique des formes fléchies et des classes d'ambiguité correspondantes.

Nous avons dû corriger manuellement un grand nombre d'erreurs d'étiquetage : *Cela/VERB_P3SG, Le/DET_SG, maçon/ADJ_SG, ...d'/DET_PL être/VAUX_INF,* par exemple. Il semble que ce tagger stochastique ne tire pas toujours parti du contexte dans lequel se trouve le mot. Lors de cette phase d'édition, nous avons également ajouté quelques distinctions supplémentaires puisque Xerox ne distingue que PREP_A et PREP_DE et regroupe les autres prépositions simplement sous PREP alors qu'elles sont souvent syntaxiquement discriminantes. À la figure 11 on peut voir un fragment du corpus étiqueté et désambiguïsé.

```
Max/NOUN_PRP a/VAUX_P3SG abaissé/PAP_SG Léa/NOUN_PRP à/PREP_A
ce/PRON qu'/CONJQUE elle/PRON demande/VERB_P3SG 100/NUM
f./NOUN_INV ./SENT

Max/NOUN_PRP a/VAUX_P3SG abaissé/PAP_SG Léa/NOUN_PRP
à/PREP_A demander/VERB_INF 100/NUM f./NOUN_INV ./SENT

Max/NOUN_PRP s'/PC est/VAUX_P3SG abaissé/PAP_SG à/PREP_A
demander/VERB_INF 100/NUM f./NOUN_INV ./SENT

Le/DET_SG vent/NOUN_SG a/VAUX_P3SG
abaissé/PAP_SG la/DET_SG température/NOUN_SG ./SENT

Le/DET_SG vent/NOUN_SG a/VAUX_P3SG causé/PAP_SG l'/DET_SG
abaissement/NOUN_SG de/PREP_DE la/DET_SG température/NOUN_SG ./SENT
...
```

FIGURE 11 : Fragment du corpus étiqueté

L'algorithme expérimental basé sur un modèle discret des n-grammes fonctionne en deux étapes. Dans la phase d'apprentissage on mémorise tous les n-grammes construits sur les catégories désambiguïsées des phrases du corpus d'entraînement. Dans la phase d'exploitation, on ne tient plus compte des catégories désambiguïsées des phrases du corpus, on consulte plutôt le lexique qui retourne la classe d'ambiguïté d'une forme fléchie et on utilise les n-grammes mémorisés lors de la phase d'apprentissage comme contraintes permettant de désambiguïser les catégories des mots de la phrase. Les résultats préliminaires montrent qu'en utilisant seulement 10 % du corpus étiqueté comme apprentissage, on obtient un taux de désambiguïsation exact à plus de 87 %. Il semble que des erreurs d'étiquetage soient la cause de ce taux de succès encore trop faible.

Segmentation de la phrase en constituants

Une fois les catégories lexicales désambiguïsées en contexte, il est facile de découper les phrases du corpus en constituants de premier niveau. L'algorithme le plus simple consiste à balayer la phrase de gauche à droite à la recherche d'un verbe conjugué, le pivot de la phrase. À partir de ce point, le segment correspondant au noyau verbal est construit en effectuant un balayage à gauche pour récupérer les clitiques, les adverbes et la négation, puis un balayage à droite pour récupérer le participe passé et les adverbes. Ensuite, on construit des segments avec le fragment de phrase qui précède le segment noyau verbal et le fragment de phrase qui le suit, borné[4] par une préposition, s'il y a lieu. À partir de ce point dans la phrase, on construit itérativement des segments prépositionnels qui sont délimités[5] à gauche et bornés à droite par des prépositions ou la fin de phrase, s'il y a lieu. La figure 12 illustre la segmentation en constituants des premières phrases du corpus.

```
1-  [Max][a abaissé][Léa][à ce qu' elle demande 100 f.]
2-  [Max][a abaissé][Léa][à demander 100 f.]
3-  [Max][s' est abaissé][à demander 100 f.]
4-  [Le vent][a abaissé][la température]
5-  [Le vent][a causé][l' abaissement][de la température]
6-  [Max][a abaissé][la toise][sur la tête][de Luc]
7-  [Max][a abaissé][la manette]
8-  [Paul][a abandonné]
...
```

FIGURE 12 : Illustration de la segmentation des phrases du corpus.

Ad hoc certes, cette méthode d'analyse a l'avantage d'être simple, robuste et facile à utiliser, du moins sur les corpus de phrases simples qui nous intéressent.

On peut constater que cette segmentation étale bien les arguments du verbe. L'algorithme de segmentation est donc une technique qui permet de mettre en relief

4 La préposition ne fait pas partie du segment en question
5 La préposition fait partie du segment en question

les cas de sous-catégorisation d'un verbe. Cependant, on observe que les **P2** qui modifient le **N2** objet direct du verbe ou qui modifie le **N2** qui est la tête du **P2** sont étalés comme si c'étaient des arguments du verbe (voir les exemples 5 et 6 de la figure 12). Tous les exemples de ce type devront être exclus d'un corpus utilisé pour l'apprentissage des cas de sous-catégorisation.

L'extraction des têtes des segments s'effectue par un simple balayage de gauche à droite de ces segments avec certains petits ajustements qui permettent de traiter des complétives et des infinitives. La figure 13 illustre les têtes extraites des exemples segmentés présentés ci-dessus.

Nous pouvons maintenant reproduire les résultats obtenus lors de l'analyse du ZYZOMYS, mais beaucoup plus rapidement.

Cependant, le problème du rattachement des **P2** demeure intégral et pour le résoudre nous devons faire appel aux restrictions de sélection, qui doivent être extraites de l'analyse d'exemples de dictionnaire ou de celle de corpus. Nous avons progressé néanmoins dans la mesure où nous avons un découpage brut des constituants.

```
[("Max" NOUN_PRP)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("Léa" NOUN_PRP)]
[("à ce qu'" PREP_A) ("elle demande 100 f." COMPL]

[("Max" NOUN_PRP)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("Léa" NOUN_PRP)]
[("à" PREP_A) ("demander 100 f." INF)]

[("Max" NOUN_PRP)]
[("s'" PC) ("est" VAUX_P3SG) ("abaissé" PAP_SG)]
[("à" PREP_A)("demander 100 f." INF)]

[("vent" NOUN_SG)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("température" NOUN_SG)]

[("vent" NOUN_SG)]
[("a" VAUX_P3SG) ("causé" PAP_SG)]
[("abaissement" NOUN_SG)]
[("de" PREP_DE) ("température" NOUN_SG)]

[("Max" NOUN_PRP)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("toise" NOUN_SG)]
[("sur" PREP) ("tête" NOUN_SG)]
[("de" PREP_DE) ("Luc" NOUN_PRP)]

[("Max" NOUN_PRP)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("manette" NOUN_SG)]

[("Paul" NOUN_PRP)]
[("a" VAUX_P3SG) ("abandonné" PAP_SG)]
...
```

FIGURE 13 : Illustration de l'extraction des têtes des segments.

Détermination des restrictions de sélection

Les restrictions de sélection sont les contraintes sémantiques qui portent sur les arguments d'une tête lexicale qui permettent de juger de la bonne formation des syntagmes construits sur ces têtes. Pour spécifier les restrictions de sélection, nous utilisons une classification de l'espace des noms qui est basée sur une structure appelée ontologie.

Choix d'une ontologie

Construite sur un petit nombre de distinctions élémentaires, une ontologie permet de discerner les classes de noms significatives du point de vue des restrictions de sélection. Dans la perspective des systèmes de types utilisés en programmation fonctionnelle ou dans la théorie *Head-Driven Phrase Structure Grammar* (HPSG) (Pollard & Sag, 1987), ces classes de noms correspondent à des types complexes définis par héritage multiple. Ces liens d'héritage permettent de généraliser facilement à partir d'un concept donné. Cette ontologie n'est pas une taxinomie dans la mesure où une classe peut être définie à partir de plusieurs classes existantes. La structure de l'ontologie prend la forme d'un treillis plutôt que celle d'une arborescence. Dans la mesure où le principe de classification utilisé fait intervenir des connaissances naïves de la réalité, c'est-à-dire du monde, cette structure de classification mérite d'être appelée une ontologie.

Nous avons effectué une comparaison de trois ontologies qui ont été proposées pour l'analyse automatique du langage naturel, les ontologies des projets CORE Language Engine et PENMAN ainsi que celle développée par Dahlgren. L'ontologie PENMAN a été développée dans le cadre d'un projet (Bateman, 1991 et 1993) visant à développer un système général de génération de textes en langue naturelle dans le paradigme des grammaires systémiques. L'ontologie développée pour le CORE Language Engine par SRI à Cambridge (Alshawi, 1992) a pour fonction la désambiguïsation syntaxique et sémantique dans un système de traitement automatique. L'ontologie qui est présentée par Dahlgren (1988 et 1993) a été développée pour un système de compréhension automatique de textes en langue naturelle. Si notre étude a pu mettre en relief de nombreuses similitudes entre ces ontologies, elle a aussi permis d'identifier plusieurs points de divergence. Nous avons adopté, de façon provisoire, l'ontologie de Dahlgren qui est reproduite à la figure 14.

ENTITY	(ABSTRACT | REAL) & (INDIVIDUAL | COLLECTIVE)
ABSTRACT	IDEAL | PROPOSITIONAL | QUANTITY | IRREAL
QUANTITY	NUMERICAL | MEASURE
REAL	(PHYSICAL | TEMPORAL | SENTIENT) & (NATURAL | SOCIAL)
PHYSICAL	(STATIONARY | NONSTATIONARY) & (LIVING | NONLIVING)
NONSTATIONARY	(SELFMOVING | NONSELFMOVING)
COLLECTIVE	MASS | SET | STRUCTURE
TEMPORAL	RELATIONAL | NONRELATIONAL
RELATIONAL	(EVENT | STATE) & (MENTAL | EMOTIONAL | NONMENTAL)
EVENT	(GOAL | NONGOAL) & (ACTIVITY | ACCOMPLISHMENT | ACHIEVEMENT)

FIGURE 14 Ontologie tirée de Dahlgren & McDowell (1986).

Extraction des restrictions de sélection

Le type associé à la tête nominale d'un syntagme est obtenu en classifiant celle-ci selon l'ontologie, une étape qui exige pour le moment une intervention de l'utilisateur.

Les types des arguments d'un même verbe sont unifiés pour obtenir un type résultant. En général, ce type résultant correspond au plus petit des majorants des types des arguments, cependant il y a lieu de dédoubler parfois la restriction de sélection, dans le cas où le type résultant est trop général par exemple.

Examinons les exemples du verbe abaisser dans le corpus (figure 13). On retrouve « Max » ou « vent » comme têtes des segments sujets et « manette », « température » et « toise » comme têtes des segments objets directs. L'ontologie classifie ces noms de la façon suivante :

```
Max          (ENTITY (REAL SENTIENT NATURAL) INDIVIDUAL)

Max          (ENTITY (REAL (PHYSICAL (NONSTATIONARY SELFMOVING)
              LIVING) NATURAL) INDIVIDUAL)

vent         (ENTITY (REAL (PHYSICAL (NONSTATIONARY SELFMOVING)
              NONLIVING) NATURAL) COLLECTIVE)

manette      (ENTITY (REAL (PHYSICAL (NONSTATIONARY NONSELFMOVING)
              NONLIVING) SOCIAL) INDIVIDUAL)

température  (ENTITY (ABSTRACT (QUANTITY MEASURE)) INDIVIDUAL)

toise        (ENTITY (REAL (PHYSICAL (NONSTATIONARY NONSELFMOVING)
              NONLIVING) SOCIAL) INDIVIDUAL)
```

L'unification de la première entrée pour « Max » (de type PERSON) et de l'entrée pour « vent » (de type FLOW_GROUP) produit le type « (ENTITY (REAL NATURAL)) » tandis que l'unification de la seconde entrée pour « Max » (de type HUMAN) et de l'entrée pour « vent » (de type FLOW_GROUP) produit le type « (ENTITY (REAL (PHYSICAL (NONSTATIONARY SELFMOVING)) NATURAL)) ». On retient donc ce second type puisqu'il est plus spécifique que le premier.

L'unification du type de « manette » (MANMADEOBJ) et de celui de « toise » (MANMADEOBJ) produit le type « (ENTITY (REAL (PHYSICAL (NONSTATIONARY NONSELFMOVING) NONLIVING) SOCIAL) INDIVIDUAL) », c'est-à-dire le même type que celui des deux arguments. Ce type caractérise, en partie du moins, le sens 1 de ce verbe dans le Grand Robert :

> ◊ 1. Faire descendre à un niveau plus bas ⇒ **Baisser** *Voulez-vous abaisser la vitre ? Abaisser qqch. en inclinant*, en penchant*.* — Arithm *Abaisser un chiffre* dans une division, écrire un chiffre du dividende à la suite du reste obtenu — Géom. *Abaisser une perpendiculaire* : mener d'un point une perpendiculaire à une ligne, à un plan.

Par contre, l'unification du type de « manette » ou de celui de « toise » avec celui de « température » (MEASUREMENT) produit le résultat « (ENTITY INDIVIDUAL) », c'est-à-dire un des types les plus vagues dans l'ontologie. Ce résultat signale qu'on doit prévoir une seconde entrée pour les restrictions de sélection du verbe abaisser qui correspond au sens 3 de ce verbe dans le Grand Robert :

> ◊ 3. Diminuer la quantité de, faire baisser. *Abaisser la température* ⇒ **Atténuer** *Abaisser le prix des denrées* ⇒ **Diminuer**. *Abaisser la voix. — Abaisser un seuil, un temps Abaisser l'âge de la retraite* : réduire le nombre d'années de travail nécessaires, dans une profession donnée, au droit à la retraite ⇒ Abaissement, cit. 1.2
> Alg *Abaisser le degré d'un polynôme, d'une équation*, le ou la réduire à un degré inférieur.

La partie analyse de cette tâche a été effectuée et sa mise en œuvre est en cours.

Conclusion

Les cas de sous-catégorisation et les restrictions de sélection sont deux informations qui ne sont que très rarement notées dans les dictionnaires usuels et qui permettent de résoudre le problème du rattachement des groupes prépositionnels dans un analyseur.

Nous avons montré comment les cas de sous-catégorisation peuvent être automatiquement extraits de l'analyse des exemples de dictionnaires sur support informatique ou de l'analyse de corpus. Pendant la période durant laquelle nous avons effectué cette recherche, notre méthodologie a évolué : d'un système écrit sur mesure pour analyser les exemples du ZYZOMYS, nous en sommes venus à utiliser certains des outils linguistiques qui sont maintenant disponibles. L'utilisation de ces outils rend le travail moins fastidieux et permet de traiter un plus grand volume de données. Dans les deux cas cependant les données présentées au programme doivent être filtrées au préalable à défaut d'avoir un informateur disponible en ligne pour résoudre l'ambiguïté du rattachement. Fondé sur les travaux de Tomita (1984), nous avons construit la maquette d'un programme qui pose des questions à l'utilisateur dans le but de lever l'ambiguïté (Bouchard & Emirkanian, 1990).

Nous avons présenté un modèle de désambiguïsation des catégories lexicales en contexte qui se fonde sur un modèle discret des n-grammes qui nous semble être un remplaçant possible des algorithmes basés sur les modèles de Markov cachés, du moins pour les corpus qui nous intéressent. Partant d'un texte catégorisé et désambiguïsé en contexte, nous avons présenté un algorithme de segmentation simple et robuste qui permet de mettre en relief les arguments du verbe.

Finalement, nous avons illustré la façon de rendre compte des restrictions de sélection à partir d'une ontologie et de l'opération d'unification des types dans le treillis de l'ontologie.

Génération de dictionnaires-machines multilingues pour la traduction automatique de diagnostics médicaux

Guy DEVILLE et Emmanuel HERBIGNIAUX

École de langues vivantes, Facultés universitaires de Namur, Belgique

1. Introduction

Le projet ANTHEM[1] est le fruit de la collaboration entre chercheurs et industriels issus de diverses disciplines : médecine, informatique et linguistique[2]. Son objectif est de développer un prototype portable d'interface en langage naturel permettant la traduction et l'encodage automatiques de diagnostics médicaux (Ceusters *et al.*, 1994).

Cet article expose l'approche adoptée dans le développement des ressources lexicales multilingues du prototype ANTHEM, approche qui a nécessité une analyse minutieuse de grands corpus multilingues représentatifs de diagnostics médicaux.

Dans la section 2, nous décrivons les travaux de mise en forme et d'échantillonnage effectués sur les corpus de diagnostics médicaux. Ensuite, nous présentons dans la section 3 le modèle de représentation sémantique élaboré sur base des observations effectuées sur les échantillons. La section 4 est consacrée à une présentation sommaire de l'architecture du prototype ANTHEM. Nous abordons dans la section 5 le développement, l'extension et la maintenance proprement dits des lexiques électroniques parallèles français et néerlandais.

En guise de conclusion, nous évoquerons dans la section 6 les perspectives futures du projet ANTHEM, telles que le raffinement du modèle, la validation interne et

1. Le projet ANTHEM « Advanced Natural Language Interface for Multilingual Text Generation in Healthcare » (LRE 62-007) est co-financé par l'Union Européenne dans le cadre du Programme LRE (Linguistic Research and Engineering).

2 Le consortium ANTHEM est coordonné par RAMIT vzw/Gent (W. Ceusters) et comprend en outre les partenaires suivants FUNDP/Namur (G. Deville), CRP-CU/Luxembourg (P. Mousel), IAI/Saarbrücken (O. Streiter), ULG/Liège (C. Gérardy), Datasoft Management NV/Oostende (J. Devlies) et l'Hôpital militaire/Bruxelles (D Penson)

sur site du prototype ou encore les possibilités d'extension vers d'autres projets de recherches connexes ou d'autres applications.

2. Corpora de diagnostics médicaux

Toute entreprise de développement d'interfaces pour le traitement du langage naturel dans le domaine des soins de santé nécessite la modélisation du langage médical. Une telle tâche de modélisation exige l'observation d'un ensemble représentatif d'expressions du sous-langage de l'application. C'est donc dans un souci de validation empirique que deux corpus de diagnostics médicaux ont été collectés dans deux types d'environnements cliniques.

Un premier corpus est constitué de diagnostics établis et enregistrés sous format électronique par l'armée belge (corpus ABL). Ce premier corpus contient un total de 227 000 diagnostics en français et en néerlandais, rédigés par des médecins militaires entre 1970 et 1993, durant leurs consultations de militaires de carrière et de civils effectuant leur service militaire.

Parallèlement au corpus ABL, un corpus de 12 671 expressions françaises et néerlandaises a également été constitué (corpus Médidoc). Ce corpus couvre approximativement la même période que le corpus ABL. Il a été réalisé en réutilisant certaines parties de dossiers médicaux électroniques produits par un programme de gestion de données médicales (Médidoc) développé par la société Datasoft Management. Ce programme est utilisé par des médecins civils dans le cadre de leurs consultations privées.

Nous disposons donc au total de 239 671 diagnostics médicaux. Cependant, dans une perspective d'analyse et de modélisation linguistique, il était impensable d'observer les deux corpus dans leur ensemble. C'est pourquoi un premier échantillon de 2 343 expressions a été créé sur base des corpus ABL et Médidoc, dans lesquels chaque diagnostic avait, au préalable, été étiqueté à l'aide d'information concernant son origine, la langue et l'année de rédaction.

Ensuite, la validité linguistique et médicale de ces expressions a été vérifiée respectivement par des linguistes et des médecins. Il en a résulté un échantillon final de 1 362 expressions valides, utilisé pour *i*) l'élaboration de lexiques et de grammaires électroniques ainsi que pour *ii*) le testing du prototype ANTHEM au fil de sa réalisation. On trouvera dans Deville et Herbigniaux (1994) une description détaillée des principes méthodologiques appliqués dans l'élaboration des corpus de diagnostics médicaux du projet ANTHEM. C'est notamment sur base de cet échantillon final que le modèle de représentation sémantique des diagnostics médicaux a été élaboré et validé. Nous décrivons brièvement ce modèle dans la section suivante.

3. Modèle de représentation sémantique

Dans la tradition des grammaires casuelles, et plus spécialement de la grammaire fonctionnelle de Dik (1989), nous appelons *formule* la représentation sémantique d'un diagnostic. Une formule est une structure qui consiste en un *prédicat* et un nombre déterminé de *termes*, arguments du prédicat. Un prédicat est une expression (le plus

souvent un groupe nominal, adjectif ou verbe) reflétant les propriétés sémantiques de ses arguments ainsi que les relations sémantiques entre ces arguments. Un terme est une expression (le plus souvent un groupe nominal ou prépositionnel) faisant référence à une entité ou un ensemble d'entités dans un univers conceptuel de sous-langage (un sous-langage étant défini ici comme un ensemble d'expressions faisant référence à un domaine conceptuel limité et défini, et utilisé dans une fonction spécifique). Par conséquent, une formule fait référence à une *configuration d'univers de sous-langage*. Une configuration d'univers de sous-langage est une constellation d'entités conceptuelles du sous-langage exprimées en termes de leurs relations mutuelles (Deville, 1989).

Dans le cadre d'ANTHEM, les entités conceptuelles du sous-langage étudié sont définies comme étant les unités minimales qui ne font pas seulement référence à des éléments atomiques (*bras*) et complexes (*paume main gauche*), mais également à des états (*inflammation*), des processus (*tomber*) et des actes (*ingestion*). Plus précisément, les termes du langage d'application d'ANTHEM font référence à des objets, et les prédicats à des états, relations, actions et procès.

3.1. Typologie de prédicats

Les prédicats sont sélectionnés parmi une liste finie de *types sémantiques*. Un type sémantique capture les propriétés sémantiques et combinatoires prototypiques qui sont partagées par un ensemble de prédicats. La plupart des types sémantiques repris dans le modèle de représentation d'ANTHEM sont hérités d'une nomenclature standard largement répandue dans le domaine de la terminologie médicale : SNOMED (*Systematized Nomenclature of Medicine*) (CAP, 1993). Cette terminologie est reprise dans le tableau 1 ci-dessous.

disdia	disease/diagnosis
morpho	morphology
funct	function
topo	topography
modifiers	modifiers & general linkage
modalities	cfr. modifiers & general linkage
livor	living organisms
chemic	chemicals, drugs ...
physic	physical agents ...
proced	procedures
soct	social context
occup	occupations
hum	human (type ajouté)
temp	temporal (type ajouté)

TABLEAU 1 : Types sémantiques du modèle ANTHEM, selon la nomenclature SNOMED International.

3.2. Système casuel

Dans une formule, la relation entre le prédicat et ses arguments est spécifiée au moyen d'un *cas*. Un cas est l'expression d'une fonction ou d'un rôle sémantique prototypique rempli par le terme d'un prédicat dans la classe sémantique dont est issu ce prédicat. La *structure casuelle* d'un type sémantique est la séquence de cas nécessaires à la définition d'un ensemble de configurations d'univers de sous-langage, tel que représenté par ce type sémantique. Ainsi, un prédicat et ses arguments font référence à une configuration d'univers de sous-langage particulier, comme illustré dans l'exemple ci-dessous, qui reprend la représentation sémantique de l'expression *rupture du ménisque interne genou droit* :

```
(Rupture [Meniscus {medial} [Knee {right}]])
```

Un type sémantique associé à sa structure casuelle fait référence, à un niveau conceptuel plus prototypique, à une classe de configurations d'univers de sous-langage, comme le montre l'exemple ci-dessous. Une telle structure de 'haut niveau' spécifie les rôles sémantiques des arguments du prédicat, en relation avec le type sémantique correspondant à ce dernier :

```
(MORPHO -loc-> [TOPO {POSIT} -loc-> [TOPO {BODSID}]])
```

Une formule peut être étendue au moyen d'arguments périphériques. Les arguments périphériques ne participent pas en tant que tels à la définition d'une configuration d'univers de sous-langage, mais en expriment les dimensions spatio-temporelles, spécifient les entités secondaires qui participent à la configuration, ou encore précisent la manière, les conditions dans lesquelles cette configuration a lieu. Contrairement aux arguments centraux, les fonctions sémantiques des arguments périphériques ne sont pas nécessaires à la définition d'un ensemble de configurations en termes d'un type sémantique et de sa structure casuelle associée. La figure 1 illustre la représentation sémantique de l'expression *rupture du ménisc· ·e interne genou droit*, selon le modèle décrit plus haut.

Pour comprendre comment *i*) le modèle décrit plus haut et *ii*) les lexiques parallèles français et néerlandais sont implémentés dans le prototype ANTHEM, il est nécessaire d'en décrire brièvement l'architecture dans la section 4.

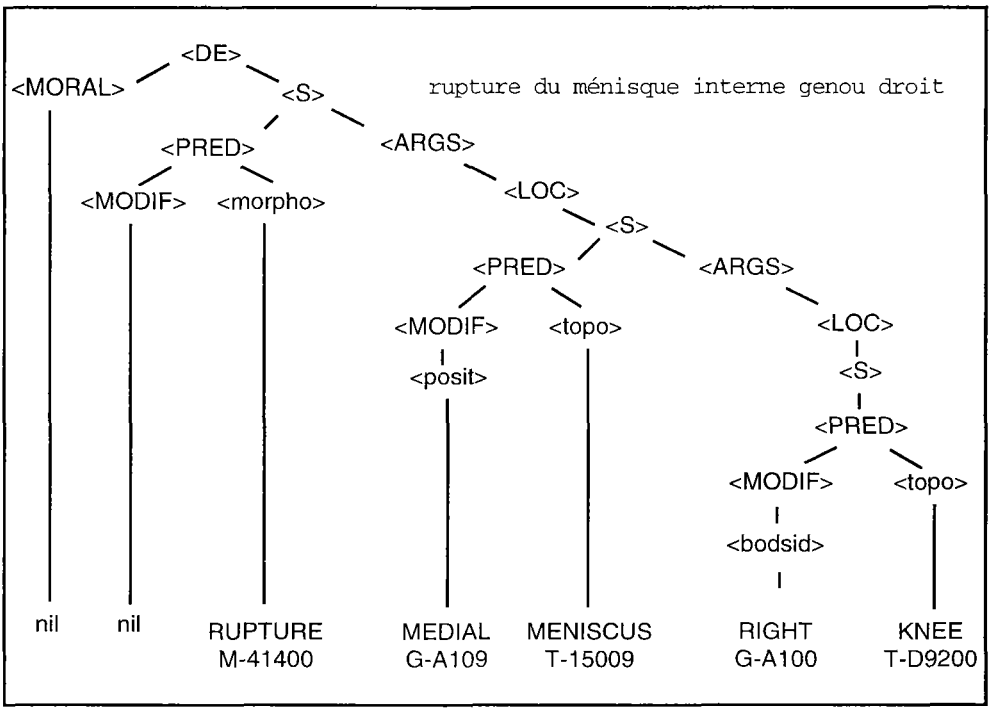


FIGURE 1: Représentation sémantique de l'expression
RUPTURE DU MENISQUE INTERNE GENOU DROIT.

4. Architecture du prototype ANTHEM

Le prototype ANTHEM est conçu pour pouvoir être intégré dans tout type d'application médicale existante, appelée application hôte. Rappelons que la fonction de ce prototype est de traduire des diagnostics médicaux exprimés en langage naturel, soit en une langue naturelle cible, un code ICD-10 ou une représentation sémantique. Pour les besoins de traduction automatique, le prototype fait appel au système de traduction CAT2, développé par l'IAI de l'Université de Saarbrücken, en marge du projet EUROTRA (Streiter *et al.*, 1994). En outre, l'encodage automatique de diagnostics vers le code ICD-10 a nécessité l'élaboration d'un système expert par le Laboratoire de Recherche en Informatique et Télématique Médicale de l'Hôpital Universitaire de Gand. Pour assurer l'interaction entre l'application hôte d'une part, et le système CAT2 et le système expert d'autre part, une interface (appelée *Application Programming Interface*) a été réalisé par le Centre de Recherche Public du Centre Universitaire de Luxembourg (Mousel et Thienpont, 1994). La figure 2 illustre l'architecture générale du prototype ANTHEM :

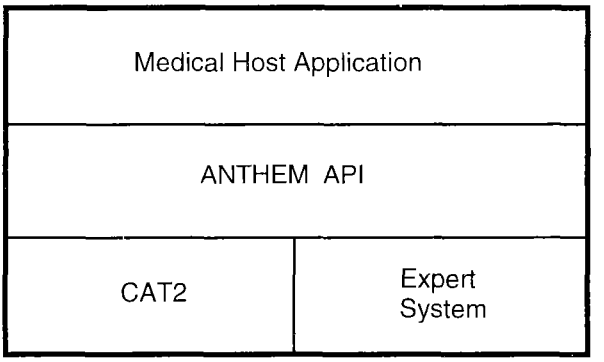


FIGURE 2 · Architecture générale du prototype ANTHEM

Avant d'aborder l'implémentation des lexiques et grammaires électroniques, nous exposons les principes de fonctionnement de CAT2. CAT2 est un système de traduction 'généraliste' basé sur l'unification de traits, et écrit en Prolog. Il nécessite *i*) la réalisation de modules de grammaire propres à chaque langue ainsi que *ii*) le développement de lexiques électroniques, également spécifiques à chaque langue. Les règles de grammaires sont en fait appliquées aux entrées conceptuelles présentes dans les lexiques et plus précisément à celles dont il est fait usage dans le diagnostic médical entré par l'utilisateur. Ces règles sont principalement de trois ordres, à savoir morphologique, syntaxique et sémantique.

Dans le cadre du projet ANTHEM, CAT2 a été adapté en vue de *i*) la traduction de diagnostics en langage naturel, ou *ii*) la transformation de ces expressions en une représentation sémantique indépendante de toute langue. C'est pour répondre à ces objectifs que des lexiques et grammaires électroniques spécifiques ont été implémentés. Au stade actuel du projet, des grammaires et des lexiques ont été produits pour le français, le néerlandais et, dans une moindre mesure, pour l'allemand. Ajoutons que l'essentiel du modèle de représentation sémantique du prototype ANTHEM est implémenté dans un module de grammaire commun à toutes les langues. Nous exposerons en détail dans la section 5 l'élaboration des lexiques électroniques dans ANTHEM.

5. Développement, extension et maintenance de lexiques électroniques

Dans cette partie, nous aborderons *i*) le développement de lexiques basé sur les corpus décrits plus haut, *ii*) les effets de l'intégration d'une composante morphologique plus élaborée dans le prototype ANTHEM sur l'extension et la maintenance des lexiques, *iii*) la typologie d'entrées lexicales actuellement en vigueur dans le projet ANTHEM et *iv*) l'extension semi-automatique du lexique, sur base de la version électronique du système de codification SNOMED International.

5.1. Développement de lexiques basé sur les corpus

Une des caractéristiques fondamentales des lexiques développés dans le cadre du projet ANTHEM est leur organisation que l'on peut qualifier de conceptuelle. En effet,

des termes comme *poumon* et *pulmonaire* sont considérés comme les réalisations substantivale et adjectivale d'un seul et même concept, celui de POUMON. Une telle approche conceptuelle des lexiques ANTHEM résulte en fait de l'organisation – également conceptuelle – du système international de codification SNOMED, utilisé comme interlangue dans notre projet.

Initialement, les lexiques français et néerlandais ont été développés sur base de quatre échantillons-tests, contenant chacun 50 expressions. Au cours de cette première phase, 215 concepts ont été implémentés dans chaque langue. Ensuite, toujours sur base d'expressions issues des corpus, le nombre d'entrées lexicales/conceptuelles a été porté à 586, dans le but de couvrir l'intégralité des 1 362 expressions sélectionnées au hasard et considérées comme valides, tant d'un point de vue linguistique que médical (échantillon final).

Notons toutefois que le lexique allemand a été élaboré et étendu sur base des entrées déjà présentes dans les lexiques français et néerlandais.

5.2. Intégration d'un composant morphologique plus élaboré

Au cours du projet, il a été décidé d'intégrer au prototype ANTHEM un module d'analyse et de génération morphologiques plus performant, dans le but de résoudre trois difficultés, à savoir *i*) l'analyse et la génération automatiques de formes plurielles (substantifs) et fléchies (adjectifs) (morphosyntaxe), *ii*) l'intégration d'informations utiles au traitement de la composition, de la négation et de la gradation (adjectifs) de termes en néerlandais (morphologie productive) et enfin *iii*) la reconnaissance automatique de variantes orthographiques en néerlandais (uniquement en analyse).

L'intégration d'un tel module nous a amené à optimiser le développement des lexiques ANTHEM. Initialement, l'information linguistiquement et médicalement relevante pour la constitution de ces lexiques était enregistrée sous forme de tables statiques. La génération des lexiques consistait alors à remplir manuellement des structures vides écrites en Prolog, à l'aide de l'information contenue dans ces tables. Outre son caractère répétitif, le principal inconvénient d'une telle approche est l'absence de lien actif et systématique entre les tables et les lexiques, la mise à jour de ces derniers nécessitant une double modification (tables et lexiques). D'autre part, toute modification affectant le format proprement-dit des entrées lexicales/conceptuelles (structures Prolog) doit être répétée quasi manuellement partout où cela s'avère nécessaire.

Afin de remédier au décalage sans cesse croissant entre les tables et les lexiques, dûs aux fréquentes modifications du lexiques sans modifications équivalentes dans les tables, nous avons procédé à l'extraction automatique de l'information relevante des lexiques, en vue de constituer une version tout à fait à jour des bases de données. Notons que cette procédure ne remédiait pas encore aux inconvénients exposés plus haut.

C'est alors que nous avons opté pour une approche en sens inverse, à savoir de constituer les lexiques de manière automatique sur base d'information linguistiquement et médicalement pertinente contenue dans des bases de données évolutives, dans la mesure où les modifications touchant aux entrées conceptuelles proprement-dites sont uniquement effectuées dans ces bases. À cette fin, une série de générateurs de

lexique ont été développés en langage de programmation AWK (Aho *et al.*, 1988), afin de pouvoir, à partir de ces bases de données, créer et adapter automatiquement des entrées conceptuelles de plusieurs types, tant en français qu'en néerlandais.

Les avantages d'une telle approche sont au nombre de quatre : tout d'abord, le développeur de tels lexiques électroniques peut se concentrer uniquement sur l'acquisition et la maintenance d'informations linguistiquement et médicalement relevantes ; ensuite, on obtient un haut degré d'adaptabilité du format d'implémentation ainsi qu'une portabilité éventuelle de l'information pertinente vers d'autres formalismes ; on peut également générer de manière quasi automatique de grandes quantités d'entrées lexicales/conceptuelles ; enfin, on peut mettre au point une série de procédures de vérification automatique de cohérence de l'information. Il est en outre possible d'effectuer rapidement divers relevés statistiques en observant les données selon différents axes. Notons que ces procédures de vérification et ces programmes de relevés statistiques sont également écrits en AWK. Dans la section suivante, nous examinons en détail l'organisation interne des lexiques électroniques d'ANTHEM.

5.3. Typologie d'entrées lexicales

Actuellement, quatre types d'entrées sont utilisés dans les lexiques ANTHEM, à savoir *i*) les termes simples, *ii*) les nombres – cardinaux et ordinaux –, *iii*) les pseudo-composés et *iv*) les *multi word units* (concepts constitués de plusieurs mots séparés par des blancs). À cette fin, quatre générateurs de lexique ont été développés pour chaque langue, en vue de générer automatiquement des entrées *i*) simples, *ii*) numérales, *iii*) pseudo-composées et *iv*) à mots multiples.

Notons encore que, parmi les entrées simples, on distingue des concepts réalisés syntaxiquement au moyen *i*) d'un ou plusieurs substantif(s), *ii*) d'un ou plusieurs adjectif(s) ou encore *iii*) d'un ou plusieurs substantif(s) et d'un ou plusieurs adjectif(s).

Parmi les entrées à mots multiples, on distingue actuellement *i*) les entrées constituées de deux mots (le plus souvent un substantif et un adjectif) et *ii*) celles constituées de trois mots (le plus souvent un substantif, une préposition et un nom propre). Nous présentons brièvement ci-dessous quelques exemples illustrant les principaux types d'entrées lexicales en français.

5.3.1. Entrées simples

Dans les exemples ci-dessous, *after_effect*, *chronic* et *fracture* sont utilisés comme aide-mémoire par le développeur. Ces étiquettes ne sont pas traitées par le système CAT2. Les valeurs *F-01460*, *G-A270* et *M-12000* (variable *lex*=) sont les codes SNOMED correspondant aux concepts de SÉQUELLE, CHRONIQUE et FRACTURE. Les valeurs *funct*, *progr* et *morpho* (variable *type*=) indiquent la catégorie sémantique – inspirée de la typologie SNOMED – à laquelle les différents concepts appartiennent. Outre la réalisation syntaxique des différents concepts en français (variable *string*=), nous trouvons aussi de l'information concernant la catégorie syntaxique (variable *cat*=), le genre (variable *gen*=) et le profile morphologique des différents mots implémentés (variable *flex*=).

```
after_effect =
{lex='F-01460',type=funct,
 known=yes,
 head={cat=n,ehead={cat=n}}}>>
({string=se1quelle,
 onset=cons,
 flex=normal,
 head={cat=n,ehead={gen=fem}}}).[].

chronic =
{lex='G-A270',type=progr,
 known=yes,
 head={cat=a,ehead={cat=a}}}>>
({string=chronique,
 onset=cons,
 flex=normal,
 head={cat=a,apos=post}}).[].

fracture =
{lex='M-12000',type=morpho,
 known=yes,
 head=({cat=n,ehead={cat=n}}
      ;{cat=a,ehead={cat=a}})}>>
({string=fracture,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}
;{string=fract,
  anasyn=ana,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}
;{string=casse1,
  onset=cons,
  flex=normal,
  head={cat=a,apos=post}}
;{string=fracture1,
  anasyn=ana,
  onset=cons,
  flex=normal,
  head={cat=a,apos=post}}).[].
```

5.3.2. Entrées numérales

Les principes exposés ci-dessus restent d'application pour les entrées de type numérales, bien que le code SNOMED soit remplacé ici par un code plus transparent (variable *lex=*). Il est à noter que les réalisations cardinales et ordinales renvoient au même concept.

```
ruleEIGHTEEN =
 {lex='18',type=quant,clex=nil,
  known=yes,
  head={cat=a,ehead={cat=a}}}>>
```

```
({string='dix-huit',
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string=xviii,
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}).[].

ruleEIGHTEENTH =
 {lex='18',type=ind,clex=nil,
 known=yes,
 head={cat=a,ehead={cat=a}}}>>
({string='dix-huitie2me',
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string='18e2me',
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string='18e',
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string=xviiie2me,
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string=xviiie,
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string='dix-huit',
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=post}}
;{string=xviii,
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=post}}).[].
```

5.3.3. Entrées pseudo-composées

Les entrées de type pseudo-composé, ainsi que celles à mots multiples présentées plus bas, sont certainement les plus intéressantes. Le principe des entrées pseudo-composées

est de rendre explicite la structure conceptuelle interne d'un mot donné. On ne procède cependant à une telle explicitation que si celle-ci est motivée tant d'un point de vue médical que linguistique. En effet, la majorité des mots d'un sous-langage donné peuvent se décomposer en une structure conceptuelle complexe. Dans le sous-langage des diagnostics médicaux, par exemple, on trouve de nombreux affixes qui ont une signification particulière à l'intérieur des mots dans lesquels ils apparaissent. Par exemple, le suffixe *-ite* (en français) ou *-itis* (en néerlandais) signifie très souvent *inflammation localisée dans/affectant 'xxx'*, 'xxx' étant exprimé dans le(s) premier(s) morphème(s) du mot.

Mais, dans le projet ANTHEM, seuls les termes nécessitant une telle explicitation de leur structure conceptuelle interne sont implémentés sous forme de pseudo-composés. Nous appelons ces termes pseudo-composés afin de les distinguer d'un autre type de composés, à savoir les composés tels que *laryngotrachéite*, dans lesquels on peut aisément reconnaître *laryngite* (ou *larynx*) et *trachéite*, et qui, dès lors, peuvent être traités automatiquement par l'analyseur morphologique du système CAT2.

Le terme *trigéminie* (inflammation du nerf facial appelé trijumeau) illustre bien le principe des entrées de type pseudo-composé. Dans cet exemple, il a fallu rendre explicite dans le lexique la structure conceptuelle de ce terme car on rencontre dans le sous-langage médical des expressions telles que *trigéminie gauche*. Or, pour interpréter correctement le terme *gauche* dans cette expression, il faut savoir que l'on réfère en fait au *nerf trijumeau gauche* (par opposition au droit), concept non réalisé grammaticalement ici, suite à un phénomène d'ellipse 'conceptuelle'. Le modificateur *gauche* ne qualifie donc pas le terme *trigéminie* au même titre que les adjectifs *aigüe*, *récidivante* ou *bénigne*, par exemple. L'entrée lexicale du concept TRIGÉMINIE aura donc la forme suivante :

```
inflammation_in_trigeminalnerve =
{lex='M-40000',type=morpho,
clex={lex='T-A8150',type=topo,
      role=loc,head={ehead={cat=n,pform=in,feature=in}}},
known=yes,
head={cat=n,ehead={cat=n}}>>
({string=trigeminie,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}).[].
```

Dans l'exemple ci-dessus, le code *M-40000* identifie le concept d'INFLAMMATION, tandis que le code *T-A8150* renvoie au concept de NERF TRIJUMEAU. Le lien entre les deux est exprimé à l'aide de la valeur *loc* (variable *role=*) ainsi que des paramètres *pform=in* et *feature=in* dont nous n'expliquerons pas le fonctionnement ici.

D'autres exemples d'entrées pseudo-composées sont : *gonarthrose* (arthrose localisée dans le genou), *tenniselbow* (inflammation localisée dans le coude), *épicondylite* (inflammation localisée dans l'épicondyle) ou encore *gonalgie* (inflammation localisée dans le genou).

5.3.4. Entrées à mots multiples

Dans cette sous-section, on trouvera, à titre d'exemple, l'implémentation des entrées à mots multiples référant aux concepts de CRÊTE ILIAQUE (substantif + adjectif) ainsi que de MALADIE DE SCHEUERMANN (substantif + préposition + nom propre).

```
crista_iliaca =
 {lex='T-1234A',type=topo,
  known=yes,
  head={cat=n,ehead={cat=n,gen=GEN,num=NUM}},
  mwu1={lex=iliaca,head={ehead={gen=GEN,num=NUM}}}}>>
({string=cre3te,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}).[].

iliaca =
 {lex=iliaca,type=mwu,role=mwu,
  known=yes,
  head={cat=a,ehead={cat=a}}}>>
({string=iliaque,
  onset=vowel,
  flex=normal,
  head={cat=a,apos=post}}).[].
```

Dans l'exemple ci-dessus (*crête iliaque*), le terme principal est *crête* (en fait 'cre3te', les accents étant implémentés sous la forme de combinaisons voyelle + chiffre). Le concept CRÊTE ILIAQUE, de type TOPO (variable *type=*), est identifié par le code SNOMED T-1234A (variable *lex=*).

Le lien entre le substantif 'crête' (*cat=s*) et l'adjectif 'iliaque' (*cat=a*) est ensuite établi à l'aide de la variable *mwu1=*, comportant entre autres, dans sa structure sous-jacente, le trait *lex=iliaca*, aide-mémoire utilisé pour désigner l'implémentation du terme 'iliaque', de type MWU (variable *type=*).

L'accord en genre et en nombre entre le substantif (*crête*) et l'adjectif (*iliaque*) est garanti par les traits *gen=GEN* et *num=NUM*, dans lesquels *GEN* et *NUM* sont des variables dont les valeurs doivent être identiques pour *crête* (terme principal) et pour *iliaque* (mwu).

D'autres exemples d'entrées à mots multiples constituées d'un substantif et d'un adjectif sont : *ongle incarné, nodule froid, corps étranger, nœud atrioventriculaire, veine saphène, tronc basilaire, globule rouge, globule blanc* ou encore *cage thoracique.*

L'implémentation du concept de MALADIE DE SCHEUERMANN illustre un second type d'entrée à mots multiples :

```
scheuermann_disease =
 {lex='D1-61210',type=disdia,
  known=yes,
  head={cat=n,ehead={cat=n}},
  mwu1={lex=de},
  mwu2={lex=scheuermann}}>>
({string=maladie,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}
;{string=syndrome,
  anasyn=ana,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=masc}}}).[].

de =
 {lex=de,type=mwu,role=mwu,
  known=yes,
  head={cat=mwu,ehead={cat=mwu}}}>>
({string=de,
  head={cat=mwu}}).[].

scheuermann =
 {lex=scheuermann,type=mwu,role=mwu,
  known=yes,
  head={cat=mwu,ehead={cat=mwu}}}>>
({string='Scheuermann',
  onset=cons,
  head={cat=mwu}}).[].
```

Dans l'exemple ci-dessus, le terme principal est *maladie/syndrome*. Le concept de MALADIE DE SCHEUERMANN, de type DISDIA (variable *type=*), est identifié par le code SNOMED D1-61210 (variable *lex=*).

Le lien entre le substantif 'maladie'/'syndrome' (*cat=s*) et les autres éléments, à savoir la préposition 'de' et le nom propre 'Scheuermann' (ayant tous deux les traits *type=mwu* et *cat=mwu*) est ensuite établi à l'aide des variables *mwu1* et *mwu2*, comportant respectivement les traits *lex=de* et *lex=scheuermann*, aides-mémoires utilisés pour désigner l'implémentation des termes 'de' et 'Scheuermann'.

Notons enfin qu'aucune variation morpho-syntaxique n'est prévue, ni pour *de*, ni pour *Scheuermann*. En outre, l'implémentation de la préposition *de* est utilisée dans plusieurs entrées à mots multiples, comme par exemple les *maladies* et/ou *syndromes de Mallory-Weiss, Becker* ou *Crohn*, la *fracture de Pouteau-Colles* ou encore la *maladie de système*, l'*accident de voiture*, la *tête de radius*, l'*accès de panique*, le *bouchon de cérumen*, l'*eczéma de contact*, la *fracture de stress*, le *mal de tête*, l'*hydrate de carbone* ou l'*acétate de désoxycorticostérone*.

5.4. Extension semi-automatique du lexique

Dans un premier temps, les lexiques français et néerlandais ont été développés dans le

but de pouvoir analyser et générer un nombre croissant d'expressions issues d'échantillons représentatifs des corpus ABL et Médidoc. C'est ainsi qu'un total de 586 entrées simples avait été implémenté.

Dans la phase d'extension intensive des lexiques ANTHEM, nous avons opté pour une certaine systématisation dans la création de nouvelles entrées. Nous avons en effet décidé d'implémenter le plus possible d'entrées simples ou atomaires appartenant essentiellement aux cinq catégories de la typologie SNOMED les plus fréquemment utilisées, à savoir DISDIA (maladies/affections), MORPHO (morphologie), TOPO (topographies/endroits du corps), GLMOD (modificateurs) et FUNCT (fonctions).

Outre le développement systématique des axes prototypiques de la nomenclature SNOMED International, le principal avantage d'une telle approche par rapport à une démarche inspirée de l'observation des corpus est évident : pour tout mot nouveau apparaissant dans un ordre aléatoire, il n'est plus nécessaire de rechercher le code correspondant au concept que ce mot réalise, ni la catégorie sémantique à laquelle ce concept appartient. Ces informations sont en effet présentes de manière systématique dans la version électronique de SNOMED dont nous sommes partis pour l'extension semi-automatique du lexique.

La phase d'extension intensive des lexiques a permis l'implémentation de 467 entrées simples appartenant à la catégorie DISDIA, 550 à la catégorie MORPHO, 342 à la catégorie TOPO, 376 à la catégorie GLMOD et 1 146 à la catégorie FUNCT, soit un total de 3 590 entrées, auxquelles se sont ajoutées 26 entrées pseudo-composées et 115 entrées à mots multiples, tant en français qu'en néerlandais.

6. Conclusion et perspectives futures

Dans cet article, nous avons exposé l'approche adoptée dans le développement des ressources lexicales multilingues du prototype ANTHEM, approche qui a nécessité une analyse minutieuse de grands corpus multilingues représentatifs de diagnostics médicaux.

Nous avons décrit les travaux de mise en forme et d'échantillonnage effectués sur les corpus de diagnostics médicaux. Nous avons également présenté le modèle de représentation sémantique élaboré sur base des observations effectuées sur les échantillons et exposé l'architecture du prototype ANTHEM. Enfin, nous avons abordé le développement, l'extension et la maintenance proprement-dits des lexiques électroniques parallèles français et néerlandais dans les différentes phases de leur élaboration, dans une perspective d'automatisation de la plupart de ces tâches.

Au travers du projet ANTHEM, nous avons tenté d'identifier les fondements d'une modélisation adéquate du langage naturel dans le domaine des soins de santé, en mettant l'accent sur la nécessité d'une validation à la fois empirique, linguistique et opératoire des modèles obtenus. Une telle approche nous semble répondre aux exigences de la multidisciplinarité évoquée plus haut : en effet, la modélisation lexicale du langage d'application d'ANTHEM *i*) est élaborée à partir d'expressions générées

dans un environnement clinique en vraie grandeur, *ii*) met en œuvre les dimensions linguistiques caractérisant la plupart des 'États de Choses' grammaticalisés dans tout diagnostic médical, et *iii*) est exprimé à l'aide d'un formalisme permettant à toute expression en langage naturel d'être directement 'calculable' par un module de traduction automatique.

Pour conclure, nous évoquons brièvement quelques perspectives futures du projet ANTHEM. Tout d'abord, le modèle de représentation sémantique du langage de l'application sera sans doute encore raffiné. Les éléments nécessaires pour ce raffinement seront obtenus essentiellement grâce aux validations internes et sur site du prototype ANTHEM, qui sera alors confronté à de nouveaux diagnostics médicaux issus des corpus ABL et Médidoc, ainsi qu'au milieu médical en vraie grandeur. Enfin, les possibilités d'extension vers d'autres projets de recherche connexes ou d'autres applications sont dès à présent envisagées avec le plus grand intérêt.

IDAREX : description formelle des expressions à mots multiples en français et en allemand dans le cadre de la technologie des états finis*

Frédérique SEGOND et Élisabeth BREIDT

Rank Xerox Research Centre, Meylan, France, et Université Tübingen, Allemagne

1. Introduction

La plupart des textes regorgent d'expressions à mots multiples. Ces expressions ne peuvent être correctement comprises, encore moins traduites, si elles ne sont pas reconnues en tant qu'unités lexicales complexes. Ces expressions, appelons-les **lexèmes à mots multiples** (LMM), englobent aussi bien les expressions idiomatiques (*se casser la tête sur quelque chose*), les proverbes (*Bien mal acquis ne profite jamais*), les constructions à verbe support (*prendre la parole*) que les collocations lexicales et grammaticales (*au vu de*). Nous présentons ici une méthode permettant de repérer de telles expressions dans un texte. Ce travail est fait pour le français et pour l'allemand. D'abord nous illustrons à l'aide d'exemples comment les LMMs, tout en obéissant à la syntaxe générale, ont un certain nombre de restrictions qui leur sont attachées en fonction des variations lexicales et/ou structurelles qu'ils autorisent : comment les LMMs sont régis par une syntaxe qui leur est propre. Nous nous proposons ensuite d'encoder le comportement intrinsèque des LMMs dans des **règles de grammaire locales**. L'implantation de ces grammaires locales est réalisée dans le cadre général des techniques à états finis (Karttunen et Yampol, 1993 et Karttunen et Beesley, 1992) à l'aide du formalisme à deux niveaux IDAREX (Segond et Tapanainen, 1995 et Tapanainen, 1994). Cet environnement, qui s'avère rapide et efficace, permet d'encoder les LMMs comme des expressions régulières. Ces règles de grammaires locales ont été utilisées dans le cadre du projet COMPASS (Adaptation de dictionnaires bilingues pour l'aide à la compréhension en ligne)[1].

* Nous tenons à remercier Jean-Pierre Chanod et Alain Lecomte pour leurs commentaires.
1. Projet de recherche LRE 62-080, financé par la CEE

2. Variabilité des LMMS

Certaines LMMs apparaissent toujours dans une forme donnée et sont donc facilement identifiables par leurs seuls éléments lexicaux. C'est le cas de *de fil en aiguille* ou de *par acquis de conscience*. Cependant la plupart des LMMs autorisent différents types de variations et de modifications[2]. Pour reconnaître de tels LMMs dans les textes il faut être en mesure d'identifier précisément celles de leurs occurrences qui dévient de la forme standard ou forme de base. Ainsi il faut être attentif aux différentes variations morphologiques, à l'ordre des mots ou encore à l'insertion de composants. Par exemple, si l'on considère l'expression « *casser sa pipe* », on s'aperçoit que le nom *pipe* ne peut être mis au pluriel, que le verbe *casser* ne peut être remplacé par un de ses synonymes (par exemple *briser*), et que la phrase ne peut être mise au passif sans que le sens de l'expression idiomatique soit perdu. Pourtant le verbe lui-même peut être conjugué. De même l'expression allemande *die Beine in die Hand nehmen* (prendre ses jambes à son cou) n'autorise aucune variation lexicale, le nom *Hand* ne peut être mis au pluriel, enfin la phrase ne peut être ni topicalisée, ni passivisée. Les exemples qui suivent montrent que l'expression idiomatique *sich über etwas den Kopf zerbrechen* (se casser la tête sur quelque chose) admet plusieurs variations. Les points de suspension indiquent les positions où l'insertion d'adjoints est possible.

Über diese Sache [...] *zerbrach* Jan *sich* schon lange [...] *den Kopf.*
Deswegen zerbrach sich Jan schon lange [...] *den Kopf.*
Sich darüber [...] *den Kopf* zu *zerbrechen*, lohnt sich nicht.
Jan *zerbricht sich* selten [...] *den Kopf über* solche Dinge.
Jan *zerbricht sich* nur *über* wenige Dinge im Leben [...] *den Kopf.*

De même on peut insérer un adverbe dans l'expression française *peser dans la balance* et obtenir ainsi les phrases suivantes : *peser lourd dans la balance* ou *peser beaucoup dans la balance*. En revanche l'insertion d'un GN dans cette phrase entraîne la perte du sens idiomatique : *peser les fruits dans la balance*. Le même type de phénomène se passe lorsqu'une expression idiomatique est incluse dans une phrase plus grande. L'expression *tomber du ciel* garde son sens idiomatique dans le cas d'une insertion d'adverbe (*ça tombe vraiment du ciel*) mais l'interprétation idiomatique échoue si l'on insert un GN comme dans la construction causative : *C'est la condensation qui fait tomber la pluie du ciel.*

Les types les plus communs de variations et de modifications pour lesquels nous donnons des exemples en section 4, sont les suivants[3] :

- **variations lexicales**
  variant lexical ; variation dans la réalisation des arguments ; échange des constituants ;

- **variations morpho-syntaxiques**
  variation en nombre ; variation casuelle (LMMs nominaux) (toujours possible en

2 Par *variations* nous entendons un échange ou une restructuration syntaxique des composants , par *modifications* nous entendons l'ajout de mots, de modifieurs, aux LMMs.
3 Voir également Fleischer (1982), Brundage *et al* (1992) et Engelke (1994)

allemand) ; forme comparative ou superlative pour un adjectif ; variation du déterminant ; variation de la personne du verbe ; variation du temps du verbe ; composition d'un élément nominal ;

- **modifications**
  modification adverbiale ; modification adjectivale ; négation ;

- **variations structurelles**
  passivisation ; topicalisation ; scrambling ; ordre des mots pour VI/V2/V finaux.

Sauf en ce qui concerne la variation de l'ordre des mots, les variations structurelles sont extrêmement restreintes. De même, la composition nominale et l'échange des constituants sont pratiquement toujours impossible. Nous supposons donc par défaut qu'aucune des variations citées précédemment n'est possible et que nous pouvons explicitement lister les exceptions. Les autres variations, telles que le nombre et le cas des noms et des adjectifs, la personne et le temps des verbes sont en revanche très courantes et seront aussi exprimées simplement dans le formalisme.

En plus de la variabilité qui doit être prise en compte dans la description des LMMs, il apparaît que certains LMMs enfreignent les règles d'accord et de gouvernement (valence du verbe déviante, nom quantifiable sans déterminant, adjectif ou préposition sans nom qui lui soit attaché). Une telle caractéristique peut être prise en compte, mais n'est pas nécessaire pour reconnaître les LMMs dans un texte.

Les exemples précédents font apparaître clairement l'insuffisance des méthodes de reconnaissance de patrons pour le repérage des LMMs. En effet, les LMMs autorisent des variations qui, pour la plupart, sont mal définies au niveau lexicographique. Une entrée classique de dictionnaire fournit généralement une seule forme pour l'expression – pas nécessairement la forme de base ou canonique –, sans aucune autre précision quant aux variations autorisées, mis à part, quelquefois, pour les variantes lexicales. Or cette information peut être encodée dans des règles de grammaire locales, règles qui ont un pouvoir d'expression plus grand que les descriptions traditionnelles.

Si on les compare à des règles de grammaire générales, les règles de grammaire locales décrivent implicitement les restrictions des LMMs. Ces restrictions énoncent les variations autorisées pour les LMMs par comparaison au cas, par défaut, où ils sont complètement figés. Dans le cas par défaut, toutes les restrictions s'appliquent, *i.e.* aucune variation n'est permise et le LMM est représenté par sa forme de surface dans laquelle tous les composants sont figés et ordonnés. Les violations des règles de grammaire standard, p. ex. constituants manquants ou accords enfreints, n'ont pas besoin d'être explicitées. Cependant elles peuvent être décrites lorsqu'elles importent pour la distinction entre sens idiomatique et sens littéral d'un LMM.

D'abord nous donnons brièvement le formalisme utilisé. Ensuite, nous décrivons, par le menu, comment sont exprimés dans les grammaires locales, les différents types de variations des LMMs.

3. IDAREX : Formalisme pour décrire la variabilité des LMMs

Les règles de grammaire locales sont écrites à l'aide du formalisme à deux niveaux

IDAREX[4] partie intégrante du compilateur à états finis développé au centre de recherches de Rank Xerox[5].

Les LMMs sont codés comme des expressions régulières en accord avec les notations décrites ci-dessous.

3.1. Les mots

Les mots sont représentés à deux niveaux : un niveau lexical et un niveau de surface. Les deux points servent de séparateur entre les deux niveaux. Il y a quatre descriptions de base possibles pour un mot.

1. :forme-de-surface
2. :forme-de-surface variable morphologique:
3. forme-de-base variable morphologique:
4. variable-classe-de-mot

Les deux premiers cas permettent la description des formes figées. Par exemple, la forme figée *pédales* dans l'expression *perdre les pédales* peut être encodée de deux manières :

1. *:pédales*
2. *:pédales Noun:*

Le troisième cas permet la description des formes variables. Dans l'exemple précédent le verbe *perdre* peut apparaître à n'importe quels temps, nombre et personne. Il est donc codé *perdre Verb:* ou la variable Verb stipule que toutes les réalisations verbales du mot *perdre* sont autorisées.

Le dernier cas permet la description de variables générales représentant des classes de mots. Ainsi nous pouvons par exemple définir une variable ADV regroupant les adverbes et les expressions adverbiales. Nous sommes maintenant en mesure d'écrire la règle de grammaire locale complète pour l'expression idiomatique précédente :

perdre Verb: ADV :les :pédales;*

3.2. Les opérateurs

Un ensemble d'opérateurs permet de combiner entre elles les descriptions des mots. Parmi eux :

- *rien* — les mots se succèdent les uns aux autres ;
- *parenthèses* () — marquent une partie optionnelle de l'expression idiomatique ;

4. **ID**ioms **A**s **R**egular **Ex**pressions.
5 Pour une description plus détaillée du formalisme nous renvoyons le lecteur à Karttunen et Yampol (1993) et Segond et Tapanainen (1995).

- *étoile de Kleene* * — marque que la chaîne qui la précède peut apparaître n'importe quel nombre de fois, y compris aucune ;
- *plus* + — marque que la chaîne qui la précède peut apparaître une ou plusieurs fois ;
- *crochets* [] — groupent une expression ;
- *barre* | — sépare les différentes possibilités ;
- *point virgule* ; — marque la fin d'une expression.

3.3. Les variables fonctionnelles

Enfin, le formalisme permet l'utilisation de variables fonctionnelles ou macros. Les lexicographes ont ainsi la possibilité de décrire, de façon compacte, des phénomènes complexes réguliers. Ces descriptions sont ensuite réécrites par le système. Nous avons utilisé des macros pour décrire, en français, les LMMs dans lesquels apparaissaient des pronoms réflexifs. Par exemple, pour décrire toutes les variations possibles, tout particulièrement les variations temporelles, de l'expression *s'aplatir comme une carpette devant quelqu'un.* Au lieu d'écrire :

REFL [être Verb: ADV* aplatir VPP: | aplatir Verb:] ADV* :comme :une :carpette (:devant NP)

nous écrivons simplement :

REFLEX(aplatir) :comme :une :carpette (:devant NP)

où `REFLEX(verbe)` est utilisé chaque fois qu'un LMM contient une construction réflexive et se réécrit en :

REFL [être Verb: ADV* verbe VPP: | verbe Verb:] ADV*

4. Règles de grammaire locales pour le français et l'allemand

Les règles de grammaire locales que nous proposons ici couvrent au maximum le niveau de la phrase. Elles sont énoncées de la façon la plus générale possible et autorisent la surgénération. Bien que des règles plus restrictives et plus spécifiques puissent être écrites elles ne sont pas nécessaires dans la mesure où l'on fait comme hypothèse de départ qu'il n'y a pas de phrase mal formée en entrée. En effet, peu importe que les règles autorisent plus de variations que celles qui apparaîtront effectivement dans les textes dans la mesure où l'on peut distinguer les emplois idiomatiques des emplois littéraux. Par exemple, étant donné que nous ne prenons pas en compte la représentation sémantique des LMMs. La règle de grammaire locale associée à l'expression française *peser dans la balance* acceptera aussi bien les phrases sémantiquement correctes comme *peser lourd dans la balance* ou *peser énormément dans la balance* que celle difficilement acceptables sémantiquement comme *peser *ardemment dans la balance.*

Certains LMMs ont une variabilité tellement productive qu'il est illusoire de vouloir en rendre compte de façon systématique. De telles variations sont, par leur na-

ture même, imprévisibles. Un exemple de cette productivité est donné par la formation *ad hoc* de mots composés ou par la combinaison de métaphores et d'expressions idiomatiques en allemand comme dans :

das bißchen Kopf, das sie noch haben, zerbrechen sie sich mit ... (ex. de Fleischer, 1982)
← *sich den Kopf zerbrechen* (se casser la tête)
\+ *Köpfchen haben/etwas im Kopf haben* (être très intelligent)

Dans ce qui suit, nous donnons pour le français et l'allemand des exemples de règles de grammaires locales écrites en IDAREX, et ce, pour chacune des variations et des modifications décrites en section 2.

4.1. Variations lexicales

Différents items lexicaux, généralement sémantiquement équivalents, peuvent être autorisés :

F : *perdre la tête/la boule/les pédales ≠ perdre la tronche*
⇒ perdre Verb: [:la :tête | :la :boule | :les :pédales]
A : *eine ruhige/sichere Hand (haben)* ((avoir) la main sûre)
≠ *eine stille Hand (haben)* ((avoir) la main tranquille)
⇒ :eine [:ruhige | :sichere] :Hand

On peut exprimer de façon analogue le fait qu'un argument interne à l'expression idiomatique admette plusieurs réalisations syntaxiques. Dans l'exemple allemand suivant, aussi bien un objet prépositionnel qu'un objet accusatif peuvent être utilisés :

A : *mit den Achseln/die Achseln zucken* (hausser les épaules)
⇒ [[:mit :den :Achseln | :die :Achseln] (ZU) zucken V: | ...]

4.2. Variations morpho-syntaxiques

La variation en nombre pour le nom peut être contrôlée par une variable morphologique (par exemple *Nsg*) associée à un nombre particulier.

F : *la politique de l'autruche ≠ la politique des autruches*
⇒ :la :politique :de :l' :autruche Nsg:
A : *grüne Welle* (l'onde verte) ≠ *grüne Wellen* (les ondes vertes)
⇒ grün A: Welle Nsg:

Dans d'autres cas la variation en nombre est autorisée, comme dans :

F : *comme un coq en pâte*
⇒ :comme un Det: coq N: :en :pâte
A : *verkrachte Existenz(en)*
⇒ verkracht A: Existenz N: (perdant (personne))

Le même phénomène se produit avec d'autres catégories syntaxiques : par exemple, l'emploi du comparatif ou du superlatif va parfois rompre le sens idiomatique, comme c'est le cas dans :

F : *faire table rase ≠ faire table plus rase*
⇒ faire Verb: ADV* :table :rase
A : *reinen Tisch machen* (faire table rase) ≠ *reinsten Tisch machen* (faire table plus rase)
⇒ [machen Vfin: (ADV* NPnom) ADV* rein Apos: :Tisch
| rein Apos: :Tisch (ZU) machen V:]
A : *jds. bessere Hälfte* (son meilleur côté) ≠ *gute/beste Hälfte* (bon/meilleur côté)
⇒ POSS gut Acomp: Hälfte N:

Alors que dans d'autre cas le sens idiomatique est conservé, et une variable morphologique moins restrictive peut alors être utilisée (par exemple `A`) :

F : *ses bons/meilleurs côtés*
⇒ POSS bon A: côté N:
A : *schlechter Scherz* (mauvaise plaisanterie) / *der schlechteste Scherz*
⇒ schlecht A: Scherz Nsg:

À un niveau plus syntaxique, certains LMMs admettent n'importe quel **déterminant** (`DET`) alors que d'autres sont plus restrictifs (p. ex. `INDEFDET`).

F : *engueuler quelqu'un comme du/un/des poisson(s) pourri(s)*
⇒ engueuler V: NP :comme INDEFDET poisson N: pourri A:
A : *von einer/der Idee durchdrungen* ((être) omnubilé par une idée)
⇒ :von DET :Idee ADV* durchdrungen A:

La plupart des LMMs verbaux permettent des **variations de personnes** ; pourtant une telle variation est interdite dans certaines expressions prédicatives figées. C'est le cas dans les exemples suivants où seule la troisième personne est autorisée :

F : *les bons comptes font les bons amis*
⇒ :les :bons :comptes faire Vpl3: :les :bons :amis
A *jdm fällt ein Stein vom Herzen* (se sentir soulagé d'un poids)
≠ *jdm fallen Steine vom Herzen* (des pierres tombent du cœur de quelqu'un)
⇒ [fallen Vsg3: (ADV* NPdat) ADV* :ein :Stein :vom .Herzen | (NPdat) ADV* :ein :Stein :vom :Herzen fallen Vsg3:]

Il arrive parfois que l'on ne puisse pas avoir de **variations temporelles**[6]. Dans ce cas, seul le niveau de surface est utilisé.

F : *qui a bu boira*
⇒ :qui :a :bu :boira

6 Cette restriction semble ne concerner que les proverbes

A : *Wasser hat keine Balken* (personne ne peut marcher sur l'eau)
≠ *Wasser hatte keine Balken* (l'eau n'a pas de poutre)
⇒ :Wasser :hat :keine :Balken

4.3. Modifications

Il est souvent possible d'insérer toute une classe syntaxique de mot. Par exemple, **l'insertion d'adverbes** (ADV) est un phénomène courant :

F : *prendre souvent le taureau par les cornes*
⇒ prendre Verb: ADV* :le :taureau :par :les :cornes
A : *Sie spitzt (plötzlich) die Ohren* ((soudain,) elle tend l'oreille)
⇒ [spitzen Vfin: (ADV* NPnom) ADV* :die :Ohren | ...][7]
(besonders) hohes Tier (un (très) grand manitou)
⇒ (ADV) hoch A: Tier Nsg:

Certains LMMs permettent la **modification adjectivale** (ADJ) comme c'est le cas dans :

F : *faire un () crochet*
⇒ faire Verb: :un ADJ :crochet
A : *seine (neugierige) Nase in etwas stecken* (mettre son () nez dans quelque chose)
⇒ [POSS (:neugierige) :Nase ADV* :in NPakk (ZU) stecken V: | ...]

Certains LMMs conservent leur sens idiomatique même sous **l'opération de négation**. Dans les exemples qui suivent ceci est assuré par la variable ADV. En allemand, la négation attributive est également possible et peut être décrite comme un cas de variation lexicale du déterminant.

F : *il ne prend jamais le taureau par les cornes*
⇒ prendre Verb: ADV* :le :taureau :par :les :cornes
A : *(nicht) das Handtuch werfen* (ne pas jeter l'éponge)
⇒ [werfen Vfin: (ADV* NPNom) ADV* :das :Handtuch | ...]
(k)eine (dicke) Lippe riskieren (risquer de dire des choses qu'il ne faut pas dire)
⇒ [[:eine | :keine] (:dicke) :Lippe ADV* (ZU) riskieren V: | ...]

4.4. Variations structurelles

La **passivisation** est prise en compte par l'ordre des mots d'un V-final et la disjonction de différents composants ordonnés différemment.

F : *crever l'abcès*
⇒ [crever Verb: ADV* :l' :abcès | :l' :abcès [avoir Vsg3: :été ADV* | être Vsg3: ADV*] :crevé]]

7. La deuxième partie de l'expression régulière qui prend en compte l'ordre des mots V1/V2 n'est pas mentionnée ici

A : *endlich wurde reiner Tisch gemacht* (finalement, table rase a été faite)
⇒ [rein Apos: :Tisch (ZU) machen V: | ...]

La **topicalisation**, le **scrambling** et l'**ordre des mots pour V1/V2/V finaux** sont tous pris en compte par la disjonction de différents composants ordonnés différemment.

F : *chercher midi à quatorze heures → Midi, il ne le cherchait pas à quatorze heures*
⇒ [chercher Verb: ADV* :midi :à :quatorze :heures | :midi NP ADV :le chercher Vsg3: :à :quatorze :heures]

A : *den Vogel abschießen* (surpasser tout le monde) → *Den Vogel dabei hat dann Jan abgeschossen*
⇒ [:den :Vogel ([:dabei | :bei NPdat]) schießen Vfin: ADV* NPnom ADV* :ab Pref2: | :den :Vogel ([:dabei | :bei NPdat]) Vaux (ADV* NPnom) ADV* abschießen V: | ...]

A : *für etw. den Kopf hinhalten / den Kopf für etw. hinhalten* (payer pour quelqu'un d'autre)
⇒ [DEFPOSS :Kopf ADV* [:für NPakk | :dafür] ADV* hinhalten V: | [:für NPakk | :dafür] ADV* DEFPOSS :Kopf ADV* hinhalten V: | ...]

Le patron général qui permet de rendre compte de la variation de l'ordre des mots en allemand est le suivant :

V1/V2 : _verbe_ Vfin: (ADV* NPnom) ADV* (_libre_comps_) _figé_
V-final : _figé_ ADV* (_libre_comps_) ADV* _verbe_ V:

où `libre_comps` représente tout complément externe à l'expression idiomatique et `_figé` représente les parties fixes de l'expression idiomatique qui reste à côté du `_verbe_`.

Les variations structurelles ne sont, pour l'instant, pas encodées pour le français alors que certaines d'entre elles le sont pour l'allemand. Cependant elles soulèvent un point intéressant quant aux règles de grammaire locales et au formalisme des états finis. C'est ce sur quoi nous nous penchons maintenant.

4.5. Pouvoir d'expression des règles de grammaire locales

4.5.1. L'ordre des mots

La plupart des variations et des modifications permettant de distinguer les LMMs de séquences totalement figées s'expriment aisément et naturellement à l'aide des règles de grammaire locales. Cependant, rendre compte de la variation de l'ordre des mots et des phénomènes qui y sont liés, comme la topicalisation et le scrambling, est une tâche plus délicate. Or tous ces phénomènes sont courants en allemand et il est donc primordial de pouvoir en fournir la description la plus appropriée possible. Ici, le pouvoir d'expression des règles de grammaire locales semble atteindre ses limites dans la mesure où il ne s'agit plus de décrire des phénomènes « locaux ».

Pour décrire tous les LMMs verbaux allemands avec IDAREX et dans le cadre de la technologie des états finis en général, il faut rendre compte de tous les agencements possibles des constituants, y compris toutes les positions où (modification externe) des adverbes peuvent être insérés. Une telle description peut être longue et laborieuse pour les verbes qui admettent des compléments au datif et à l'accusatif, et ce plus particulièrement, lorsque topicalisation et scrambling sont autorisés. Non seulement de telles expressions sont pénibles à lire et à décrire, mais en plus la compilation des réseaux qui leur sont associés est longue. D'autre part, la définition de variables pour les éléments pouvant être insérés nécessite une description partielle de la syntaxe de l'allemand, tout particulièrement pour ce qui concerne les constructions GN et GP.

Une alternative pragmatique et moins coûteuse du point de vue informatique consiste à regrouper dans une même variable ANY tous les constituants pouvant apparaître entre le verbe et ses compléments figés, y compris les arguments externes non idiomatiques. Bien que, dans certains cas, cette approche conduise à l'identification erronée de patrons, elle est assez fiable pour avoir été utilisée avec succès dans le cadre de COMPASS.

4.5.2. L'accord

Pour certains LMMs il est nécessaire de contrôler l'accord entre les différents constituants. Ainsi l'expression française *casser sa pipe* perd son sens idiomatique si le réflexif n'est pas à la même personne que le sujet. Un exemple analogue est donné par l'expression allemande *seine Meinung ändern* (changer d'avis). L'accord est marqué explicitement dans les règles (variable particulière) bien que l'implantation n'en fasse, pour le moment, pas l'usage.

Les deux phénomènes décrits précédemment, l'ordre des mots et l'accord, n'appartiennent plus à la classe des phénomènes locaux. Idéalement un mécanisme plus puissant serait nécessaire pour décrire leur régularité d'une façon générale. Pourtant, notre approche a été en mesure de couvrir au moins les cas les moins complexes.

5. Les applications d'IDAREX

Identifier les LMMs est essentiel pour tout traitement du langage naturel basé sur des informations lexicales : ceci est vrai pour des applications comme la concordance ou l'indexation intelligente, la traduction automatique, ou la consultation automatique de dictionnaires. Dans cet article, nous avons montré comment leur description peut être faite à l'aide d'IDAREX, en construisant des règles de grammaire locales exprimées sous la forme d'expressions régulières dans le formalisme des états finis.

Les règles de grammaires locales ici décrites sont utilisée dans LOCOLEX[8], l'outil automatique d'aide à la compréhension développé chez Rank Xerox, et utilisé

8. Pour une description plus complète de LOCOLEX le lecteur pourra consulter Bauer, Segond et Zaenem (1995).

dans le cadre du projet COMPASS. Son principal propos est de fournir à l'utilisateur une consultation automatique en contexte d'un dictionnaire de compréhension bilingue. Imaginons, par exemple, un anglophone ayant une certaine connaissance de l'allemand qui lit un texte électronique en allemand et à qui il manque certains mots du texte. Lorsqu'il clique sur un mot inconnu COMPASS renvoie non pas à l'entrée du dictionnaire dans son entier mais uniquement à la partie de l'entrée pouvant aider à la compréhension du mot dans ce contexte bien précis. Par exemple, COMPASS donnera uniquement les traductions associées à la partie du discours appropriée ou, dans le cas d'un LMM, la traduction associée à l'expression. Pour réaliser cela nous avons enrichi des dictionnaires électroniques[9] avec des règles de grammaire locales.

Un certain nombre d'applications existantes peuvent être améliorées grâce à l'utilisation de règles de grammaire locales. Ainsi les grammaires des dates peuvent améliorer les performances des systèmes optiques de reconnaissance des caractères.

Plus généralement, les règles de grammaire locales sont utiles à l'analyse syntaxique, p. ex. la description d'expressions adverbiales complexes telles que les dates en français (*le lundi 21 août au matin*)[10] ou toute expression n'obéissant pas à la syntaxe générale. Nombreux sont les cas où l'analyseur syntaxique échouera tout simplement parce qu'incapable d'analyser correctement le LMM inclus dans une phrase plus large. Par exemple, en allemand, la syntaxe générale demande qu'un nom quantifiable soit précédé d'un déterminant. Cette règle est enfreinte dans le LMM *von Haus aus* (originale).

La prochaine étape pour l'amélioration de l'analyse syntaxique consiste à incorporer les règles de grammaire locales dans un composant syntaxique général.

En ce qui concerne la technique que nous utilisons, le formalisme à deux niveaux dans un système à états finis ainsi qu'IDAREX, elle a l'avantage de fournir une représentation compacte. Comme nous l'avons vu elle donne la possibilité de définir des variables générales telles que « n'importe quel adverbe » (ADV) ou des variables morphologiques plus spécifiques telles que « seulement la troisième personne singulier du verbe » (Vsg3). Cela soulage le lexicographe de la pénible tâche d'expliciter toutes les formes possibles. Les variables fonctionnelles, ou macros, permettent d'exprimer des généralisations pour des patrons qui sont attachés à toute une classe de mots. D'autre part, les deux niveaux nous permettent d'exprimer des faits soit sur la forme de surface, soit sur la forme lexicale. Ainsi lorsque l'on veut exprimer qu'une forme est figée on n'utilise que le niveau de surface, évitant par là même, de se perdre dans tous les traits du niveau lexical.

Cette technologie permet d'effectuer des opérations sur les réseaux engendrés par les expressions régulières : addition, soustraction, intersection et composition. Bien que nous n'ayons pas encore utilisé cette possibilité dans notre travail, elle donne un grand pouvoir d'expression. Imaginons, par exemple, que nous souhaitions

9 Dans ce projet nous avons utilisé les dictionnaires suivants : le *Dictionnaire anglais-français Oxford-Hachette* (1994) et le *Dictionnaire allemand-anglais Harper-Collins* (1991). Nous sommes particulièrement reconnaissantes aux éditeurs qui ont accepté de nous fournir ces dictionnaires à des fins de recherche.

10 Un traitement analogue pour les adverbes de date en français est proposé par Denis Maurel (1993).

prendre en compte la sémantique des LMMs et pour ce faire nous ayons besoin de restreindre les règles écrites. Une possibilité qui s'offre à nous consiste à écrire de nouvelles expressions régulières et de soustraire les réseaux ainsi générés à ceux construits précédemment. De telles expressions régulières pourraient ainsi décrire la compatibilité entre les classes sémantiques des noms et des adjectifs.

Lexicographie bilingue informatisée au quotidien : témoignage du rédacteur face à l'écran

Thomas SZENDE et Dominique RADANYI

INALCO, Paris et CIEH, Université de Paris III, France

• *Abstract* •

The only Hungarian-French dictionary in use today is obsolete, i.e. *its "hungarocentrism" is obvious ; it is full of archaic meanings, lacks semantic indicators, and so on. It seemed essential to create a new dictionary both for Hungarian and French language communities, taking into account their particular needs, and to give every possible information – semantic, grammatical, stylistic and even cultural – on the lexical units and their considerations in discourse. Compiling this dictionary required the organization of a technical structure allowing it not to be a static product but a tool open to reactualization and research procedures. This paper introduces the computer tools used during the various steps of the compiling work and which will eventually be used in the electronical version of the dictionary. 1) data collecting ; headword listing (WORD-CRUNCHER) ; 2) article editing (WRITER STATION) ; 3) dictionary database (PAT).*

Nous allons présenter quelques aspects informatiques et pratiques de la préparation d'un nouveau dictionnaire hongrois-français[1].

À l'initiative du professeur *Jean Perrot*, directeur du Centre Interuniversitaire d'Études Hongroises à l'Université de Paris III, deux équipes lexicographiques ont été constituées : une en Hongrie à l'Université de Szeged, sous la direction de *Miklós Pálfy* réalisant la partie français-hongrois et l'autre à Paris, au sein du CIEH pour la partie hongrois-français sous la direction *de Thomas Szende*. L'édition de l'ouvrage serait confiée aux *Éditions Akadémiai Kiadó* (Budapest) et aux *Éditions le Robert* (Paris).

Les seuls ouvrages de référence qui existent actuellement, le dictionnaire d'*Au-*

1. Nous tenons à remercier tous nos collaborateurs (*Joëlle Dufeuilly, Viktória Eröss, Károly Ginter, Emilie Molnos, Jean-Léon Muller et Péter Zimonyi*), et plus particulièrement *Chantal Philippe* qui outre ses activités de lexicographe prend en charge la gestion informatique des fichiers au sein de l'équipe parisienne

rélien Sauvageot, publié dans les années trente et celui de *Sándor Eckhardt*, publié dans les années cinquante, sont largement dépassés pour des raisons évidentes. Les dernières générations de hungarophones étudiant le français et de francophones étudiant le hongrois sont largement tributaires de ces deux ouvrages : nous leur devons de bonnes observations sur les deux lexiques mais aussi des erreurs de traduction qu'il faut corriger et des absences douloureuses qu'il faut combler.

Dès le début des travaux, il nous est apparu fondamental qu'un nouveau dictionnaire :

- reflète fidèlement l'état actuel des deux langues dans leurs différents registres ;
- soit conçu en fonction des besoins des deux communautés linguistiques, hungarophone et francophone ;
- donne un maximum de renseignements sémantiques, grammaticaux, stylistiques et même culturels sur les unités lexicales retenues et leur fonctionnement dans le discours.

De plus, le contraste entre une langue comme le français et le hongrois, langue agglutinante, appartenant à la famille des langues finno-ougriennes, renforçait pour nous le besoin d'un dictionnaire nouveau, réalisé à partir de cette approche lexicographique particulière.

Un dictionnaire bilingue n'est pas la description lexicale exhaustive de deux états de langue, ni la reproduction fidèle des innombrables réalisations concrètes des discours en langue source et en langue cible. Il a pour fonction de mettre en parallèle les lexiques des deux langues, en apportant à l'utilisateur, à travers un nombre limité d'exemples pertinents, le moyen de produire des énoncés les plus naturels possible et d'éviter au maximum des erreurs d'interprétation.

À cette fin, l'ouvrage envisagé devra enregistrer un vocabulaire général d'environ 40 000 à 50 000 mots ; il sera complété par des lexiques spécialisés. Il va de soi que la nomenclature de ce dictionnaire contiendra obligatoirement un noyau de mots usuels, pleinement représentatif du fonds culturel, mais aussi de très nombreux termes illustrant la richesse et la diversité des langues actuelles constamment nourries des apports nouveaux de la civilisation.

Ancrée, engluée dans de longues traditions, la lexicographie se renouvelle aujourd'hui grâce à l'ordinateur[2]. Elle absorbe l'outil informatique à plusieurs niveaux ; aussi bien au niveau de l'analyse préliminaire du matériel de référence qu'à celui de la constitution du texte dictionnairique proprement dit et de la production de textes imprimés par la composition automatisée et programmée.

2 P Imbs l'a déjà constaté il y a un quart de siècle « Dans l'état actuel du monde, il nous a semblé que la machine était encore serve du livre et non pas son substitut elle déleste l'homme de tâches serviles, notamment dans le domaine de la documentation, qu'elle aide à maîtriser lorsqu'elle est. . surabondante et constamment foisonnante , elle peut aussi l'aider à poser et à résoudre des problèmes de nature quantitative ou même qualitative, en lui fournissant p ex des listages qui, pour grossière que soit leur approche, n'en facilitent pas moins les analyses fines, qui sont l'essence même de la science exacte » *Trésor de la Langue Française*, CNRS, 1971, p XIII.

La mise à profit de l'informatique suppose une analyse rigoureuse et exhaustive de la démarche méthodique du lexicographe, analyse qui non seulement tient compte mais encore explique la totalité des décisions prises par le lexicographe au cours de sa production[3]. L'informatique permet d'enregistrer l'ensemble des articles du dictionnaire, de les stocker dans une base de données, de prévoir et d'harmoniser toute intervention du lexicographe.

Les différentes structures d'articles, ainsi que les variantes probables doivent être identifiées, classifiées et répertoriées ; toutes les sections du dictionnaire sont ainsi « étiquetées ». Ce travail doit aboutir à la rédaction d'une grammaire formelle décrivant la structure du dictionnaire avec des codes identificateurs de données à chaque changement typographique, ce qui donne une structuration arborescente des articles.

En standardisant les différents champs constitutifs de chaque type d'article, il devient possible de faire appel à des procédures de recherche et d'interrogation.

L'idée même d'aide informatique à la rédaction d'un dictionnaire bilingue repose sur le principe d'une décomposition de l'ensemble du processus lexicographique en opérations et en parcelles d'opérations.

Autrement dit, on précise non seulement le contenu du dictionnaire mais aussi, et parallèlement, la structure à donner à ce contenu.

L'élaboration de notre dictionnaire par les deux équipes a nécessité la mise en place d'une structure technique et informatique permettant :

- d'une part, que le manuscrit en gestation et, plus tard, l'ouvrage soient l'objet d'une mise à jour permanente et,
- d'autre part, qu'une version électronique puisse être envisagée.

L'ensemble de programmation doit permettre de prendre en charge les exigences spécifiques d'un dictionnaire bilingue, susceptible d'assurer la cohérence et l'homogénéité de la nomenclature et d'automatiser au mieux chacune des démarches que comporte la rédaction d'un tel ouvrage.

À cette fin, nous exploitons quotidiennement deux outils informatiques conçus pour des PC (retenus après de nombreux essais infructueux avec différents logiciels).

- Pour la collecte de données et l'établissement de la nomenclature : *WORDCRUNCHER* qui permet :

  - la recherche des éléments en contexte dans le corpus (ce contexte peut se réduire à quelques lignes ou être élargi à une page-écran dont la ligne centrale contient l'occurrence recherchée) ;

3. « L'ordinateur étant idiot, on ne peut pas s'en servir sans tout lui expliquer, ce qui suppose un immense travail métalexicographique jamais encore fourni.» F J Hausmann, « La métalexicographie à l'échelle mondiale », *Coloquio de Lexicografía*, Verba, Anexo 29, Universidade de Santiago de Compostela, 1988, pp 79-109

- de connaître d'une manière très précise les différentes nuances d'un terme ou de cerner l'emploi des termes que l'on ne trouve pas dans les dictionnaires existants ; et
- de générer automatiquement des tableaux de fréquences lexicales.
- Pour la saisie et la rédaction des articles : *WRITER STATION* sur lequel nous reviendrons ultérieurement.

Nous avons également à notre disposition le logiciel *PAT* pour toute consultation de la base de données réunissant l'ensemble des articles déjà rédigés par les deux équipes. Ce logiciel est accessible à l'Institut de Linguistique de l'Académie des Sciences de Budapest. Nous devons en effet la mise au point de cet ensemble informatique à *Júlia Pajzs*, chercheur dans cet institut.

Notre méthode, décrite sous tous ses angles dans un protocole de rédaction, se perfectionne sans arrêt à l'usage. Nous l'avons voulue à la fois stricte et souple afin qu'elle soit valable pour le plus grand nombre de cas possibles.

Les spécifications de la rédaction ne pouvaient pas être fixées *a priori*. Seul un processus itératif d'essais pouvait permettre, en partant de spécifications initiales forcément approximatives, d'aboutir à des spécifications adéquates. La rédaction de plusieurs centaines d'articles tests, entre 1991 et 1993, a mis en évidence les insuffisances de la « grammaire » auxquelles il a fallu remédier.

La grammaire en question fait appel au langage standard et généralisé de marquage qui porte le nom de SGML.

Le travail de rédaction proprement dit a commencé ainsi en septembre 1993 et continue grâce à une équipe stable composée de linguistes bilingues et de traducteurs.

Voici très rapidement la manière dont nous avons organisé le travail de rédaction.

Dans une première phase, les rédacteurs hungarophones préparent la nomenclature comportant :

- les unités lexicales destinées à apparaître dans le corps du dictionnaire comme vedettes ;
- les exemples qui devraient illustrer leur emploi et leur place dans le discours ;
- les locutions figées les plus courantes autour du mot-vedette ;
- les indications sémantiques en langue hongroise nécessaires à la distinction des sens ;
- les marques d'emploi relatives au domaine et registre d'utilisation.

Dans une deuxième phase, les rédacteurs francophones :

- sélectionnent les équivalents les plus pertinents ;
- proposent une traduction des exemples et des locutions figées ;
- ajoutent s'il y a lieu des indications sémantiques et les marques d'emploi pour la partie française ;
- suggèrent des modifications dans la structure des articles en fonction des critères sémantiques du français.

Dans une troisième phase, les rédacteurs hungarophones et francophones réexaminent ensemble chaque projet d'article et adoptent une version quasi définitive.

La connaissance de trois langues est indispensable pour la rédaction de ce dictionnaire bilingue : langue source (hongrois) + langue cible (français) + langue du logiciel.

En effet, l'exploitation de ce logiciel n'étant pas une affaire d'informaticiens, il est crucial que des non-spécialistes comme nous puissent la maîtriser sans trop de difficulté.

Le système a dû être organisé de façon à ce que le commun des mortels puisse le contrôler, quelles que soient sa langue maternelle et ses compétences en informatique.

Grâce à WRITER STATION, nous bénéficions d'un tel système, fondé sur le dialogue, dans un environnement interactif, convivial, adapté à nos besoins ponctuels.

Les yeux fixés sur le masque visible dans la partie inférieure de l'écran, le claviste, qui est toujours lexicographe en même temps, dispose d'un menu pour chaque étape de son travail rédactionnel.

Voici la structuration la plus simple du cheminement informatique : il s'agit d'un article ne comportant qu'une vedette hongroise, sa catégorie grammaticale et un seul équivalent français avec son genre grammatical :

ARTICLE/ARTICLE RENVOI
|
ENT
|
VEDETTE
\
FIN VDT
\
FIN ENT/PHON/**CATÉGORIE**/RCT/DDS/RDL/LIG
|
FIN CGR/RCT/DDS/RDL/LIG
|
FIN ENT/RCT/DDS/RDL/LIG
/
FIN ART/BLS/BGR/LOC
|
FIN BLS/GEN/BEQ/DDS/RDL/LIG/IDS/RCT/EXP
\
ÉQUIVALENT
|
FIN EQV
|
GENRE/DDS/RDL/LIG/IDS/FIN BEQ/EQV
|
FIN GEN
\
DDS/RDL/LIG/IDS/FIN BEQ/EQV
\
FIN BLS/RCT/BEQ/EXP/LOC
|
FIN ARTICLE

Guidé par WRITER STATION tout au long de la saisie de l'article, le lexicographe n'est concerné que par le haut de l'écran où s'effectue son travail.

Mais, comme la pratique lexicographique le montre, il arrive rarement qu'à un terme A de la langue source corresponde précisément un terme B de la langue cible. Étant donné cette complexité de la réalité linguistique, nous avons recours à différents types d'usages du logiciel. En voici quelques exemples présentés sous leurs deux aspects :

– non formaté avec les jalons textuels qui ouvrent et ferment les champs successifs ; et
– formaté, tels qu'ils apparaîtront dans la version papier du dictionnaire.

L'article *alátét* ne comporte qu'un seul bloc sémantique, à l'intérieur duquel des indications sémantiques distinguent les différents équivalents :

<ART><ENT><VDT>alátét </VDT><CGR>n </CGR></ENT><BLS><IDS>(íróasztal) </IDS><BEQ><EQV>sous-main </EQV><GEN>m; </GEN></BEQ><BEQ> <IDS>(csavar) </IDS><EQV>rondelle </EQV><GEN>f; </GEN></BEQ><BEQ> <IDS>(pohár) </IDS><EQV>dessous </EQV><GEN>m </GEN><EQV>de verre; </EQV> </BEQ><BEQ><IDS>(edény) </IDS><EQV>dessous-de-plat </EQV><GEN>m; </GEN></BEQ><BEQ><IDS>(étkészlet) </IDS><EQV>set </EQV><GEN>m </GEN>(de table) </BEQ></BLS></ART>

alátét *n* **1** (*íróasztal*) sous-main *m* ; (*csavar*) rondelle *f* ; (*pohár*) dessous *m* de verre ; (*edény*) dessous-de-plat *m* ; (*étkészlet*) set *m* (de table)

L'article *alatti* se divise en plusieurs blocs sémantiques ; les divisions sont justifiées par des indications sémantiques et illustrées par des exemples :

<ART><ENT><VDT>alatti </VDT><CGR>adj </CGR></ENT><BLS><IDS> (hely) </IDS><EXP> kép ~ szöveg </EXP><TRD>légende <IDS>(d'un dessin, d'une photo); </IDS></TRD><EXP>vminek a víz ~ része </EXP><TRD>la partie immergée de qc; </TRD><EXP>Béke utca 5. (szám) ~ ház </EXP><TRD>l'immeuble du/situé au 5 rue Béke; </TRD><EXP>az elsõ pont ~ rendelkezések </EXP> <TRD>les dispositions mentionnées au point un</TRD></BLS><BLS><IDS>(idõ) </IDS><EXP>óra ~ beszélgetés </EXP><TRD>bavardage pendant les cours; </TRD> <EXP>a háború ~ nélkülözések </EXP><TRD>les privations de la guerre </TRD></BLS><BLS><IDS>(szint) </IDS><EXP>tíz fok ~ hõmérséklet </EXP> <TRD>température inférieure à dix degrés; </TRD><EXP>két perc ~ eredmény </EXP> <TRD> un temps inférieur à deux minutes </TRD></BLS></ART>

alatti *adj* **1** (*hely*) **kép ~ szöveg** légende (*d'un dessin, d'une photo*) ; **vminek a víz ~ része** la partie immergée de qc; **Béke utca 5. (szám) ~ ház** l'immeuble du/situé au 5 rue Béke; **az elsõ pont ~ rendelkezések** les dispositions mentionnées au point un **2** (*idő*) **óra ~ beszélgetés** bavardage pendant les cours; **a háború ~ nélkülözések** les privations de la guerre **3** (*szint*) **tíz fok ~ hőmérséklet** température inférieure à dix degrés; **két perc ~ eredmény** un temps inférieur à deux minutes

L'article *barátkozik* est du même type mais comporte en plus des renseignements grammaticaux (rection) et stylistiques ; dans le premier bloc, nous avons indiqué par des chevrons que l'équivalent n'était qu'approximatif ; parmi les exemples on trouve également une locution figée précédée d'un marquage spécifique, le dièse :

<ART><ENT><VDT>barátkozik </VDT><CGR>v intr </CGR></ENT><BLS><RCT> ~ vkivel </RCT><BEQ><EQV><se faire un/des ami(s)> : </EQV></BEQ><EXP> könnyen ~ </EXP><TRD>se lier facilement; </TRD><EXP>nehezen ~ </EXP> <TRD>avoir du mal à se faire des amis; </TRD><TRD>être peu liant; </TRD> <EXP>csak lányokkal ~ </EXP><TRD>il n'a que des amies filles; </TRD><EXP> mív vészekkel ~ </EXP><TRD>fréquenter des artistes </TRD><LFG># <EXP> fy vel-fával ~ <RDL>péj </RDL></EXP><TRD>il fraye/se lie avec n'importe qui </TRD></LFG></BLS><BLS><RCT> ~ vmivel </RCT><EXP> ~ a gondolattal, hogy </EXP><TRD>se faire à l'idée que </TRD></BLS></ART>

barátkozik *v intr* **1** ~ **VKIVEL** <se faire un/des ami(s)> : **könnyen** ~ se lier facilement; **nehezen** ~ avoir du mal à se faire des amis; être peu liant; **csak lányokkal** ~ il n'a que des amies filles; **művészekkel** ~ fréquenter des artistes # **fűvel-fával** ~ ***péj*** il fraye/se lie avec n'importe qui **2** ~ **VMIVEL** ~ **a gondolattal, hogy** se faire à l'idée que

Les deux articles *bár* présentent un cas d'homonymie. D'autre part, l'un des deux comprend plusieurs blocs grammaticaux :

<ART><ENT><VDT>1 bár </VDT><CGR>n </CGR></ENT><BLS><BEQ><EQV>bar </EQV><GEN>m; </GEN><EQV>boîte </EQV><GEN>f </GEN><EQV>de nuit </EQV></BEQ></BLS></ART>

1 bár *n* **1** bar *m*; boîte *f* de nuit

<ART><ENT><VDT>2 bár </VDT></ENT><BGR><CGR>conj </CGR><BLS> <BEQ><EQV>bien que +subj; </EQV><EQV>quoique +subj; </EQV><EQV>encore que +subj; </EQV><EQV>alors que : </EQV></BEQ><EXP> ~ nem szép, mégis sokan udvarolnak neki </EXP><TRD>bien qu'elle ne soit pas (vraiment) belle/une beauté, elle est très courtisée; </TRD><EXP>segítek rajta, ~ nem érdemli </EXP> <TRD>je l'aide, quoiqu'il ne le mérite pas; </TRD><EXP> ~ részletes, mégsem teljes a felsorolás </EXP><TRD>bien que détaillée, la liste est incomplète </TRD> </BLS></BGR><BGR><CGR>adv </CGR><BLS><BEQ><EQV>pourvu que +subj; </EQV><EQV>si seulement : </EQV></BEQ><EXP> ~ így lenne ! </EXP> <TRD>pourvu que cela se passe ainsi !; </TRD><EXP> ~ ne tette volna ! </EXP><TRD>si seulement il n'avait pas fait cela ! </TRD></BLS></BGR></ART>

2 bár I *conj* **1** bien que +subj; quoique +subj; encore que +subj; alors que : ~ **nem szép, mégis sokan udvarolnak neki** bien qu'elle ne soit pas (vraiment) belle/une beauté, elle est très courtisée; **segítek rajta,** ~ **nem érdemli** je l'aide, quoiqu'il ne le mérite pas; ~ **részletes, mégsem teljes a felsorolás** bien que détaillée, la liste est incomplète **II** *adv* **1** pourvu que +subj; si seulement : ~ **így lenne !** pourvu que cela se passe ainsi !; ~ **ne tette volna !** si seulement il n'avait pas fait cela !

À l'instar du phénomène appelé *clôture lexicale* (Kittredge, 1983) – qui signifie grosso modo que le nombre de nouveaux termes rencontrés dans une nouvelle page

diminue rapidement et tend vers zéro ou une valeur très faible quand les pages augmentent –, on peut parler d'une *clôture rédactionnelle* : au fur et à mesure que le temps passe, se profilent des préférences lexicales, grammaticales, conceptuelles courantes dans la mise en rapport des deux langues.

Le rédacteur prend conscience de sa marge de manœuvre à l'intérieur du système.

Ainsi, notre pratique quotidienne de WRITER STATION révèle différentes fonctions à l'usage.

- Une première fonction pourrait s'appeler *traitement de texte amélioré*. Le logiciel nous permet tout rajout, reformulation et correction, et améliore les conditions et la vitesse d'exécution des opérations.
- Une deuxième fonction concerne *l'optimisation des manipulations* les plus diverses : à chaque étape, la situation du lexicographe est définie. On obtient ainsi des gains de qualité et de productivité considérables.
- Une troisième fonction, *l'aide documentaire* permet l'homogénéité des exemples et des traductions, notamment par le biais de procédures, telles que : consultation, recherches, analyse et extraction d'informations.
- Une quatrième fonction, *le contrôle*, intervient en cas de dérapage ou d'oubli : un clignotement en jaune dans la fenêtre texte rappelle l'utilisateur à l'ordre et permet ainsi une maîtrise permanente de la rédaction : le logiciel reconnaît facilement qu'un article est non conforme et les erreurs de manipulation n'empêchent pas le déroulement normal de la procédure rédactionnelle.
- Une cinquième et dernière fonction, *la gestion*, autorise l'appréciation des volumes de textes traités et la constitution rationnelle des fichiers (Gouadec, 1994 : 59-74).

Une telle saisie des articles aboutit à un *système clos et ouvert* à la fois :

- le système est clos dans la mesure où tout est répertorié : on prend conscience des éventuels écarts ou informations déficitaires, et on est appelé à combler les manques ;
- le système est ouvert car l'informatisation permet toutes sortes de réorganisations cohérentes des informations données et autorise à tout moment des radioscopies insolites.

Une autre spécificité de la saisie informatique des articles est ce qu'on pourrait nommer *appel de l'équivalent* : une fois le champ ouvert, le lexicographe est tenu d'obéir au système et de compléter le canevas préétabli en fournissant une traduction, même si rien ne garantit que l'emploi qui sera fait dans les textes des mots enregistrés sera limité aux équivalents et exemples présentés.

Ceci implique une constante nécessité de réagir. Dans l'idéal, une traduction devrait rendre le sens exact, mais aussi la connotation, telle ou telle allusion ou référence culturelle, ou tels effets pouvant se situer au niveau du signifiant, comme des allitérations, par exemple. À supposer que, pris isolément, chacun de ces aspects est traduisible, tout n'est cependant pas transmissible. En principe, il est exclu qu'il y ait dans les ressources de la langue cible une équivalence où se retrouvent justement tous les aspects qui coïncident dans l'unité lexicale de la langue source.

De cette façon, nous espérons réduire au minimum les échecs de consultation, les faux-sens, les contresens et les recherches infructueuses.

L'ordinateur est un dispositif qui aide la rédaction, mais l'activité traduisante du lexicographe demeure traditionnelle : il oublie les signifiants de la langue source, tout en retenant les éléments de signification pour pouvoir les faire réapparaître grâce à des signifiants nouveaux dans la langue cible. La machine se contente d'assister la traduction humaine. C'est toujours le lexicographe qui sélectionne, traduit, agence, même si les sources à sa disposition sont plus que jamais multiples. Pondérer, filtrer, nuancer, il n'y a que le linguiste et le traducteur qui puissent le faire.

D'autre part, il nous a été impossible de prévoir et de faire apparaître tous les aspects de l'organisation interne des articles. Le masque est figé, alors que la langue est extensible.

Un des enseignements principaux que l'on tire de la pratique lexicographique est que les langues naturelles sont d'une texture à défier les formalisations et les systématisations les plus ingénieuses. En effet, les langues naturelles sont faites :

a) d'analogies et d'anomalies ;
b) de polymorphies et de polysémies ;
c) de redondances et de déficiences ;
d) d'explicitations et d'implications ;
e) de constantes et de variantes.

Le rédacteur doit donc compenser en permanence les défaillances du système. Il intervient à la fois en amont (il faut s'assurer que les articles qu'il crée soient traitables par la machine) et en aval : il faut rectifier les erreurs de l'ordinateur. Il y a nécessairement des formes « interdites » que la machine considère comme une anomalie. Il est donc nécessaire d'adapter les articles aux contraintes du logiciel.

Sans vouloir être en totale contradiction avec l'esprit même de ce colloque, nous devons avouer que la rédaction d'un dictionnaire bilingue reste un travail éminemment artisanal.

On dit souvent que la situation de tout traducteur est, par essence, exceptionnelle. De la même façon, nous avons observé que chaque cas lexicographique est unique. Les informaticiens ont résolu de nombreux problèmes posés par la technique lexicographique, mais pas les problèmes posés par les langues elles-mêmes.

Cependant, l'informatique a un impact extrêmement fort sur la réflexion méthodologique. Grâce à elle, notre discours sur le dictionnaire et notre pratique de la lexicographie bilingue sont devenus algorithmiques.

Abréviations :

ART = ARTICLE
ENT = ENTRÉE
BEQ = BLOC ÉQUIVALENT

BGR = BLOC GRAMMATICAL
BLS = BLOC SÉMANTIQUE
CGR = CATÉGORIE GRAMMATICALE
DDS = DOMAINE DE SPÉCIALITÉ
EQV = ÉQUIVALENT
EXP = EXEMPLE
GEN = GENRE
IDS = INDICATION SÉMANTIQUE
LFG = LOCUTION FIGÉE
LIG = LIMITATION GÉOGRAPHIQUE
LOC = LOCUTION FIGÉE
PHON = TRANSCRIPTION PHONÉTIQUE
RCT = RECTION
RDL = REGISTRE DE LANGUE
TRD = TRADUCTION
VDT = VEDETTE

Orientation de combinants dans les langues de spécialité : comparaison entre l'anglais et le français

Patricia Thomas et Frank Knowles

Terminologue indépendante, Cranleigh et Aston University, Birmingham, Grande-Bretagne

Introduction

Cette analyse vise à établir ce que nous nommons l'*orientation* d'une collocation. Le terme *orientation* est pertinent pour deux raisons : premièrement, du point de vue dictionnairique, le terminologue doit indiquer en entrée une partie de la collocation et, deuxièmement, l'entrée est, pour le terminologue, le principal point de référence lors de la collecte des données. Les groupes d'utilisateurs doivent pouvoir sélectionner les données selon leurs besoins. Étant donné le nombre fini d'unités lexicales qui composent une phrase, c'est donc la deuxième tâche qui est plus difficile à cause du nombre infini d'utilisateurs (même de groupes d'utilisateurs).

1. L'orientation – comment la définir et l'identifier ?

L'orientation a pour but de trouver l'entrée d'une phrase afin que le terminologue puisse l'inclure dans le dictionnaire. Ceci permet au chercheur, ou au lecteur, de retrouver la phrase dès la première recherche. Un tel choix est basé sur le **contexte** de la phrase ; en ce qui concerne la construction **verbe + proposition**, c'est ou le verbe ou la proposition qui peut être choisi comme entrée. Si la proposition est un terme composé, il faut identifier d'abord l'orientation de celui-ci selon le contexte, qui dépend à son tour des besoins de l'utilisateur. La tâche nécessite alors deux analyses.

2. Le combinant

La collocation que nous avons examinée à fond est celle du **verbe + substantif**. Cette construction est certes fort utile aux traducteurs et aux chercheurs écrivant des articles dans une langue étrangère.

La construction verbe + substantif a été nommée *combinant* en langues de spécialité. Le terme *combinant* permet une plus grande variation que le terme *collocation* du point de vue de la syntaxe (par exemple, certains adjectifs et adverbes font partie intégrale d'une phrase, tandis que d'autres, soi-disant « libres », ne le font pas, p. ex. *the vaccine was genetically-engineered* mais *the woman was attractively dressed*).

3. Les corpora

L'étude se fonde sur l'analyse de deux corpora de sciences biologiques, l'un en français et l'autre en anglais, constitués de livres didactiques et de comptes rendus de colloques sur la virologie et la bactériologie, émanant de grandes maisons d'édition et d'organisations internationales telles que l'OMS et l'OCDE. Chacun des corpora contient un demi-million de mots. Les domaines traités sont donc très restreints et les textes ne sont pas des chimères prototypiques. Cependant, étant donné que ces domaines sont en plein essor, les textes contiennent beaucoup de néologismes. Les analyses effectuées dans les deux langues ont permis de constater, d'un point de vue statistique, les conclusions d'orientation terminologique ou dictionnairique des combinants.

4. L'analysateur de textes

L'analyse des textes a été effectuée au moyen de l'Aston Text Analyser (ATA[1]). Cet analyseur fournit des listes de fréquences de mots ainsi que des concordances qui donnent quatre mots de chaque côté du mot-clé, triés par ordre alphabétique à droite ou à gauche.

Pour l'ATA, ce qui est important, c'est le *profil synoptique* dans lequel le mot-clé est indiqué par un astérisque. Dans le tableau 1, les trois colonnes à gauche et à droite du mot-clé contiennent les mots qui apparaissent en positions -3 à -1 et +1 à +3 respectivement, en ordre décroissant de fréquence.

Du profil synoptique on peut passer directement à la section du texte dans laquelle se trouve le mot-clé ; ces quelques lignes de texte vont donc au-delà du niveau de la phrase.

5. Comparaison des verbes « de spécialité » en anglais et en français

Le pourcentage total de tous les verbes dans les corpora – et non seulement ceux qui sont représentatifs du domaine – dépasse les 8 %. Il convient de noter la rareté des verbes dits « de spécialité » dans les deux langues ; en anglais, la proportion n'est que 0,02 % du nombre total de mots. On appelle « verbes de spécialité » ceux qui n'existent que dans un domaine spécifique ou dont l'usage est significativement plus fréquent que d'ordinaire. Dans les deux corpora, se trouvent 28 verbes de spécialité anglais et 24

1 Développé par le Dr Peter Roe, le professeur Frank Knowles et leurs collègues à l'Université d'Aston de Birmingham en Grande-Bretagne, avec leur partenaire MS Technology A/SA à Copenhague au Danemark.

français en hapax. Le tableau 2 indique la fréquence des verbes de spécialité dans les sciences biologiques en anglais et en français ; ce qui est intéressant, c'est que le plus grand nombre de verbes dans leur forme non lemmatisée apparaît moins de cinq fois. En plus, en anglais on trouve quelques occurrences de verbes composés (par exemple *to phase-vary*). Ces verbes sont en général des verbes dénominalisés et, étant donné le progrès scientifique rapide et continu dans ces domaines, ils sont souvent des néologismes qui finissent par être adoptés dans la langue écrite.

6. Critères d'analyse et exemples

Une analyse a été faite des combinants, verbe + substantif, qui se présentent plus de trois fois (chiffre purement arbitraire), sans tenir compte de la position du substantif par rapport au verbe. Des substantifs peuvent se trouver en positions +1, +2, +3 et +4. La place du verbe est même parfois plus éloignée.

Des exemples sont donnés du verbe transitif *express* en anglais et du verbe transitif/réflexif (*s'*)*exprimer* en français (tableaux 3a et 3b).

7. La valence des verbes de spécialité

Nous avons incorporé la théorie de valence à notre recherche sur l'orientation de la collocation. Dans cette théorie, le verbe est considéré comme le noyau de la phrase, les substantifs et d'autres éléments sont en second lieu. Quoique la valence soit considérée principalement comme structure syntaxique, elle comprend néanmoins des restrictions sémantiques et, en plus, elle a l'avantage d'opérer dans les limites de la phrase. Il est intéressant de noter que des différences au niveau de la valence peuvent se trouver entre les verbes de spécialité et les verbes en langue générale, p. ex. *l'agrégat cristallise* (et non pas « se cristallise ») ; *patients present with* (symptômes). Les deux exemples ont un réflexif qui n'est pas exprimé mais qui est sous-entendu.

8. L'orientation basée sur les actants et les circonstants

L'identification des actants et des circonstants en théorie de la valence peut fournir une aide importante lorsqu'on veut constater l'orientation d'un combinant. Les circonstants se présentent souvent sous forme de syntagmes prépositionnels qui peuvent être facultatifs, dont on peut faire abstraction sans que la phrase perde son sens. Tous les combinants ont des verbes avec un actant obligatoire mais des circonstants facultatifs.

Il en résulte trois catégories de combinants auxquelles il faut accorder des rôles de *base* et de *collocataire* pour arriver à l'orientation. Premièrement, si on passe de la forme active à la forme passive, c'est l'actant obligatoire en première position de valence qui est la base, et qui fournit donc l'orientation du combinant, tandis que le verbe est la partie collocataire (tableau 4a). Deuxièmement, aux cas où la nominalisation nécessite un verbe support ayant peu de valeur sémantique, l'orientation reste sur le substantif en deuxième position de valence, car c'est celui-ci qui contient la plus

grande partie de l'information de la phrase et qui est la base du combinant (tableau 4b). Troisièmement, il est possible que les verbes avec un réflexif qui peut être sous-entendu ont le verbe comme base du combinant, le sujet ou le pronom réflexif devenant les collocataires (tableau 4c).

Conclusion

Ce travail sert à fournir des critères pour l'identification des *bases* et des *collocataires* dans les combinants des langues de spécialité, afin de faciliter le travail du terminologue. En plus, il vise à établir les fondations des *dictionnaires de collocations*.

Annexe

Span -3	**Span -2**	**Span -1**	**Type**	**Span +1**	**Span +2**	**Span +3**
2 Pir-46	2 JRS4	4 and	**expresses**	8 a	2 F	2 2
2 a	2 cells	4 that	**expresses**	5 the	2 and	2 F
2 generally	2 contain	4 which	**expresses**	2 It	2 functional	2 Pheromone
2 perhaps	2 encodes	2 constitutively	**expresses**	2 pheromone	2 genes	2 Staphylococcus
2 shown	2 more	2 efficiently	**expresses**	2 protein	2 inhibitor	2 amino
2 the	2 respectively	2 importantly	**expresses**	2 recombinant	2 multi-functional	2 protein
2 vaccinia	2 single	2 pLRO49	**expresses**	1 CR2	2 stable	2 under
2 which	2 vector	2 plr	**expresses**	1 authentic	1 Thirty-three	2 urease
1 MVA	1 Raji	1 OECD	**expresses**	1 both	1 Cells	1 As
1 Recombinant	1 The	1 presumably	**expresses**	1 its	1 RPV	1 H
1 bodies	1 VV	1 simultaneously	**expresses**	0 -	1 fusion	1 HA
1 line	1 and	0 -	**expresses**	0 -	1 haemagglutinin	1 Proceedings
1 recombinant	1 developed	0 -	**expresses**	0 -	1 immunising	1 antigen
1 selection	1 recombinant	0 -	**expresses**	0	1 influenza	1 glycoprotein
1 tested	1 that	0 -	**expresses**	0 -	1 recombinant	1 hemagglutinin
1 virus	1 vaccine	0 -	**expresses**	0 -	1 sincere	1 thanks
1 was	1 virus	0 -	**expresses**	0 -	1 the	1 were

TABLEAU 1 Exemple de profil synoptique du verbe anglais *expresses* utilisant l'Aston Text Analyser (ATA)

Fréquence	Nombre total de verbes **anglais** de spécialité (toutes formes morphologiques)	Nombre total de verbes **français** de spécialité
> 200	3	0
200 - 101	3	0
100 - 81	3	2
80 - 61	9	0
60 - 41	14	8
40 - 21	28	23
20 - 16	22	8
15 - 11	19	15
10 - 6	21	18
5 - 1	140	77

TABLEAU 2 : Fréquence des verbes spécialisés dans les sciences biologiques en anglais et en français.

Corpus examples of *express* + *cell(s)*

express	*packaging cells. A major*	(+2)
express	*inefficiently in cells in*	(+3)
express	*on their cell surface*	(+3)

Corpus examples of *express* + *gene(s)*

express	*genes from other organisms*	(+1)
express	*any genes for foreign antigens*	(+2)
express	*this gene during infection*	(+2)
express	*the IFN-gene*	(+3)
express	*the reporter gene.*	(+3)
express	*heterologous protein-specifying genes*	(+3)
express	*viral and recombinant genes.*	(+4)

Corpus examples of *express* + *polysaccharide(s)*

	(reported to) express polysaccharide were examined under	(+1)
express	*polysaccharide. Microbial polysaccharides have*	(+1)
express	*polysaccharide. Microbial polysaccharides have*	(+3)
express	*two distinctly different polysaccharide phenotypes*	(+4)

Corpus examples of *express* + *protein(s)*

express	*protein F even under*	(+1)
express	*protein F normally (compare)*	(+1)
express	*either protein F or*	(+2)
express	*heterologous protein-specifying genes for*	(+2)
express	*some proteins such as*	(+2)
express	*these proteins in L.*	(+2)
express	*multiple M-like proteins;*	(+3)
express	*the gE protein, and*	(+3)
express	*a plasmid encoded protein*	(+4)

Corpus examples of *express* + *sequence(s)*

express	*retroviral sequences One reason*	(+2)
express	*such sequences when placed*	(+2)
express	*these sequences Polytropic viruses*	(+2)
express	*VL30 sequences at high*	(+2)
express	*VL30 retroviral-related sequences*	(+3)

Corpus examples of *express* + *phenotype(s)*

express	*the ropy phenotype. Commercial*	(+3)
express	*the ropy phenotype expressed*	(+3)
express	*the ropy phenotype This*	(+3)
express	*two distinctly different polysaccharide phenotypes*	(+5)

TABLEAU 3(a) : Exemples du corpus du verbe *express* + propositions qui se rencontrent plus de 3 fois. La position de la proposition suivant le verbe est entre parenthèses

Forme transitive

un virus vivant atténué	exprimant	la GP160, une protéine d'enveloppe du HIV	(+2)
la création de bibliothèques d'ADN	exprimant	les gènes qui codent pour ces protéines immunogènes	(+2)
toutes les cellules nerveuses	expriment	la béta-galactosidase	(+2)
le dosage est	exprimé en	indiquant la quantité	(-2)
concentration de ractopamine intacte	exprimée	en chlorhydrate de ractopamine	(-2)
bacillus megaterium	exprimée	dans Bacillus subtilis	(-2)
bacillus stearothermophilus	exprimée	dans Bacillus subtilis	(-2)
la quantité de diazinon,	exprimée	en g/kg.	(-1)
référence dont la dureté,	exprimée	en carbonate de calcium	(-1)
matière caséeuse peut être	exprimée	des lésions	(-4)

Forme réflexive

on va savoir comment	s'exprime	un message génétique	(+2)
les médecins	s'expriment	peu	(-1)
les symptômes gravissimes	s'expriment	tôt dans l'enfance	(-2)
le gène marqueur	s'y exprimait	rapidement chez un homme	(-2)

TABLEAU 3(b) : Exemples du corpus français du verbe « (s')exprim* » qui paraissent plus de 3 fois. Les chiffres entre parenthèses indiquent la position avant ou après le verbe selon la (non-) réflexivité
* = *joker*

Exemple du corpus *Enzymes*	**SLOT 2** **Valence** position 1 **(BASE DU COMBINANT)**	**Actant** obligatoire
were encoded	**SLOT 1** **Verbe de spécialité** **(PARTIE COLLOCATAIRE)**	
by three genes	**SLOT 3** **Valence** position 2	**Actant** obligatoire circonstant facultatif
for sugar metabolism	**SLOT 4** **Valence** position 3	**Circonstant** facultatif

TABLEAU 4a · Base et collocation d'un verbe transitif au passif.
La base est une première position de valence.

Exemple du corpus *Fowlpox virus recombinant*	**SLOT 2** **Valence** position 1	**Actant** obligatoire
confers	**SLOT 1** Verbe support **(PARTIE COLLOCATAIRE)**	
protection	**SLOT 3** **Valence** position 2 **(BASE DU COMBINANT)**	**Actant** obligatoire
in chickens	**SLOT 4** **Sous-valence** position 1 au substantif en position de valence 2	**Actant** obligatoire *ou* **Circonstant** facultatif

TABLEAU 4b : Base et collocation d'un verbe support. La base est en deuxième position de valence.

Exemple du corpus	**SLOT 2**	
Le gêne marqueur (s'y)	**Valence** position 1 **(PARTIE COLLOCATAIRE)**	**Actant** obligatoire
exprimait	**SLOT 1** **Verbe réflexif** **(BASE DU COMBINANT)**	
rapidement	**SLOT 3** **Valence** position 2	**Circonstant** facultatif
chez un homme	**SLOT 4** **Sous-valence** position 1 au substantif en position de valence 1	**Actant** obligatoire *ou* **Circonstant** facultatif

TABLEAU 4c : Base et collocation des verbes réflexifs (parfois sous-entendus).

ACABIT : une maquette d'aide à la construction automatique de banques terminologiques

Béatrice DAILLE[1]

Université de Nantes, IRIN, Nantes, France

1. Introduction

Une banque terminologique contient le vocabulaire d'un domaine technique : les termes. Ce vocabulaire technique comprend des unités lexicales simples et des unités lexicales complexes. Parmi ces unités lexicales complexes, les noms composés sont les plus nombreux. Benveniste (1966) les a baptisés « synapsies » et les a caractérisés par un certain nombre de propriétés d'ordre morphosyntaxique et sémantique. La synapsie serait, toujours d'après Benveniste, la formation de base des nomenclatures techniques.

Le travail d'élaboration d'une banque de terminologie est un travail difficile, long et qui demande à la fois des connaissances linguistiques et terminologiques. L'enjeu est donc de fournir des outils permettant d'aider à la création de ces banques, ou quand elles existent déjà, de pouvoir les valider. Il existe deux techniques principales de dépouillement terminologique : une technique structurelle fondée sur une analyse syntaxique plus ou moins poussée de l'énoncé et une technique statistique et numérique qui décèle les associations préférentielles présentes dans les corpus.

En ce qui concerne la technique structurelle, nous pouvons citer les travaux de David et Plante (1990) et Bourigault (1992). Le logiciel TERMINO présenté dans David et Plante est un système de reconnaissance des synapsies dont les fondements théoriques s'inscrivent dans la théorie X-barre. Le module de repérage des synapsies opère sur un texte non préalablement étiqueté, à l'inverse de notre travail et de celui de Bourigault. Ce module est partie intégrante du module d'analyse syntaxique et s'appuie sur une décomposition des synapsies en tête (nom) et expansion(s) (adjectif,

1. Université de Nantes - IRIN, 2, rue de la Houssinière, 44072 Nantes Cedex 03 email : Béatrice.Daille@irin.univ-nantes.fr

groupe prépositionnel ou encore nom). Ces règles de dépistage s'appuient sur une description des marques syntaxiques de frontières et des structures grammaticales admissibles en exploitant des informations morphologiques sur la catégorisation grammaticale des mots.

À la différence de TERMINO qui effectue une analyse syntaxique de la phrase, le logiciel LEXTER, présenté dans Bourigault (1992), utilise des techniques d'analyse syntaxique locale par patron de surface. Il ne s'agit plus d'implémenter une grammaire complexe des termes mais : « de s'appuyer sur des connaissances *en négatif* concernant les configurations grammaticales dont on sait qu'elles ne peuvent pas être des constituants de termes (verbe, conjonction, pronom, etc.) ». Le texte pré-étiqueté est donc découpé en syntagmes nominaux grâce au repérage de leurs frontières potentielles. Ces groupes nominaux, dits « maximaux », ainsi que les sous-groupes qui les constituent, sont des candidats termes qu'il faudra soumettre à un terminologue. Ces deux approches, malgré la différence de leur complexité d'analyse, sont structurelles et ne permettent pas d'obtenir une liste ordonnée des candidats termes.

L'autre approche est l'approche statistique. Cette approche très prisée outre-Atlantique a donné d'excellents résultats dans le domaine du traitement du langage naturel, principalement pour la reconnaissance de la parole et pour l'assignation d'étiquettes grammaticales. Dans le cadre de l'aide à la construction de dictionnaires monolingues, l'application de modèles statistiques sur des textes fournit des informations quantitatives et qualitatives sur les affinités lexicales que peuvent présenter certains mots entre eux. Par exemple, Church et Hanks (1990) sur l'anglais, Calzolari et Bindi (1990) sur l'italien se sont intéressés aux cooccurrences lexicales mises à jour par l'utilisation d'une mesure proche du concept d'« information mutuelle », le « score d'association » (*association ratio*). Smadja et McKeown (1990), à partir d'un texte étiqueté et de l'utilisation d'une mesure similaire au score d'association, recensent les cooccurrences lexicales et les expressions figées et les intègrent, après un filtrage *a posteriori*, dans un dictionnaire utilisé par un programme de génération. À ce jour, ce sont les seuls qui produisent une application pratique de ces informations lexicales extraites automatiquement d'un corpus. Dans le domaine plus spécifique de l'acquisition de terminologies, il faut mentionner les bons résultats obtenus avec le système ANA développé par Enguehard (1992). Grâce à l'exploitation d'une liste donnée *a priori* de concepts pertinents du domaine et la mise en œuvre d'heuristiques statistiques finement ajustées, ce système extrait des concepts d'un texte avec un bon taux de précision, sans effectuer d'analyse linguistique.

Le problème principal avec l'une ou l'autre de ces approches est le « bruit ». En effet, les critères morphosyntaxiques ne permettent pas véritablement de différencier groupes nominaux libres et termes, et les cooccurrences extraites grâce à des méthodes statistiques relèvent d'associations diverses. Rappelons que parmi les cooccurrences extraites par le modèle statistique de Lafon (1984) se trouvent des associations sémantiques, des associations fonctionnelles parmi lesquelles on rencontre des noms composés ou des termes, et des associations impossibles à caractériser.

ACABIT est un logiciel de dépouillement terminologique, chargé de préparer la tâche du terminologue en lui proposant une liste ordonnée de « candidats termes », c'est-à-dire des noms composés les plus représentatifs du domaine à ceux qui le sont le moins. Il utilise des méthodes statistiques qui sont tout à fait adaptées à ce genre de

tâche puisque leurs analyses de corpus de grande taille fournissent des résultats inaccessibles à un observateur humain ou à un analyseur syntaxique et permettent de recueillir des observations générales. Il guide ces modèles statistiques sur les cooccurrences que nous voulons extraire, les termes, et évite le plus possible la prise en comptes des autres types de cooccurrences. ACABIT procède en deux étapes : d'abord il filtre les séquences morphosyntaxiques qui caractérisent les « termes de base » grâce à des grammaires locales (voir section 2.2.), puis il utilise un modèle statistique pour distinguer parmi ces cooccurrences lesquelles sont le plus probablement des termes.

2. Données linguistiques

De manière à déterminer sur quelles séquences d'unités lexicales nous allons appliquer nos mesures statistiques, nous allons utiliser les résultats d'une étude linguistique sur les structures morphosyntaxiques des termes rencontrés soit dans des corpus, soit dans des banques terminologiques existantes. Cette étude va nous permettre de dégager des spécifications linguistiques pour les termes que nous utiliserons pour établir des filtres linguistiques.

2.1. Spécifications linguistiques

Les termes sont majoritairement des unités lexicales complexes de type nominal. Nous avons voulu vérifier cette affirmation en étudiant une banque terminologique multi-domaines d'environ 800 000 termes. Il nous est impossible pour des raisons de confidentialité d'invoquer le nom de cette banque, nous la nommerons donc BANQUE tout au long de cet article.

Une première étude statistique simple portant sur la longueur des termes de BANQUE montre que 85 % de ceux-ci sont de longueur > 1 et confirme donc le fait que les termes soient majoritairement des unités lexicales complexes. Il reste donc à démontrer que ces unités lexicales complexes sont effectivement majoritairement de type nominal et d'identifier les structures morphosyntaxiques les plus représentées.

Notre deuxième étude concerne donc la représentativité des structures morphosyntaxiques des termes de BANQUE. En effet, si l'on considère les termes comme une sous-classe des noms composés, ceux-ci peuvent être caractérisés par certaines propriétés morphologiques ou/et syntaxiques mises à jour dans des études sur la composition nominale (présentées dans Gross *et al.*, (1986) ; Noailly, (1990), etc.). Plus précisément, les termes peuvent être classés en fonction de leur structure morphosyntaxique, N Adj, N1 *de* N2, etc., et s'adaptent donc à la typologie plus générale des noms composés du français élaborée par Mathieu-Colas (1988). Nous avons donc décidé d'assigner sa catégorie grammaticale à chacune des unités lexicales contenues dans les termes de BANQUE. Nous avons utilisé un logiciel développé par l'équipe de recherche en TAO du CITI[2] qui utilise le *Dictionnaire morphologique du français*

2 Centei for Information Technology Innovation (CITI), 1575 Boulevard Chomedey, Laval (Québec), Canada H7V 2X2

(*DMF*) et assigne toutes les catégories grammaticales possibles d'une unité lexicale reconnue. Les mots du *DMF* provenant essentiellement des dictionnaires *Petit Robert* et *Petit Larousse*, ils appartiennent principalement à la langue courante. La proportion des mots de BANQUE inconnus du *DMF* est de l'ordre de 30% : une grande partie du vocabulaire technique n'est donc pas reconnu. Au problème des mots inconnus se rajoute celui des ambiguïtés grammaticales. Les programmes d'étiquetage automatique tel que celui de Foster (1991) ou de El-Bèze (1993) nécessitent un contexte phrastique et ne sont pas véritablement performants sur des mots ou des courtes séquences de mots isolés. Nous avons donc développé un programme de désambiguïsation à base de règles à partir des catégories grammaticales proposées par le *DMF*. Nous avons aussi inclus dans ce programme un module chargé du traitement des unités lexicales inconnues à l'intérieur d'un terme. Nous ne détaillerons pas plus ici ce programme qui fera l'objet de publications ultérieures. Après l'application de notre programme, la proportion d'unités lexicales inconnues à l'intérieur de termes de structure complexe tombe à 3%. Il nous est donc possible d'évaluer la proportion de termes composés de plusieurs unités lexicales complexes de type nominal : celle-ci est de l'ordre de 95 %.

L'étude des termes présents dans BANQUE a donc bien démontré que ceux-ci sont pour leur grande majorité des unités lexicales complexes de type nominal.

2.1.1. Les termes de base

L'étiquetage grammatical des termes de BANQUE nous permet de classer ceux-ci en fonction de leur structure morphosyntaxique. Le tableau ci-dessous (tableau 1) présente les structures des termes les plus fréquentes (à l'exception de la structure N1 N2) ainsi que leur fréquence et leur représentativité par rapport au nombre total de terme de longueur > 1 égal à 738 072 dans BANQUE.

Structures morphosyntaxiques	**Nombres**	**%**
N Adj	182 267	25
N1 Prep N2	170 710	23
N1 Prep Det N2	46 967	6
N1 N2	14 895	2
Total	**414 839**	**56**

TABLEAU 1. Structures morphosyntaxiques des termes de base de BANQUE.

De cette étude statistique des structures morphosyntaxiques des termes, il apparaît que les termes de longueur 2, où seules sont prises en compte les unités lexicales pleines tels que les noms, les adjectifs et les adverbes séparés par des blancs dans l'écriture, sont de loin les plus nombreux. L'approche statistique exigeant une bonne représentation du nombre d'échantillons, ACABIT se concentre sur l'extraction des termes de longueur 2, appelés « termes de base », et qui s'apparient à l'une des structures morphosyntaxiques suivantes :

N Adj : *indicateur environnemental*
N1 Prep N2 : *protéine de poissons*
N1 Prep Det N2 : *chimioprophylaxie au rifampine*
N1 N2 : *bague étalon.*

2.1.2. Les termes ternaires et n-aires

Les termes de longueur > 2 sont généralement moins représentés dans les textes que les termes de base. Dans BANQUE, ils représentent pourtant environ 40 % des termes de longueur > 1. Nous pouvons néanmoins affirmer que la majorité de ces termes de longueur > 2 sont créés récursivement à partir des termes de base. Nous avons identifié deux opérations qui permettent de passer d'un terme de base à un terme de longueur > 2 : l'insertion et la juxtaposition. L'insertion, à la différence de la juxtaposition, modifie la structure morphosyntaxique du terme de base.

2.1.2.1. Insertion

Un terme obtenu par insertion est construit à partir soit d'un terme de base lorsque celui-ci est modifié, soit de deux termes de base lorsqu'il y a substitution. Dans les deux cas, la structure morphosyntaxique du terme de base est altérée.

2.1.2.1.1. Insertion de modifieurs

Ce sont principalement les adjectifs et les adverbes qui peuvent s'insérer à l'intérieur d'un terme de base ; les adjectifs dans la structure N1 Prep N2 ou N1 Prep Det N2 et les adverbes à l'intérieur d'une structure N Adj :

N1 Prep N2 → N1 **Adj** Prep N2 : *charge* ***corporelle*** *d'équilibre*
N Adj → N **Adv** Adj : *fer* ***non*** *hémique*

2.1.2.1.2. Substitution

La substitution se définit ainsi : étant donné un terme de base, l'une des unités lexicales pleines de ce terme est remplacée par un autre terme de base dont la tête est cette unité lexicale. Par exemple, dans la structure N1 Prep1 N2, N1 peut être remplacé par un terme de structure N1 Prep2 N3, pour former un surcomposé de structure N1 Prep2 N3 Prep1 N2 : par exemple, le nom *réseau* dans le terme *réseau à satellite(s)* est remplacé par le terme de base *réseau de transit* pour former le surcomposé *réseau de transit à satellite(s).*

La substitution se différencie de la juxtaposition (présentée ci-dessous) car :

– elle demande obligatoirement l'emploi de deux termes de base, dans l'exemple ci-dessus *réseau à satellite* et *réseau de transit* ;
– elle brise la structure interne de l'un des deux termes : dans notre exemple, la structure de *réseau à satellite* est altérée.

2.1.2.2. Juxtaposition

Un terme obtenu par juxtaposition est construit à partir d'un terme de base. Nous avons distingué deux sortes de juxtaposition : la surcomposition et la modification.

2.1.2.2.1. Surcomposition

La juxtaposition utilise au minimum un terme de base et se caractérise par les propriétés suivantes :

– les éléments de la structure du ou des termes de base restent solidaires ;
– lorsqu'un nom simple se juxtapose à un terme de base, c'est le plus souvent le nom simple qui précède le terme ;
– la juxtaposition s'effectue par l'intermédiaire d'une préposition ;
– les enchevêtrements à l'intérieur de la structure juxtaposée ne réfèrent pas à des termes de base. Cette propriété est illustrée dans les exemples qui suivent, où les termes de base apparaissent entre crochets :

N1 Prep1 [N2 Adj] (longueur 3)
dispositif d'[éclairage ultraviolet]
Avec *dispositif ultraviolet* qui n'est pas un terme de base.

[N1 Adj1] Prep1 [N2 Prep2 Det N3] (longueur 4)
[*accès multiple*] *avec* [*assignation à la demande*]
Ni *accès avec assignation*, ni *accès à la demande* ne sont des termes de base.

2.1.2.2.2. Postposition de modifieurs

Les termes de longueur ≥ 3 obtenus par post-modification sont les plus représentés dans BANQUE. Les adjectifs et les groupes prépositionnels adverbiaux sont les principaux modifieurs des termes unaires ou binaires qui sont à l'origine de nouveaux termes :

[N1 Adj1] **Adj2** (longueur 3) : [*production primaire*] ***épondique***
[N1 Prep N2] **Adj** (longueur 3) : [*cessation d'emploi*] ***forcée***
[N1 Adj1] [**Prep Adj N**] (longueur 4) : [*câble(s) sous-marin(s)*] [***à large bande***]

Dans un texte technique, il est difficile de déterminer si une séquence morphosyntaxique pouvant caractériser un terme de longueur > 2 obtenu par insertion ou juxtaposition est ou n'est pas un terme. Pour certaines séquences, l'introduction d'une abréviation permet d'entériner leur statut terminologique : comme par exemple l'abréviation *BLU* associée à la séquence *bande latérale unique*, obtenue par juxtaposition et plus précisément par post-modification du terme de base *bande latérale* par l'adjectif *unique* – adjectif par ailleurs très fréquent en français et qui ne porte aucune marque de technicité. Néanmoins dans la plupart des cas, il est impossible de déterminer si une séquence morphosyntaxique d'un terme de base auquel s'est appliquée l'opération de juxtaposition ou d'insertion réfère ou non à une notion du domaine. Nous avons donc décidé de ne pas trancher et de nous concentrer sur les termes de base.

Une fois ceux-ci identifiés, les « nouveaux termes » obtenus par juxtaposition pourront être reconnus à l'aide d'un programme de mise en évidence des variations terminologiques comme, par exemple, le logiciel FASTR de Jacquemin (1994).

Cependant, même en nous restreignant à l'extraction des termes de base, il nous faut prendre en compte leurs variantes.

2.1.3. Les variantes des termes de base

Il existe cinq catégories principales de variantes : les abréviations, les variantes orthographiques, les variantes morphosyntaxiques, les variantes syntaxiques et les variantes elliptiques. Les abréviations qui sont extraites des corpus par des heuristiques ne sont pas décrites ci-dessous.

2.1.3.1. Variantes orthographiques

Les variantes orthographiques d'un terme de base sont principalement de trois types :

– variation en nombre de N2 normalement invariable dans les structures N1 Prep N2 : *réseau(x) à satellite* ou *réseau(x) à satellites* ;
– l'un des composants du nom composé à plusieurs graphies possibles : ***S**ervice national* ou *service national* ;
– caractère optionnel du trait d'union dans la structure N1 N2 : *mode-paquet* ou *mode paquet.*

2.1.3.2. Variantes morphosyntaxiques

Les variantes morphosyntaxiques d'un terme de base sont principalement de trois types :

– simplification de la structure du terme binaire par l'effacement de la préposition ou/et du déterminant qui apparaît à l'intérieur de celui-ci : *tension d'hélice* = *tension hélice ;*
– relation de synonymie entre deux structures de nom composé qui diffèrent seulement par l'une de leurs unités lexicales pleines : *réseau commuté* ou *réseau à commutation* ;
– variation de la préposition : *réseau pour données* → *réseau de données.*

2.1.3.3. Variantes syntaxiques

Les variantes syntaxiques d'un terme de base sont principalement de deux types :

– insertion d'un modifieur à l'intérieur d'un terme de base : *réseau numérique* → *réseau* ***entièrement*** *numérique ;*
– coordination de deux termes de base ; *élévateurs de fréquence* + *abaisseurs de fréquence* → *élévateurs* ***et*** *abaisseurs de fréquence.*

2.1.3.4. Variantes elliptiques

Un terme peut être évoqué par une forme elliptique où une ou plusieurs de ses unités lexicales non grammaticales ont disparu. Pour les termes de base, c'est principalement l'élément de queue qui disparaît : *débit* sera employé à la place de *débit binaire*.

Cette étude linguistique a montré que, d'une part, les termes de base sont les plus représentés dans BANQUE et que, d'autre part, la majorité des termes de longueur > 2 sont construits à partir de termes de base. La décision de se concentrer sur les termes de base est donc linguistiquement motivée. À cette motivation linguistique, il faut ajouter l'argument statistique de leur bonne représentation numérique dans les corpus. Il nous reste donc à expliquer comment ACABIT extrait d'un corpus ces termes de base.

2.2. Filtres linguistiques

Nous nous trouvions devant le choix suivant : soit nous isolions, grâce aux mesures statistiques, les collocations du corpus puis nous appliquions des filtres linguistiques (à la manière du travail présenté dans Smadja et McKeown (1990)), soit nous appliquions d'abord les filtres linguistiques et ensuite les mesures statistiques. C'est cette dernière solution que nous avons adoptée. En effet, la première solution demandait l'utilisation d'une fenêtre de taille fixe ; si vous utilisez une fenêtre de petite taille, vous perdez les occurrences des termes de base modifiés par un modifieur inséré et les structures de termes de base coordonnées ; si vous prenez une fenêtre de taille plus importante, vous obtenez beaucoup de mauvaises séquences. Dans les deux cas, et même avec un filtrage linguistique postérieur, les comptes de fréquences étaient erronés. L'utilisation de filtres linguistiques avant l'application de mesures statistiques est donc la solution retenue.

ACABIT filtre les termes binaires en utilisant leurs structures morphosyntaxiques. Notre programme nécessite donc en entrée un corpus nettoyé où chaque unité lexicale a reçu son étiquette grammaticale et son étiquette morphologique (lemme). ACABIT prend en entrée uniquement jusqu'à maintenant des corpus étiquetés et lemmatisés par les programmes d'assignation d'étiquettes grammaticales et morphologiques d'IBM-France développé par l'équipe de recherche sur la reconnaissance de la parole (se référer, par exemple, aux travaux de El-Bèze (1993)).

2.2.2. Programme d'extraction et de relevé des fréquences

Les termes binaires sont considérés comme des cooccurrences particulières qui possèdent les propriétés linguistiques décrites ci-dessus : ils se définissent par rapport à leur structure morphosyntaxique ; ils ont la propriété de donner naissance à de nouveaux termes ; ils admettent des variantes. Une cooccurrence qui caractérise un terme binaire répond aux conditions suivantes :

1) elle est orientée et suit l'ordre linéaire du texte ;
2) elle met en jeu deux unités lexicales pleines ;
3) elle doit apparaître dans l'une des structures morphosyntaxiques des termes binaires.

Le relevé des fréquences des occurrences des termes de base candidats est essentiel puisque ce sont ces fréquences qui sont les paramètres des mesures statistiques. Un mauvais relevé des fréquences entraînera des résultats statistiques faux ou non pertinents pour notre application. Nous prenons en compte dans ces automates les variantes graphiques et morphosyntaxiques à l'exception des variantes elliptiques, ainsi que les variations syntaxiques qui affectent les termes de base lors des opérations de coordination et d'insertion de modifieurs. Nous avons vu en section 2.1.2. qu'un terme de longueur > 2 pouvait être obtenu à partir d'un terme de base par insertion ou juxtaposition. Si l'opération de juxtaposition ne pose pas de problème à l'extraction des termes de base puisqu'elle ne modifie pas la structure interne de celui-ci, ce n'est pas le cas de l'opération d'insertion. Une séquence comme *antenne parabolique de réception* peut référer soit à un terme de longueur 3 obtenu par insertion (soit par insertion de modifieur ou par substitution), soit à un terme de base *antenne de réception* modifié par l'adjectif *parabolique*. Si d'un côté, nous ne souhaitons pas extraire de termes de longueur > 2, d'un autre côté, nous ne pouvons pas ignorer cette variation syntaxique. ACABIT prend donc en compte l'insertion possible de modifieurs dans les structures N Adj, N1 Prep N2 et N1 Prep Det N2. L'autre variation syntaxique qui est prise en compte dans notre programme est la coordination. La coordination de deux termes de base ne produit pas, en général, un nouveau terme. Ainsi, une séquence comme *équipements de modulation et de démodulation* est considérée comme équivalente à la séquence : *équipements de modulation et équipements de démodulation.*

Ces choix nous conduisent à extraire des termes de longueur > 2.

Deux automates ont donc été écrits : le premier regroupant les types élémentaires : N1 de (Det) N2 (*signal de raccrochage*), N1 à (Det) N2 (*tube à ondes*), N1 Prep N2 (*multiplexage par répartition*), N1 N2 (*voie support*)[3] ; le second pour le type élémentaire N Adj (*dissipation thermique*).

Une séquence morphosyntaxique reconnue par l'un des automates constitue une occurrence d'un couple appartenant à un de nos deux patrons : N Adj et N1 (Prep (Det)) N2. Un couple est constitué des deux lemmes qui correspondent aux deux extrémités lexicales de la séquence ; par exemple, pour le type élémentaire N1 (Prep (Det)) N2, le couple (satellite, orbite) pourra correspondre au séquences suivantes : *satellite sur orbite, satellites sur orbite, satellites en orbite, satellite mis en orbite*. Chaque séquence relevée est accompagnée de son schéma morphosyntaxique et de sa position dans le corpus (fichier, phrase, position dans la phrase).

3. Statistiques lexicales

ACABIT dans un deuxième temps utilise les résultats d'une évaluation de différentes mesures de statistique lexicale. Cette évaluation a permis de découvrir la meilleure mesure pour notre application, c'est-à-dire celle qui assigne un score élevée aux séquences les plus susceptibles de constituer des termes parmi notre liste de candidats.

3 Les parenthèses à l'intérieur des structures morphosyntaxiques indiquent le caractère optionnel d'une ou de plusieurs étiquettes grammaticales

Les caractéristiques numériques calculées par ACABIT ont chacune un rôle particulier : les fréquences sont les paramètres du critère d'association retenu ; le critère d'association mesure la force du lien entre les deux lemmes du couple. D'un point de vue statistique, les deux lemmes qui forment un couple sont considérés comme deux variables qualitatives dont il s'agit de tester la liaison. Les données se représentent sous la forme d'un tableau croisé, appelé tableau de contingence et défini à partir des comptes précédents.

Un tableau de contingence est associé à chaque couple de lemmes (L_i, L_j) :

	L_j	$L_{j'}$ avec $j' \neq j$
L_i	a	b
$L_{i'}$ avec $i' \neq i$	c	d

Les valeurs a, b, c et d résument les occurrences d'un couple :

a = le nombre d'occurrences du couple (L_i, L_j) ;
b = le nombre d'occurrences des couples où L_i est le premier élément d'un couple et L_j n'est pas le second ;
c = le nombre d'occurrences des couples où L_j est le second élément du couple et L_i n'est pas le premier ;
d = le nombre d'occurrences de couples où ni L_i ni L_j n'apparaissent.

La somme (a + b + c + d), notée N, est le nombre total d'occurrences de tous les couples trouvés pour un patron morphosyntaxique.

La littérature statistique regorge de mesures destinées à « tester l'indépendance » ou à « mesurer la liaison » ou encore à « mesurer le degré de similitude ou d'affinité » entre deux variables régies par un tableau de contingence. Dans Daille *et al.* (1995), nous avons évalué une dizaine de mesures dont : le score d'association, proche du concept d'information mutuelle, introduit par Church et Hanks (1990) :

$$IM = \log_2(a / (a+b)(a+c)) \qquad (1)$$

le coefficient du Φ^2 proposé par (Gale et Church, 1991) :

$$\Phi^2 = (ad - bc)^2 / (a+b)\,(a+c)\,(b+c)\,(b+d) \qquad (2)$$

ou encore le coefficient de vraisemblance présenté par (Dunning, 1993) :

$$\begin{aligned} Loglike = {} & a \log a + b \log b + c \log c + d \log d - (a+b) \log (a+b) \\ & - (a+c) \log (a+c) - (b+d) \log (b+d) - (c+d) \log (c+d) \\ & + (a+b+c+d) \log(a+b+c+d) \qquad (3) \end{aligned}$$

Cette évaluation a démontré que la fréquence d'un couple est un très bon indicateur de son caractère terminologique. Ce résultat a contredit de nombreux travaux ré-

cents dans le domaine de l'extraction de ressources lexicales, qui proclamaient que l'information mutuelle donnait de meilleurs résultats que la fréquence (voir par exemple, Church et Hanks (1990)). Néanmoins, le classement proposé par la fréquence intégrant très rapidement du bruit, *i.e.* des couples qui ne réfèrent pas à des termes, nous avons choisi de ne retenir que le coefficient de vraisemblance (formule 3). Le coefficient de vraisemblance sélectionne les termes du domaine en leur attribuant une valeur forte : l'amplitude des valeurs dépend du nombre d'occurrences du couple ; plus le couple est fréquent plus la valeur du coefficient de vraisemblance tend à être élevée et ce indépendamment du nombre de couples extraits.

4. Expérimentation

ACABIT a été appliqué à deux corpus : *Le manuel des télécommunications par satellite (MTS)* (200 000 mots) et *Le livre bleu des communications (LBC)* (800 000 mots). Il en a extrait des couples pour les deux patrons N1 (Prep (Det)) N2 et N Adj. Une occurrence d'un couple correspond à une cooccurrence où les deux éléments du couple entrent dans un de ces deux patrons syntaxiques. Les tableaux 2 résument les fréquences des cooccurrences exprimées en nombre de couples ; ainsi pour le corpus *MTS* et le patron N Adj, nous avons relevé 4 483 couples dont 3 144 n'ont qu'une occurrence, 655 deux occurrences et 684 plus de deux occurrences.

MTS	1 occurrence	2 occurrences	plus de 2 occurrences	total
N Adj	3 144	655	684	4 483
N1 (Prep (Det))N2	6 834	1 503	1 616	9 953

LBC	1 occurrence	2 occurrences	plus de 2 occurrences	total
N Adj	5 201	1 507	2 113	8 821
N1 (Prep (Det))N2	12 167	3 481	6 288	21 936

TABLEAUX 2 : Nombres de cooccurrences extraites.

ACABIT applique ensuite aux couples, dont le nombre d'occurrences est au moins égal à deux, le critère d'évaluation retenu par notre évaluation. Nous obtenons donc en sortie de notre programme une liste ordonnée de couples ; chaque couple représentant un concept possible du domaine. Nous donnons dans les tableaux 3 et 4 qui concernent respectivement les patrons N1 (Prep (Det)) N2 et N Adj, les valeurs les plus élevées du coefficient de vraisemblance pour nos deux corpus. Nous utilisons les notations suivantes : Logl pour le coefficient de vraisemblance (*Loglike*) et Nbc pour le nombre d'occurrences du couple.

Sous chaque couple, nous trouvons toutes les variantes présentées en section 2.1.3. rencontrées dans nos corpus ainsi que leur localisation.

Couples de structure N1 (Prep (Det)) N2 dans ***MTS***	Séquence la plus fréquente	**Logl**	Nbc
(largeur, bande)	*largeur de bande* (197)	**1328**	223
(température, bruit)	*température de bruit* (110)	**777**	126
(bande, base)	*bande de base* (142)	**745**	145
(amplificateur, puissance)	*amplificateur(s) de puissance* (137)	**728**	137
(temps, propagation)	*temps de propagation* (93)	**612**	94
(règlement, radiocommunication)	*règlement des radiocommunications* (60)	**521**	60
(produit, intermodulation)	*produit(s) d'intermodulation* (61)	**458**	61
(taux, erreur)	*taux d'erreur* (70)	**420**	70
(mise, œuvre)	*mise en œuvre* (47)	**355**	47
(télécommunication, satellite)	*télécommunication(s) par satellite* (88)	**353**	99
(bilan, liaison)	*bilan(s) de liaison* (37)	**344**	55
Couples de structure N1 (Prep (Det)) N2 dans ***LBC***	Séquence la plus fréquente	**Logl**	Nbc
(canal, sémaphore)	*canal / canaux sémaphore(s)* (1188)	**5738**	1188
(accusé, réception)	*accusé de réception* (558)	**3983**	592
(système, signalisation)	*système(s) de signalisation* (82)	**2417**	85
(complément, étude)	*complément d'étude* (242)	**1985**	245
(point, sémaphore)	*point(s) sémaphore(s)* (677)	**1822**	679
(intervalle, temps)	*intervalle(s) de temps* (249)	**1782**	251
(trame, sémaphore)	*trame(s) sémaphore(s)* (354)	**1444**	354
(signal, fin)	*signal / signaux de fin* (385)	**1407**	391
(sous-système, utilisateur)	*sous-système utilisateur* (195)	**1226**	195
(bout, bout)	*bout en bout* (136)	**1155**	137
(contrôle, continuité)	*contrôle(s) de continuité* (171)	**1116**	171

TABLEAU 3 · Classement des couples de structure N1 (Prep (Det)) N2 proposé par ACABIT.

Couple de structure N Adj dans ***MTS***	**Logl**	Nbc	Couple de structure N Adj dans ***LBC***	**Logl**	Nbc
(station, ternen)	**2934**	750	(équipement, terminal)	**1425**	275
(débit, binaire)	**716**	134	(considération, général)	**1385**	25
(accès, multiple)	**605**	105	(service, supplémentaire)	**1275**	340
(voie, téléphonique)	**512**	118	(télégraphie, harmonique)	**1250**	152
(liaison, montant)	**457**	88	(étude, ultérieur)	**1171**	169
(liaison, descendant)	**408**	77	(caractère, graphique)	**1112**	19
(secteur, spatial)	**341**	79	(entité, fonctionnel)	**999**	19
(service, fixe)	**326**	66	(centre, international)	**964**	325
(lobe, latéral)	**299**	40	(adresse, complet)	**874**	183
(faisceau, hertzien)	**244**	35	(effet, local)	**865**	169
(puissance, surfacique)	**205**	35	(station, mobile)	**855**	164

TABLEAU 4 : Classement des couples de structure N Adj proposé par ACABIT

Rares sont les couples n'admettant aucune variante. Nous donnons ci-dessous quelques exemples de couples accompagnés des occurrences extraites :

(demande, trafic) : *demande de trafic, demandes en trafic, demande réelle en trafic.*

(liaison, satellite) : *liaison par satellite, liaisons par satellite, liaisons (très rapides + numériques + téléphoniques nationales) par satellite, liaisons numériques par satellites, liaisons satellite, liaisons entre satellites.*

(signal, fin) : *signal de fin, signaux de fin, signal (local + national + valide + périodique) de fin, signal émis à des fins, signal numérique utilisé à des fins*[4].

(ligne, abonné) : *ligne d'abonné, lignes d'abonné, ligne de l'abonné, lignes de l'abonné, ligne d'abonnés, lignes des abonnés, ligne(s) (téléphonique(s) + numériques(s) + analogique(s)) d'abonné, ligne(s) (numérique(s) + analogique(s)) de l'abonné, lignes et services d'abonné.*

Ces quelques exemples montrent que les modifieurs pouvant s'insérer à l'intérieur d'un terme binaire sont en nombre réduit. L'enregistrement de ces modifieurs, comme des autres altérations que subit la structure de base, y compris les différentes flexions rencontrées, n'est pas une tâche insurmontable surtout si celle-ci est effectuée automatiquement. Ces informations lexicales sont présentes sous l'entrée de chaque couple et pourront donc être directement intégrées dans une banque terminologique.

4 Les occurrences *signal émis a des fins* et *signal numérique utilisé à des fins* illustrent un problème de notre approche quantitative : nous n'avons aucune assurance qu'un couple regroupe des cooccurrences désignant un seul et unique concept ou encore, comme ici, des cooccurrences qui soient toutes valides.

5. Conclusion

Nous avons présenté une nouvelle approche pour l'extraction automatique de terminologies monolingues qui allie informations linguistiques et mesures statistiques. Cette méthode nous a permis d'extraire automatiquement un certain nombre de candidats termes du domaine classés selon leur pertinence terminologique à partir de corpus préalablement étiquetés et lemmatisés. Tous ces candidats termes sont accompagnés de leurs variantes morphologiques et de certaines de leurs variantes morpho-syntaxiques, ainsi que des modifieurs qui altèrent leurs structures. ACABIT permet donc une amélioration sensible des performances dans l'aide à la construction de banques terminologiques à partir de corpus et permet d'établir l'efficacité de l'enrichissement des systèmes statistiques par la linguistique. Cette méthode a été étendue à l'extraction de termes bilingues à partir de corpus alignés phrase à phrase (voir nos travaux dans Daille *et al.* (1994)).

Conception et exploitation d'un logiciel d'extraction de termes : problèmes théoriques et méthodologiques[1]

Didier BOURIGAULT

Centre d'analyse et de mathématiques sociales, (Unité mixte EHESS-CNRS-Paris Sorbonne), et EDF-Direction des études et recherches, Clamart, France

1. Introduction

La conception, la réalisation et l'utilisation d'un système automatique d'extraction terminologique conduisent à aborder sous un angle nouveau les questions théoriques et méthodologiques de la terminologie. Dans cet article, nous exposons l'état de notre réflexion sur quelques-unes de ces questions. Cette réflexion est à la fois exigée et nourrie par nos recherches sur la conception et la réalisation du logiciel d'aide au dépouillement terminologique Lexter (Logiciel d'EXtraction de TERminologie), ainsi que par les expériences d'utilisation effective de Lexter dans divers projets de recherche et développement à la Direction des études et recherches d'Électricité de France. Nous ne décrivons pas le logiciel en tant que tel dans cet article. Nous renvoyons le lecteur intéressé par les techniques de Traitement Automatique des Langues Naturelles (TALN) implémentées dans le logiciel à des publications antérieures (Bourigault, 1993a, 1993b, 1994, 1995).

La discipline terminologique hérite d'une définition du terme, établie dans le cadre de la Théorie Générale de la Terminologie fondée par Eugen Wüster en 1931, qui confère au terme le statut de symbole d'une notion. Même si elle fait parfois l'objet de débats, cette définition recueille un certain consensus dans la communauté des chercheurs en terminologie. Dans la section 2 de cet article, après un rapide survol historique (section 2.1.), nous exposons en quoi notre expérience de conception et d'utilisation d'un logiciel d'extraction de termes nous amène à participer à une critique constructive de cette définition (section 2.2.).

1 L'auteur remercie Benoît Habert (ELI, École normale supérieure de Fontenay Saint-Cloud) pour ses conseils et commentaires

Dans la section 3, nous donnons une caractérisation de la notion de *candidat terme*, en décrivant les critères de validité syntaxique et d'autonomie discursive, à partir desquels sont établies les règles opératoires de dépistage implémentées dans notre système automatique d'extraction de termes. La section 4 est consacrée aux aspects méthodologiques concernant la conception et la réalisation d'un logiciel d'extraction de terminologie.

2. Définition du terme : problèmes théoriques

2.1. Survol historique

En France, É. Benveniste et L. Guilbert sont parmi les premiers linguistes à s'être intéressés aux termes techniques des vocabulaires spécialisés. Dans son article sur la composition nominale, Benveniste (1966) entreprend de donner un statut à une nouvelle forme de composition, à la base de toutes les nomenclatures techniques, et il propose le terme nouveau de « synapsie ». Il caractérise la synapsie par un ensemble de sept traits, qui sont tous de type morpho-syntaxique, à l'exception du dernier, qui en appelle à une caractérisation sémantique de la synapsie : « le caractère unique et constant du signifié ». Cette dernière caractéristique est mentionnée au même niveau que les autres. Mais il apparaît clairement à la lecture de l'article que Benveniste la considère comme primordiale : « C'est toujours et seulement la nature du désigné qui permet de décider si la désignation syntagmatique est ou n'est pas une synapsie ». Cependant, Benveniste ne mentionne pas une éventuelle spécificité sémantique de la synapsie par rapport au mot de la langue. Il n'indique pas si la fonction de désignation complète et unique de la synapsie en fait une entité linguistique d'un type différent. Dans un article antérieur, où il complétait les thèses de Saussure sur l'arbitraire du signe, Benveniste (1939) avait insisté sur le fait que le lien entre le signifiant (image acoustique) et le signifié (représentation mentale) n'était pas arbitraire mais, au contraire, nécessaire. À la lecture conjointe de ces deux articles, il n'est pas aisé de cerner la position de Benveniste sur les compatibilités ou rapports entre le caractère nécessaire du lien entre signifiant et signifié dans le signe linguistique, le caractère arbitraire de la relation entre le signe linguistique et l'objet extralinguistique et la désignation complète et unique de l'objet par la synapsie.

Dans un travail de grande envergure, Guilbert (1965) décrit comment s'est formé le vocabulaire spécifique de l'aviation, entre les années 1861 et 1890. Il identifie et étudie de façon approfondie quatre formes de néologisme dans la formation du vocabulaire de l'aviation : néologisme morphologique, néologisme sémantique, néologisme grammatical, néologisme syntagmatique. Cette dernière forme, qui est la plus productive, contribue à la création d'unités lexicales complexes. La notion d'unité lexicale complexe recouvre partiellement la notion de synapsie de Benveniste, mais Guilbert propose une caractérisation plus détaillée, et différente sur certains points. Guilbert mentionne trois critères qui distinguent selon lui l'unité lexicale complexe du groupement syntagmatique du discours : la stabilité du rapport syntagmatique au plan du discours, la stabilité du rapport de signification entre l'unité syntagmatique et un signifié unique, la fréquence d'emploi qui stabilise à la fois le lien syntagmatique et le rapport de signification. Le second critère est, pour Guilbert, essentiel. À l'instar de Benveniste, il considère que la caractérisation essentielle de l'unité lexicale complexe est « d'essence sémantique ». La caractérisation de l'unité lexicale complexe, par op-

position au syntagme du discours, est la « constance » (ou la « permanence ») du rapport de signification entre l'unité syntagmatique et un signifié unique. Dans la conclusion de sa thèse, Guilbert adopte une position plus claire que celle de Benveniste sur la spécificité sémantique de l'unité lexicale complexe. La caractérisation qu'il en a proposée le conduit à mettre en doute la conformité du signe ainsi conçu avec le signe linguistique saussurien : le signe linguistique constitué par l'unité lexicale complexe « tendrait à se confondre avec le pur symbole qui n'a de contenu sémantique, toujours et en toutes circonstances, que celui qui lui a été préalablement conféré ».

De nos jours, la doctrine terminologique moderne prend appui sur les travaux fondateurs de Wüster développés dans le cadre de sa Théorie Générale de la Terminologie. Dans le cadre de la Théorie Générale de la Terminologie de Vienne, parue en 1931, Wüster avait proposé un modèle du terme qui tentait de concilier les théories de Saussure sur le signe linguistique et le triangle sémiotique « classique », proposé par un certain nombre d'auteurs et dont les sommets représentent le symbole, la notion et l'objet. Un certain consensus règne actuellement dans la communauté des terminologues sur la définition du terme. La définition donnée par H. Felber (1987), de l'École de Vienne, dans son manuel de terminologie, fait autorité : « Un terme est un symbole conventionnel représentant une notion définie dans un certain domaine de savoir ». Prise littéralement, cette définition du terme consacre une rupture entre le terme et le mot de la langue.

2.2. « Revisiter » la doctrine terminologique

Le point qui nous semble d'abord critiquable dans la définition du terme imposée par la doctrine terminologique est que cette définition participe d'une conception de la terminologie qui donne la primauté aux concepts sur leurs expressions linguistiques, ces dernières n'étant considérées que comme de simples symboles censés représenter de façon univoque ces notions. Cette vision mécaniste du couplage entre le terme et la notion s'est imposée dans le cadre intellectuel de l'universalisme et de l'empirisme logique, que le monde scientifique a depuis largement remis en cause (Slodzian, 1994, 1995). Cette vision ignore la complexité des phénomènes langagiers qui sont à l'œuvre dans les textes spécialisés, comme dans tous les types de textes, et donc interdit une analyse sémantique productive de ce type de textes. Nous souhaitons conjuguer nos efforts à ceux des chercheurs en linguistique, en socio-linguistique, en épistémologie, ainsi qu'en terminologie, qui visent à « revisiter » la doctrine terminologique, pour proposer une approche renouvelée de la terminologie qui intègre la terminologie au sein de la linguistique (de spécialité). Nos arguments prennent leur source dans la réflexion théorique qu'exigent la conception et la réalisation d'un outil automatique d'analyse de textes pour l'extraction de terminologie, ainsi que son utilisation dans des contextes applicatifs réels.

Sur le plan pragmatique, la conception doctrinaire du terme et de la terminologie se heurte à certaines réalités de la pratique terminographique, qui doit répondre à des besoins nouveaux. Ces besoins naissent de la demande plus forte pour une maîtrise des terminologies spécialisées et pour leurs utilisations dans des contextes de plus en plus diversifiés (Condamines, 1995). La normalisation n'est pas l'objectif prioritaire de l'activité terminologique dans les entreprises et les grandes organisations. Les besoins se situent d'abord au niveau de la description. Les types de « produits ter-

minologiques » que les terminographes vont avoir à réaliser sont de plus en plus variés. À côté des applications traditionnelles que constituent les bases de données terminologiques multilingues pour la traduction et les recueils de définitions pour l'enseignement, on voit apparaître de nouveaux types d'applications. C'est ainsi que, à la Direction des études et recherches d'EDF, nous menons diverses expériences dans lesquelles le logiciel d'aide au dépouillement terminologique Lexter est utilisé pour la réalisation de produits terminologiques de types divers. Ces projets s'inscrivent principalement dans le champ de la Gestion Électronique de Documents (GED). Nous concevons des systèmes de consultation de documentation technique, dans lesquels la terminologie joue un rôle central, sous la forme soit d'index terminologique structuré, soit de modèle terminologico-conceptuel.

Les besoins accrus en accès aux bases textuelles de plus en plus volumineuses feront certainement émerger la nécessité d'autres types de produits terminologiques. Pour toutes ces applications, le critère de définition du terme comme symbole d'une notion dans un domaine, déjà critiquable sur le plan théorique, s'avère non opératoire sur le plan empirique. En effet, l'expérience montre que, selon le type de produit terminologique à construire, les éléments lexicaux retenus, pour un même domaine et à partir d'un même corpus, seront différents.

Il convient donc d'abandonner l'approche néopositiviste qui pose la préexistence *a priori* d'un système de concepts que la terminologie aurait pour charge de dévoiler, pour adopter une démarche constructiviste et fonctionnelle plus propre à une approche linguistique de l'analyse des textes spécialisés, comme celle proposée par Lerat (1995). Il faut distinguer la notion théorique de *concept* dans les sciences cognitives, comme constituant fondamental de la pensée et des croyances, et la notion opératoire de *concept* en linguistique de spécialité. Dans le cadre qui nous concerne ici, à savoir la construction d'un modèle terminologico-conceptuel d'un domaine ou d'une activité, le concept est un construit. Il est le résultat produit par une analyse sémantique réglée d'un corpus constitué et pour une utilisation identifiée. Notre position sur ce point s'appuie sur les propositions de Rastier (1987, 1994, 1995). Construire un modèle terminologico-conceptuel, c'est produire un artefact, dans un code sémiotique particulier, celui des modèles de représentation des connaissances développés par l'intelligence artificielle. Le passage au concept est donc le résultat d'un travail de modélisation. La linguistique de spécialité doit s'intéresser aux conditions de ce passage, et ici elle doit se rapprocher de cette branche de l'intelligence artificielle qui s'intéresse à l'acquisition et à la modélisation des connaissances (Bourigault et Lépine, 1995 ; Bachimont, 1995).

3. La notion de candidat terme

3.1. De la notion de terme à celle de candidat terme

Le contexte de la conception et de la réalisation d'un logiciel d'aide au dépouillement terminologique est le suivant : il s'agit de concevoir un programme informatique qui recevra en entrée un corpus de textes techniques portant sur un domaine (quelconque), et qui devra en sortie proposer des mots ou des séquences de mots, extraites de ce corpus, qui pourront être retenus par un analyste humain en charge de la construction d'un produit terminologique (lui aussi quelconque). Nous imposons donc au système

une double contrainte de généralité, sur le domaine et sur le produit terminologique à construire.

La mise au point contrôlée d'un outil informatique d'aide au dépouillement terminologique exige une caractérisation linguistique rigoureuse du type de séquences, que nous désignerons sous le nom de « candidats termes », que ce système va avoir à extraire par analyse du corpus traité. La critique de la définition du terme, esquissée dans la section précédente, ne remet pas en cause la possibilité d'une telle caractérisation, mais au contraire la rend possible. Il s'agit de s'appuyer sur le constat selon lequel tout lecteur, averti ou non, est capable de relever dans les textes spécialisés des séquences de mots que son intuition pousserait à qualifier de « termes », ou, de façon plus prudente d'« unité lexicale complexe », d'« unité polylexicale », de « lexie complexe », ou encore de « synapsie ». Dans le cadre de la conception d'un outil d'aide au dépouillement terminologique, nous avons cherché à formaliser autant que possible les bases sur lesquelles repose ce jugement interprétatif, de façon à en déduire des règles opératoires de dépistage.

Nous caractériserons un candidat terme comme toute séquence *attestée dans le corpus* traité qui vérifie les contraintes de *validité syntaxique* et d'*autonomie discursive*. Cette caractérisation est sujette à discussion. Comme nous le verrons dans les deux sections suivantes, aucun des deux critères mentionnés ne va sans poser de sérieuses questions, à la fois sur le plan théorique de l'analyse linguistique et le plan pratique de l'implémentation informatique. Néanmoins, l'utilité de cette caractérisation est de fournir un cadre d'analyse pour guider la recherche de règles opératoires, implémentables dans un système d'analyse automatique de textes.

3.2. Le critère de la validité syntaxique

Le premier critère de caractérisation du candidat terme est celui de la validité syntaxique : un candidat terme doit correspondre à une séquence syntaxiquement valide dans la proposition de laquelle il a été extrait. Nous illustrons ce critère à l'aide des quelques exemples donnés dans le tableau 1.

L'exemple 1 pose le problème classique du rattachement des adjectifs et des groupes prépositionnels. La description syntaxique du groupe révèle que la séquence **file de filtration** n'est pas un bon candidat terme au sens du critère de la validité syntaxique, bien qu'elle corresponde à un patron terminologique valide ('nom préposition nom'). Le sous-groupe **filtration iodée** est un bon candidat terme.

L'exemple 2 illustre le problème posé par les propriétés de sous-catégorisation des adjectifs. La description syntaxique du groupe révèle que la séquence **capteur sensible** n'est pas un bon candidat terme au sens du critère de la validité syntaxique, bien qu'elle corresponde à un patron terminologique valide ('nom adjectif'). Cela ne signifie pas que la collocation qui associe l'unité lexicale « capteur » à l'unité lexicale « sensible » ne soit pas digne d'intérêt, mais pour conserver un contrôle sur les traitements effectués par le système, il est nécessaire de contraindre celui-ci à ne chercher à extraire que des séquences syntaxiquement valides.

description syntaxique :	[1]	*file de filtration iodée* [file de [filtration iodée]] (-) **file de filtration**
description syntaxique :	[2]	*un capteur sensible à une élévation de température.* un [capteur [sensible à une élévation de température]] (-) **capteur sensible**
description syntaxique :	[3]	*On installe le câble contre le coffret de décharge.* On installe [le câble] [contre le coffret de décharge] (-) **câble contre le coffret de décharge**
description syntaxique :	[3bis]	*On installe une protection contre les grands froids.* On installe [une protection contre les grands froids] (+) **protection contre les grands froids**

TABLEAU 1 : Illustration du critère de la validité syntaxique. Sous chaque exemple figure sa description syntaxique (partielle) correcte On en déduit les séquences qui vérifient (+) ou ne vérifient pas (-) le critère de la validité syntaxique

Les exemples 3 et (3bis) illustrent le problème posé par les propriétés de sous-catégorisation des noms. La description syntaxique de l'exemple (3) révèle que la séquence **câble contre le coffret de décharge** n'est pas un bon candidat terme au sens du critère de la validité syntaxique, alors que dans l'exemple (3bis), qui présente une configuration analogue en terme de succession de catégorie grammaticale (un nom suivi de la préposition « contre » et de l'article défini), la description syntaxique révèle que la séquence **protection contre les grands froids** constitue elle un bon candidat terme. Dans ce dernier cas, le nom « protection » est construit avec un complément introduit par la préposition « contre », alors que dans l'exemple (3), cette préposition introduit un complément du verbe « raccorde » de la proposition principale.

3.3. Le critère de l'autonomie discursive

Le second critère de caractérisation du candidat terme est celui de l'autonomie discursive : un candidat terme doit posséder une autonomie discursive vis-à-vis du contexte textuel duquel il a été extrait. Ce critère vient réduire encore le champ des possibles. Parmi les groupes syntaxiquement valides, il faut éliminer ceux qui n'ont pas d'autonomie contextuelle, c'est-à-dire ceux qui ne peuvent être interprétés sans le contexte textuel antérieur ou postérieur. Ceux-là ne peuvent en aucun cas être extraits de leur contexte textuel pour être intégrés comme entrée autonome dans un quelconque produit terminologique.

Nous illustrons ce critère à l'aide des quelques exemples donnés dans le tableau 2, qui concernent tous l'article défini en français.

En discours, l'article défini est passible de diverses valeurs sémantiques, qui ne

sont pas toutes compatibles avec une extraction hors contexte. Dans les exemples (4) et (5), l'article défini, avec les valeurs anaphorique et cataphorique ne peut être correctement interprété sans l'accès aux contextes textuels antérieur et postérieur. Les séquences **intégrité de la paroi** ne sont pas de bons candidats termes au sens du critère de l'autonomie discursive.

Par contre, dans l'exemple (6), l'article défini a la valeur d'unique. Le corpus traité constitue la description d'une tranche nucléaire générique, d'un type bien défini. Dans ce cas, le groupe « la tranche » supporte toujours la même interprétation (« la tranche générique » dont on donne la description dans cette documentation), et peut donc être extrait de tout contexte textuel en tant que constituant de candidat terme. Dans cet exemple, la séquence **niveau de puissance de la tranche** est un bon candidat terme au sens du critère de l'autonomie discursive. L'exemple (7) illustre la valeur de générique qui elle aussi est compatible avec une extraction hors contexte.

[4] *Le circuit d'aspersion est installé contre une paroi de confinement. Il a pour rôle de maintenir l'intégrité de* ***la*** *paroi.*	
valeur de l'article défini :	anaphorique
	(-) **intégrité de la paroi**
[5] *Le circuit d'aspersion a pour rôle de maintenir l'intégrité de* ***la*** *paroi contre laquelle il est installé.*	
valeur de l'article défini :	cataphorique
	(-) **intégrité de la paroi**
[6] *Ce système règle le niveau de puissance de* ***la*** *tranche.*	
valeur de l'article défini :	unique
	(+) **niveau de puissance de la tranche**
[7] *sensibilité à* ***la*** *chaleur*	
valeur de l'article défini :	générique
	(+) **sensibilité à la chaleur**

TABLEAU 2 . Illustration du critère de l'autonomie discursive. Sous chaque exemple figure la valeur sémantique de l'article défini. On en déduit les séquences qui vérifient (+) ou ne vérifient pas (-) le critère de l'autonomie discursive.

4. Aspects méthodologiques de conception : approche expérimentale

Étant donné la caractérisation que nous venons de donner du candidat terme, il apparaît clairement qu'un système automatique de repérage de candidats termes doit mettre en œuvre une analyse de type syntaxique, comme l'ont déjà souligné David et Plante (1990). Le critère de la validité syntaxique exige que le système soit en particulier doté de règles d'analyse capables de résoudre au mieux les problèmes de rattachement dans les situations ambiguës. Le critère de l'autonomie pose des problèmes

d'implémentation encore plus délicats. L'essentiel de nos efforts actuels de développement du logiciel Lexter concerne ce point. Nous ne décrivons pas dans cette section les techniques de Traitement Automatique des Langues Naturelles mises en œuvre dans Lexter, nous exposons quelques points de méthode concernant les phases de conception et de réalisation d'un tel système.

Il s'agit d'implémenter des règles de dépistage qui extraient du corpus d'apprentissage des séquences de mots qui satisfassent autant que possible les critères de la validité syntaxique et de l'autonomie discursive. La tâche de conception et de réalisation d'un système automatique de dépistage de candidats termes exige que soient menées de front deux types d'activité : une activité de recherche théorique sur la caractérisation linguistique du terme et sur son fonctionnement discursif, et une activité d'implémentation informatique de règles de dépistage dans l'outil d'analyse automatique. Ces deux activités se développent conjointement dans une démarche dialectique très fructueuse. L'analyse théorique guide la réalisation informatique, tout en profitant de ses résultats. Les deux activités se nourrissent mutuellement. La linguistique joue un rôle de régulation, avant, pendant et après la réalisation informatique :

- en amont de la conception du système, une analyse linguistique du fonctionnement du terme en discours est nécessaire pour établir les principes de base optimaux de l'analyse automatique ;
- pendant la mise au point du système, les règles d'analyse effectivement implémentées sont établies dans une démarche linguistique expérimentale privilégiant le test sur corpus ;
- en aval, la validation de l'outil et l'édification d'une méthode d'utilisation de l'outil par un expert humain procède encore d'une analyse linguistique et ergonomique de l'activité terminologique.

L'activité de dépouillement terminologique est une activité d'analyse conceptuelle d'un domaine, et donc en ce sens une activité « hautement » intellectuelle. Doter une machine de règles de dépistage de candidats termes dans un texte relève d'une certaine gageure. Mais l'analyse linguistique, à laquelle se subordonne l'implémentation informatique, permet de gérer et de maîtriser le processus d'approximation que constitue l'établissement de règles opératoires pour une machine. En ce sens, elle révèle et assume les limites des capacités d'une machine à travailler sur le sens.

5. Conclusion : linguistique de corpus

La démarche de conception d'un logiciel d'aide au dépouillement terminologique est donc par essence de type expérimental. Analyse théorique, investigation et expérimentation sur corpus sont menées de pair. Le corpus joue un rôle de pivot dans la démarche :

(i) en tant qu'objet d'analyse pour le système ;
(ii) en tant que source d'information pour le système (Lexter est doté de procédures dites « d'apprentissage endogène » qui lui permettent d'acquérir par lui-même certaines informations syntaxiques de sous-catégorisation dont il a besoin pour effectuer une analyse syntaxique précise) ;
(iii) en tant qu'élément de base du dispositif expérimental.

Ce dernier aspect est essentiel. Une fois donnés les principes généraux de conception, nous avons progressivement élaboré les techniques d'analyse et les règles des différents modules du système en associant réflexion linguistique et validation par test sur corpus. L'analyseur sert d'outil d'investigation dans les corpus (on peut parler dans ce cas d'analyse linguistique assistée par ordinateur). Il s'agit alors de concilier les visées de l'analyse linguistique qui met en exergue les phénomènes marginaux et donne au contre-exemple un pouvoir de remise en cause, et les contraintes de la réalisation informatique, qui privilégient les phénomènes de masse. L'expérimentation sur corpus est une activité qui exige patience et rigueur, et qui peut être parfois fastidieuse. Parce qu'elle dévoile toujours des problèmes nouveaux, la confrontation avec le corpus est à la fois décourageante et passionnante.

Amélioration automatique incrémentale de dictionnaires bilingues utilisant un corpus monolingue

Kumiko TANAKA et Violaine PRINCE

Université de Tokyo, Japon et LIMSI-CNRS, Paris, France

1. Introduction

Pour développer automatiquement des dictionnaires électroniques bilingues, il faut aligner des corpus bilingues. Mais, réciproquement, l'alignement de corpus ne peut se faire sans l'aide de dictionnaires bilingues relativement complets (Utsuro *et al.*, 1994). Cette interdépendance entre l'alignement des corpus et la construction de dictionnaires électroniques bilingues est une des raisons pour lesquelles l'identification et la mise à jour de dictionnaires bilingues reste un problème difficile. L'objectif que nous nous proposons d'atteindre par le biais de l'algorithme présenté dans cette contribution est de transformer le problème de la mise en correspondance bilingue, dont la difficulté vient d'être citée, en un problème monolingue de recherche d'un ensemble de mots (M) sémantiquement proches de l'entrée lexicale originelle. Ce résultat peut ensuite être transféré dans le cadre bilingue, par transfert des équivalents des mots de M comme équivalents de l'entrée originelle.

Dans des travaux préalables de génération de dictionnaires bilingues par l'intermédiaire d'une troisième langue jouant le rôle de pivot, Tanaka et Umemura (1994) considèrent que les mots ayant des sens multiples dans la langue pivot transportent des équivalents qui ne sont pas des candidats pertinents. Néanmoins, ces candidats parasites peuvent être écartés en procédant à une consultation inverse du dictionnaire. Cette hypothèse peut ici être exprimée différemment : le problème de la correspondance de mots entre langues peut être rapporté à un problème intralinguistique (dans la langue pivot), problème de mesure de proximité sémantique entre deux mots. Dans ce même ordre d'idée, nous nous proposons – au lieu d'utiliser une troisième langue pivot et de transporter des termes parasites pour les écarter ensuite – de nous appuyer sur une étude de proximité sémantique grâce au traitement d'un corpus monolingue d'une part, et aux informations de synonymie et de proximité morphologique conte-

nues dans des lexiques électroniques de la langue source (LS) d'autre part, dans le but de compléter automatiquement des dictionnaires bilingues.

Pour cela, nous nous donnons trois types d'informations monolingues :

- des heuristiques calculant une certaine proximité sémantique grâce à la présence de similarités morphémiques (très indicatives de l'identité de racine en japonais) ;
- la présence de synonymes fournis par le lexique de la LS ;
- des valeurs de cooccurrence établies à partir de grands corpus (aussi en LS).

En pratique, le fait d'utiliser les corpus comme sources de relations sémantiques donne un aspect incrémental à notre algorithme de raffinement de dictionnaires. En effet, les corpus rendent compte des évolutions sémantiques et de la dynamique de la langue, ce qui est important pour la mise à jour des dictionnaires électroniques bilingues. Nous pensons que lorsque les dictionnaires sont incomplets ou comportent des équivalences qui ne sont plus forcément à jour, des extraits actualisés de discours que sont les corpus contemporains sont probablement les sources les plus riches pour transformer ces dictionnaires. De plus, il existe de nombreuses méthodes relativement éprouvées de calcul de la similarité sémantique dans les corpus, ce qui en fait une source aisément exploitable. Enfin, si les corpus sont spécialisés dans un domaine, on peut se servir de cette particularité pour spécifier plus précisément la terminologie en vigueur afin de produire des dictionnaires de spécialité.

Si nous avons choisi le japonais comme LS et l'anglais comme langue cible (LC), c'est pour les raisons suivantes :

- le japonais utilise les idéogrammes kanji et la formation de mots à partir d'idéogrammes se fait de telle sorte que la reconnaissance des racines morphologiques est évidente à mettre en œuvre et fournit beaucoup de renseignements ;
- le japonais est très différent des langues indo-européennes et on ne peut pas jouer sur la proximité des lexies pour deviner le sens, comme on peut le faire entre les langues à base latine ;
- un bon dictionnaire électronique bilingue japonais-anglais existe, ce qui permet de tester l'algorithme dont le but est d'être ensuite appliqué au français pour lequel il n'existe malheureusement pas de bons dictionnaires bilingues avec le japonais. Le bon dictionnaire servira de structure témoin, et pour simuler une situation de dictionnaire incomplet, nous « dégraderons » le dictionnaire d'origine (c'est-à-dire que nous en produirons une version délibérément amoindrie) de manière à voir dans quelle mesure l'algorithme et l'usage de corpus nous permettent d'au moins restaurer le bon dictionnaire d'origine. Par la suite, si l'algorithme s'avère bon, il sera facile de l'appliquer au dictionnaire japonais-français, qui sera effectivement enrichi par cette manœuvre.

Dans la section suivante, nous expliquons la méthode sous-jacente à notre algorithme en l'illustrant par un exemple. La section 3 reprend l'algorithme de manière formelle et montre le principe incrémental de la méthode. La section 4 décrit les données ; et la section 5 analyse les résultats obtenus.

Dans ce qui vient, les lexies japonaises sont translittérées en romain (alphabet romain) en *italique* avec chaque idéogramme kanji séparé par un tiret («-»). Nous avons

associé le sens de chaque forme translittérée entre parenthèses. Les termes anglais sont en courier. Les traductions françaises des termes anglais sont fournies entre parenthèses consécutivement à ces derniers.

2. Présentation générale de l'algorithme

Pour illustrer notre méthode nous avons pris le cas du mot japonais *ken-kyuu* (recherche). Nous avons commencé par produire un dictionnaire dégradé japonais-anglais, c'est-à-dire une dictionnaire dans lequel la lexie *ken-kyuu* n'est rattachée qu'aux deux mots anglais research (recherche) et work (travail). Nous souhaiterions pouvoir récupérer des entrées telles que search (recherche dans le sens de quête, ou enquête), investigation qui sont proches du terme anglais research (termes qui sont en fait des synonymes du mot français « recherche »).

Une manière de faire consiste à utiliser un corpus aligné et de compter les mots qui cooccurrent, dans les deux langues, avec *ken-kyuu*. Une autre manière consiste à s'appuyer sur des renseignements en provenance d'un thésaurus dans la LS. À partir d'un dictionnaire japonais, nous avons obtenu les informations suivantes : *ken-kyuu* est relié aux mots *chou-sa* (investigation) et *tan-kyuu* (quête, enquête). Nous pouvons donc raisonnablement penser que, dans un dictionnaire de qualité moyenne, *chou-sa* a des équivalents tels que search et investigation et que *tan-kyuu* serait relié avec research et search. Ces associations sont représentées dans le graphe de la figure 1.

À partir de cela, nous définissons la notion de **correspondance** entre deux mots, comme étant tout arc du graphe permettant de relier un mot avec un autre. Les équivalents d'une entrée lexicale peuvent être redéfinis comme des mots dans la LC, reliés à elle par des correspondances. Nous définissons de même la notion de **similarité**, comme étant la relation entre les mots de la LS en correspondance avec notre entrée lexicale. Ces mots sont alors appelés des **similaires**.

Dans notre exemple, le mot anglais search possède deux mots en correspondance avec lui, chacun passant par le biais de *chou-sa* d'une part, et de *tan-kyuu* d'autre part. De plus, le mot investigation possède une correspondance passant par le biais de *chou-sa*. De ce fait, le mot anglais search devrait avoir un lien plus fort avec *ken-kyuu* que n'en a le mot investigation[1]. Il faut aussi remarquer que dans la mesure où research reçoit des arcs aussi bien depuis *tan-kyuu* que depuis *ken-kyuu*, dès lors, *ken-kyuu* est plus fortement lié à research que dans la version originelle.

La manière la plus simple de définir l'importance d'une relation entre une entrée lexicale et ses équivalents consiste à pondérer les mots qui sont en correspondance avec elle. Par conséquent, nous appellerons le poids d'une correspondance (PC) entre deux mots *a* et *b* une valeur qui sera désignée par la formule *w(a, b)*.

1 Ce qui voudrait dire qu'il faudrait privilégier l'équivalent français « recherche » à l'équivalent français « investigation »

Les PC entre une entrée lexicale et ses équivalents sont initialement calculés à partir du dictionnaire dégradé. Ce poids est soit égal à l'inverse du nombre d'équivalents, ou bien peut dépendre de l'ordre des équivalents apparaissant dans le dictionnaire (on fait l'hypothèse que l'équivalent le plus fort est celui qui apparaît en premier, et ainsi de suite). Les valeurs données dans la figure 1 illustrent notre mode de calcul de *w(a, b)*.

Les PC entre l'entrée lexicale et ses similaires peuvent être judicieusement calculés à l'aide de trois heuristiques :

- la similarité morphologique (graphèmes communs en japonais) ;
- la coprésence de synonymes dans le thésaurus en LS ;
- les valeurs de cooccurrence de termes, valeurs obtenues dans de grands corpus en LS.

En ce qui concerne les aspects morphologiques, *tan-kyuu* possède un graphème (qui est aussi un morphème) commun avec *ken-kyuu* qui est *kyuu*. Nous définissons alors ce que nous appelons le **score morphologique** comme étant le nombre de kanjis (morphographèmes) communs entre deux termes de la LS. Par exemple, lorsque l'on compare *chou-sa* et *gaku-jutsu-chou-sa* il existe deux kanjis communs *chou* et *sa*, et donc le score morphologique est de 2. Dans le cas de notre exemple, la première ligne de la table 1 est ainsi obtenue.

Le deuxième nombre signifiant est le nombre de synonymes de l'entrée lexicale considérée, dans le lexique en LS. Dans notre expérience, *ken-kyuu* possède deux synonymes : *tan-kyuu*, *gaku-mon*. Dès lors, le poids de *tan-kyuu* augmente, mais pas celui de *chou-sa*. C'est ainsi que nous avons calculé les valeurs indiquées dans la deuxième ligne de la table 1.

Le troisième nombre que nous considérons est une valeur de cooccurrence dans un corpus. La manière la plus simple d'évaluer la cooccurrence est d'utiliser le principe d'information mutuelle de Church[2], valeur obtenue à partir de corpus en LS. La quatrième ligne de la table 1 est calculée à partir d'un corpus de 33 méga-octets, avec une fenêtre de 1 024 octets.

Les deux premiers nombres dénotent des informations statiques, alors que le troisième varie en fonction de la taille et de la nature du corpus.

En terme de méthode, nous commençons par transformer le dictionnaire à l'aide des poids d'origine statique, puis nous le transformons à nouveau à partir du poids en provenance du corpus. Ensuite nous normalisons chaque ligne de la matrice des poids en divisant chacune des valeurs qu'elle contient avec la somme des valeurs (norme

2 La valeur d'information mutuelle (Church, 1990) entre deux mots *a* et *b* (tous deux en LS) est obtenue par la formule suivante

$$\log_2 \frac{P(a, b)}{P(a)P(b)}$$

où P(a) est la probabilité de trouver le mot *a* dans le corpus

pondérée). C'est ainsi que nous engendrons les valeurs des lignes 1 et 3 de la matrice 1. Les lignes ainsi obtenues sont ensuite additionnées, et les sommes sont stockées dans la quatrième ligne, laquelle est de nouveau normalisée, et la valeur normale est stockée dans la cinquième ligne. Cette dernière ligne reflète les PC des similaires d'une entrée lexicale dont une illustration est fournie en figure 1.

Le raffinement du dictionnaire peut être défini comme la réévaluation des PC pour chaque équivalent d'une entrée lexicale. Cela peut se faire à partir de la somme des PC originels, *modulo* les poids obtenus par les trois scores (morphologique, synonymes et corpus), et cela, par le biais de la procédure suivante :

$$w_{new}(x_i, y_j) = r(w(x_i, y_j) + \sum_{\forall k \in D} w(x_i, x_k) w(x_k, y_j)) \quad (1)$$

où $r = 1.0/\sum w_{new}(x_i, y_j)$. Par exemple :

w(*ken* — *kyuu*, `research`)
= *r* x (0.9 + 0.75 x 0.2 + 0.25 x 0.6)
w(*ken* — *kyuu*, `work`) = *r* x 0.1
w(*ken* — *kyuu*, `search`) = *r* x (0.25 x 0.4 + 0.75 x 0.3)
w(*ken* — *kyuu*, `investigation`) = *r* x (0.75 x 0.8)

où *r* est la somme de *w*(*kenkyuu*, `research`), *w*(*kenkyuu*, `work`) et *w*(*kenkyuu*, `search`). Le nouveau graphe avec les nouveaux PC des équivalents de *ken-kyuu* est représenté dans la figure 2. Le dictionnaire s'est enrichi de nouvelles correspondances, dans la mesure où ce nouveau graphe possède plus d'arcs et de nœuds que le précédent, et cela, à partir de connaissances en LS.

Pour modifier les PC avec les poids obtenus à partir du corpus, nous appliquons la même procédure que pour les scores statiques, en utilisant les PC de la table 1 associés aux similaires de l'entrée lexicale. Le PC pour chaque paire de mots, dans une langue donnée, est impossible à calculer, puisque théoriquement la matrice des co-occurrences potentielles est de dimension 10^9. Par conséquent, pour restreindre la charge calculatoire, nous avons choisi de retenir les lexies japonaises qui appartiennent à au moins un des deux ensembles suivants :

- l'ensemble des synonymes de toutes les entrées, lexicales que nous considérons, synonymes se trouvant dans le lexique en LS que nous possédons ;
- les termes japonais obtenus en recherchant dans le dictionnaire anglais-japonais pour obtenir les équivalents anglais.

	tan-kyuu	*chou-sa*
morphème	1,00	0,00
synonyme	1,00 0,50	1,00 0,50
morph + syn	1,50 0,75	0,50 0,25
corpus	11,18 0,57	8,48 0,43

TABLEAU 1 · Nombres discriminants pour le poids de correspondance.

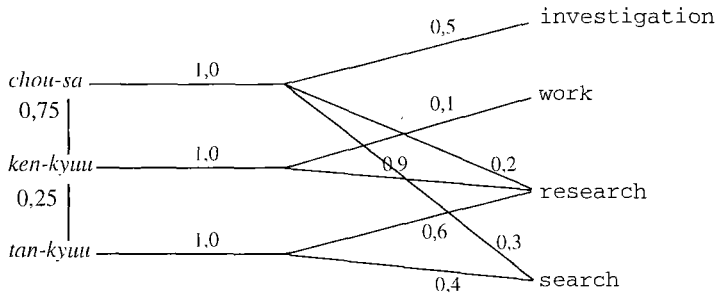


FIGURE 1 : Structure représentant les connaissances d'origine

En conclusion de cette présentation générale de notre méthode, nous dirons, pour résumer, que l'algorithme proposé réduit la problématique bilingue (c'est-à-dire « est-ce que deux mots dans des langues différentes ont un sens similaire ou pas ? ») en une problématique monolingue (c'est-à-dire « est-ce que deux mots, dans une même langue, ont un sens similaire ou pas ? »). Le problème ainsi transposé peut être ensuite restitué dans le cadre bilingue d'origine en recalculant les poids de correspondance entre une entrée lexicale en LS et ses équivalents en LC, ces poids étant modifiés par des scores dénotant des connaissances morphologiques et sémantiques (synonymes), puis normalisés. Dans la section suivante, nous décrivons formellement les différents pas de la méthode que nous avons suivie.

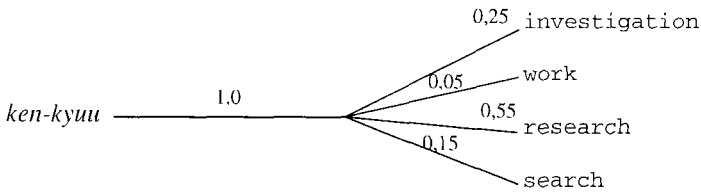


FIGURE 2 : Poids des correspondances affectés par les connaissances morphologiques et sémantiques.

3. Description formelle de la méthode

3.1. Matrices de correspondance bilingue et monolingue

Les mots de la LS sont notés *s* et ceux de la LC sont notés *t*. Nous définissons une matrice de correspondance bilingue, symbolisée par *BCM* (pour *Bilingual Correspondence Matrix*) dont chaque (*i*, *j*)-th élément est noté $w(s_i, t_j)$. Comme nous l'avons précédemment expliqué, la matrice est initialisée à partir du dictionnaire bilingue d'origine, dont on sait qu'il est de qualité médiocre (incomplet ou éventuellement erroné). Dans cette matrice, on fait l'hypothèse que le mot s_i possède *n* équivalents notés t_k (k=1,...,n). Il existe deux façons de calculer la pondération initiale des équivalents. Soit on décide d'affecter une équiprobabilité à l'ensemble des termes, ce qui donnera une valeur initiale de $1/n$ à chaque élément $w(s_i, t_j)$, soit on décide de tenir compte de l'ordre de présentation des équivalents dans le dictionnaire initial et on affecte au *i*ème équivalent un poids égal à $1-(2i/(n(n+1)))$ (la somme d'un rang est 1,0).

Nous définissons ensuite une matrice de cooccurrence monolingue, symbolisée par *MCM* (pour *Monolingual Cooccurrence Matrix*) dans laquelle chaque élément (*i*, *j*)th est noté $w(s_i, s_j)$. Cet élément désigne les valeurs de pondération des similaires obtenues à partir des trois nombres discriminants représentant des connaissances lexicales (morphologiques, sémantiques et cooccurrence) tels que nous les avons présentés dans la section précédente.

Cette même section avait décrit la procédure locale de transposition de la problématique bilingue en monolingue, et de restitution dans le cadre d'origine, en l'illustrant par l'exemple de l'entrée lexicale japonaise *ken-kyuu*. En pratique, cette procédure se définit comme une multiplication de matrices et ce, par le biais de la formule suivante :

$$(MCM + E) \times BCM$$

E est la matrice unité (neutre de la multiplication). Elle est ajoutée pour bien marquer l'état originel de la matrice de correspondance bilingue, c'est-à-dire les correspondances originelles entre une lexie source *s* et une lexie cible *t*. Cette multiplication des correspondances est exactement ce qui est montré dans la formule 1.

Notons que l'addition et la multiplication de matrices sont toujours suivies par une normalisation, puisque les valeurs correspondent à des poids. Remarquons de même que la matrice résultante a la même dimension que *BCM*. Certains éléments ini-

tialement évalués à 0 (simplement parce qu'ils n'étaient pas représentés dans la pondération d'origine) reçoivent ensuite des poids non nuls, par addition des PC des équivalents.

3.2. Incrémentalité

Parmi les heuristiques considérées, nous avions déjà mentionné que le score morphologique et le nombre des synonymes étaient des informations statiques et relativement stables, alors que les valeurs de cooccurrence de termes dans des corpus apparaissent comme dynamiques et sans stabilité *a priori*, puisqu'elles peuvent varier en fonction de la nature du corpus et de sa taille. La matrice *BCM* obtenue après pondération par le biais du score morphologique et du nombre des synonymes est notée MCM_s. La matrice obtenue après pondération par le score de cooccurrence est notée MCM_c, où le terme *c* représente le corpus considéré.

Dans la précédente section, nous avons montré que cette transformation matricielle était réalisée en deux étapes. Effectivement, la première étape consiste à calculer MCM_s, c'est-à-dire à incrémenter le dictionnaire bilingue (la matrice *BCM* d'origine) à partir des sources d'information monolingues et statiques. La deuxième étape calcule la matrice MCM_c à partir de *BCM* et des valeurs de cooccurrence dépendant des propriétés du corpus *c*. En d'autres termes, cette dernière matrice va ajouter à la matrice origine des relations sémantiques dérivées d'un usage actuel de la LS.

Le raffinement final du dictionnaire s'obtient alors par la formule :

$$(MCM_c + E) \times (MCM_s + E) \times BCM$$

En pratique, quand nous parlons de processus incrémental, c'est pour désigner explicitement notre méthode d'obtention de la matrice MCM_c. Celle-ci n'est pas calculée directement, mais de manière récursive sur les corpus eux-mêmes, et ce, dans le but d'atteindre tout de même un minimum de stabilité des relations sémantiques dans les corpus. Il existe deux manières de faire varier l'impact du corpus et de réaliser l'incrémentation :

- soit on considère que la matrice MCM_c utilisée dans la procédure de raffinement définie ci-dessus est elle-même la résultante (le produit) de plusieurs matrices MCM_{ci} et donc qu'elle se définit comme la matrice obtenue sur l'ensemble des corpus d'expérience. Dans ce cas, le membre $(MCM_c + E)$, où *c* représente l'union de tous les corpus, de la formule de raffinement n'est multiplié qu'une seule fois ;
- pour chaque corpus d'expérience, nous calculons un *MCM* local, et ce dernier est appliqué incrémentalement au *BCM* courant, dans la formule de raffinement. Ce qui fait que nous obtenons des dictionnaires plus ou moins raffinés selon les corpus. Le développement de la formule donne :

$$(MCM_{c1} + E) \times \ldots \times (MCM_{c0} + E) \times (MCM_s + E) \times BCM$$

Dans les deux cas, nous n'avons pas la possibilité de démontrer qu'il existe une

distributivité de la loi de composition sur les corpus, c'est-à-dire que :

$$MCM_{c} \times MCM_{c'} = ?MCM_{(c+c')}$$

Cette distributivité voudrait dire que les corpus ont un effet additif, ce qui n'est pas une hypothèse forcément raisonnable en terme de relations de cooccurrence tout au moins. Nous avons choisi, pour cette première expérience, de réaliser une incrémentation du premier type, c'est-à-dire avec une matrice réalisée sur l'ensemble des corpus et multipliée une seule fois, en se réservant pour un futur proche la mise en œuvre du second type et la comparaison des deux techniques. Il est certain que l'incrémentalité à partir de corpus peut produire des effets secondaires : si, par exemple, le dictionnaire d'origine est très pauvre (comme ce sera malheureusement le cas pour notre dictionnaire japonais-français pour lequel toute notre expérience a été menée), et s'il est entraîné incrémentalement sur des corpus homogènes et spécifiques, alors il aura tendance à refléter cette spécificité. Par exemple, si nous entraînons notre dictionnaire dégradé avec des corpus scientifiques uniquement, il est très probable que nous obtiendrons une version scientifique du dictionnaire, ce qui dans certains cas, peut devenir fort utile pour créer des dictionnaires bilingues de spécialité.

corpus	occurrences	couverture
corpus6	3005549	49173
corpus5	1200000	45023
corpus4	672378	39850
corpus3	300000	27476
corpus2	150000	19883
corpus1	75000	13841

TABLEAU 2 : Corpus utilisés dans l'expérience.

4. Les données de l'expérience

Les dictionnaires japonais-anglais et anglais-japonais qui nous ont servi ici sont Ichikawa (1990) et Koine (1990). Le nombre d'entrées lexicales dans le japonais-anglais est 47 808 et nous en comptons 28 168 dans la version anglais-japonais. Le lexique électronique japonais provient de Takebe *et al.* (1976). Il contient 27 145 entrées. Pour les besoins de l'expérience, nous avons créé le dictionnaire dégradé de la manière suivante : les noms communs sont extraits de façon semi-automatique depuis chacun des deux dictionnaires bilingues et sont inclus dans un dictionnaire qui va grosso modo inclure des correspondances un à un entre les termes. De plus, le dictionnaire comprend des erreurs (contresens, absurdités).

Les corpus que nous avons choisis sont extraits des archives du journal *ASAHI* (l'équivalent japonais du journal français *Le Monde*) et forment environ 33 mégaoctets de texte. Le corpus de notre expérience correspond à la totalité du corpus dis-

ponible. Les données sont lemmatisées par le système JUMAN, et nous recueillons les noms et les verbes. Le nombre d'occurrences de chaque lexème retenu et le taux de couverture linguistique sont fournis dans le tableau 2. Pour fournir un cadre à la fonction incrémentale citée dans la section précédente, nous avons coupé l'ensemble des données en six (numérotés de 1 à 6). Les corpus 1 à 5 sont obtenus à partir des différentes parties principales du corpus6 (qui, lui, correspond à la totalité du texte). Par exemple, corpus5 correspond à la première partie de corpus6 et corpus4 à celle de corpus5 et ainsi de suite. Le coefficient d'information mutuelle est calculé avec une fenêtre de 1 024 octets.

Dans notre expérience, nous nous sommes restreints à cent entrées lexicales choisies dans une liste des mots les plus fréquents apparaissant dans le corpus japonais. Les cent mots que nous avons choisis appartiennent au sous-ensemble des mots les plus fréquents obéissant aux conditions suivantes :

- mots possédant au moins deux kanjis ;
- mots apparaissant avec une grande fréquence dans corpus2.

En pratique, bien que les dictionnaires bilingues entre le japonais et l'anglais soient réputés être de bonne qualité, ils ne sont pas exempts d'erreurs. Cependant, nous avons estimé cette qualité comme suffisante pour servir de témoin. Donc, pour mettre en évidence les effets de notre raffinement, les équivalents dans le dictionnaire japonais-anglais des cent mots choisis ont été délibérément dégradés, en supprimant au hasard une bonne moitié, et en ajoutant des traductions inappropriées. Par exemple, *ken-kyuu* était traduit à l'origine par `research`, `work`, `study`, `investigation`, `inquiry`, `examination`, ce qui correspond à un éventail tout à fait honorable d'équivalents. `Study`, `work`, `examination` ont été volontairement remplacés par `tweed` (tweed), `torrid` (torride) et `stover` (étuve, fourneau ou poêle, mais avec une erreur).

5. Analyse de l'expérience et de ses résultats

5.1. L'exemple de *ken-kyuu*

Le nombre de mots japonais qui étaient reliés à *ken-kyuu*, dans nos corpus est de 78. Pour chacun de ces mots, nous avons trouvé de 3 à 7 équivalents en anglais, dans les dictionnaires considérés. En pratique, environ 400 mots anglais étaient concernés par l'expérience. Tous les équivalents qui ont obtenu un score supérieur à 0,01 pour leur poids de correspondance après raffinement utilisant le corpus6 (la totalité des données) sont les suivants :

```
inquiry(0,24), research(0,12), work(0,10), examination(0,075),
exploration(0,074), study(0,034), investigation(0,030), line(0,029),
specialty(0,017), experiment(0,017), question(0,016), test(0,014),
wicker(0,014), labour(0,013), fabrication(0,011), machinery(0,011),
trial(0,011), snapshot(0,011), wonder(0,010), trade(0,010)
```

Remarquons donc que dans les textes considérés, c'est la notion d'enquête (`inquiry`) qui apparaît comme la plus fréquente pour traduire notre terme original de *ken-kyuu*, ce qui est normal pour un corpus journalistique (et qui explique donc le report de la

spécificité du corpus dans le dictionnaire raffiné). Il est ensuite suivi de « recherche » (`research`) pour indiquer la recherche scientifique, lequel est renforcé par le terme suivant immédiatement : « travail » (`work`). Remarquons aussi que les termes `study` (étude), `examination` (examen) et `investigation` (investigation) arrivent en bon rang. Le bruit est en revanche représenté par des termes tels que `line` (ligne), `wicker` (osier), `trade` (profession, commerce), et `snapshot` (instantané) quoique ce dernier terme puisse être relié à la notion dominante d'enquête (policière, juridique,...).

corpus	les 7 mots les plus forts
corpus6	`inquiry research work examination exploration study investigation`
corpus5	`research inquiry work exploration examination line investigation`
corpus4	`inquiry work exploration examination research study line`
corpus3	`inquiry work exploration examination study investigation line`
corpus2	`inquiry examination research work line study question`
corpus1	`examination research line work study inquiry investigation`
morph + syn	`research inquiry line work examination preview study`

TABLEAU 3 Les sept meilleurs équivalents de *ken-kyuu*.

L'évolution incrémentale du résultat après usage de chaque corpus est fournie dans le tableau 3 pour les sept équivalents anglais de *ken-kyuu* les plus importants en terme de poids relatif. Des équivalents assez peu adéquats tels que `line` (ligne) et `preview` (avant-première, annonce) sont apparus dans le champ des sept meilleurs mots après application de MCM_s (matrice des informations d'origine statique, notée « morph + syn » dans le tableau), mais leurs poids tendent à diminuer au fur et à mesure que les corpus considérés s'élargissent. Cela semble indiquer qu'au moins sur les premiers pas de l'algorithme, la qualité de la distribution est proportionnelle à la taille du corpus. En fait on s'aperçoit que les équivalents se stabilisent à partir du corpus4, mais c'est leur ordre qui change. On s'aperçoit aussi que le terme parasite `line` disparaît au niveau du corpus6, qui ne retient plus que des équivalents appropriés.

5.2. Quelques résultats statistiques concernant les 100 mots les plus fréquents

Nous définissons une proportion commune entre deux dictionnaires comme étant le

nombre total d'équivalents (toutes entrées lexicales confondues) divisé par le nombre total d'entrées lexicales communes. Nous notons cette proportion EF (pour *Equivalence Fraction*). Pour chaque mot en LS les équivalents qui ont des poids supérieurs à 0,001 sont conservés, et pour cet ensemble d'équivalents, nous calculons le rappel de la manière suivante : **rappel** est l'EF du dictionnaire dégradé en entrée (noté dictionnaire1) et du dictionnaire témoin japonais-anglais qui sert de référence, noté dictionnaire2. On remarquera qu'à l'origine il y a exactement 45,5 % d'équivalences communes, telles qu'elles apparaissent dans le tableau 4.

corpus	rappel
corpus6	61,8 %
corpus5	63,1 %
corpus4	62,5 %
corpus3	62,8 %
corpus2	55,1 %
corpus1	60,2 %
morph + syn	53,3 %
dégradé	45,5 %

TABLEAU 4 Valeurs de rappel et incrémentalité

Chaque valeur de rappel de ce tableau est calculée après chaque pas dans la procédure incrémentale. Le dictionnaire qui résulte du pas d'incrémentation est comparé avec le dictionnaire témoin.

Le dictionnaire final, résultant de l'application de la procédure incrémentale avec corpus6, fournit un rappel de 61,8 %, ce qui veut dire que le dictionnaire dégradé par nous a évolué depuis une couverture de 45,5 % des équivalents, jusqu'à couvrir environ les deux tiers des équivalents du dictionnaire de référence.

Ainsi, la valeur ajoutée par notre algorithme en terme de rappel seulement est de 16,3 % avec un corpus de 33 méga-octets, qui est de taille modeste lorsque l'on considère des données écrites en kanji, et qui ne provient que de journaux. Cette valeur n'est pas obligatoirement faible dans la mesure où, d'une part effectivement, le corpus n'est pas si grand, et d'autre part, l'usage courant de la langue, tel qu'on peut le voir dans les journaux, n'utilise pas forcément la totalité des associations entre un mot et ses similaires, loin s'en faut. Il existe des synonymes très techniques, et on ne les trouvera pas forcément dans les journaux alors qu'on peut les trouver dans le dictionnaire. Remarquons par ailleurs que pour le mot *ken-kuyu*, l'algorithme a restitué la totalité des équivalents cités : cela signifie que pour des mots de ce type, les corpus journalistiques fournissent une assez bonne couverture en terme d'associations sémantiques, alors que cela peut ne pas être le cas pour d'autres termes, mêmes lorsqu'ils font partie des cents mots les plus employés. On s'aperçoit donc très rapidement que ce sont surtout les résultats qualitatifs de notre algorithme qui sont les plus importants. Outre

ce que nous venons de dire, nous avons obtenu de nouveaux équivalents qui n'étaient pas enregistrés dans le dictionnaire de référence anglais-japonais et qui sont montrés en annexe de cette contribution. Cela signifie notamment que nous sommes en mesure d'enrichir des bases de connaissances lexicales bilingues dynamiquement, avec des associations qui ne sont pas obligatoirement présentes dans des dictionnaires d'usage très général.

A	B	pourcentage
Dégradé	syn + morph	56,1 %
syn + morph	corpus1	94,1 %
corpus1	corpus2	67,2 %
corpus2	corpus3	99,7 %
corpus3	corpus4	91,9 %
corpus4	corpus5	88,4 %
corpus5	corpus6	88,4 %

TABLEAU 5 Convergence de la méthode

Remarquons que la fonction de rappel n'est pas monotone. Ainsi, les valeurs obtenues pour les corpus 3, 4 et 5 sont meilleures qu'avec le dernier corpus, qui est aussi le plus grand en volume. Ce peut être un effet de seuil : à partir d'une certaine taille, plus un corpus est grand, plus certains mots peuvent s'y éparpiller, car les corpus ne sont pas homogènes. Parmi les cent mots sélectionnés, certains ont obtenu une fréquence d'occurrence inférieure à 0,01.

On peut aussi remarquer que le meilleur pourcentage de rappel est fourni au niveau du corpus5. C'est ce qui nous a amenées à nous poser la question de la convergence éventuelle de la méthode. Pour estimer cette convergence, le tableau 5 montre la EF entre le dictionnaire obtenu après l'étape n (appelé A) de l'algorithme et le dictionnaire résultant de l'étape $n + 1$ (appelé dictionnaire B). Les poids correspondants aux critères statiques (score morphologique et de synonymie) modifient l'état du dictionnaire, qui semble alors acquérir une quantité appréciable d'information. En fait, le tableau 4 montre que déjà près de la moitié du gain quantitatif de l'algorithme est réalisé au niveau de l'adjonction de l'information statique. Le corpus commence à influencer le résultat à partir du corpus2. Les effets des corpus 3 et 4 demeurent faibles, alors qu'ils se potentialisent au niveau du corpus5 tout en se stabilisant. Il semble que, de manière globale, 10 % environ des équivalents sont modifiés par le corpus, même lorsque l'on considère le plus important de tous, c'est-à-dire le corpus6. Il est raisonnable de penser que la méthode se stabilise au niveau du corpus5, et que le dictionnaire obtenu après cette étape est le meilleur possible, ce qui conduit à arrêter le processus à ce niveau et à considérer un gain quantitatif de 18 % avec l'algorithme, gain dont on est sûr de la qualité.

5.3. Conclusion sur les résultats

Pour résumer les propos précédents, nous pouvons dire que l'algorithme présenté a permis de concrétiser les points suivants :

- un dictionnaire dégradé (ou de mauvaise qualité) peut gagner numériquement 20 % de bonnes équivalences en plus par la méthode incrémentale ;
- le meilleur score de l'algorithme n'apparaît pas au niveau du corpus le plus grand, mais dans un corpus de taille suffisante assurant la meilleure distribution des occurrences ;
- la fonction de rappel n'est pas monotone, mais semble se stabiliser, ce qui est quand même un indice de sa fiabilité ;
- l'information statique (c'est-à-dire les connaissances obtenues à partir de dictionnaires et de lexiques monolingues en LS) demeure fondamentale, dans la mesure où elle fournit à elle seule la moitié du gain quantitatif et qualitatif de l'algorithme. On ne peut donc pas en faire l'économie, et se contenter de raffiner des dictionnaires exclusivement à partir de corpus ;
- néanmoins, l'information dynamique extraite de morceaux de discours garde tout son intérêt puisqu'elle complète judicieusement cet ensemble d'informations statiques, en fournissant l'autre moitié du gain de raffinement, conséquent à l'application de l'algorithme sur le dictionnaire ;
- la méthode ainsi décrite dote le dictionnaire résultant d'équivalents qui n'étaient pas présents dans le dictionnaire de référence.

Il ne faut cependant pas perdre de vue que notre algorithme possède des limites et nécessite des améliorations, notamment sur les points suivants.

- Le premier point discutable est le choix de la formule de calcul des poids. Ici, les entrées possèdent systématiquement un poids unitaire qui se divise sur les différents arcs afférents : c'est donc une méthode à somme constante. La valeur maximale de 1 est ajoutée (par addition avec la matrice unitaire) pour ne pas avoir d'explosion combinatoire. L'inconvénient d'un tel calcul est que l'on perd la symétrie entre deux entrées qui sont en correspondance. Donc LS et LC ne sont pas interchangeables. Ainsi, dans une situation où le japonais est LS et l'anglais est LC et que l'on raffine le dictionnaire dans le sens japonais-anglais, on ne peut pas simultanément le raffiner dans l'autre sens, en utilisant un corpus en langue anglaise. De plus, les entrées varient avec le nombre de leurs équivalents, dès lors, on ne peut pas associer de prime abord le même poids initial de 1 à toutes les entrées. En pratique, il faudrait pondérer une entrée par rapport à l'ensemble des entrées du dictionnaire.
- Le second point concerne le sort des équivalents dont les scores sont bas. L'algorithme rejette ces équivalents. Cependant, certains peuvent être utiles parce que leur association en contexte avec les entrées concernées indique qu'il existe une relation sémantique d'association, qui n'est pas forcément une relation de synonymie, et qui peut intervenir dans une représentation sémantique conceptuelle de ces mots. On peut supposer que les équivalents dont les poids sont les plus forts sont en relation forte avec l'entrée, que ce soit une relation d'hyperonymie-hyponymie, ou en relation de synonymie. On peut aussi considérer que les équivalents à poids plus faibles ne sont pas absents de l'attribution du sens à un mot, et qu'ils seront alors utiles pour compléter d'une façon pertinente une résolution

de polysémie en contexte. Mais cette réflexion concernant la possibilité de bâtir une véritable représentation conceptuelle à partir d'aspects quantitatifs sur des données fait l'objet d'un autre travail que nous avons entrepris de réaliser pour l'amélioration de dictionnaires à partir de réseaux sémantiques bilingues.

6. Conclusion générale et perspectives

Dans cette contribution, nous avons proposé une méthode de raffinement d'un dictionnaire bilingue de qualité médiocre, à partir de données monolingues fournies dans la langue source. Cette méthode est dite incrémentale, en ce sens qu'elle définit récursivement des matrices-dictionnaires sur des pondérations de termes cooccurrants dans des corpus de la langue source, dont la taille varie à chaque pas de la méthode. Nous avons mis en commun des données aussi bien statiques (lexiques monolingues de la langue source) que dynamiques (associations en discours de lexies sur lesquelles des hypothèses de correspondance sémantique sont faites).

De manière formelle, l'algorithme est centré autour d'une multiplication de deux matrices correspondant l'une à un dictionnaire monolingue, enrichi par des informations morphosémantiques issues d'un lexique, et dynamiquement augmentées par un traitement de corpus, et l'autre à un dictionnaire bilingue dénotant les relations entre les lexies en langue source et leurs équivalents en langue cible.

L'algorithme a été testé avec un dictionnaire japonais-anglais dégradé par suppressions d'une moitié des équivalents et par adjonction de correspondances erronées. Nous avons aussi utilisé un corpus de 33 méga-octets des archives du journal *ASAHI*, et nous avons sélectionné une liste de cent mots qui ont servi à mener l'expérience. Les résultats montrent que l'algorithme a pu restituer la moitié des équivalents manquants, et a introduit de nouveaux équivalents pertinents (voir l'annexe).

Trois voies de recherche ultérieures peuvent être envisagées. Premièrement, rendre symétrique l'algorithme de raffinement qui comme nous l'avons mentionné, donne des résultats dépendant du rôle de chaque langue considérée, ce qui permettra d'utiliser un corpus monolingue dans chacune des deux langues, indifféremment. Deuxièmement, vérifier l'impact des caractéristiques du corpus (taille et nature) pour bien maîtriser les propriétés de l'incrémentalité. Troisièmement, tester l'algorithme avec des corpus typés tels que des corpus scientifiques ou littéraires pour produire des correspondances bilingues plus spécifiques. De toutes façons, quelle que soit la voie qui sera explorée en premier, nous fondons de grands espoirs sur une amélioration des processus automatiques d'extraction de correspondances bilingues. Cela nous mènera assez rapidement vers notre objectif final, qui est la réalisation d'un bon dictionnaire électronique japonais-français et français-japonais, dont le contenu est actualisé par les corpus de discours contemporains.

Remerciements

Les auteurs remercient le Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur (LIMSI) du CNRS, et son directeur Joseph Mariani, pour la mise à disposition de l'environnement de test, l'accueil et les conseils. Le Dr K. Umemura de

Nippon Telegraph and Telephone a gracieusement fourni les données électroniques et nous lui exprimons notre gratitude. Nous remercions de même Dr S. Hayamizu pour ses précieux conseils et ses commentaires intéressants. Le Dr Utsuro de AIST a eu la bonté de nous offrir l'accès à l'analyseur morphologique du japonais JUMAN.

Annexe

japonais	équivalents en anglais
kei-kaku	aim intent meaning purpose idea predestination intention design sense bet schedule message program
jou-hou	communication correspondence mass_communication transmission liaison intercourse junction news telegraph intelligence contact epistle missive letter chou-sa examination exploration test trial research inquiry checkup proof interrogatory audit probation search check hearing medical
hen-ka	mutation alteration demolition cataclysm diversity shift transformation metamorphosis defeat warp transmutation distance switch wreck swing transfiguration vicissitude shuffle
i-ken	apprehension attitude posture understanding judgement position anxiety idea capture bearing attest antenna setup set stance amyloid concern collar assessment unrest care fear view opinion verdict solicitude worry capacity
kou-zou	system scheme dot acer assemble making constitution taylor composition texture frame stout tow structure construction tidal conformation make-up
sho-ri	treatment disposal proceeding cure disposition remedy arrangement
jo-sei	feminine female womankind hen she bitch cow woman daughter
nin-ki	popularity star epidemic boom fashion report notoriety repute reputation currency name child whisper mode

Conception d'un dictionnaire terminologique et phraséologique trilingue anglais/français-arabe dans le domaine de l'optique

Xavier LELUBRE

Université Lumière Lyon-2, France

• *Abstract* •

This is a project of a dictionary aimed at translators, essentially from French or English into Arabic language. This is why it associates to each concept its definition, the set of terms relative to it in each precited language On the other hand this dictionary will offer for each term a cluster of phraseologisms relative to it

It will also take under consideration the problems of synonymy, with a special consideration for Arabic optic terms

Introduction

Le point examiné ici est relatif aux choix et aux contraintes qui président à l'établissement d'un dictionnaire spécialisé trilingue français-anglais-arabe dans le domaine des interférences lumineuses, sous-domaine de l'optique.

Un travail de ce type doit prendre en considération le concept de *langue de spécialité*. Du point de vue qui nous intéresse, quatre composantes contribuent à l'établissement du texte de spécialité :

i. la composante terminologique : la terminologie de la spécialité concernée, c'est-à-dire les dénominations des unités référentielles constituant le domaine traité ;

ii. la composante phraséologique : les phraséologies relatives à cette terminologie, c'est-à-dire les contextes préférentiels d'occurrence des termes dans les textes de spécialité ;

iii. la composante brachygraphique : les signes brachygraphiques ;

iv. la composante discursive : les expressions propres aux types de discours intervenant dans les textes de la spécialité concernée.

Il convient aussi de prendre en compte d'une part les différents lieux des textes de spécialité (en physique, le texte de spécialité est susceptible de se dérouler selon différents moments tels que le corps du texte, y compris les notes de bas de page, le texte des figures (légendes sous la figure ; textes dans la figure) et illustrations, le texte des exercices, le texte des résumés, le texte des titres, les « textes brachygraphiques » (les formules), le texte de notices historiques, etc.) et d'autre part les différents niveaux de spécialisation (publications destinées aux spécialistes, manuels destinés aux apprenants, ouvrages de vulgarisation).

Un texte de spécialité n'est généralement pas homogène ; des discours d'autres spécialités viennent y contribuer (par ex., discours mathématique, notice historique dans un texte d'optique).

Quelle est l'incidence de ces différents aspects sur l'entreprise dictionnairique, c'est-à-dire quels sont donc les éléments de la langue de spécialité qui devront être représentés dans le dictionnaire, et comment vont-ils être représentés ? Elle dépend de toute évidence de la finalité qu'on lui assigne. Il s'agit ici d'un dictionnaire trilingue, et il est – tout naturellement – conçu comme un outil pour la traduction[1].

À titre d'exemple, voici un texte relatif aux interférences lumineuses[2] :

> « C'est dans la région M1 M2 où se superposent les faisceaux diffractés que l'on peut observer les franges d'interférences. En un point quelconque M de l'écran P, la différence de marche δ est (fig 2 b)
> (8) $\delta = [..]$;
> en considérant les triangles A1A2H et COM, et en posant [..], on a :
> (9) $\delta \# [..]$
> [..]
> En première approximation, les franges sont des droites parallèles et équidistantes. Elles sont dirigées perpendiculairement au plan de la figure 2 b. La formule (10) montre que les maximums de lumière, c'est-à-dire les franges brillantes, sont donnés par
> (11) $\delta / \lambda = [..]$.»

Pour traiter un texte de ce type, le traducteur a besoin des éléments suivants :

– **termes** (*faisceaux diffractés, franges d'interférences, franges, écran, différence de marche, triangle, droites parallèles, droites équidistantes, plan, maximum, lumière, franges brillantes*) ;

---

1 Remarquons au passage que le sens le plus fréquent de la traduction va du français ou de l'anglais vers l'arabe, ces deux langues étant aujourd'hui, comme on le sait, les deux principales langues – et de loin – de référence en ce qui concerne les domaines scientifiques, dans le Monde arabe. C'est la raison d'ailleurs pour laquelle un tel ouvrage se doit d'être au moins trilingue, avec le français et l'anglais comme « langues de référence » : ne tenir compte que de l'une de ces deux langues, à l'exclusion de l'autre, constituerait un vice rédhibitoire (ne serait-ce, nous y reviendrons, que parce que les terminologies scientifiques arabes contemporaines se constituent la plupart du temps sur la base des terminologies françaises et anglaises).

2 Maurice Françon, « interférences lumineuses », in *Encyclopaedia Universalis* (vol 9, 1968 : 2). C'est nous qui soulignons.

– connaissance de la façon dont ces termes fonctionnent en contexte, c'est-à-dire les **phraséologismes** au sein desquels ils interviennent (« se superposent les faisceaux diffractés », « observer les franges d'interférences », ...) ;

– signes **brachygraphiques** (d pour la différence de marche, l pour la longueur d'onde, M1M2, A1A1H, ...) ;

– **expressions des discours** organisant le texte (discours mathématique : « en un point quelconque de... », « en considérant les triangles », « en posant », « on a : », « la formule... montre que », « les maximums [..] sont donnés par : ». Physique : « c'est dans la région... que l'on peut observer... », « en première approximation ». Didactique/ terminologique : « les maximums de lumière, c'est-à-dire les franges brillantes ».

1. L'établissement du dictionnaire

L'étude terminologique préalable a porté sur la délimitation du domaine, la constitution d'un corpus, par dépouillement des textes en français, en anglais et surtout en arabe.

Nous avons pris comme base scientifique des textes, français et anglais, de manuels universitaires. Quant aux termes arabes, nous les avons pris essentiellement de manuels arabes qui ont été à notre disposition, tout en tenant compte des lexiques – lexiques bilingues ou trilingues –, sur lesquels bien souvent l'on peut faire de nombreuses réserves des points de vue lexicographique, terminologique et scientifique.

1.1. Organisation générale

Nous avons établi l'arbre (un arbre possible) du domaine et adopté à partir de cet arbre une classification systématique[3]. Chaque unité référentielle est repérée par un indice. Tout terme est repéré par un indice, qui est celui de l'unité référentielle qu'il dénomme, complétée par un code relatif à la langue :

xxFA5/F	frange centrale
xxFA5/E	central fringe
xxFA5/A11	/hudbat markazijjat/
xxFA5/A12	/hudb markazijj/

Les grandes divisions du sous-domaine des interférences lumineuses que nous avons dégagées sont les suivantes :

3. Il s'agit d'un système mixte de classification, hiérarchique et à facettes. Il utilise un ensemble de caractères alphanumériques (Lelubre, 1992 : 141-142).

Division A	(types d'interférences)
Division D	(nature ondulatoire de l'interférence)
Division F	(franges)
Division H	(éléments physiques qui interviennent)
Division K	(interférence en lumière polarisée)
Division U	(dispositifs d'obtention)
Division V	(interféromètres)
Division W	(pièces optiques utilisées pour les dispositifs interférentiels)
Division X	(applications)

1.2. Appareil lexicographique

Il est classique. En particulier, il convient de fournir à l'utilisateur les éléments non prédictibles, comme, pour l'indication du genre pour les substantifs français, de la forme au pluriel pour les substantifs arabes, etc.

Quant aux définitions, chacun des termes qui s'y trouve doit être à son tour une vedette du lexique[4].

2. Traitement des faits de variation

2.1. Prise en compte des phénomènes de synonymie

Les termes synonymes, pour une langue donnée, sont répertoriés par un indice placé après le code de la langue. Les faits de variations sont de nature et de degré variable (Lelubre, 1992 : 381-416) ; cela peut aller de la simple variante orthographique jusqu'à des termes exprimant pour la même unité référentielle des traits de substance différents. Nous en avons tenu compte de manière simplifiée, traitant de manière différente les seconds :

– indice à deux chiffres dans le cas de synonymie proche
xxVA3 /F11 interféromètre de Fabry-Pérot
xxVA3/F12 interféromètre de Pérot et Fabry

– variation de l'indice en premier rang pour une synonymie plus éloignée
xxFD9 /F1 facteur de visibilité /E1 visibility factor
xxFD9/F2 efficacité lumineuse /E2 luminous efficiency

Par ailleurs, il convient de noter que des termes ambigus ont des indices différents, apparaissant à des endroits différents de l'arbre du domaine.

4. En fait, si le domaine de spécialité traité par un dictionnaire est étroit, nombre de ces termes ne peuvent figurer comme vedette

2.2. Indication des « aires d'emploi » des termes

La question de la variation en ce qui concerne les termes arabes est particulièrement importante. L'un des problèmes épineux de la terminologie scientifique arabe est dû à ce fait de dispersion, dont les causes sont multiples mais dont deux sont à signaler : la multiplicité des foyers de création terminologique, joint au fait qu'aucun d'eux ne domine véritablement les autres, et l'existence de deux langues fournisseuses de terminologie, le français et l'anglais, dont l'on peut dire que chacune est pratiquée à l'exclusion de l'autre dans telle ou telle aire du Monde arabe.

De fait, on peut dire qu'il existe quelques grandes tendances : la terminologie majoritairement utilisée en Syrie – c'est le seul pays où l'enseignement des matières scientifiques se fait en arabe, y compris dans le supérieur – dont bon nombre d'enseignants ont travaillé dans d'autres pays arabes, la terminologie utilisée en Egypte – dont on connaît le poids dans le Monde arabe – et dont de nombreux enseignants ont eux aussi travaillé dans les autres pays arabes. De plus il existe aussi des organismes inter-arabes, comme l'Union des Physiciens, et le Bureau de Coordination pour l'arabisation, qui a organisé des congrès d'unification terminologique, qui produisent des listes de plusieurs centaines de termes « unifiés » en cette occasion, et publient des lexiques spécialisés – de valeur lexicographique et terminographique souvent discutable – c'est le cas pour le domaine de la physique. Une étude serait à mener sur l'influence réelle de ces efforts normatifs[5].

Un dictionnaire qui ne se veut pas normatif se doit de fournir à son utilisateur les variantes qui existent, pour chacune des trois langues, et de lui indiquer aussi les aires d'emploi de chacune d'entre elles. Cela est nécessaire quand on traduit vers l'arabe, car il va de soi que la terminologie adoptée doit être cohérente de ce point de vue.

Nous avons codé pour de nombreux termes arabes leur aire d'emploi (C pour les termes communs à tous les pays, qui ne posent de ce point de vue aucun problème, U pour les termes « unifiés », avec éventuellement des indices en cas de variantes officiellement admises, R pour des termes répandus, pouvant coexister avec des variantes, S pour les termes plus spécifiquement « syriens », E pour les termes « égyptiens »...). On peut de toutes façons indiquer aussi, pour chaque terme arabe, les références des textes où ils ont été trouvés (nous avons codé ces ouvrages en fonction des critères suivants : pays d'origine, type d'ouvrage, niveau, domaine et sous-domaine, et enfin, un numéro d'ordre ; par exemple : SUPO2 Syrie Universitaire Physique Optique 2. Il n'y figure pas d'indication d'ordre diachronique).

3. Traitement des phraséologismes

3.1. Caractérisation et modélisation des phraséologismes

On entend ici par *phraséologisme*, tout syntagme non terminologisé, comprenant au

5 La question de la *norme* se pose de manière différente selon les domaines. Pour ce qui est de la terminologie française de la physique, c'est essentiellement au sein du monde des spécialistes – enseignement supérieur, recherche, revues scientifiques – que la norme s'établit. Notons en particulier le rôle joué par les revues scientifiques de haut niveau, qui donnent le ton, non seulement en matière terminologique, mais encore qui imposent un style de rédaction donné. Rien de tel en arabe, même s'il existe des revues scientifiques arabes, mais qui sont des revues de vulgarisation, fussent-elles de bon niveau. Les chercheurs publient davantage en anglais ou en français.

moins un terme, d'étendue supérieure au terme et inférieure à la phrase, et par ailleurs de « fréquence non négligeable » dans les textes de spécialité.

Il s'agit, on le voit, d'une définition empirique, exprimant la notion intuitive de phraséologisme comme entourage préférentiel d'un ou de plusieurs termes. De fait, tout syntagme du type évoqué dans cette définition a vocation à constituer un phraséologisme. Quant au critère de fréquence, il relève d'un tout autre plan.

Nous travaillons actuellement sur la question des phraséologismes tels qu'ils se présentent dans les textes de physique, et nous sommes loin d'en avoir établi une liste qui ne pourrait être véritablement significative que si elle s'appuyait sur un dépouillement quantitativement important de textes de la spécialité étudiée. Si l'établissement de la terminologie d'un domaine relève d'une démarche fondamentalement onomasiologique, l'établissement de la phraséologie correspondante relève d'une démarche sémasiologique, pour laquelle le texte est le matériau de départ et les termes sont les signaux d'existence.

Nous donnons ci-dessous quelques phraséologismes – en français – relatifs au terme *frange (d'interférence) /hudbat/* (pl. */hudb/*) :

. nous obtenons des franges très étroites
. observe des franges d'interférence à la *surface* de la lame
. franges d'interférence produites par une *double source*
. nous pouvons mesurer l'*interfrange*
. on appelle *interfrange* la *distance* qui sépare deux franges voisines

L'on peut considérer les phraséologismes sous deux plans :

– le plan de la forme, le plan syntaxique : le phraséologisme est composé d'une *relation* mettant en rapport des *arguments*, que l'on peut représenter sous la forme d'un *vecteur phraséologique* :

. nous obtenons des franges très étroites
<obtenir 1 2 | nous; franges très étroites>
<obtenir 1 2 | [humain]; franges >

. franges étroites
<étroit 1 | franges>

Les phraséologismes sont susceptibles de s'enchâsser :

<obtenir 1 2 | nous; franges très étroites>
<obtenir 1 2 | nous; <étroit 1 | franges> >

La représentation phraséologique ci-dessus est susceptible de représenter des phraséologismes comme :

. on obtient des franges étroites
. l'obtention de franges étroites
. les franges étroites obtenues

. (pour) l'obtention de franges étroites
. des franges étroites s'obtiennent
. (on peut) obtenir des franges étroites

. des franges étroites
. des franges qui sont étroites
. ces franges sont étroites
. l'étroitesse des franges

Ces phraséologismes ont des structures syntaxiques de surface qui sont différentes, mais dont la représentation en termes de relation et arguments est la même[6].

À l'inverse la possibilité d'avoir diverses réalisations de la représentation phraséologique – en contexte, faisant donc intervenir d'autres lexies et d'autres contraintes syntaxiques – pourront être différentes selon la langue (ainsi en arabe, on passe facilement de la forme verbale conjuguée à l'« infinitif » correspondant, pouvant être utilisé de manière beaucoup plus souple qu'en français).

.* notre obtention de franges étroites
./HuSu:l-u-na: ' ala: hudb-in Dajjiqat-in/

– le plan du fond, d'ordre sémantique : le phraséologisme exprime un type de relation référentielle qu'une unité référentielle d'un certain type entretient avec d'autres, de nature généralement différente, et, si l'on peut dire, avec le reste de l'univers (physiciens compris !).

Argument 1	**Relation**	Argument 2	
nous	**obtenons**	des *franges*	très étroites
[humain] expérimentateur	réussir à atteindre (un résultat (*Petit Robert*)	phénomène physique observable, s'étendant dans l'espace de manière discontinue et dénombrable	qualification relative à une entité à deux dimensions

6 Le terme *frange* est ici employé au pluriel. L'ensemble des phraséologismes relatifs à un terme n'est pas forcément le même selon que le terme est employé au singulier ou au pluriel (quand le pluriel est possible), quand le terme est déterminé par l'article ou pas, etc. Un exemple, *source cohérente / source_s cohérente_s* (ces dénominations sont en fait métonymiques : ce sont les *rayonnements* – vus comme grandeurs sinusoïdales –, émis par ces sources qui sont cohérentes, c'est-à-dire qui ont des relations de phase constantes au cours du temps (Mathieu *et al.*, 1985 : 81)). Une source cohérente est telle que tous les rayons émis par cette source sont cohérents. Deux ou plusieurs sources cohérentes (cohérentes entre elles) sont telles que les rayons émis par elles sont cohérents. En définitive, si la cohérence concerne les rayons lumineux, dans l'emploi le plus fréquent, deux *sources cohérentes* sont cohérentes entre elles mais ne sont pas par elles-mêmes chacune une *source cohérente*.

		Argument 1	**Relation**
		des *franges*	très **étroites**
		phénomène physique observable, s'étendant dans l'espace de manière discontinue et dénombrable	qualification relative à une entité à deux dimensions

3.2. Traitement lexicographique des phraséologismes

Sur le plan lexicographique, la question se pose de savoir quels sont les phraséologismes qui doivent être insérés dans le dictionnaire spécialisé.

Par exemple, si nous considérons les deux termes suivants, dont l'un est générique et l'autre spécifique, l'ensemble des phraséologismes repérés pour chacun d'eux, *frange* et *frange centrale* – la *frange centrale* étant une *frange* particulière dans un *système de franges* – est-il le même ? Certainement pas : si les (ou des) *franges* et **la** *frange centrale* se déplacent, s'observent, sont situées sur une surface, par contre, on dénombre, on compte des *franges*, mais pas la *frange centrale*...

Trois critères semblent pouvoir être pris en considération pour le classement dans l'article du dictionnaire relatif à un terme, les phraséologismes qui lui sont associés :

– d'ordre formel : nombre d'arguments, type de constructions syntaxiques.

– d'ordre sémantique : les types de relation pouvant exister entre des types d'unités référentielles de nature donnée. On peut envisager une grille du type :

```
production/fabrication – lieu/déplacement – temps/évolution
- ...
```

– d'ordre statistique : fréquence des occurrences.

Si le critère statistique est pertinent pour le choix des phraséologismes retenus, l'ordre sémantique paraît s'imposer pour l'ordre de présentation des phraséologismes, du fait que le dictionnaire est trilingue.

Conclusion

Un dictionnaire comme celui esquissé ici concerne un domaine très étroit. Si l'utilisateur est préoccupé avant tout de choix de termes, de questions terminologiques, il doit trouver, dans un ouvrage de ce genre, les termes dont il a besoin, avec en particulier leurs variantes, avec leur aire d'emploi.

Mais l'aide au rédacteur ou au traducteur que doit apporter un dictionnaire termino-phraséologique plurilingue de ce genre est en relation avec les types de discours de spécialité qui viennent interférer dans les textes de la spécialité concernée. Or de ce point de vue l'intérêt d'un lexique qui ne traite que d'un domaine très étroit est très limité, du fait justement qu'il ne traite pas de la variété de ces discours de spécialité qui concourent, chacun avec sa terminologie et sa phraséologie au texte de la spécialité.

Ce travail ne peut être conçu que comme la partie d'un tout bien plus vaste, qui serait un dictionnaire termino-phraséologique de la physique.

Si la détermination des termes d'une langue relatifs à un domaine s'appuie tout naturellement sur la connaissance des unités référentielles qui le constituent, il n'en est pas de même pour les phraséologismes qui, eux, ne peuvent être mis au jour qu'à partir des textes de spécialité et donc de corpus significatifs quantitativement. Cette exigence implique le recours à une procédure de traitement automatique de ces textes, pour repérer les occurrences et les contextes de chaque terme, ce qui est nécessaire pour donner à l'expression « fréquence non négligeable » d'occurrence de phraséologismes évoquée plus haut une signification pertinente[7].

C'est à cette condition que pourra être achevé un véritable dictionnaire de spécialité relatif au domaine de l'optique.

Annexe 1

phraséologisme	Source	Argument 3	Argument 2	Argument 1	Relation
naHsulu`ala: hudb-in Dajjiqat-in zudd-an	EUPO1:365		hudb Dajjiqat zidd-an	/naHnu/	HaSala 1 `ala · 2
nous obtenons des franges très étroites			**franges très étroites**	**/nous/**	**obtenir 1 2**

Exemple de relevé de phraséologisme avec formalisation sous forme de vecteur phraséologique

7 Cela implique que les termes du domaine soient au préalable stockés dans une base de données informatique. Nous disposons de procédures de traitement automatique – essentiellement au niveau morphologique – de textes arabes non voyellés (voir, par ex. Dichy et *al*., 1989). Il reste à établir les procédures permettant la reconnaissance d'unités terminologiques complexes.

xxF	*Division Franges*								
xxF0	MKF 200 franges d'interférences «On nomme *système de franges d'interférences* une famille de courbes tracées sur une surface en tous points de laquelle se produisent des phénomènes d'interférence. Sur chaque courbe ou *frange*, la *différence de marche* entre les vibrations interférentes est constante et on peut caractériser chaque frange par un nombre, l'*ordre d'interférence* p = δ/λ, λ est la longueur d'onde des ondes qui interfèrent. (...) «Le *facteur de contraste* (facteur de visibilité en Optique) a pour expression V = (IM-Im)/(IM+Im).» V.Pitt 151 fringes «Circular or rectilinear patterns of alternate light and dark or of colour, produced by *interference* or *diffraction* of light.»	xxF0/F	frange	xxF0/E	fringe	xxF0/A11	hudb /pl ʔahdāb	S U	IUPO1 27 MLS1 MLS2 MSP3 195 SSP2 158 SUPO2 lex SUPO6 252 UL1opt
xxF0		xxF0/F	frange	xxF0/E	fringe	xxF0/A12	hudbat /pl hub	E	EUP1 630 EUPO1 364 LUP2 263 SUPO5 lex
xxF0		xxF0/F	frange	xxF0/E	fringe	xxF0/A2	$arīT/pl ʔa$rīTat	¤	RUP2 119
xxF01		xxF01/F	frange_s d'interférence	xxF01/E	interference fringe_s	xxF01/A11	ʔahdāb (ʔa)t-tadāxul (pl)	S U	IUPO1 24 MSP3 197 SSP2 157 SaUPO1a 288
xxF01		xxF01/F	frange_s d'interférence	xxF01/E	interference fringe_s	xxF01/A12	hudb (ʔa)t-tadāxul	E	ELPN1 EUP1 629 LUP2 256 UL1opt ULP1b
xxF01		xxF01/F	frange_s d'interférence	xxF01/E	interference fringe_s	xxF01/A13	ʔa$rīTat (ʔa)t-tadāxul (pl)	¤	RUP2 119

Génération automatique de néologismes arabes à partir des règles de formation de mots

Hussein HABAILI et Mohamed BEN AHMED

Laboratoire de Recherche en Informatique Arabisée et Documentique Intégrée, (RIADI), Tunis, Tunisie

• *Abstract* •

This paper enters within the theoretical framework of lexicalist hypothesis which preconizes an autonomous morphological component based on the lexicon, responsible of flexion, derivation and composition.

This research allowed to specify with details the nature of lexical rules permitting the derivation of words departing from basic words. Such relationship is achieved according to a set of word formation rules. These rules are characterized by their logical and mathemathical formalism. They can be computerized and naturally applied to Arabic morphological processes such as flexion, derivation and composition.

Indeed, their application in Arabic morphology is of a great importance for lexical neology, mainly while creating new words in technical and scientific domains.

A generator is designed to generate automatically new words for different grammatical categories (substantives, adjectives, action names, etc.) departing from a set of word formation rules set up for the standard Arabic.

Through a freak of word derivation and composition departing from basic words (by prefixation, suffixation and composition), the generator thus allows to update and enrich the Arabic dictionary with neologisms.

The generator is linked to a dictionary including basic words, roots, prefixes and suffixes with grammatical information associated with words and a set of word formation rules.

1. Introduction

L'hypothèse lexicaliste de Chomsky (1970) n'est pas l'unique proposition dans le cadre de la grammaire générative qui préconise la séparation des **règles de formation**

de mots (**RFM**) de la syntaxe. En effet, Kiefer (1973)[1] a proposé une composante morphologique autonome responsable de **la flexion**, de **la dérivation** et de **la composition** comme le montre le schéma suivant :

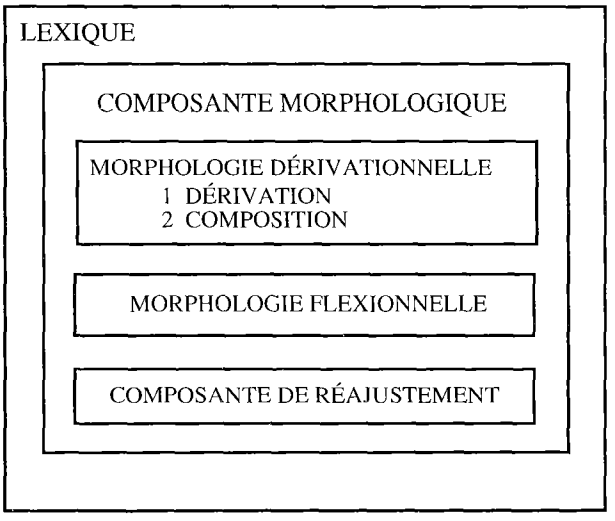


FIGURE 1 : La composante morphologique

Selon Kiefer la composante morphologique s'applique exactement après les transformations syntaxiques et avant la phonologie. Cet ordre est motivé, en un sens qu'il permet d'introduire les informations transformationnelles aux règles morphologiques.

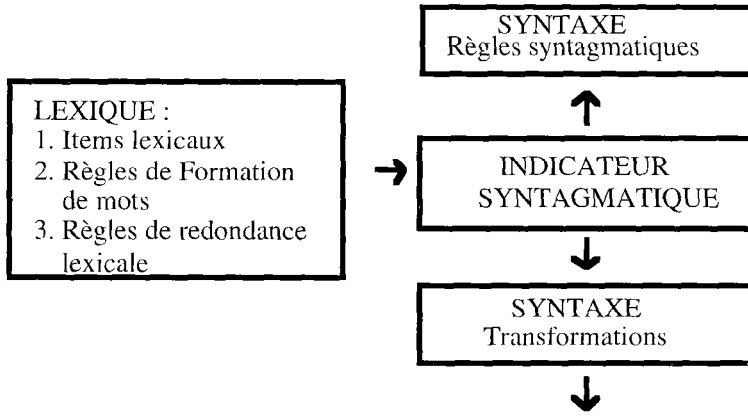
Ce point de vue est développé dans des études subséquentes dans le même cadre théorique de l'hypothèse lexicaliste originelle par certains linguistes (Siegel, 1974 ; Aronoff, 1976 ; Allen, 1978 ; Strauss, 1979, 1982 ; Habaili, 1990).

En outre, ces études ont permis de spécifier, d'une façon détaillée, la nature des règles lexicales qui permettent de dériver des mots à partir des mots de base. Cette relation s'accomplit en partie par un ensemble de règles de formation de mots : (WFR, cf. spécialement Aronoff, 1976), qui permettent d'attacher un affixe à la base.

Malgré les différences de détails de ces hypothèses toutes les études citées s'accordent sur un principe fondamental du lexicalisme selon lequel **la morphologie dérivationnelle** est construite au niveau du lexique.

L'organisation de la grammaire qui rendra compte de ce principe sera la suivante :

1 F. Kiefer, « Morphology in Generative Grammar », Gross et al. (Eds), *The Formal Analysis of Natural Languages*, The Hague, Mouton, 1973



Pour répondre à la question qui consiste à savoir si les règles de formation de mots s'appliquent dans le lexique, nous passons en revue les différentes positions de certains linguistes.

En 1965, N. Chomsky définit le lexique ainsi :

(A) lexicon... is simply an unordered list of all lexical formatives. More precisely, the lexicon is a set of lexical entries, each lexical entry being a pair (D,C), where D is a phonological distinctive feature matrix "spelling" a certain lexical formative and C is a collection of specified syntactic features (a complex symbol).

Dans son hypothèse prélexicaliste Chomsky considère les formants individuels comme provenant du lexique et insérés dans la syntaxe... Ces propriétés des formants lexicaux qui ne sont pas prévisibles sont considérées comme essentiellement idiosyncrasiques et doivent être associées aux formants dans le lexique.

Halle (1973) a observé que les mots flexionnels peuvent avoir une certaine idiosyncrasie. En effet, certaines formes flexionnelles ont une signification idiosyncrasique, d'autres sont des exceptions à des règles phonologiques régulières et d'autres, sont accidentellement absents des paradigmes. Pour Halle, de telles formes doivent entrer dans le lexique.

Allen (1978) a examiné des mots composés. Elle les considère comme ayant une signification idiosyncrasique. Elle arrive à la conclusion que ces mots composés qui ne sont pas sémantiquement prévisibles sont lexicalisés et leur sens n'est pas dérivable, mais sont listés simplement dans le lexique.

Ce faisant, la « **condition d'idiosyncrasie** » pour entrer dans le lexique dépasse les formants lexicaux pour inclure d'autres éléments qui peuvent figurer dans le lexique :

LEXIQUE

1. RACINES ET AFFIXES
2. MOTS DÉRIVÉS
3. MOTS FLEXIONNELS
4. MOTS COMPOSÉS
5. PHRASES SYNTAXIQUES

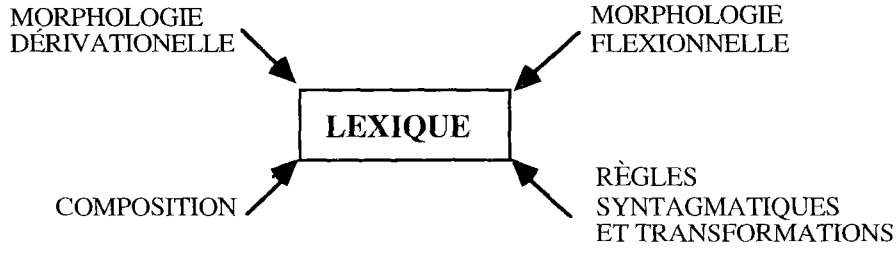
Ceci nous ramène à nous poser de nouveau la question qui consiste à déterminer la place des différents types de règles génératives dans la grammaire, en partant de la formulation originelle de l'hypothèse lexicaliste. En effet, nous pouvons nous demander pour quelle raison nous pouvons affirmer que les règles qui engendrent des entrées lexicales (mots dérivés...) sont considérées comme partie intégrante du lexique ; tandis que les règles relevant de la morphologie flexionnelle, de la composition et de la syntaxe ne le sont pas ?

Si les règles de la **morphologie dérivationnelle** sont situées dans le **lexique** pour la simple raison qu'elles sont capables d'engendrer des mots dérivés, la même raison est aussi valable pour les règles qui peuvent engendrer des entrées lexicales et pour cela ces dernières doivent figurer aussi dans le lexique.

Halle (1973) a raisonné en ce sens : il est arrivé à la conclusion selon laquelle les règles de la morphologie flexionnelle sont situées à la base du lexique, étant donné qu'il existe des mots flexionnels situés dans le lexique en raison de leur idiosyncrasie. Cependant, faut-il placer aussi les règles syntaxiques dans le lexique puisqu'il existe aussi des phrases syntaxiques idiosyncrasiques (les idiomes) ?

Si nous considérons toutes les règles génératives (structurales) de la grammaire comme faisant partie du lexique nous détruisons pour cela la séparation des règles de formation de mots de la syntaxe, sans oublier que l'essence même de l'hypothèse lexicaliste déjà atteinte consistait à déplacer les RFM de la composante syntaxique au lexique.

Pour résoudre ce problème, il faut considérer les règles de la morphologie dérivationnelle aussi régulières (sémantiquement et phonologiquement) que les autres règles génératives. Ce faisant, nous pouvons donc penser que chaque classe de règles peut engendrer des mots ou expressions qui peuvent devenir des items lexicaux :



En outre, pour marquer d'une façon précise la séparation de la morphologie de la syntaxe, il faut distinguer entre deux types de sous-grammaire connus sous les noms de *WORD-GRAMMAR* et de *SENTENCE-GRAMMAR*.

En effet, la première comprend la morphologie dérivationnelle, la morphologie flexionnelle et la composition et ne peut engendrer que des catégories lexicales (N, V, Adj...) ; tandis que la seconde qui comprend les structures de phrases et les transformations constitue un champ d'application des règles phonologiques.

2. Principes théoriques pour l'élaboration des règles de formation de mots

2.1. Les règles de formation de mots

La phonologie lexicaliste a pour objet l'établissement d'un principe de base pour l'explication morphologique en phonologie. Elle doit aussi déterminer le rôle des traits morphologiques et des structures de mots dans la forme et l'application des règles phonologiques.

La phonologie lexicaliste s'inscrit dans le cadre de l'hypothèse lexicaliste (Chomsky, 1970). Cependant, elle est à la recherche d'une morphologie autonome dont les structures de mots sont justifiées à la base à partir de la distinction des morphèmes.

La phonologie lexicaliste a donc un double objectif. Le premier consiste à élaborer et à justifier une théorie autonome et interne de formation des mots. Le second cherche à analyser et à expliquer comment les structures de mots engendrées par la théorie de formation de mots influencent la forme et l'application des règles phonologiques.

Considérons la représentation de la relation qui existe entre une suite terminale **t** et une suite non-terminale **A** :

(1)
A
|
t

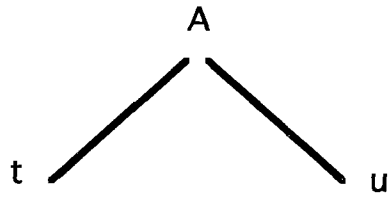
À partir de la représentation (1), nous pouvons formellement exprimer le fait que /kataba/ est un verbe, et que /baabun/ est un nom :

(2)

V	**N**
|	|
/kataba/	/baabun/

En supposant qu'il existe une possibilité logique qui consiste à ajouter une suite terminale **u** à droite de **t** dans (1), ceci nous donne les trois possibilités suivantes :

(3)

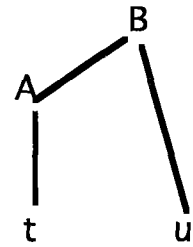


Dans cette représentation, la suite terminale **u** s'adjoint à **t** sans suite non terminale additionnelle. C'est ainsi que la structure de la représentation (3) est équivalente à **[t + u]A**, et la dérivation est :

[t]A → [t + u]A.

La deuxième possibilité est celle qui adjoint **u** à **[t]A** avec une construction simultanée d'un nouveau nœud **B** de la façon suivante :

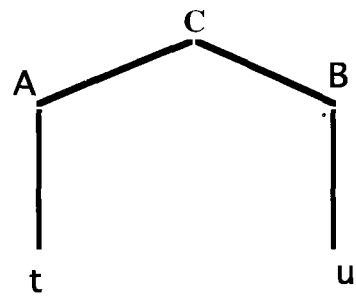
(4)



La dérivation de (4) est donc : **[t]A → [[t]A + u]B**. Si l'on compare les représentations (3) et (4) et leur dérivation, l'on constate que la dérivation de (4) diffère de celle de (3), étant donné que deux suites non terminales résultent d'un seul input initial.

La troisième possibilité est celle qui adjoint **u** à une suite non terminale qui possède à elle seule :

(5)



La dérivation de (5) devient alors :

[t]A → [[t]A + [u] B]C.

Grâce à cette distinction binaire entre le vocabulaire terminal et non terminal, les représentations (3), (4) et (5) constituent les seules possibilités où une suite terminale

u devient à droite d'une suite terminale **t** qui est un **A**. C'est ainsi que ces trois représentations constituent la division fondamentale de la formation des mots : **la flexion**, **la dérivation**, et **la composition**.

Un second principe de la théorie de la formation des mots est celui qui considère la flexion[2] comme n'altérant jamais les catégories lexicales à la base. Un nom fléchi demeure un nom et un verbe fléchi demeure un verbe, etc.

Par contre, la morphologie dérivationnelle peut en principe altérer les catégories lexicales à la base.

D'autre part, la représentation linéaire équivalente à (4) et (5) est (6) :

(6) a : [t + u] : Flexion
(6) b : [[t] + u] : Dérivation
(6) c : [[t] + [u] : Composition

Cette présentation linéaire montre clairement que les distinctions peuvent être faites par une inclusion sélective des crochets et par un seul opérateur de concaténation « + » dans la description structurale (**DS**) des règles phonologiques. C'est ainsi, que pour réécrire une règle phonologique qui s'appliquerait uniquement à une suite à l'intérieur du morphème + **u**, nous omettons le symbole de frontière de morphème « + » et les crochets « [ ] », de la description structurale de la règle :

(7) DS: t u
1 2

Cependant, une règle phonologique qui a comme déterminant un segment flexionnel doit être réécrite avec « + » dans la description structurale :

(8) DS: t + u
1 2 3

Par contre, une règle dont le déterminant est un segment dérivationnel doit être réécrite avec un seul crochet dans la description structurale :

(9a) DS: t] u
1 2 3

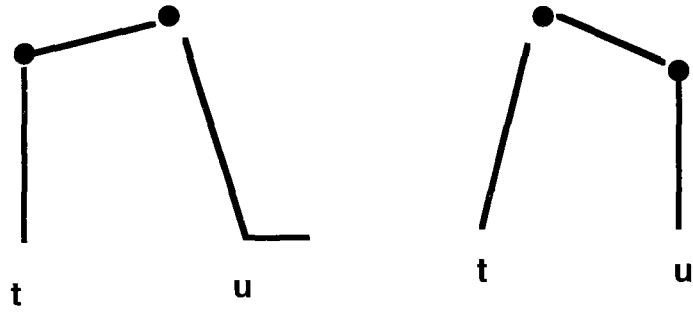
(9b) DS: t [u
1 2 3

Il faut préciser que (9a) s'applique à des formes suffixales, tandis que (9b) s'applique à des formes préfixales :

---

2 Mark Aronoff définit la dérivation et la flexion en ces termes « there are traditionally two types of morphological phenomena, derivational and inflectional. The distinction is delicate, and sometimes elusive, but nonetheless important. Inflection is generally viewed as encompassing the "purely grammatical" markers, those for tense, aspect, person, number, gender, case, etc. Within a lexicalist, theory of syntax (cf. Chomsky, 1970), inflectional morphemes would be dominated by the node X, and perhaps higher nodes (cf. Siegel, 1974), while derivational morphemes would be dominated by the node X. Derivational morphology is thus restricted to the domain of lexical category » (Mark Arnoff, *Word Formation in Generative Grammar*, Cambridge, The MIT Press, 1976, p. 2)

(10a) (10b)



Étant donné les définitions formelles de la flexion, la dérivation et la composition, il n'existe pas d'autres façons pour interpréter la description structurale de la représentation (9).

Enfin, quand une règle phonologique s'applique à des mots composés, sa description structurale spécifiera les descriptions gauche et droite des crochets comme dans (11) :

(11) DS: t] [u
1 2 3 4

Ce système proposé présente des critères définitoires formels qui permettent de se référer à des classes naturelles qui ont les descriptions structurales suivantes :

(12) a : DS: t (]) + u
b : DS: t + ([) u
c : DS: t <]> + <[> u
d : DS: t] ([) u
e : DS: t (]) [*u*
f : DS: t (]) + ([) u
g : DS: t (]) (+) ([) u

Ces descriptions structurales peuvent être développées pour s'appliquer respectivement aux classes naturelles suivantes :

(12')a. Formes suffixées et fléchies :
t] + u
t + u
b. Formes préfixées et fléchies :
t + [u
t + u
c. Formes composées et fléchies :
t] + [u
t + u
d. Formes composées et suffixées :
t] [u
t] u

e. Formes composées et préfixées :
t] [u
t [u
f. Formes composées, dérivées et fléchies :
t] + [u
t + [u
t] + u
t + u
g. Formes composées, dérivées, fléchies et formes internes du morphème :
t] + [u
t + [u
t] + u
t + u
t u

Ces représentations développées montrent que dans (12'd) et (12'e) le symbole « + » des frontières de morphème est omis de leur description étant donné que sa présence à l'input est entièrement prédictible de l'occurrence obligatoire d'au moins un crochet.

Cependant, dans les autres descriptions structurales où les crochets sont tous optionnels, la référence à des segments flexionnels exige l'inclusion du symboles « + ». C'est ainsi, que pour faciliter l'écriture des règles phonologiques conditionnées par les mots structurés, nous adoptons la convention qui considère un terme non segmental (+, [, ou]) dont la présence est entièrement prédictible, comme redondant et par conséquent peut être omis de la description structurale de la règle.

3. Définition formelle des règles de formation de mots

Les abréviations ci-dessous ont les significations suivantes : **t** et **u** désignent des suites terminales, et **A**, **B**, **C**, représentent les catégories grammaticales :

Suffixation flexionnelle	:	[t]A → [t + u]A
Préfixation flexionnelle	:	[t]A → [u + t]A
Suffixation dérivationnelle	:	[t]A → [[t]A + u]B
Préfixation dérivationnelle	:	[t]A → [u + [t]A]B
Composition (par adjonction à droite)	:	[t]A → [[t]A + [u]B]C
Composition (par adjonction à gauche)	:	[t]A → [[u]B + [t]A]C
Dérivation nulle	:	[t]A → [[t]A]B

4. Les règles de formation de mots en arabe

4.1. La dérivation

Pour répondre aux exigences de la vie moderne, l'arabe standard forme un grand nombre d'adjectifs relatifs (**Adj.Rel.**) et de noms abstraits (**NAb.**) par suffixation de la semi-consonne /+ **j** / ou /+**ijj** /. Cette formation obéit à des règles de formation de

mots. En effet, l'arabe moderne forme la plupart de ces adjectifs à partir d'un substantif au singulier (après suppression éventuelle du suffixe final /+**at** /) ou au pluriel sans que sa structure soit modifiée :

a) Adjectifs relatifs :

L'arabe standard moderne fait un usage très fréquent de ces noms de « relation », la « ***nisba*** » des grammairiens arabes (ou relatifs) en / **+ii** /. Ils sont généralement construits à l'aide des « règles de formation de mots » (RFM) de suffixation dérivationnelle du type :

[t]A → [[t]A + u]B

ou **t** : est une suite terminale, **u** : est aussi une suite terminale et **A**, **B** sont des catégories :

RFM_1 :	[t]N →	[[t	]N	+ ii]Adj.	Rel.	
		[[ʒadal	]N	+ ii]Adj.	Rel.	« dialectique »
		[[ɻamal	]N	+ ii]Adj.	Rel.	« pratique »
		[[waðiif	]N	+ ii]Adj.	Rel.	« fonctionnel »
		[[raʔiis	]N	+ ii]Adj.	Rel.	« principal »

Quelques adjectifs relatifs sont construits sur un nom d'action (**N.Act.**) et le suffixe /**+ii**/ :

RFM_2 :	[t]N Act. →	[[t	]	N Act.	+ ii]Adj.	Rel.	
		[[taqadum	]	N Act.	+ ii]Adj.	Rel.	« progressiste »
		[[taɻaawun	]	N Act.	+ ii]Adj.	Rel.	« coopératif »
		[[ʔinhizaam]		N Act.	+ ii]Adj.	Rel.	« défaitiste »

Certains adjectifs relatifs sont construits sur des mots composés (**MC**) et le suffixe / **+ii** / :

RFM_3 :	[t]MC →	[[t	]MC	+ ii]Adj.	Rel.	
		[[laaʃuɻuur	]MC	+ ii]Adj.	Rel.	« inconscient »
		[[raʔsmaal	]MC	+ ii]Adj.	Rel.	« capitaliste »

D'autres adjectifs relatifs sont construits sur des emprunts (**EM**) et le suffixe / + **ii** / :

RFM_4 :	[t]EM →	[[t	]EM	+ ii]Adj.	Rel.	
		[[kahrabaaʔ	]EM	+ ii]Adj.	Rel.	« électrique »
		[[sinimaaʔ	]EM	+ ii]Adj.	Rel.	« cinématographique »

Les noms propres (**NP**) peuvent aussi former des adjectifs relatifs :

RFM_5 :	[t]NP →	[[t	]NP	+ ii]Adj.	Rel.	
		[[naasir	]NP	+ ii]Adj.	Rel.	« nassérien »

Les adjectifs relatifs construits sur des néologismes (**NE**) du type :

RFM_6 : [t]NE →	[[	t	]NE	+ ii]Adj. Rel.		
	[[ʔalmaan		]NE	+ ii]Adj. Rel.	« laïque »	
	[[qablaan		]NE	+ ii]Adj. Rel.	« a priori »	
	[[baʔdaan		]NE	+ ii]Adj. Rel.	« a posteriori »	

Enfin, les adjectifs relatifs construits sur des particules (**P**) :

RFM_7 : [t]P →	[[	t	]P	+ ii]Adj. Rel.	
	[[ɣajr		]P	+ ii]Adj. Rel.	« altruiste »

b) Les noms abstraits :

L'arabe standard moderne forme un grand nombre de noms abstraits (**NA**) qui correspondent généralement à des noms abstraits français en / **+té** / ou / **+ism** / à l'aide du suffixe / **+ijja** / et qui ont une grande variété d'origines.

Certains sont construits sur des participes passifs (**PP**) à l'aide des règles de formation de mots de suffixation dérivationnelle.

RFM_8 : [t]PP →	[[	t	]PP	+ ijja]NA	
	[[masʔuul		]PP	+ ijja]NA	« responsabilité »
	[[maʃruuʕ		]PP	+ ijja]NA	« légitimité »
	[[mawḍuuʕ		]PP	+ ijja]NA	« objectivité »

D'autres noms abstraits sont construits sur des élatifs (**E**) :

RFM_9 : [t]E →	[[	t	]E	+ ijja]NA	
	[[ʔaɣlab		]E	+ ijja]NA	« majorité »
	[[ʔaqall		]E	+ ijja]NA	« minorité »

Certains noms abstraits sont construits sur des particules (**P**) :

RFM_{10} : [t]P →	[[	t	]P	+ ijja]NA	
	[[kamm		]P	+ ijja]NA	« quantité »
	[[kajf		]P	+ ijja]NA	« modalité »

Parfois les noms abstraits sont construits sur des mots composés (**MC**) :

RFM_{11} : [t]MC →	[[	t	]MC	+ ijja]NA	
	[[laadiin		]MC	+ ijja]NA	« irreligiosité »
	[[laaʃajʔ		]MC	+ ijja]NA	« nulleté »
	[[laanihaaʔ		]MC	+ ijja]NA	« infinité »

Cependant, la majorité de ces noms abstraits (**NA**) est formée sur des noms substantifs (**N**) :

RFM_{12} : [t]N →	[[	t	]N	+ ijja]NA	
	[[ʔinsaan		]N	+ ijja]NA	« humanité » (qualité)
	[[baʃar		]N	+ ijja]NA	« humanité » (genre)

[[ʒins	]N	+ ijja	]NA	« nationalité »
[[ʔaxlaaq	]N	+ ijja	]NA	« moralité »
[[ʃaxṣ	]N	+ ijja	]NA	« personnalité »

D'autres noms abstraits sont formés plutôt sur des adjectifs (**Adj.**) :

RFM_{13} : [t]Adj. →	[[t	]Adj.	+ ijja	]NA	
	[[ħurr	]Adj.	+ ijja	]NA	« liberté »
	[[laaʔiik	]Adj.	+ ijja	]NA	« laïcité » (fr.)
	[[ʕalmaa	]Adj.	+ ijja	]NA	« laïcité »

Quant aux noms abstraits arabes qui correspondent aux noms abstraits français en **/+ism/**, ils ont aussi une grande variété d'origines, certains sont construits sur une particule comme (**P**) :

[[ɣajr]P + ijja]NA « altruisme »

D'autres sur un pronom (**PRO**), comme :

[[ʔanaa]PRO +[n]+ ijja]NA « égoïsme »

D'autres sont formés sur des emprunts (**EM**) :

RFM_{14} : [t]EM →	[[t	]EM	+ ijja	]NA	
	[[qajṣar	]EM	+ ijja	]NA	« césarisme »
	[[barlamaan	]EM	+ ijja	]NA	« parlementarisme »
	[[maarks	]EM	+ ijja	]NA	« marxisme »

Parfois aussi sur des mots composés (**MC**) :

RFM_{15} : [t]MC →	[[t	]MC	+ ijja	]NA	
	[[raʔsmaal	]MC	+ ijja	]NA	« capitalisme »

Cependant, la majorité des noms abstraits sont formés sur des noms (ou substantifs) :

RFM_{16} : [t]N →	[[t	]N	+ ijja	]NA	
	[[ʔinsaan	]N	+ ijja	]NA	« humanisme »
	[[taʔθiir	]N	+ ijja	]NA	« impressionnisme »
	[[ħizb	]N	+ ijja	]NA	« sectarisme, partisme »
	[[ðaat	]N	+ ijja	]NA	« subjectivisme »
	[[ʔiʃtiraak	]N	+ ijja	]NA	« socialisme »
	[[qawm	]N	+ ijja	]NA	« nationalisme »
	[[waṭan	]N	+ ijja	]NA	« patriotisme »

En guise de conclusion sur la dérivation, nous pouvons récapituler les règles de formation de mots de suffixation dérivationnelle du type :

[t]A → [[t]A + u]B :

RFM_1 :	[t]N	→	[[	t	]N	+ ii	]Adj.	Rel.
RFM_2 :	[t]N Act.	→	[[	t	]N Act.	+ ii	]Adj.	Rel.
RFM_3 :	[t]MC	→	[[	t	]MC	+ ii	]Adj.	Rel.
RFM_4 :	[t]EM	→	[[	t	]EM	+ ii	]Adj.	Rel.
RFM_5 :	[t]NP	→	[[	t	]NP	+ ii	]Adj.	Rel.
RFM_6 :	[t]NE	→	[[	t	]NE	+ ii	]Adj.	Rel.
RFM_7 :	[t]P	→	[[	t	]P	+ ii	]Adj.	Rel.
RFM_8 :	[t]PP	→	[[	t	]PP	+ ijja	]NA	
RFM_9 :	[t]E	→	[[	t	]E	+ ijja	]NA	
RFM_{10} :	[t]P	→	[[	t	]P	+ ijja	]NA	
RFM_{11} :	[t]MC	→	[[	t	]MC	+ ijja	]NA	
RFM_{12} :	[t]N	→	[[	t	]N	+ ijja	]NA	
RFM_{13} :	[t]Adj.	→	[[	t	]Adj.	+ ijja	]NA	
RFM_{14} :	[t]EM	→	[[	t	]EM	+ ijja	]NA	
RFM_{15} :	[t]MC	→	[[	t	]MC	+ ijja	]NA	
RFM_{16} :	[t]N	→	[[	t	]N	+ ijja	]NA	

4.2. La composition

La composition consiste à former un seul mot à partir de deux ou plusieurs mots réunis. Le véritable composé construit un mot nouveau (à sens nouveau) et l'on perd la conscience linguistique des composants.

Comme l'arabe standard moderne est défavorisé sur ce point, il lui est arrivé de former des mots composés à l'aide de deux mots brefs selon la règle de formation de mots composés (RFMC) :

RFMC1 : [t]A	→	[[	t	]A	+	[	u	]B	]MC	
		[[barr		]A	+	[maaʔii		]Adj. Rel.	]MC	« amphibie »
		[[ʒaw		]A	+	[maaʔii		]Adj. Rel.	]MC	« hydravion »

4.2.1. La préfixation

L'arabe standard moderne forme un certain nombre de mots composés par préfixation, selon les règles de formation de mots composés.

a) Les mots composés à préfixe / laa + / :

Le préfixe / laa +/ exprimant la négation permet la formation d'environ une cinquantaine de mots composés, surtout nominaux et rarement verbaux. Certains sont des termes abstraits ou philosophiques formés selon la règle de composition à gauche :

RFMC2 : [t]A	→	[[u	]B	+	[	t	]A	]MC	
		[[laa	]Préf.	+	[mubaalaat		]N	]MC	« négligence »
		[[laa	]Préf.	+	[diinii		]Adj.R.	]MC	« irreligieux »
		[[laa	]Préf.	+	[ʔiraada		]N	]MC	« automatisme »
		[[laa	]Préf.	+	[ʃuʕuur		]N	]MC	« l'inconscient »

D'autres mots composés sont des termes politiques administratifs ou sociologiques et sont formés selon la même règle :

RFMC2' : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[laa	]Préf.	+	[niðaam	]N	]MC	« anarchie »
	[[laa	]Préf.	+	[muħaaraba	]N	]MC	« non belligérance »
	[[laa	]Préf.	+	[tamarkuz	]N	]MC	« déconcentration »
	[[laa	]Préf.	+	[diimuqraaṭi	]N	]MC	« antidémocratie »

Certains mots composés sont des termes scientifiques ou techniques :

RFMC2" : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[laa	]Préf.	+	[ʒafn	]N	]MC	« ablépharique »
	[[laa	]Préf.	+	[ʒinsii	]Adj.R.	]MC	« asexué »
	[[laa	]Préf.	+	[silkii	]Adj.R.	]MC	« sans fil »
	[[laa	]Préf.	+	[ʔinʃiṭaar	]N	]MC	« non fission »

b) Les mots composés à préfixe / ʃibhu + / :

Le préfixe /ʃibhu+/ entre en composition avec le sens de « semblable à », quasi, para, proto, etc. Ces mots composés sont formés à l'aide de la même règle de composition à gauche :

RFMC3 : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[ʃibhu	]Préf.	+	[ʒaziira	]N	]MC	« presqu'île »
	[[ʃibhu	]Préf.	+	[ṭibii	]Adj.R.	]MC	« paramédical »

Ce préfixe se présente tantôt soudé au mot :

RFMC4 : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[ʃib	]Préf.	+	[billawr	]N	]MC	« cristalloïdes »

c) Les mots composés à préfixe /ɣajru +/ :

Le préfixe / ɣajru+ / est assez productif en arabe standard moderne. Il correspond fréquemment aux préfixes français : **in+**, **a+**, **non**, **de+**, et parfois même, **extra**. Il forme des mots composés selon la règle de composition à gauche :

RFMC5 : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[ɣajru	]Préf.	+	[ʔinsaanii	]N	]MC	« inhumain »
	[[ɣajru	]Préf.	+	[mutabalwar	]Adj.R.	]MC	« amorphe »
	[[ɣajru	]Préf.	+	[ʕaadii	]N	]MC	« extraordinaire »

d) Les mots composés avec / ʕadam + / :

Le préfixe / ʕadam+ / s'emploie en arabe standard moderne généralement devant un substantif en rapport d'annexion déterminé par l'article pour indiquer : l'absence, la privation.

Il correspond souvent aux préfixes français : **a+**, **in+**, **non**, **de+**, **me+**, ils forment

des mots composés selon la même règle de composition à gauche :

RFMC6 :	[t]A	→	[[u]B	+	[t]A	]MC		
			[[ʢadam	]Préf.	+	[muʔaaxaða	]N	]MC	« indulgence »
			[[ʢadam	]Préf.	+	[ħadd	]N	]MC	« non-détermination »
			[[ʢadam	]Préf.	+	[ʔirtijaaħ	]N	]MC	« mécontentement »
			[[ʢadam	]Préf.	+	[ʔittifaaq	]N	]MC	« désaccord »

e) Les mots composés avec /ḍidd +/ :

La préposition /ḍidd+/ rend l'idée de contre, /anti+/ en arabe standard moderne. Elle permet la formation de mots composés selon la règle de composition à gauche :

RFMC7 :	[t]A	→	[[u]B	+	[t]A	]MC		
			[[ḍidd	]Préf.	+	[ʔal ʃawsasa	]N	]MC	« contre espionnage »
			[[ḍidd	]Préf.	+	[ʔal dabbaat	]N	]MC	« antichars »

f) Les mots composés avec / suuʔ / :

Le préfixe / suuʔ +/ se présente comme le premier terme d'un rapport d'annexion. Il correspond en français à « mauvais », ou préfixes : **/mal+, in+, mès+ /** : il forme des mots composés à l'aide de la même règle de composition à gauche :

RFMC8 :	[t]A	→	[[u]B	+	[t]A	]MC		
			[[suuʔ	]Préf.	+	[ʔal baxt	]N	]MC	« malchance »
			[[suuʔ	]Préf.	+	[ʔal ħaal	]N	]MC	« mauvais état »
			[[suuʔ	]Préf.	+	[ʔal fahm	]N	]MC	« incompréhension »
			[[suuʔ	]Préf.	+	[ʔaltafaahum	]N	]MC	« mésentente »
			[[suuʔ	]Préf.	+	[ʔal tanḍiim	]N	]MC	« vices de structure »

Pour conclure sur ce point, nous pouvons dire qu'il existe deux types de composition en arabe standard. Une **composition à droite** du type :

[t]A → [[t]A + [u]A]MC

et une **composition à gauche** du type :

[t]A → [[u]B + [t]A]MC

qui se présentent de la façon suivante :

RFMC1 :	[t]	→	[[t	]A	+	[u	]B	]MC	SUFF. DÉRIV.
RFMC2 :	[t]A	→	[[ɭaa	]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC3 :	[t]A	→	[[ʃibhu	]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC4 :	[t]A	→	[[ʃib	]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC5 :	[t]A	→	[[ɣajru	]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC6 :	[t]A	→	[[ʢadam	]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC7 :	[t]A	→	[[ḍidd	]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC8 :	[t]A	→	[[suuʔ	]B	+	[t	]A	]MC	PRÉF. DÉRIV.

5. Conclusion

Cette recherche, qui s'inscrit dans le cadre de l'hypothèse lexicaliste, préconise une composante morphologique autonome, dont le siège est le lexique, responsable de la flexion, de la dérivation et de la composition.

Cette étude a permis de spécifier d'une façon détaillée la nature des règles lexicales qui permettent de dériver des mots à partir des mots de base. Cette relation s'accomplit par un ensemble de règles de formation de mots. Ces règles se caractérisent par leur formalisme logico-mathématique. Elles sont informatisables et s'appliquent d'une façon naturelle à des processus morphologiques de l'arabe tels que la flexion, la dérivation et la composition.

En effet, leur application en morphologie arabe est d'une grande importance pour la néologie lexicale, surtout lors de la création de mots nouveaux dans les domaines scientifiques et techniques.

Un générateur est conçu pour engendrer automatiquement des mots nouveaux pour différentes catégories grammaticales (substantifs, adjectifs, noms d'action, etc.) à partir d'un ensemble de règles de formation de mots élaborées pour l'arabe standard.

Par un jeu de dérivation et de composition de mots à partir des mots de base (par préfixation, suffixation et composition), le générateur permet ainsi de mettre à jour et d'enrichir le dictionnaire arabe par des néologismes.

Le générateur est relié à un dictionnaire contenant les mots de base, racines, préfixes et suffixes, avec des informations grammaticales associées aux mots ; et à un ensemble de règles de formation de mots.

Lexicographie berbère. Construction des formes de mot et classification des entrées lexicales

Miloud TAIFI

Université de Fès, Maroc

À la mémoire de mon ami et collègue Kaddour CADI

État des lieux

La langue berbère occupe un vaste espace allant de l'oasis de Siwa en Égypte jusqu'en Afrique noire (Niger, Mali et Burkina Faso) en passant par le Maghreb qui constitue véritablement le fief du berbère, de par le nombre très important des populations berbérophones en Algérie et surtout au Maroc. La langue berbère est constituée de plusieurs dialectes ou supra-systèmes qui s'étendent sur des zones géographiques plus ou moins étanches : on dénombre ainsi le touareg (dans les régions sud-sahariennes algériennes, au Mali et au Niger), le tachelhiyt, le tamazight et le tirifiyt au Maroc, le kabyle, le hchaouit et le tamzaybit en Algérie (cf. Galand, 1988 : 207-242).

Les études sur le berbère sont très anciennes, mais les véritables traités de grammaire et les premiers recensements de vocabulaire datent de la deuxième moitié du XIX^{e} siècle. Le premier lexique bilingue berbère/français (dialectes Algérie) est publié en 1844 par Venture de Paradis. Ont été édités ensuite plusieurs travaux de lexicologie et des inventaires de vocabulaire dont les plus importants, en nous astreignant aux dialectes marocains, sont les travaux de lexicologie de Destaing sur vocabulaire tachelhiyt en 1920 et de Laoust sur les mots et choses berbères, l'étude Loubignac sur le berbère des Zaïan et Ait-Sgougou en 1924 et celle de Mercier sur le dialecte des Ait-Izdeg en 1937.

Mais le premier véritable dictionnaire est sans doute celui de Charles de Foucauld : *Dictionnaire touareg-français* (*dialecte de l'Ahaggar*) publié en 1951. Est paru ensuite, en 1982, le *Dictionnaire kabyle-français* de Jean-Marie Dallet, à titre posthume. Le *Dictionnaire mozabite-français* de Jean Delheure vient en 1984, augmenter les travaux lexicographiques berbères. Le dernier travail dans le domaine est à ce jour, le *Dictionnaire tamazight-français* (*parlers du Maroc central*) que j'ai publié

moi-même en 1992. D'autres travaux lexicographiques de grande envergure sont actuellement en cours de réalisation dans le cadre de la préparation de thèses de doctorat.

Les quatre importants dictionnaires cités, recouvrant différents dialectes de la langue berbère constituent une somme considérable de données lexicales et assoient, du point de vue méthodologique, une tradition lexicographique. Ils ont en effet tous adopté, avec quelques options et amendements partiels, la classification par racines, sacrifiant ainsi aux exigences mêmes de la morphologie du berbère qui construit les formes de mot en associant les racines et les schèmes.

Les premières monographies de morphologie berbère ont abordé la construction des formes de mot à travers la morphologie des langues romanes, notamment le français. Plusieurs auteurs ont ainsi essayé de retrouver dans le berbère les procédés de dérivation affixale dont la segmentation isole les bases lexématiques et les morphèmes affixaux (préfixes, infixes et suffixes). Mais bien vite l'application des principes d'analyse valables pour les langues romanes s'avéra impropre à la langue berbère. On découvrit en effet que le berbère appartient à la famille chamito-sémitique et qu'il fallait par conséquent chercher du côté de la morphologie du sémitique.

L'appartenance et l'apparentement du berbère à la famille des langues chamito-sémitiques sont fondés sur plusieurs aspects communs et suffisants pour justifier, du point de vue linguistique, les rapprochements entre le berbère et le sémitique et, du point de vue méthodologique, l'application des mêmes paramètres d'analyse et de description (cf. Galand, 1979a : 463-478). D'où le transfert de la racine sémitique au domaine lexical berbère et l'adoption, dans la pratique lexicographique, de la classification par racines (cf. Cohen, 1993 : 161-175). Mais ce transfert, justifié par la même morphologie du berbère, fait apparaître d'innombrables problèmes tant théoriques que pratiques. C'est de ces problèmes que traitera cette communication.

Construction des formes de mot en berbère

Racines et schèmes

Les formes de mot en berbère sont toutes des formes construites par l'association de deux constituants formels : une racine et un schème. Le premier constituant représente le lexique, le second la morphologie ou plus exactement la grammaire. La racine est généralement définie comme un groupe de consonnes se présentant dans un ordre impératif et qui constitue l'invariant formel d'un paradigme lexical ; le schème comme une structure formelle comportant des éléments vocaliques et/ou consonantiques et assignant des places destinées à être occupées par les radicales de la racine. Le schème porte théoriquement un sens grammatical, puisqu'il catégorise les formes de mot construites en différentes parties de discours, comme le montre la figure suivante (cf. Chaker, 1984 : 136).

RACINES LEXICALES INDIFFÉRENCIÉES

⇓

FORMES DE MOT

⇓	⇓	⇓
Marques verbales	Marques nominales	Marques zéro
VERBES	NOMS	DÉTERMINANTS AUTONOMES
– formes simples	– substantifs	CONNECTEURS
– formes complexes	– adjectifs	– prépositions
	– numéraux	– coordonnants
	– pronoms	– conjonctions

Ainsi, par exemple, la racine MGR porteuse du sens « moisson », est commune à toutes les formes de mot suivantes attestées en berbère.

Formes verbales : *mger, mgir, megger, ttumger, ttumgir, ttumgar.*
Formes participiales : *imgern, imgirn, mgernin, ittumgern, ittumgirn, ttumgernin.*
Formes nominales : *amgar, imgarn, amggar, imeggarn, amg^{w}er, imeg^{w}ran, tamg^{w}ert, timegrin.*

Pour montrer cette communauté formelle, nous allons dégager le schème de chaque forme de mot en remplaçant chaque radicale de la racine par le symbole C que nous notons avec un trait souscrit à chaque fois que la radicale est tendue dans la forme de mot : C̲. Nous obtenons ainsi :

– CCC, CCIC, C̲C̲C, ttuCCC, ttuCCiC, ttuCCaC
– iCCCn, iCCiCn, CCCnin, ittuCCCn, ittuCCiCn, ttuCCCnin
– aCCaC, iCCaCn, aC̲C̲ar, iC̲C̲aCn, aCCC, iCCCan, taCCCt, tiCCCin.

Comme le montre cet exemple, la formation du mot se fait par dérivation associative qui consiste en l'insertion des radicales d'une racine dans les places vides du schème. Une telle insertion est régie par des règles morphologiques dépendant des associations phonétiques et/ou sémantiques permises par la langue. Pour les premières, certains voisinages de phonèmes sont neutralisés de par la nature articulatoire de ces derniers. Ainsi les suites consonantiques k/G, G/K, X/γ, γ/X, Q/γ, γ/Q sont rares, sinon exclues parce qu'elles sont imprononçables. Pour les secondes, c'est l'ordre linéaire des radicales qui confère à la racine son sens lexical : la racine MGR, notée ci-dessus, rend la notion de « moisson », invariant sémantique que se partagent toutes les formes de mot qui dérivent de la racine MGR.

Le changement de la racine par permutation des radicales engendre d'autres suites qui peuvent être soit attestées et constituer le chef de file d'un paradigme lexical, soit non attestées et donner lieu à des créations nouvelles (néologismes ou monstres linguistiques). La permutation de MGR fournit six racines trilitères dont une seule est non attestée (dans le dictionnaire de Taifi, 1992).

MGR : « notion de moisson »
MRG : « notion d'amour »
GMR : « deux notions » : 1) « chasse »; 2) « cheval »

GRM : « trois notions » : 1) « ronger, grignoter »; 2) « qui a une seule corne, qui est sans cornes (ovin) »; 3) « saint, marabout ».
RMG*
RGM : « notion d'insulte, de malédiction ».

L'ordre RMG n'est donc pas exploité par la morphologie du berbère (on notera du moins qu'en Kabyle RMG avec R emphatique rend le sens de « tonner » : Dallet, 1982 : 726). Les autres racines donnent lieu à des familles lexicales dont chacune est constituée d'un certain nombre de formes de mot ; chaque famille forme un champ morpho-sémantique. Ainsi la racine GMR dont le contenu lexical relève de la notion de « chasse » fournit le champ morpho-sémantique suivant comportant 17 formes de mot attestées :

verbes :	*gmer, gmir, gemmer, ttugmer, ttugmir, ttugmar*
participes :	*igmern, igmirn, gmernin, ittugmern, ittugmirn, ttugmernin*
noms :	*tagemrawt, tigemrawin, tanegmart, anegmar, inegmarn*

construites respectivement sur les schèmes suivants :

– CCC, CCIC, C<u>C</u>C, ttuCCC, ttuCCIC, ttuCCaC
– iCCCn, iCCiCn, CCCnin, ittuCCCn, ittuCCiCn, ttuCCCnin
– taCCCawt, tiCCCawin, tanCCaCt, anCCaC, inCCaCn

De même, le champ morpho-sémantique de la racine RGM relative au domaine notionnel de « insulte et malédiction » est composé de 20 formes de mot :

verbes :	*rgem, rgim, reggem, tturgem, tturgim, tturgam, mergam, ttemergam*
participes :	*irgemn, irgimn, rgemnin, itturegmen, itturgimn, tturegmnin, mergamnin, ttemergamnin*
noms :	*argam, irgamn, tareggimt, tirggam*

dont les schèmes se présentent ainsi :

– CCC, CCIC, C<u>C</u>C, ttuCCC, ttuCCiC, ttuCCaC, mCCaC, ttmCCCam
– iCCCn, iCCiCn, CCCnin, ittuCCCn, ittuCCiCn, ttuCCCnin, mCCaCnin, ttmCCaCnin
– aCCaC, iCCaCn, taC<u>C</u>iCt, tiC<u>C</u>aC

Les trois champs morpho-sémantiques présentés répondent à la définition des racines et des schèmes donnée par Cantineau (1950 : 74) et souvent citée pour décrire la construction lexicale dans les langues appartenant à la famille chamito-sémitique.

> Chaque mot a sa racine et son schème; on pourrait comparer le vocabulaire à un tissu dont la trame serait l'ensemble des racines attestées dans la langue et la chaîne l'ensemble des schèmes existants. Chaque point d'intersection de la chaîne, et de la trame, serait un mot, car tout mot est entièrement défini sans ambiguïté par sa racine et son schème, tout schème de son côté fournissant des mots à différentes racines et la plupart des racines fournissant des mots de différents schèmes.

Cette métaphore de tisserand n'explique cependant pas tout. Si le schème relève de la morphologie et constitue un cadre formel prêt à accueillir les radicales de la racine, celle-ci, par contre, n'est pas suffisamment et clairement définie. Premièrement, la racine est-elle exclusivement consonantique, ou bien y a-t-il lieu de considérer certains segments vocaliques comme radicales dans les cas où ils sont constants et ne subissent pas de changement ou d'effacement ? Deuxièmement, est-ce que la racine est tout simplement un groupe d'éléments commun à une série de formes de mot, ou bien est-ce un signifiant doté de signifié précis ? Autrement dit, la racine est-elle seulement une unité formelle, ou une unité formelle et sémantique. Apporter des réponses à ces questions par l'analyse lexicologique, est un préalable à toute pratique lexicographique berbère, et aussi, dans certaines mesures, à celle de l'arabe.

Racine : consonnes et voyelles

Si le critère qui préside à l'établissement d'une racine dans une famille lexicale est l'invariabilité de ses radicales dans tous les lexèmes construits, il n'y a pas lieu d'exclure les éléments vocaliques qui répondent à ce critère. Si certaines voyelles sont constantes, elles ne peuvent appartenir qu'à la racine et non aux schèmes. La définition donnée par Meillet (cité par Cohen, 1993 : 162) corrobore ce point de vue : « Un mot " appartient " à une racine, il fait partie d'un ensemble de mots ayant en commun un groupe de phonèmes auquel est associé un certain sens général. » Ainsi les radicales d'une racine, selon l'auteur, sont des phonèmes, ceux-ci pouvant être soit consonantiques ou vocaliques, l'essentiel étant leur régularité dans tous les mots appartenant à une même famille lexicale. Il faut ajouter que le critère de la constance n'est valide que si les voyelles occupent toujours la même place dans les schèmes.

Ceci amène Cohen (1993 : 162), commentant la citation de Meillet, à définir la racine ainsi : « la racine est une séquence ordonnée de phonèmes qui constitue la totalité des éléments communs à un ensemble dérivatif », en remarquant que si la racine a été toujours considérée exclusivement consonantique dans les langues chamito-sémitiques, ce n'est qu'un fait d'observation : les consonnes sont en effet (surtout en arabe classique, pris comme référence) beaucoup plus sujettes à la constance que ne le sont les voyelles ; mais ceci ne justifie pas l'exclusion des voyelles constantes de la racine. Cohen propose par conséquent, pour le berbère, de conférer aux voyelles régulières le statut de radicales et d'en tenir compte dans la pratique lexicographique. Donnons un exemple pour illustrer ce point de vue.

Soit les formes de mot suivantes :

aḍar	« pied »	schème : iCaC
iḍarn	« pieds »	schème : iCaC (n)
taḍartt	« petit pied, pieds d'enfant »	schème : (t)aCaC(tt)
tiḍarin	« petits pieds, pieds d'enfant »	schème : (t)iCaC(in)

Comme on le constate, ces formes se partagent la séquence -ḍar- qui sera ainsi la racine CVC = ḌAR. Les éléments mis entre parenthèses sont les marques du genre et du nombre. Par contre, dans le paradigme suivant, aucune voyelle n'est constante en occupant la même place dans les différents schèmes :

afus	« main »	schème : aCuC
ifassen	« mains »	schème : iCaC(n)

tafustt	« petite main, main d'enfant »	schème : (t)aCuC(t)
tifassin	« petites mains, mains d'enfant »	schème : (t)iCaC (in)

Dans ce cas, seules les consonnes constitueront la racine, à savoir CC : FS, puisque les lexèmes dérivés n'ont en commun aucune voyelle constante dans la même position.

L'intérêt de la promotion des voyelles constantes au statut de radicales, permettra, selon Cohen (1993 : 161-175), de pouvoir distinguer les racines homophones, surtout les monolitères et les bilitères, en réduisant leur nombre par l'isolement de celles qui comportent une ou plusieurs voyelles constantes. Ainsi, par exemple, au lieu de sept racines consonantiques homophones ḌR fournissant chacune un paradigme lexical (cf. Taifi, 1992 : 91-93), il n'y en aura que quatre, si l'on classe à part celles qui contiennent un élément vocalique régulier. Voici les données :

ḌR$_1$:
Sens général : « descendre, baisser (intransitif) ».
Formes de mot :
Verbes : *ḍer, tḍar/ḍḍar, ḍir, sḍer, sḍir, ttesḍar, ttusḍer, ttusḍir, ttusḍar.*
Participes : *iḍern, iḍirn, ḍernin, iḍḍarn, ḍḍarnin, isḍern, isḍirn, sḍirnin, ittesḍarn, ttesḍarnin, ittusḍern, ittusḍirn, ttusḍarnin.*
Noms : *taḍuri, taḍurin, asḍar, isḍarn.*

ḌR$_2$:
Sens général : « salir, souiller » .
Formes de mot :
Verbes : *aḍer, uḍer, uḍir,ttaḍer, tyiḍer, tyaḍar, myaḍar, temyaḍar.*
Participes : *yuḍern, yuḍirn, uḍernin, ityiḍern, tyiḍernin, ityaḍarn, tyaḍarnin, myaḍarnin.*
Noms : *iḍer.*

ḌR$_3$:
Sens général : « être sourd » .
Formes de mot :
Verbes : *ḍurḍer, tḍurḍur, ḍurḍir.*
Participes : *iḍurḍern, iḍurḍirn, ḍurḍernin, itḍurḍurn, tḍurḍurnin.*
Noms : *aḍerḍur, iḍerḍurn, taḍerḍurt, tiḍerḍurin, tiḍerḍert.*

ḌR$_4$:
Sens général : « nuire, faire mal »
Formes de mot :
Verbes : *ḍerra, tḍerra/ḍḍerra, tuḍerra, ttuḍerra, mḍerra, temḍerra.*
Participes : *iḍerran, ḍerranin, itḍerran, tḍerranin, ituḍerran, tuḍerranin, mḍerranin, temḍerranin.*
Noms : *aḍerra, lmaḍerra, lmaḍerrat, ḍḍarar.*

ḌR$_5$:
Sens général : « pied »
Formes de mot :
Noms : *aḍar, iḍarn, taḍartt, tiḍarin.*

ḌR$_6$:
Sens général : « maïs »
Formes de mot :
Noms : *ḍḍra, aḍḍra, aḍḍraten, taḍḍrat, taḍḍratin.*

Les racines 5 et 6 sont constituées de consonnes et de voyelles, celles-ci étant constantes, et se présentent ainsi : ḌAR, ḌRA. La différence de position (médiane # finale) permet donc de distinguer par la constance vocalique les deux racines à l'origine homophones. Ces racines ne seront plus alors considérées comme bilitères, mais s'ajouteront à l'ensemble des trilitères, puisqu'elles sont formées de trois radicales. Le traitement lexicographique placera ces nouvelles formes à leur place dans l'ordre alphabétique : ḌAR, ḌR (1, 2, 3, 4,), ḌRA.

Le second exemple illustre la même procédure ; cette fois, le paradigme des racines homophones est plus fourni : il comporte dix formes semblables :

LS$_1$:
Sens général : « se vêtir, s'habiller ».
Formes de mot :
Verbes : *lsi, lsa, lessa, lsi, ttulsa, ttulsi, ssels, sselsi, sselsa, mselsa, ttemselsa.*
Participes : *ilsan, ilsin, lsanin, ilessan, lessanin, ittulsan, ittulsin, ttulsanin, isselsen, isselsin, sselsanin, mselsanin, ttemselsan.*
Noms : *melsiwt, melsiwat, timelsit, timelsa, assels, isselsan, aselsu, iselsa.*

LS$_2$:
Sens général : « tondre ».
Formes de mot :
Verbes : *lles, telles, llis, ttulles, ttullas.*
Participes : *illsen, illisn, llesnin, ittellesn, ittullsen, ittullasn, ttullasnin.*
Noms : *talasa, talusi, ulus, ilis, ilisn, tilist, tilisin, amlas, imlasn, amlus, imlas, imlusen.*

LS$_3$:
Sens général : « souiller, salir ».
Formes de mot :
Verbes : *lles, telles, llis, ulus.*
Participes : *illesn, illisn, llesnin, itellesn, tellesnin.*
Noms : *ulus, ulusn.*

LS$_4$:
Sens général : « être obscur, sombre; faire noir ».
Formes de mot :
Verbes : *lles, telles, llis, ssuls, ssulus.*
Participes : *illesn, illisn, llesnin, issulsen, ssulsnin.*
Noms : *tallest, tillas.*

LS$_5$:
Sens général : « recommencer, refaire, répéter ».
Formes de mot :

Verbes : *als, ttals, ulis.*
Participes : *yulsen, yulisn, ulesnin, ittalsen, ttalesnin.*
Noms : *alas, ils.*

LS_6 :
Sens général : « être écarté de la succession du grand-père par ses oncles paternels, par suite de la mort de son père (petit-fils) ».
Formes de mot :
Verbes : *als, uls, ttels, ulis.*
Participes : *yulsen, yulisn, ulesnin, ittalsen, ttalesnin.*
Noms : *ulus, amalas, imalasn.*

LS_7 :
Sens général : « repas de la fin de la matinée ».
Formes de mot :
Noms : *allas, allasn.*

LS_8 :
Sens général : « frère/sœur du mari (pour la femme) ».
Formes de mot :
Noms : *alus, ilusn, talust, tilusin.*

LS_9 :
Sens général : « langue (organe et idiome) ».
Formes de mot :
Noms : *ils, alsiwn, tilset, tilsatin.*

LS_{10} :
Sens général : « mousse de savon ».
Formes de mot :
Noms : *alus, ilusa.*

Dans ce deuxième exemple, nous avons donc dix racines homophones LS. Mais si l'on tient compte des voyelles constantes, les racines 7, 8 et 9 doivent être exclues de cet ensemble : les formes de mot de LS_7 comportent, en effet, un /a/ constant, de même celles de LS_8 et de LS_{10} un /u/ régulier. Les trois nouvelles racines ainsi dégagées sont LAS, LUS et LUS devenant des trilitères, de par l'élément vocalique introduit. La nouvelle classification par ordre alphabétique sera : LAS, LS (1, 2, 3, 4, 5, 9) et LUS. Comme on l'aura remarqué, si l'homophonie de LS est ainsi réduite, il y a cependant création d'un autre cas de ressemblance formelle entre LUS < LS_8 et LUS < LS_9. On aura remarqué aussi que la voyelle a plus de chances de rester constante lorsque la racine ne fournit que très peu de formes de mot, c'est le cas effectivement des racines LAS (2 dérivés), LUS (4 dérivés) et LUS (2 dérivés).

Si la constance vocalique a, sans doute, un statut théorique qui exige l'insertion des voyelles dans l'armature des racines, répondant aux définitions de Meillet et de Cohen, il n'en demeure pas moins que son application dans la pratique lexicographique pose, du moins pour le berbère, plus de problèmes qu'elle n'en résout, car les voyelles sont plus alternantes que constantes. Et ceci pour plusieurs raisons :

1– Le nombre des voyelles en berbère est de trois phonèmes : /a/, /i/ et /u/. Ce sont celles-là qui constituent le triangle vocalique de base. Les trois voyelles connaissent cependant des allophones dus à certains environnements consonantiques emphatiques ou vélaires qui imposent une plus grande aperture vocalique. Ces allophones, n'apparaissent donc que lors de la construction des formes de mot ou des séquences syntagmatiques.

Sachant que plus un ensemble est réduit, plus les éléments qui le composent sont fréquemment utilisés, les trois voyelles basiques doivent donc nécessairement, et le plus souvent, alterner pour différencier les schèmes qui accueillent les racines. Une telle différenciation formelle des unités lexicales à travers les schèmes n'est possible en berbère que si les voyelles constitutives des schèmes alternent. C'est ce qui explique, sans doute, que la constance vocalique est très rare, sinon impossible, dans des paradigmes lexicaux comportant plusieurs formes de mot. C'est le cas notamment de ceux qui sont construits à partir des racines verbo-nominales : exemple de la racine FD (Taifi, 1992 : 103-104) qui fournit treize formes verbales (simples et complexes) et sept formes nominales (sans compter les formes participiales de chaque verbe) ; de même, la racine SY (Taifi, 1992 : 663-664) fournit vingt-trois formes verbales (simples et complexes) et huit formes nominales. Étant donné que chaque forme de mot doit s'articuler sur un schème spécifique, pour éviter une homophonie excessive, la morphologie berbère recourt aux voyelles, et comme celles-ci ne sont que trois, le jeu formel d'alternance, par des opérations de commutation (remplacement d'une voyelle par une autre) et de permutation (changement de position) est la seule voix de salut.

C'est ce rôle morphologique (laborieux sans conteste : elles ne sont que trois !) assuré par les voyelles qui a fait dire, avec raison, à André Basset (1929 : XXV) qu'en berbère, « la voyelle s'affirme par ailleurs comme un élément morphologique pour qu'on puisse lui attribuer pareille valeur même là où elle forme avec des éléments consonantiques, un ensemble invariable ». La valeur dont parle Basset est celle de l'appartenance de la voyelle à la racine. Il faut noter qu'une telle constatation n'exclut pas la constance vocalique. Mais celle-ci ne peut être observée que pour des racines à paradigme lexical réduit, notamment les racines exclusivement nominales ou celles, très peu nombreuses, qui fournissent des outils grammaticaux (connecteurs, conjonctions, prépositions...)

2 – L'alternance intervient, par ailleurs, dans la conjugaison des verbes selon les personnes et les valeurs aspectuelles (cf. Galand, 1984 : 304-315) et affecte essentiellement les initiales et les finales des formes verbales ; exemple : le verbe construit à partir de la racine F et signifiant « trouver » se conjugue ainsi :

Aoriste : *af-x, t-af-d, y-af, t-af, n-af, t-af-m, taf-mt, af-n, af-nt*
Accompli : *ufi-x, t-ufi-d, y-ufa, t-ufa, n-ufa, t-ufa-m, t-ufa-mt, ufa-n, ufa-nt*
Accompli négatif : *ufi-x, t-ufi-d, y-ufi, t-ufi, n-ufi, t-ufim, t-ufimt, ufi-n, ufi-nt*
Inaccompli : *ttafa-x, ttafa-d, ittafa, t-tafa, n-ttafa, ttafa-m, ttafa-mt, ttafa-n, ttafa-nt.*

Les voyelles assurent donc, dans la conjugaison, un rôle morphologique et leur alternance est un critère de distinction des thèmes verbaux. On aura ainsi, en éliminant les indices de personne (pronoms personnels) et le formant du schème de l'inaccompli -tt/t, quatre formes du même verbe : *af, ufi, ufa, afa*, dans lesquelles seule la consonne radicale F est constante : aC, uCi, uCa, aCa.

3 – L'alternance vocalique caractérise aussi les formes nominales à cause de l'opposition État libre/État d'annexion. : les noms au singulier ayant un /a/ à l'initiale et les noms féminins ayant (ta) subissent des changements dans certains contextes syntagmatiques : la voyelle /a/ est réalisée /u/ pour le singulier et elle est effacée pour le féminin. Ainsi, par exemple, le nom *argaz* « homme » est réalisé *urgaz* (a > u) lorsque le nom est postposé au verbe et a la fonction de complément explicatif : *argaz, i-rwel*, « l'homme, il s'est enfui » en contraste avec *i-rwel urgaz*, « il s'est enfui, l'homme ». De même *tameṭṭuṭṭ* « femme » devient *tmeṭṭuṭṭ* (a > Ø) dans le même contexte syntaxique : *tameṭṭuṭṭt, t-rwel*, « la femme, elle s'est enfuie » et *t-rwel tmeṭṭuṭṭ*, « elle s'est enfuie, la femme ».

4 – L'alternance vocalique est due aussi aux changements morphologiques relatifs à la catégorie du nombre, surtout pour les pluriels dits « internes » : exemple : singulier : *amazir* « campement » , pluriel : *imizar* (a > i, a > i, i > a) ; singulier : *adaku* « sandale », pluriel, *iduka* (a > i, a > u, u > a).

5 – Les semi-voyelles sont parfois réalisées en voyelles correspondantes (y >i, w > u,) ce qui complique davantage la reconstitution des voyelles radicales (cf. Taifi, 1990b : 219-232).

6 – L'alternance, sans être d'origine structurale, affecte les voyelles différemment selon les dialectes berbères, elle est, dans ce cas, une marque de l'habitus articulatoire d'un groupement géo-linguistique : le processus de dialectalisation de la langue berbère a eu comme conséquence l'accentuation des particularismes phonétiques concernant aussi bien les phonèmes consonantiques que vocaliques : exemple de différence dialectale due à l'alternance vocalique : le terme qui signifie « ficelle » est *ifili* dans un parler, mais *ifilu* dans un autre, la finale étant un /i/ dans l'un et un /u/ dans l'autre ; de même dans *abaxxu/abuxxu*, « insecte » et *ablullu/iblelli*, « papillon de nuit », il y a alternance vocalique dialectale.

En conclusion à cette présentation, il semble que l'introduction des radicales vocaliques dans la reconstitution des racines n'est pas efficace et rentable dans un travail lexicographique, dans la mesure où la constance vocalique ne concerne que très peu de racines et que son utilisation n'a pas, par conséquent, une grande incidence méthodologique quant à la différenciation des racines homophones.

Racine : tension et réduplication des radicales consonantiques

La lexicographie berbère adoptant la classification par racines est confrontée à deux autres phénomènes qui caractérisent certaines racines. Le premier concerne la tension des radicales et le second leur réduplication. Pour le premier cas, la question (pour le lexicographe) est la suivante : faut-il considérer le trait articulatoire de tension comme critère de distinction des racines et noter par conséquent les radicales tendues ? Ou bien considérer qu'une même consonne tendue (simple) ou non tendue est en fait une seule radicale. Illustrons le phénomène.

Soit les formes de mot : *abrid, iberdan* « chemin(s) » construites à partir de la racine BRD, et les formes de mot *aberrad, iberradn* « théière(s) » dérivées aussi de la

racine BRD, mais dont la deuxième radicale est tendue : BRRD. La tension est donc un trait différentiel qui distingue les deux racines, comme elle l'est aussi dans *aglu* « gésier » en opposition à *agella* « tenture ». Théoriquement, les deux racines (à radicale non tendue # à radicale tendue) ne peuvent être traitées comme homophones et doivent donc être classées séparément. La tension serait ainsi réductrice d'homophonie.

Les faits ne sont pas cependant aussi simples. Il faut noter qu'en berbère, la tension a un double rôle : elle est d'abord un trait phonologique distinguant des paires minimales : *ilis* « toison » # *illis* « sa fille », *kes* « paître » # *kkes* « enlever » ; mais elle est aussi un formant du schème et assure donc un rôle morphologique : l'inaccompli est rendu en berbère, soit par des schèmes comportant le formant consonantique t/tt à l'initiale : *afd* > *ttafd* « s'en aller », *af* > *ttafa* « trouver », *ddu* > *teddu* « partir » ..., soit par la tension de l'une des radicales de la racine : *rzu* > *rezzu* « chercher », *mger* > *megger* « moissonner », *γmes* > *γemmes* « couvrir » ...

Le même trait articulatoire est utilisé en outre dans la formation des dérivés nominaux : de nombreux noms d'agent et noms qualificatifs sont construits à partir des racines trilitères sur le schème aCC̲aC dans lequel la deuxième radicale est tendue sans qu'elle le soit forcément dans les autres formes de mot fournies par les mêmes racines (voir Taifi, 1989 : 872-926) : ainsi des racines MDY, NBḌ, ZDM, sont dérivés respectivement les verbes *mdey* « guetter », *nbeḍ*, « commander » et *zdem*, « chercher du bois », dans lesquels les radicales sont non tendues. Par contre, les noms d'agent correspondants ont la deuxième radicale tendue : *amedday*, *anebbaḍ*, *azeddam*. De même, le schème aCC̲aC constitue le cadre de formation pour certains noms qualificatifs : BXN > *bxin* « être noir et *abexxan* « noir » ; WSR > *wsir* « être vieux » et *awessar* « vieux » , LWγ > lwiγ, « être mou » et *alegg*^{w}*aγ* « mou » (*ww* > *gg*^{w}).

Ces exemples montrent donc que la tension relève plus de la morphologie, puisqu'elle est formant du schème, que du lexique. Un tel constat est suffisant pour soutenir que toutes les formes de mot avec une consonne tendue ne proviennent pas automatiquement de racines à radicale tendue. Ce point de vue est corroboré d'ailleurs par l'apparition aléatoire de la tension : dans des cas où elle ne joue aucun rôle morphologique, elle peut être en effet le résultat d'une assimilation phonétique : *tirnit* < *tirrit* « victoire », *anli* < *alli* « cerveau », ou tout simplement gratuite, imposée par l'arbitraire du signe, et n'affectant que certaines formes de mot d'une même famille lexicale : *lles* « tondre » mais *ilis* « toison » et *talasa* « action de tondre » ; *ffeγ* « sortir » mais *ufuγ* « action de sortir » .

La pertinence phonologique de la tension qui en aurait fait une caractéristique de la racine et, par conséquent, une propriété du lexique, est neutralisée par son rôle morphologique et grammatical plus dominant en berbère, et aussi, si l'on s'en tient seulement aux structures des mots, c'est-à-dire à l'arbitraire des signifiants, à l'aléatoire des régularités formelles des familles lexicales. Il est donc plus économique, dans un travail lexicographique, de ne pas tenir compte de ce trait dans l'établissement et la classification des racines dans un dictionnaire.

Le second fait lexical est la réduplication des radicales consonantiques, que nous illustrons par les exemples suivants : les formes de mot *adrar* « montagne » et *adlal* « longue tresse de cheveux » comportent respectivement trois consonnes DRR et DLL

qui constituent les racines à partir desquelles elles sont construites. On remarque que la deuxième et la troisième radicale sont de même nature consonantique, il y a donc réduplication d'une radicale. Faut-il donc, dans la classification par racines, tenir compte de la réduplication ou non ? Les formes de mot *adrar* et *adlal* seront-elles répertoriées respectivement sous les racines DRR et DLL ou bien tout simplement sous DR et DL. Dans le premier cas, les racines sont des trilitères et dans le second des bilitères.

La réduplication des radicales se présente dans deux cas de figure : la racine redupliquée reste invariable et commune à toutes les unités lexicales qu'elle informe. Ainsi la racine QQŠ, rendant le sens de « épier (à travers une ouverture) », fournit les formes verbales *qiqš, tqiqiš, tuqiqš, tuqiqiš, mqiqiš, ttemqiqiš*, les participes *iqiqšn, itqiqišn, tqiqišnin, ituqiqšn, ituqiqišn, mqiqišnin, ttemqiqišn*, et les formes nominales *aqiqš, iqiqšen*. Dans cette famille lexicale, la racine dont la première radicale est redupliquée, constitue la base commune à toutes les formes de mot construites.

Il nous semble, dans ce cas, en nous tenant seulement au niveau méthodologique, qu'il n'y a aucun intérêt pour le lexicographe, à distinguer, dans la procédure de classification, entre les racines redupliquées et celles qui ne le sont pas. Autrement dit, il serait plus économique, pour éviter une dispersion exagérée des formes, de répertorier la racine ci-dessus à l'adresse QŠ et non à QQŠ. Le même principe de classification peut être appliqué aussi pour des racines qui connaissent une double réduplication, procédé utilisé en berbère généralement dans un but expressif (cf. Azougarh, 1992 : 114-135). Les racines dγDγ et FLFL fournissent, la première, avec le sens de « être contusionné, meurtri », les verbes *deγdeγ, dγdiγ, tdeγdaγ, sdeγdeγ, sdeγdaγ, sdeγdiγ*, les participes *ideγdγen, deγdeγnin, isdeγdeγn, sdeγdeγnin* et les noms *adeγdeγ, asdeγdeγ* ; la seconde, avec le sens de « déborder (liquide qui bout) », les verbes *flufel, ttefluful*, les participes *ifluflen, flufelnin, itteflufuln, tteflufulin* et le nom *aflufel*. Ces racines illustrent la double réduplication $C_1C_2C_1C_2$ et seront donc classées comme bilitères C_1C_2 : Dγ et FL.

Une telle option doit être cependant amendée dans les cas où les radicales redupliquées ne sont pas limitrophes. Les racines quadrilitères suivantes comportent des radicales redupliquées : ŠNŠL ayant le sens de « secouer violemment » et fournissant les verbes *šenšel, šenšil, tšenšil, ttušenšel, ttušenšil, ttušenšal*, les participes *išenšeln, išenšiln, šenšelnin, šenšalnin, tšenšalnin*, et les noms *ašenšel, išenšeln* et SKSW avec le sens de « regarder, voir » et étant la racine des verbes *seksiw, sseksiw*, des participes *iseksiwn, seksiwnin, sseksiwnin* et du nom *aseksiw*. Les deux racines ont la structure $C_1C_2C_3C_4$ dans laquelle les deux radicales C_1 et C_3, quoique identiques, sont séparées cependant par une autre radicale et ne sont donc pas limitrophes. Pour ces cas, il y a nécessité procédurale, afin d'identifier ces racines, de les classer sous leur forme initiale, c'est-à-dire comme quatrilitères : ŠNŠL et SKSW et non pas comme des trilitères ŠNL* et SKW* car il n'y a aucun indice qui permettra d'identifier correctement les formes de mot construites à partir de leur racine d'origine.

Le deuxième cas de figure concerne la variabilité de la réduplication des radicales dans le paradigme lexical ; certaines formes comportent deux radicales identiques, et d'autres une seule, tendue ou non tendue. Examinons quelques exemples : soit la famille lexicale suivante dont l'invariant sémantique est la notion de « mastication » : les verbes *fezz, fezzi, fezza, fezzi, tefzaz, ttufezzi, ttufezza, ttufezzi*, les par-

ticipes : *ifeẓẓen, ifeẓẓin, itefẓaẓn, tefẓaẓnin, ittufeẓẓen, ittufeẓẓin, ittufẓaẓn, ttufẓaẓnin*, les noms *afẓaẓ, ifẓaẓn, tuffiṭ, tuffaẓ, uffaẓ, uffaẓn*. À examiner ces différentes unités lexicales, l'on constate que la deuxième consonne radicale ẓ est réduplicquée dans quelques formes et elle est unique, tendue (*feẓẓ*) ou non tendue (*tuffiṭ*), dans d'autres. Pour ces cas, et ils sont nombreux dans le lexique berbère, faut-il relever la racine bilitère Fẓ ou la racine réduplicquée : la trilitère Fẓẓ. Il nous semble que, du point de vue méthodologique et dans un souci de simplification de la classification, il est préférable de reconstruire, pour ces cas, les racines à radicale unique, car la réduplication n'ayant pas de rôle morphologique régi par des règles de grammaire est aléatoire et fluctuante, dépendant de l'arbitraire des signes.

Nous avons examiné, dans ce qui précède, quelques faits formels de la racine en considérant la nature et les combinaisons des radicales. Nous avons proposé quelques options méthodologiques qui doivent présider à la classification par racines en lexicographie berbère. Nos propositions sont essentiellement dictées par le principe d'économie et de simplicité et ne sont pas toutes justifiées par une quelconque théorie de la racine.

Racine : forme et sens

La racine n'est pas cependant une simple forme basique d'un paradigme lexical et la considérer comme telle est sans intérêt pour le lexicographe désireux de fournir des informations sur l'organisation et les structurations morpho-sémantiques de la langue dont il confectionne le dictionnaire. La racine ne sera, dans cette perspective, qu'un simple indicateur d'ordre et de regroupement aberrants. Donnons un exemple pour montrer une telle aberration : les unités lexicales suivantes comportent toutes l'invariant consonantique BD pris comme racine (n'est fourni ici qu'un seul élément, avec son premier sens, de chaque champ morpho-sémantique) :

bedd : « se lever, se dresser »
bdu : « commencer, débuter »
abda : « toujours, tout le temps »
abadu : « canal d'irrigation »
biddi : « obligatoirement, nécessairement »
lebda : « feutre (étoffe) ».

Signalons d'abord que le (L) initial de *lebda* est un article défini de l'arabe que le berbère garde lorsqu'il emprunte à l'arabe, mais ce morphème perd sa fonction en berbère.

Il est tout à fait évident qu'il n'est pas raisonnable, du point de vue lexicographique, de classer les formes de mot données en exemple sous une même racine. Un tel regroupement transgresse la règle sémantique relative aux affinités de sens. En effet, aucune relation sémantique n'est possible à établir entre les différentes formes de mot construites à partir de la séquence commune BD, il y a donc lieu de les dégrouper bien qu'elles soient liées dans leur forme par la même charpente consonantique. On considérera ainsi que nous avons affaire non pas à une racine unique, mais à plusieurs BD_1, BD_2, BD_3, BD_4, BD_5, BD_6, qui sont homophones mais différenciées quant à leur sens général.

Le critère sémantique est donc important pour pouvoir donner à la racine un statut théorique et pratique acceptable en lexicographie berbère et pour qu'elle puisse être utilisée comme paramètre de classification. Le critère sémantique n'est cependant pas facile à circonscrire : selon quelle analyse peut-on différencier sémantiquement des racines homophones ? Autrement, qu'appelle-t-on un sens général ? Est-ce une notion suffisamment large pour intégrer diverses acceptions et nuances sémantiques ? Ou bien est-ce, au contraire, un sens unique ?

De telles questions redoutables n'ont, malheureusement, pas été l'objet de recherche en linguistique berbère. Mais, indépendamment de problèmes de sémantique théorique, le lexicographe est confronté à des options méthodologiques imposées par les lois du genre. Il nous semble qu'il y a deux positions extrêmes qui doivent être évitées si l'on veut que le travail lexicographique ne soit pas en contradiction avec les données de la langue. (cf. Taifi, 1988 : 15-26).

Premièrement la racine ne doit pas être considérée comme un simple ensemble invariable de consonnes commun à une série de formes de mot. Une telle option permettra certainement l'éradication de l'homophonie des racines, mais celles-ci n'auraient plus aucun statut théorique et aucune signification. Elles ne peuvent être utilisées comme principe de classification du lexique, car les regroupements des sous-entrées, c'est-à-dire des formes de mot, sous une racine, seront hétérogènes et ne constitueront plus des champs morpho-sémantiques. Or, un dictionnaire, qui est, sans doute d'abord, un travail sur le lexique, doit nécessairement tenir compte des structurations lexicales de la langue qu'il recense et présenter aux consultants de telles structurations. Définir la racine comme simple forme et l'utiliser comme telle dans la classification lexicographique aboutirait donc à établir des ensembles lexicaux composés d'éléments qui n'ont pas de relations sémantiques.

La seconde option extrême consiste à distinguer une racine pour chaque effet de sens délimité, c'est-à-dire à attribuer aux unités lexicales une stricte monosémie, en neutralisant les caractéristiques polysémiques du lexique. Ainsi, la forme de mot *ifri* ayant quatre significations apparentées : « grotte », « caverne », « gîte » et « terrier » sera classée sous quatre racines différentes FR_1, FR_2, FR_3 et FR_4 puisqu'à chaque sens unique doit correspondre, selon l'option de la monosémie stricte, une racine. Une telle procédure aboutirait, on s'en doute, à l'éclatement et à l'éparpillement du lexique en augmentant excessivement le nombre des racines.

La position intermédiaire, celle adoptée par les quatre dictionnaires précédemment cités, tient compte des deux aspects définitionnels de la racine : la forme et le sens. Mais dans la pratique, les critères de différenciation ne sont pas toujours suffisamment clairs pour permettre une rigoureuse distinction et circonscription des effets de sens. En fait, c'est l'éternel et fameux phénomène de l'opposition homophonie/polysémie qui resurgit à chaque fois, phénomène caractérisant toutes les langues du monde, mais plus épineux pour la lexicographie berbère à cause justement de l'organisation du lexique en racines et schèmes et aussi, dans l'état actuel de la linguistique berbère, de l'indigence des travaux en lexicologie et en sémantique.

Conclusion

Nous avons essayé, dans cette communication, d'exposer un certain nombre de problèmes de méthodologie en lexicographie berbère. Nous avons soutenu, en filigrane, que la dictionnairique berbère ne peut faire l'économie de la racine comme principe de classification, car ce principe est imposé par la morphologie de la langue. La construction des formes de mot se fait en effet par dérivation associative, insérant les racines dans des schèmes. La classification par racines bute, toutefois, sur des difficultés d'application et de procédure. Nous avons proposé des solutions à certaines de ces difficultés, solutions dictées par notre pratique. Mais il en reste d'autres, celles concernant, notamment, l'altération des racines et la dispersion des schèmes que nous avons décrites dans des articles précédents (cf. Taifi, 1990a : 219-232 et Taifi, 1990b : 92-110). L'objectif de cette contribution est de participer à l'élaboration de la métalexicographie berbère.

Informatisation du *Dictionnaire explicatif et combinatoire* : le projet NADIA-DEC

Gilles SÉRASSET

GETA-CLIPS-IMAG (UJF & CNRS), Grenoble, France

Présentation du projet

Le projet NADIA-DEC est basé sur les travaux de définition d'un système universel de bases de données lexicales multilingues au laboratoire GETA-CLIPS et sur les travaux de définition du *Dictionnaire explicatif et combinatoire du français contemporain* au laboratoire GRESLET.

Ce projet vise la création d'une version informatisée du *Dictionnaire explicatif et combinatoire du français contemporain* (DEC). Cette version devra contenir l'ensemble des informations présentes dans le DEC sous une forme aussi structurée que possible. Elle s'appuie donc sur le système de gestion de bases lexicales multilingues SUBLIM défini au GETA-CLIPS.

Ce projet répond à plusieurs motivations de la part de chacun des partenaires. D'une part, il permet de tester le système SUBLIM en l'utilisant pour un dictionnaire mettant en œuvre des structures complexes. D'autre part, les informations contenues dans le DEC présentent une richesse que l'on ne trouve dans aucun autre dictionnaire informatisé. Enfin, la mise en œuvre de ce projet suppose la création d'outils informatiques simplifiant la gestion d'un tel dictionnaire.

Pour atteindre ces différents objectifs, nous souhaitons non seulement informatiser une version du DEC, mais aussi informatiser sa chaîne de production afin de faire en sorte que le DEC existe d'abord sous forme informatique puis sous une forme imprimable. La méthodologie du projet est donnée par la figure 1.

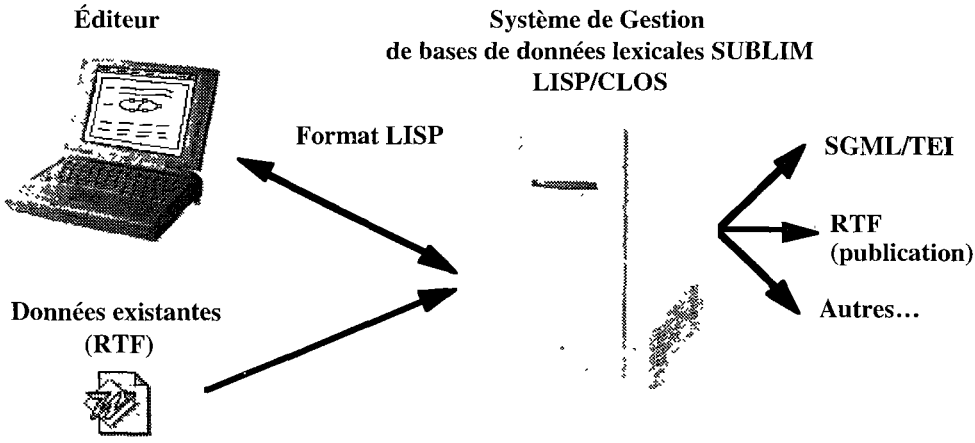


FIGURE 1 . Méthodologie du projet NADIA-DEC.

Ainsi, le travail à effectuer se décompose en différentes étapes :

- création d'un éditeur informatique spécialisé pour le DEC ;
- création (en utilisant le système SUBLIM) d'un système de gestion des données lexicales. Les vérifications des différentes contraintes sur les données seront vérifiées à ce niveau ;
- réalisation d'un mécanisme d'import des données existantes, actuellement au format RTF (*Rich Text Format*) ;
- réalisation d'un module d'export vers différents formats utilisables informatiquement (SGML/TEI, format LISP...) et pour la publication papier (RTF, MIF...).

Cet article présente la structure informatique utilisée à la fois pour l'éditeur et pour le système central de gestion des données lexicales.

Le projet SUBLIM

L'étude détaillée des travaux menés dans le cadre de la construction de systèmes de bases lexicales multilingues ou monolingues informatisées nous a permis de faire les constatations suivantes :

- de nombreux projets existent, qui présentent des particularités intéressantes. Il est donc vain de proposer un nouveau système qui ne puisse prendre en charge les méthodes de ces divers projets ;
- les projets multilingues existants figent en général l'approche lexicale que l'on peut utiliser pour réaliser le multilinguisme. Ainsi, le projet Esprit Multilex (par exemple) ne permet l'utilisation que d'une approche bilingue et ne peut donc satisfaire les lexicologues tentés par des approches interlingues ;
- les projets existants offrent parfois au lexicologue la possibilité de personnaliser les structures linguistiques utilisées dans le lexique, mais de manière insuffisante. Ainsi, si, dans le projet MULTILEX, il est possible de définir ses

propres traits linguistiques, il est impossible de s'affranchir de l'utilisation d'une structure de traits typés pour coder les informations lexicales.

Partant de ces constatations, Sérasset (1994) propose un système informatique de gestion de bases lexicales multilingues qui réponde à ces différentes critiques. Ainsi, ce système peut satisfaire les lexicologues souhaitant coder des informations lexicales complexes utilisant des structures informatiques diverses (arbres, automates, graphes, structures de traits, ensembles...) dans le cadre de bases lexicales monolingues ou multilingues, basées sur les approches par dictionnaires de transfert ou par dictionnaires interlingues.

Pour atteindre cet objectif, le système SUBLIM demande au lexicographe de définir la structure linguistique du dictionnaire qu'il souhaite informatiser par deux langages ayant des objectifs différents.

- Dalex (Définition de l'Architecture LEXicale) permet de définir les dictionnaires mis en œuvre dans la base lexicale et leurs liens. Il permet, entre autres, de définir l'approche lexicale utilisée (interlingue ou par transfert).
- Darling (Définition de l'ARchitecture LINGuistique) permet, pour chaque dictionnaire, de définir les structures informatiques utilisées dans les entrées. Le lexicographe peut utiliser de nombreuses structures prédéfinies comme les arbres, graphes, automates, structures de traits...

Nous ne donnerons pas dans cet article plus de détails sur ces langages. Le lecteur en trouvera néanmoins des exemples d'utilisation dans la suite de cet article.

Le *Dictionnaire explicatif et combinatoire du français contemporain*

Une unité de ce dictionnaire (lexème) est un sens de mot ou de locution. Cette unité lexicale est associée à une unité morphologique, à une définition, à d'éventuelles connotations, à un régime, à des exemples, et à des fonctions lexico-sémantiques. Nous lui affectons de plus un numéro de sens qui l'identifie parmi les différents sens d'une entrée.

Un lexème peut aisément être codé comme une structure de traits :

```
(def-linguistic-class lexème
  (feature-structure
    ((numéro           numéro)
     (UMorph           UMorph)
     (définition       définition)
     (connotations     connotations)
     (régime           régime)
     (exemples         exemples)
     (lexico-sem-fns   lex-sem-fns))
))
```

Une unité morphologique comprend une forme graphique et des informations morphologiques. Elle peut être reliée à plusieurs lexèmes qui lui sont associés de manière arborescente :

CŒUR, nom, masc.

I.1a.	Organe principal de la circulation sanguine d'une personne... [*le cœur de* ***Jean***]
1b.	Organe principal de la circulation sanguine d'un animal... [*le cœur de lion*]
2.	Produit alimentaire ... [*le cœur de veau*]
3.	Partie de la poitrine d'une personne ... [*Il a serré son fils sur son cœur*]
4a.	Organe imaginaire des sentiments ... [*Le cœur espère toujours*]
4b.	Organe imaginaire de l'intuition ... [*Son cœur le lui dit*]
5a.	... propriété de la personnalité ... [*un cœur de glace*]
5b.	Personne possédant le cœur I.5a [*Vous devez la vie à un noble cœur, à un homme vaillant*]
II.1a.	Partie principale d'une unité fonctionnelle... [*le cœur du bateau*]
1b.	Élément principal [*le cœur du problème*]
2a.	Partie centrale d'un espace... [*le cœur du royaume*]
3.	Objet... ayant la forme du cœur I.1a [*un cœur en papier*]
4.	Une des quatre couleurs 2 des cartes à jouer... [*l'as de cœur*]
III.	Organe imaginaire des nausées ... [*Cette senteur lui tournait le cœur*]

Aussi, nous définirons une unité morphologique comme un arbre portant des informations morphologiques à la racine et des lexèmes sur les feuilles.

```
(def-linguistic-class UMorph
  (tree :root     Morphological-information
        :leaves lexème))
```

L'information morphologique associée à la racine de cet arbre ne comporte qu'une graphie, une catégorie, un genre et un nombre.

```
(def-linguistic-class Morphological-information
  (feature-structure
     ((graph       string)
      (catégorie   cat)
      (genre       gnr)
      (nombre      nbr))))

(def-linguistic-class cat
  (one-of (nom verbe adjectif adverbe... )))
(def-linguistic-class gnr
  (one-of (masculin féminin)))
(def-linguistic-class nbr
  (one-of (singulier pluriel)))
```

Une définition du DEC n'est pas une simple chaîne de caractères :

I.1a. *Cœur de X* = Organe principal de la circulation sanguine d'une personne X qui se trouve dans la partie centrale du corps II.1d de X et qu'on représente symboliquement comme ayant la forme ♡.

Mis à part le fait que l'on y trouve une image, on peut remarquer qu'elle se compose de deux parties principales. La première (indiquée en italiques) présente un usage du lexème dans une locution où les différents arguments du prédicat représenté sont indiqués sous forme de variable. La seconde est une explicitation du sens du lexème, elle réutilise les variables de la première partie. On remarque aussi qu'elle fait référence à des lexèmes définis par ailleurs dans le dictionnaire (corps II.1d).

Nous simplifierons cette structure en la décomposant simplement en deux chaînes de caractères, l'une contenant la forme du prédicat, l'autre contenant sa définition :

```
(def-linguistic-class définition
  (feature-structure
    ((prédicat  string)
     (explicite string))))
```

Après cette partie de définition, on trouve éventuellement une partie consacrée aux connotations :

Connotations

1) Cœur I.1a est le siège des sentiments [voir CŒUR I.4a].

2) Cœur I.1a est le siège de l'intuition [voir CŒUR I.4b].

3) Cœur I.1a qui bat 1 représente la vie [voir les phrasèmes correspondants dans CŒUR I.1a].

Cette partie se présente comme une liste de connotations. Chacune est donnée sous forme de chaîne de caractères faisant référence à au moins un lexème. Il est donc intéressant, dans une version informatique de ce dictionnaire, de conserver à la fois la connotation sous forme de chaîne de caractères et sous forme d'un ensemble de liens vers d'autres lexèmes :

```
(def-linguistic-class connotations
  (set-of connotation))
(def-linguistic-class connotation
  (feature-structure
    ((texte    string)
     (réfère-à (set-of ((link :target lexème)))))))
```

À la suite de ces éventuelles connotations, on trouve le régime du prédicat. Ce régime donne les informations sur les différentes réalisations syntaxiques des arguments du prédicat. Le régime est donné sous forme de tableau dont les colonnes correspondent aux arguments et les lignes aux différentes réalisations. Certaines combinaisons ainsi établies étant non valides, on en reprend ensuite la liste, en indiquant leur impossibilité. On reprend aussi un certain nombre de ces combinaisons pour en donner des exemples (l'exemple suivant est tiré de l'entrée « enseigner 1 ») :

1. *X enseigne Y à Z* = X, censé avoir la qualification professionnelle dans le domaine Y, cause que Z apprenne III.1b Y en transmettant, méthodiquement et dans un cadre officiel, à Z des connaissances (portant sur) Y ou des techniques (portant sur) Y [≅ $CausConv_{21}$(*apprendre III.1b*)].

Régime

1 = X	2 = Y	3 = Z
1. N	1. N 2. *à* V_{inf}	1. *à* N 2. **rare** N

1) $C_{2\,2}$ sans $C_{3\,1}$
2) $C_2 + C_{3\,2}$: *impossible*

$C_1 + C_2$: *Pierre enseigne la grammaire <la couture >/ à faire cela*

$C_2 + C_2 + C_3$: *Pierre enseigne la grammaire à ses élèves*

La structure correspondant à cette partie est beaucoup plus compliquée que celle des parties précédentes. En effet, cette présentation n'est que le reflet, imprimable, d'une structure complexe où l'on retrouve l'ensemble des combinaisons possibles de réalisations d'arguments. On peut donc représenter cette partie de deux manières :

- en restant proche de sa forme papier. On a alors un tableau et une liste des combinaisons impossibles ;
- en représentant cette structure de manière plus abstraite. On peut ainsi la représenter par un automate dont chaque chemin forme une combinaison valide.

Si l'on choisit la seconde solution, le régime donné en exemple sera représenté par l'automate donné en figure 2.

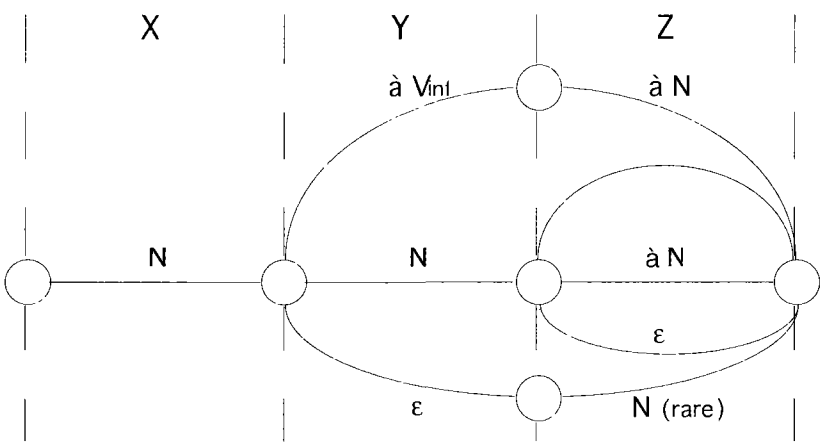


FIGURE 2 : Régime d'enseigner 1, sous forme d'automate.

Pour exprimer cette solution, nous utiliserons la structure logique d'automate prédéfinie dans le système SUBLIM où l'on peut définir des contraintes sur les classes

acceptables en décoration des différents éléments de l'automate (arcs, nœuds ou états, états initiaux, états finals).

Ainsi, la structure du régime s'exprimera sous forme d'une structure de traits regroupant l'automate des régimes, l'ordre d'apparition des arguments et l'ensemble des exemples :

```
(def-linguistic-class régime
  (feature-structure
    ((automate       automate-régime)
     (argument-order (list-of (string)))
     (exemples       exemples-régime))))
(def-linguistic-class automate-régime
  (automaton :arcs réalisation-argument))
(def-linguistic-class exemples-régime
  (set-of ((feature-structure
            ((réalisations (list-of (string)))
             (exemple      string))))))
```

La partie la plus importante de ce dictionnaire réside dans l'ensemble des fonctions lexicales du lexème. Leur meilleur définition est donnée, en première partie du DEC, par l'auteur, Igor Mel'čuk :

> Les fonctions lexicales (FL) présentent l'ensemble de la cooccurrence lexicale restreinte intéressant le lexème considéré. Elles constituent une innovation lexicographique qui permet de décrire d'une façon systématique un vaste ensemble de locutions plus ou moins figées qui ne sont quand même pas des expressions idiomatiques *stricto sensu*. Il s'agit, par exemple, des locutions comme une FERME intention, une résistance ACHARNÉE, un argument DE POIDS, un bruit INFERNAL, un désir ARDENT, une envie FOLLE, une règle STRICTE, une vérité INCONTESTABLE, où des adjectifs bien spécifiques doivent être employés avec les différents noms pour exprimer la même idée d'intensification. Comme autre exemple de locution de ce type, on peut citer les expressions DONNER une leçon, FAIRE un pas, COMMETTRE un crime, PORTER une accusation, etc., où des verbes sémantiquement vides (ou presque vides) différents doivent être choisis en fonction du nom d'action pour lier le nom d'agent en tant que sujet grammatical au nom d'action en tant que complément d'objet direct.

L'écriture générale d'une FL est de la forme : **f**(X) = Y, où **f** est la FL, X est son argument (un lexème ou bien une locution), et Y est la valeur de la FL **f** pour cet argument, c'est-à-dire l'ensemble des expressions linguistiques qui peuvent exprimer le sens ou le rôle syntaxique donné (noté par **f**) auprès de X.

Comme ce dictionnaire est imprimé, les expressions linguistiques sont données sous une forme linéaire :

MÉPRIS, nom, masc.

I. Attitude émotionnelle défavorable... [*le mépris pour ce corrupteur*] [...]

Fonctions lexicales

$Caus_3Func_1$	: engendrer [ART ~ chez N] [*La familiarité engendre le mépris*]
$Caus_{(3)}Func_1$	: apprendre, inculquer [ART ~ à N] [*Jean inculque à ses étudiants le mépris de l'hypocrisie; Son attitude partiale envers ses employés apprend à ces derniers le mépris de leur chef*]
$Caus_{(2/3)}Func_1$	: inspirer [ART ~ à N] [*Cet événement inspire aux travailleurs le mépris de leur patron; L'argent inspirait à ce philosophe un tel mépris qu'il a donné son héritage à son frère; L'hypocrisie de Jean leur inspirait un profond mépris*]

Mais la structure interne de ces expressions linguistiques est un arbre syntaxique donnant la construction de ces expressions linguistiques et de l'argument X pour réaliser la fonction **f**.

Ainsi, la structure interne de $Caus_3Func_1$(Mépris I) est l'arbre donné ci-dessous :

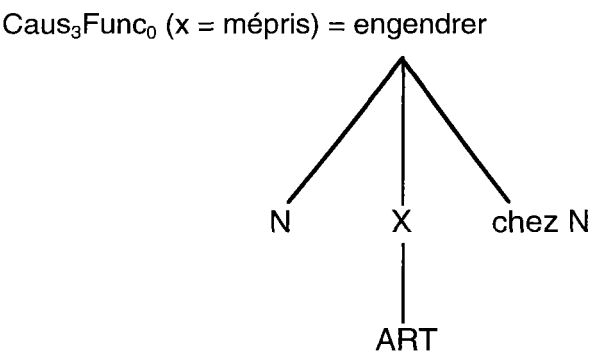


FIGURE 3 : Structure interne d'une expression linguistique, valeur de fonction lexicale.

Une fonction lexicale représente donc un lien entre un lexème et une expression linguistique complexe comportant d'autres lexèmes. Aussi, la valeur de ces fonctions lexicales peut être représentée comme un ensemble d'arbres dont certains nœuds sont des variables, et d'autres sont des lexèmes.

Il faut aussi représenter les fonctions lexicales. En effet, s'il y a un nombre limité de fonctions lexicales de base, on trouvera des fonctions composées dans les différents articles de dictionnaire.

Prenons un exemple : les fonctions **Oper**$_1$, **Oper**$_2$... ont pour valeur les verbes sémantiquement vides qui prennent le nom du premier, deuxième... actant comme sujet grammatical et C_0 (leur argument) comme complément d'objet principal :

Oper$_1$*(attention)* = *faire*
Oper$_2$*(attention)* = *attirer*
Oper$_1$*(conseil)* = *donner*
Oper$_2$*(conseil)* = *recevoir*
Oper$_1$*(aide)* = *prêter, accorder*
Oper$_2$*(aide)* = *recevoir*

La fonction **Caus** représente la notion : « faire en sorte que quelque chose ait lieu ». Elle s'emploie le plus souvent en combinaison avec d'autres fonctions. Ainsi, si **Oper**$_1$*(désespoir)* = *éprouver, ressentir, avoir*, **CausOper**$_1$*(désespoir)* représente « faire en sorte que quelqu'un éprouve du désespoir ». Donc **CausOper**$_1$*(désespoir)* = *pousser, réduire* [*qqn au désespoir*], *jeter* [*qqn dans le désespoir*], *frapper* [*qqn de désespoir*].

Il n'est donc pas possible de représenter chaque fonction lexicale comme un attribut dans une structure, puisque la possibilité de composition entraîne toute une combinatoire des fonctions lexicales. Nous les représenterons donc par la structure logique de base *function* prédéfinie dans le système SUBLIM, où l'on peut contraindre les classes acceptables pour les différents éléments de fonction (nom, arguments, valeur). Ainsi, la structure correspondant aux fonctions lexicales peut s'exprimer comme suit :

```
(def-linguistic-class lex-sem-fns
  (set-of (lex-sem-fn)))
(def-linguistic-class lex-sem-fn
  (function   :label nom-FL
              :arguments FL-arg
              :value expression-linguistique))
```

Pour représenter la composition de fonctions, on peut autoriser l'utilisation d'une fonction lexicale en argument d'une fonction lexicale. Néanmoins, la valeur de la fonction intermédiaire (si elle existe) n'est pas nécessairement pertinente. Seuls les noms des fonctions composées sont porteurs d'information. Aussi, le plus simple est d'autoriser une valeur complexe comme nom de la fonction. Nous définirons donc un label de fonction comme une liste (ordonnée) de noms de fonctions de base.

```
(def-linguistic-class nom-FL
  (list-of (nom-FL-base)))
```

Le nom d'une fonction de base est donné par un identificateur de la fonction (une chaîne de caractères) et par un éventuel numéro de l'actant sur lequel elle opère :

```
(def-linguistic-class nom-FL-base
  (feature-structure
    ((fonction string)
     (actant   integer))))
```

L'argument de la fonction est un lexème. Le fait d'indiquer cet argument est re-

dondant puisque cette fonction est définie à l'intérieur de la structure du lexème (il sera néanmoins représenté afin de simplifier les mécanismes d'extraction d'information).

```
(def-linguistic-class FL-arg lexème)
```

L'expression linguistique valeur de la fonction est représentée sous forme d'arbre (comme nous l'avons indiqué plus haut). Les nœuds de cet arbre sont soit des lexèmes, soit des variables. Pour simplifier, nous noterons les variables comme des chaînes de caractères.

```
(def-linguistic-class expression-linguistique
  (tree :nodes (one-of (lexème string))))
```

La définition de la structure linguistique du DEC, même simplifiée, illustre parfaitement le besoin ressenti par les linguistes de pouvoir mélanger différentes structures logiques dans une seule et même structure linguistique. Le fait de proposer différentes structures logiques permet au linguiste de manipuler des concepts proches de ceux utilisés dans sa théorie. Cela permet de simplifier le travail du linguiste en lui permettant de rester à un niveau d'abstraction très utile lorsqu'il souhaite implémenter une théorie complexe.

Conclusion et perspectives

L'étape de formalisation de la structure du DEC est nécessaire pour son informatisation. On remarque que ce dictionnaire met en œuvre de nombreuses structures informatiques de base (arbres, listes, structures de traits, etc.) et nécessite donc un environnement permettant de gérer un tel ensemble de structures hétérogènes. Cette constatation nous conforte dans l'idée que les futurs systèmes de gestion de bases lexicales devront être à même de gérer de telles structures.

Le projet NADIA-DEC va permettre différentes études :

- multilinguisation du DEC : le DEC est un ensemble de dictionnaires monolingues, mais il se base sur une théorie qui s'applique à l'ensemble des langues étudiées jusqu'alors. Il semble donc souhaitable de mettre en correspondance les différents lexèmes des différentes langues disponibles, afin d'étudier les problèmes éventuels d'une multilinguisation du DEC ;
- vérification automatique de cohérence : pour la sortie de chaque volume du DEC, les auteurs doivent vérifier de nombreuses contraintes de cohérence sur les différents articles. Certaines de ces contraintes (comme la vérification de la numérotation des lexèmes, de certaines contraintes de formatage...) peuvent être prises en charge par le gestionnaire de bases lexicales. Nous étudierons les contraintes à vérifier et un environnement permettant de les spécifier et de les vérifier ;
- récupération des informations existantes : l'ensemble des informations déjà publiées n'est disponible que sous forme de fichiers de traitement de texte. Nous souhaitons étudier les problèmes d'import de ces entrées existantes, en les généralisant afin de définir un environnement d'import de différentes sources existantes dans une base lexicale.

En résumé, le projet NADIA-DEC ne vise pas seulement le développement d'une version informatique du DEC, mais permet aussi, de mettre en pratique, sur un cas réel et exigeant les idées de généricité développées au GETA dans le cadre du système SUBLIM. Il permettra de plus de développer d'autres aspects de recherche dans le domaine de la lexicographie (multilinguisation du DEC) et dans le domaine de l'informatique linguistique (développement d'environnements de manipulation de données linguistiques, etc.).

Pour terminer, je tiens à remercier l'AUPELF-UREF pour son soutien à cette action de recherche partagée entre le GETA-CLIPS (institut IMAG, Université Joseph Fourier - Grenoble I) et le GRESLET (Département de linguistique et de traduction, Université de Montréal).

La formalisation des collocations pour le traitement automatique du langage naturel : le modèle des fonctions lexicales syntagmatiques[1]

Agnès TUTIN

URA SILEX, Université de Lille III, Villeneuve d'Ascq, France

Introduction

Le modèle des Fonctions Lexicales, qui appartient au *Dictionnaire explicatif et combinatoire* (DEC), constitue un formalisme séduisant pour le traitement des collocations :

- Les Fonctions Lexicales (FLs désormais) font partie d'une théorie linguistique complète dont le DEC est un composant essentiel (Mel'čuk, 1981; Mel'čuk et Polguère, 1987). La théorie Sens-Texte apparaît bien adaptée au traitement automatique du langage et les FLs ont déjà été exploitées pour des applications dans le domaine de la traduction assistée par ordinateur (Heylen et Maxwell, 1994) et en génération de textes (Iordanskaja, Kim et Polguère, 1995).

- Le modèle des FLs fournit une formalisation à la fois sémantique et syntaxique : les collocations sont décrites par le biais d'une étiquette syntactico-sémantique, par exemple la FL **Magn** signifie ‹très›, ‹intensément› et elle produit des modifieurs (des adjectifs ou des adverbes selon la partie du discours du mot d'entrée).

- Le modèle est « génératif » : les FLs sont combinables (par exemple, **AntiMagn** ‹peu intensément› est construite à partir de **Anti** et de **Magn**), même si la syntaxe et la sémantique des règles de combinaison apparaissent parfois confuses.

- Le modèle a déjà été utilisé sur un très large échantillon de données : les trois volumes du *DECFC* (Mel'čuk *et al.*,1984, 1988, 1992).

1. Un grand merci à Alain Polguère qui a relu attentivement cette version et m'a bien entendu signalé quelques incohérences.

Dans cette étude, nous essayons d'approfondir le formalisme des Fonctions Lexicales syntagmatiques, dans la perspective d'adapter ce modèle à des applications de traitement automatique en français, complétant ainsi le défrichage entamé par Alonso Ramos (1993). Pour cette tâche, nous nous basons sur les trois volumes du *Dictionnaire explicatif et combinatoire du français contemporain* (Mel'čuk *et al.*, 1984, 1988, 1992), le DEC français, qui constituent déjà une riche base de recherche. Notre étude a pour objet de déterminer dans quelle mesure le formalisme actuel doit être remanié et approfondi et elle vise à proposer des modes de description adaptés pour le TALN.

1. Les fonctions lexicales syntagmatiques du DEC

1.1. Une brève définition de la notion de collocation[2]

Conformément aux perspectives du DEC, nous définissons une collocation comme un **syntagme semi-figé** ou semi-phrasème (par exemple, *gros fumeur, prendre un bain,...*) dans lequel le sens d'une unité lexicale A (la base) est transparent, alors que le sens de l'autre unité lexicale B (le collocatif) est déterminé par la présence de A. Les collocations s'opposent, d'une part, aux phrasèmes (par exemple, *prendre le taureau par les cornes*) pour lesquelles le sens du tout n'est pas déductible (ou difficilement) du sens des parties, et d'autre part, aux syntagmes standard (*un chanteur gros, un taureau méchant*) dont l'interprétation découle du sens des composants. En d'autres termes, la façon dont un mot est lexicalisé (le *collocatif*) dépend du mot auquel il est rattaché sémantiquement (la *base*). Ainsi, la lexie utilisée pour exprimer l'intensité dans le contexte de *fumeur* est *gros*. Dans cette approche, la collocation n'est pas définie par une fréquence statistique ; la récurrence de certains couples de mots est simplement une conséquence du phénomène sémantique de la cooccurrence restreinte[3].

1.2. Différents types de Fonctions Lexicales

Les collocations ont été formellement décrites d'un point de vue syntaxique et sémantique au moyen des FLs et plus exactement, par le biais des FLs syntagmatiques. Une FL est une fonction, au sens mathématique, qui associe un ensemble d'unités lexicales à une unité lexicale. Dans le cas des FLs syntagmatiques, le mot-clé peut être considéré comme la base, et la(les) valeur(s), le(s) collocatif(s). Les FLs les plus productives et les plus universelles ont été étiquetées et une cinquantaine de FLs standard a été proposée. Par exemple, la FL **Magn** (sens d'intensité) est appliquée à *fumeur* pour produire *gros*, alors que la FL $\mathbf{Oper_1}$ est appliquée à *bain* pour produire *prendre*. Cela est codé de la façon suivante :

Magn(*fumeur*) = *gros*
$\mathbf{Oper_1}$(*bain*) = *prendre*

2 Pour une discussion approfondie sur la notion de collocation, voir Heid (1994).
3. Cf. Smadja et McKeown (1991).

Toutes les FLs ne sont pas utilisées pour décrire des relations de collocation. En utilisant des critères de cooccurrence et des critères sémantiques, trois types de FLs peuvent être dégagés :

- **les FLs syntagmatiques** comme **Magn** ou $\mathbf{Oper_i}$ sont utilisées pour formaliser les collocations ;

- **les FLs paradigmatiques** sont utilisées pour associer une unité lexicale qui partage avec d'autres unités lexicales un composant de sens non trivial. La(es) valeur(s) et le mot-clé ne forment habituellement pas un syntagme. Ainsi, $\mathbf{S_{loc}}$ (nom typique pour le lieu) et $\mathbf{S_3}$ (nom typique pour le troisième actant) sont des relations paradigmatiques : $\mathbf{S_{loc}}$(*patiner*) = *patinoire* et $\mathbf{S_3}$(*conférence*) = *assistance, audience* ;

- **les FLs mixtes** partagent les caractéristiques des deux types précédents[4]. Elles sont utilisées pour associer un mot-clé avec un ensemble de lexèmes qui partagent une composante sémantique, mais la valeur et le mot-clé peuvent aussi se combiner pour former un syntagme, par exemple **Mult**(*bateau*) = *flotte* ('nom standard pour un ensemble de') ou **Cap**(*lycée*) = *censeur* ('la tête de').

1.3. Les combinaisons de FLs

De plus, les FLs peuvent être combinées. Trois types de combinaisons peuvent être réalisées.

- **Les FLs composées** sont combinées de façon similaire aux compositions de fonctions mathématiques. La valeur finale est produite par le biais des valeurs intermédiaires des fonctions intermédiaires. Par exemple, pour calculer la valeur de $\mathbf{S_0}$(**Gener**(*étuver*)), on cherchera d'abord la valeur de **Gener**(*étuver*) = *cuire*, puis de $\mathbf{S_0}$(*cuire*) = *cuisson*. La composition est bien entendu non commutative[5].

  Pour l'étude des collocations, les FLs composées ne paraissent pas pertinentes[6]. Tout d'abord, elles sont avant tout utilisées pour décrire des fonctions paradigmatiques. Par ailleurs, la question de l'encodage des FLs composées se résout d'elle-même, puisque les compositions peuvent être calculées de façon régulière à partir des FLs simples.

- **Les configurations de FLs** ont un mot-clé en commun. Les FLs sont liées par un signe «+ », qui est ici commutatif.

  Par exemple, **Magn + Func**$_0$(*tempête*) = *faire rage* signifie que *faire rage* est la valeur quand *tempête* est le sujet grammatical et que la tempête est intense.

---

4. Plus exactement, les FLs considérées comme mixtes sont en fait des FLs pour lesquelles dans le cas de certaines valeurs (*essaim* pour **Mult**(*abeille*), par exemple), on peut avoir une cooccurrence facultative du mot-clé et de la valeur. Cela n'est pas le cas pour l'ensemble des valeurs produites pour une fonction mixte donnée, mais pour un sous-ensemble restreint. De ce fait, la notion de fonction mixte n'est pas opérationnelle car elle conduit à des interprétations erronées dans le décodage. Dans un contexte donné, il faut en effet pouvoir décider de façon claire si l'interprétation de la FL est syntagmatique ou paradigmatique.

5. **Gener**($\mathbf{S_0}$(*étuver*)) ne produit aucune valeur parce que $\mathbf{S_0}$(*étuver*) ne produit lui-même pas de valeur.

6. Notre intention n'est pas de dire qu'elles ne sont d'aucune utilité pour le TALN. Au contraire, nous sommes persuadée qu'elles peuvent être utilisées avec profit pour générer des expressions de référence appropriées (Alonso Ramos, Tutin et Lapalme, 1995).

Les configurations devraient aussi être prises en compte pour les relations de collocations, mais elles apparaissent relativement rares, et elles ne seront pas davantage examinées ici.

- **Les fonctions complexes** sont des combinaisons qui ne peuvent pas être décomposées. La valeur n'est pas obtenue par une combinaison régulière, à la différence des compositions. Par exemple, la valeur de **AntiMagn**(*prix*) [= *modique*] ne peut pas être déduite de **Magn**(*prix*) [= *élevé*]. *Modique* en effet n'est pas dans l'absolu un antonyme de *élevé*, mais seulement lorsqu'il apparaît en cooccurrence avec *prix*. Quand une combinaison est syntagmatique, il s'agit généralement d'une FL complexe, alors que les fonctions complexes peuvent elles aussi être paradigmatiques.

  Une fonction complexe n'est donc pas une composition : en fait, une FL complexe est une FL résultant elle-même d'une combinaison. Dans l'exemple de **AntiMagn, Magn** peut être considérée comme une fonction, alors que **Anti** peut être considérée comme une « fonctionnelle », cette dernière étant définie comme une fonction de second ordre s'appliquant elle-même à des fonctions.

  Nous pouvons ainsi définir une **fonctionnelle lexicale** comme une fonction qui s'applique à une fonction simple ou complexe pour produire une fonction complexe (le mécanisme est récursif). La FL complexe s'applique elle-même à une unité lexicale pour produire des unités lexicales.

  De nombreuses fonctions sont à la fois des FLs et des fonctionnelles lexicales, mais lorsque ce cas se produit, elles ne partagent pas nécessairement exactement les mêmes caractéristiques. Par exemple, **Magn** est une FL syntagmatique simple (**Magn**(*pluie*) = *torrentielle*) et est aussi une fonctionnelle utilisée pour produire des FLs complexes où elle exprime l'intensification de la fonction sur laquelle elle porte (**Magn** signifie 'très, intensément', alors que **MagnManif** signifie 'se manifester avec intensité' et doit être considéré comme un tout[7]).

Pour l'implémentation des collocations en TALN, la sémantique et la syntaxe des fonctions complexes doivent être définies de façon détaillée. À l'heure actuelle, un certain flou règne chez les rédacteurs du DEC, dépositaires cependant d'une expertise précieuse en matière de FLs. La méconnaissance des heuristiques employées et des règles de combinaison constitue une entrave à la formalisation et l'implémentation des FLs en TALN.

Une première étape dans cette direction serait d'établir une description détaillée des FLs simples et des fonctionnelles. Cette tâche devrait permettre dans un second temps d'élaborer une grammaire formelle des FLs complexes.

2. Une description détaillée des FLs syntagmatiques : les FLs simples et les fonctionnelles utilisées pour construire les FLs syntagmatiques

2.1. Les FLs syntagmatiques simples

2.1.1. Un inventaire des FLs syntagmatiques

Tout d'abord, un inventaire des FLs syntagmatiques doit être constitué, tâche qui

7. On pourra noter que la nature syntagmatique de la FL simple **Magn** n'influe pas ici dans la combinaison.

peut être effectuée à partir des trois volumes du DECFC. Ce travail, qui peut paraître absolument trivial, s'est en fait révélé plus complexe que prévu.

La nature paradigmatique ou syntagmatique de la FL peut habituellement être déduite des définitions données dans le DEC (par exemple, le DEC 2 : 91-94) et des exemples. Ainsi, **Oper**$_i$ et **Magn** sont clairement définies comme syntagmatiques :

- « **Oper**$_0$, **Oper**$_1$, **Oper**$_2$ ... : verbe sémantiquement vide qui prend le pronom personnel (*il*) ou le nom du premier, deuxième,... actant de la situation C_0 comme son sujet grammatical (SG), et le mot-clé C_0 comme son complément d'objet (CO) principal.

  Oper$_0$(*vent 1.1*) = *faire*
  Oper$_1$(*attention*) = *faire*
  Oper$_2$(*attention*) = *attirer*
  ... » (DEC 2 : 93)

- « **Magn**: ‹très›, ‹intense/intensément›, ‹à un degré élevé›.

  Magn(*mémoire 1.1*) = *prodigieuse, excellente, étonnante, d'éléphant*
  Magn(*remercier*) = *vivement, chaleureusement, de tout cœur*
  ... » (DEC 2 : 92)

Néanmoins, pour un certain nombre de FLs, le statut paradigmatique/syntagmatique demeure peu clair. Par exemple, il est difficile de décider d'après les exemples relevés si **Mult** (‹un ensemble régulier de›) est syntagmatique ou paradigmatique. En fait, nous avons relevé trois cas avec cette FL.

1. La valeur ne peut pas être utilisée sans le mot-clé (en conservant le sens de **Mult**) : **Mult** est ici clairement syntagmatique :

   Mult(*brebis*) = *troupeau*.

   Sans le mot-clé, la valeur a un sens beaucoup plus générique que celui qui est encodé par **Mult**.

2. La valeur ne peut pas apparaître en cooccurrence avec le mot-clé : **Mult** est clairement paradigmatique.

   Mult(*bleu*^{2}) = *bleusaille*.

   Dans ce cas, l'association avec le mot-clé est agrammaticale : **une bleusaille de bleus*.

3. La valeur peut apparaître soit en cooccurrence avec le mot-clé, soit sans celui-ci : **Mult** est une FL mixte.

   Mult(*abeille*) = *essaim*;

   essaim est ici un synonyme plus large pour *essaim d'abeilles*[8].

8. *Essaim* peut être considéré comme ayant pour prototype : *essaim d'abeilles*. Ce n'est pas du tout le cas avec *troupeau* et *troupeau de brebis*.

À l'heure actuelle, la codification pour **Mult** n'est donc pas cohérente, puisque **Mult** encode à la fois des relations paradigmatiques et/ou syntagmatiques. Nous proposons de considérer cette fonction comme une fonction syntagmatique. Nous la noterons de la façon suivante, en utilisant le signe de fusion[9] lorsque la FL est utilisée de façon paradigmatique :

Mult(*brebis*) = *troupeau*
Mult(*bleu*^{2}) = // *bleusaille*
Mult(*abeille*) = *essaim*; // *essaim*

La même ambiguïté apparaît dans le DEC avec quelques FLs comme $\mathbf{A_0}$, $\mathbf{A_1}$, $\mathbf{Adv_i}$, **Gener, Sing.**

2.1.2. Paramètres descriptifs

Pour décrire formellement les FLs syntagmatiques, nous proposons un ensemble de paramètres. Les valeurs obtenues pour ces paramètres varient selon la langue et les exemples fournis dans cette section sont tirés du français[10].

Une FL donnée peut s'appliquer à de nombreuses parties du discours[11]. Puisque la syntaxe de la fonction peut légèrement varier selon ce paramètre (la partie du discours de la valeur varie souvent en fonction de la partie du discours du mot-clé), nous pensons que les FLs doivent être définies selon la partie du discours de la base.

Les paramètres descriptifs engloberont :

- **La relation Syntaxique Profonde**[12] **(RSyntP) apparaissant entre la base et le collocatif** : s'agit-il d'une relation attributive ou actantielle ? Par exemple, **Magn** est une relation attributive (un modifieur), alors que **Mult** est une relation actantielle (la valeur prend la base comme complément de nom).

- Le type de **réalisation de surface pour la valeur** : les parties du discours et le type de constituant généré par la FL. Ces valeurs peuvent être : des lexèmes (a), des phrasèmes (b) ou des constituants libres (c).

  (a) **Magn**(*pluie*) = *torrentielle, diluvienne*
  (b) **Magn**(*chanter*) = *à tue-tête*
  (c) **Magn**(*manger*) = *comme un ogre*

  Il est crucial de distinguer les lexies (lexèmes et phrasèmes) des constituants « libres », car ces derniers ne possèdent pas d'entrée propre dans le DEC[13].

9. Une fonction fusionnée est une fonction paradigmatique dont la valeur est équivalente à l'ensemble mot-clé + valeur. La fusion est encodée avec le signe « // »
10. La description proposée ici complète celle qui a déjà été proposée dans Alonso Ramos (1993) et Alonso Ramos et Tutin (1995)
11. Par exemple, **Magn** s'applique à des noms, des verbes, des adjectifs et des adverbes
12. Ce niveau de description est propre à la théorie Sens-Texte (voir par exemple Mel'čuk, 1981).
13. Strictement, comme cela est souligné par Alonso Ramos et Mantha dans ce volume, toutes les lexies correspondant à des collocatifs ne doivent pas faire l'objet d'entrée « sémantique » dans le DEC. Ainsi, *noir* dans *café noir* ne peut pas être considéré comme un collocatif nécessitant une entrée propre : ici, le collocatif *noir* emprunte sa forme et certains aspects sémantiques à l'adjectif *noir*. Tout collocatif doit néanmoins faire l'objet d'une entrée « morphologique » qui indiquera le mode de flexion, les caractéristiques distributionnelles, etc.

- Le type d'***indices*** et d'**exposants** que chaque fonction peut avoir, et dans quelle mesure ils sont facultatifs ou obligatoires.

  Parmi les indices, on trouve des indices actantiels (sémantiques ou syntaxiques) qui indiquent l'actant concerné (par exemple, $\mathbf{Oper}_1$, $\mathbf{Magn}_1$), des indices ensemblistes ($\mathbf{Syn}_{\supset}$, $\mathbf{Ver}_{\cap}$, ...), des indices sémantiques ($\mathbf{Magn}_{\text{'conséquence'}}$).

  Parmi les exposants, on relève : des exposants sémantiques ($\mathbf{Magn}^{\text{temp}}$, $\mathbf{Magn}^{\text{quant}}$), et des exposants d'intensité ($\mathbf{Real}^{\text{II}}{}_3$, $\mathbf{Real}^{\text{I}}{}_3$).

Dans le tableau ci-dessous, quelques FLs portant sur les noms sont examinées selon ces paramètres :

FL	**Définition**	**Partie du discours du collocatif**	**Indices et exposants**	**RSyntP**	**Exemples**
Magn	'très', 'intense/ intensément/ à un degré élevé'	- adjectif ou locution adjectivale - syntagme prépositionnel	- indice sémantique facultatif - indice actantiel facultatif - exposant sémantique facultatif	Magn ←--ATTR--X	*attention soutenue feu d'enfer bruit à crever les tympans*
Epit	modifieur vide sémantiquement	- adjectif ou locution adjectivale - syntagme prépositionnel	Aucun	Epit ←--ATTR--X	*quenotte petite numéro appel*
Real	verbe qui signifie 'réaliser', qui prend le mot-clé comme premier complément et le nom du ième actant comme sujet grammatical	- verbe ou locution verbale	- indice actantiel obligatoire - exposant d'intensité facultatif	Real_i --II--→ X	$\text{Real}^{\text{I}}{}_3$ (*conseil I I*) *accepter* $\text{Real}^{\text{II}}{}_3$ (*conseil I I*) *suivre*

TABLEAU 1 · Description de quelques FLs syntagmatiques portant sur des noms

2.2. Les fonctionnelles

Comme on l'a déjà signalé, les fonctionnelles lexicales sont des fonctions qui s'appliquent à des fonctions simples ou complexes pour produire des fonctions complexes (le mécanisme est récursif). Les fonctionnelles peuvent avoir le même nom que quelques FLs simples paradigmatiques ou syntagmatiques, auxquelles elles empruntent leur sens, mais pas leurs caractéristiques de surface, alors que certaines fonctionnelles comme **Caus** n'ont pas de FL simple équivalente. Nous nous intéresserons seulement ici aux combinaisons engendrant des FLs complexes syntagmatiques.

Pour les fonctionnelles, les paramètres suivants devront être examinés.

- Le fait que la fonctionnelle est une **fonctionnelle régente** ou une **fonctionnelle modificatrice**.

  Dans une FL complexe, toutes les FLs n'ont pas le même statut. La **fonc-**

tionnelle régente (voir Alonso Ramos, 1993[14]) détermine habituellement :

1 - la partie du discours auquel la FL complexe s'applique ;
2 - la partie du discours que la FL complexe produit ;
3 - la nature syntagmatique/paradigmatique de la FL complexe selon le statut des FLs auxquelles elle s'applique.

Par exemple, dans la fonction complexe **AntiMagn**, **Anti** ne modifie pas nature de **Magn** qui reste un modifieur. Au contraire, la fonctionnelle **Caus** est une fonction régente parce qu'elle modifie la structure actantielle de la FL affectée.

- L'**ensemble des FLs simples et complexes** auxquelles les FLs peuvent s'appliquer. Cet ensemble de FLs peut lui-même être un ensemble de sous-ensembles. Par exemple, on utilisera le sous-ensemble des verbes supports (**Oper$_i$, Func$_i$, Labor$_{ij}$**).

- Les **types d'indices et d'exposants** que les fonctionnelles peuvent prendre. Leur usage peut ici être assez différent de l'usage que l'on repère sur les FLs simples. Par exemple, **Magn**, comme fonctionnelle, ne semble pas prendre d'indices actantiels.

- Le **type de fonction complexe** (nominale, verbale, adjectivale) produite par la fonctionnelle.

Dans le tableau suivant, nous décrivons quelques fonctionnelles à l'aide des paramètres proposés plus haut.

14. La notion de fonction régente est différente chez Alonso Ramos (1993) qui n'effectue pas de distinction entre fonction et fonctionnelle. Pour elle, la fonction régente est la fonction qui impose à la combinaison sa signification et/ou sa valeur syntaxique. Ainsi, dans **AntiMagn, Magn** est la fonction régente, dans **S$_0$Oper$_1$, S$_0$** est la fonction régente.

Notre objectif étant de produire, entre autres, des règles de réécriture, nous n'examinons que la fonction qui apparaît à gauche de la combinaison, le mécanisme de réécriture étant bien entendu récursif.

Fonctionnelle	Définition	Type de fonctionnelle	FLs sur lesquelles porte la fonctionnelle	Indices et exposants	Exemples
Incep	'Commencer à'	Fonctionnelle modificatrice qui n'affecte pas les actants de la valeur produite par la FL sur laquelle porte la fonctionnelle.	- Verbes supports : **Func$_i$**, **Oper$_i$**, **Labor$_{ij}$**. - Verbes de réalisation: **Fact$_i$**, **Real$_i$**, **Labreal$_{ij}$**. - Verbes d'expression : **Involv**, **Manif**, **Degrad**, **Son**, **Excess**. - Verbalisateur : **Pred** - FLs complexes verbales.	Aucun	**IncepOper$_1$** *(alphabétisation)* = *entreprendre* **IncepReal$_3$** *(école I.1a)* = *entrer* **IncepInvolv** *(fureur 3b)* = *se déchaîner* **IncepPredPlus** *(chagrin I 1)* = *s'accroître, augmenter, grandir* **IncepProxOper$_1$** *(objection 1)* = *concevoir*
S_0	Nominalisation	Fonctionnelle régente. Prend une FL verbale simple ou complexe en entrée.	- Verbes supports : **Func$_i$**, **Oper$_i$**, **Labor$_{ij}$**. - Verbes de réalisation: **Fact$_i$**, **Real$_i$**, **Labreal$_{ij}$**. - Verbes d'expression: **Involv**, **Manif**, **Degrad**, **Son**, **Excess**. - Verbalisateur : **Pred**. - FLs complexes verbales	Aucun	**S$_0$Oper$_2$** *(appel téléphonique)* = *réception* **S$_0$Real$_2$** *(numéro de téléphone)* = *composition* **S$_0$Son**(*tempête 1*) = *hurlement* **S$_0$PredPlus$_1$**(*paie 1*) = *augmentation, hausse, majoration*] **S$_0$Caus$_1$Func$_0$** *(scénario 1)* = *écriture, rédaction*

TABLEAU 2 : Paramètres descriptifs pour les fonctionnelles.

3. Une grammaire des FLs

Une fois que la description des FLs simples et des fonctionnelles est achevée, une grammaire des FLs complexes peut être élaborée. Le but de cette grammaire « formelle » est double.

- Permettre une vérification des combinaisons dans un éditeur électronique (voir la contribution d'Alain Polguère dans Mel'čuk, Clas et Polguère, 1995), de façon à faciliter le travail du rédacteur dans l'encodage des FLs, ce qui devrait lui permettre de se concentrer davantage sur les questions lexico-sémantiques que sur les codifications formelles.

- Permettre l'implémentation des FLs complexes dans un système de traitement automatique qui garantisse la cohérence des informations lexicales.

Les règles peuvent être facilement implémentées à l'aide d'une grammaire hors-contexte. Nous présentons ci-dessous un extrait incomplet d'une telle grammaire pour les FLs complexes. Les signes suivants sont utilisés : « , » pour la séquence, « | » pour un « ou » exclusif, « ? » pour un composant facultatif. Les règles apparaissent en gras, les commentaires en caractères romains.

ind_act --> "1" | "2" | "3".; un indice actantiel est "1", "2" ou "3".
exp_int --> "I" | "II" | "III".; un exposant est "I", "II" ou "III".
verbe_supp --> (("Oper" | "Func"), ind_act) | ("Labor", ind_act, ind_act).; un verbe support est soit Oper, soit Func suivis d'un indice actantiel, soit Labor suivi de deux indices actantiels.
verbe_real --> (("Real" | "Fact"), exp_int?, ind_act) | ("Labreal", exp_int?, ind_act, ind_act). ; un verbe de "réalisation" est soit Real, soit Fact éventuellement suivi d'un exposant de réalisation et d'un indice actantiel obligatoire, soit Labreal suivi éventuellement d'un exposant de réalisation et obligatoirement de deux indices actantiels.
fl_phasale --> "Incep" | "Fin" | "Cont".; les fonctionnelles exprimant une phase sont soit "Incep", soit "Fin", soit "Cont".
fls_verbale --> fl_phasale?, (verbe_supp | verbe_real). ; une FL syntagmatique verbale (c-à-d, qui produit un verbe) est un verbe_supp ou un verbe_real éventuellement précédés d'une fonctionnelle phasale.
fls_nominale --> s0, fls_verbale.; une FL syntagmatique verbale peut être une fonctionnelle nominale suivie d'une FL syntagmatique verbale.

FIGURE 1 : Un extrait de la grammaire formelle des FLs.

Selon cette grammaire « jouet », $\mathbf{S_0IncepReal^{II}_3}$ serait analysée comme une fonction syntagmatique nominale, alors que $\mathbf{Incep_2S_0Real^{I}}$ serait considérée comme agrammaticale. Par ailleurs, comme la grammaire ne peut être seulement construite à partir des données observées dans les trois volumes du DEC, la syntaxe et la sémantique des FLs devront être bien maîtrisées.

4. Les propriétés syntaxiques des collocations

Pour le TALN, l'information syntaxique et sémantique concernant les collocations devra être extrêmement détaillée. Bien que le DEC semble très cohérent et très formalisé pour le lecteur humain (au risque d'être parfois, avouons-le, d'une lecture rebutante), l'information syntaxique concernant les collocations n'apparaît pas suffisamment détaillée.

4.1. Le régime des collocations[15]

Le régime des lexies apparaissant comme bases dans les collocations est très exhaustivement détaillé dans le DEC. Tel n'est cependant pas le cas pour la collocation elle-même dont la structure actantielle est assez superficiellement codée. Or, la structure actantielle de la collocation n'est pas systématiquement déductible du régime de la base. Ainsi, les collocations verbales incluant des noms, en particulier les constructions à verbe support, ne prennent pas nécessairement les mêmes actants que le nom qui est à la base de cette construction :

15 Le régime est la structure actantielle des prédicats.

- de nouveaux actants peuvent apparaître, d'autres disparaître ;
- les actants peuvent être réalisés en surface pas des prépositions différentes.

Par exemple, dans le DEC 1, l'entrée ENSEIGNEMENT **1a** (p. 93) contient les collocations suivantes:

Oper$_1^{\text{actual}}$: donner, dispenser, **litt** prodiguer
Oper$_1^{\text{usual}}$: être [dans l'~]

Le régime de ENSEIGNEMENT **1a** comporte trois actants : l'enseignant (actant 1), la matière (actant 2), les étudiants ou élèves (actant 3). Les valeurs pour **Oper**$_1^{\text{actual}}$ acceptent les trois actants, alors que **Oper**$_1^{\text{usual}}$ n'accepte que le premier :

Léa dispense un enseignement de sémantique aux étudiants de troisième année.
*Léa est dans l'enseignement de la sémantique aux étudiants de troisième année.

Pour qu'un système de génération de textes puisse utiliser la collocation *être dans l'enseignement*, le lexique devrait « savoir » que les actants 2 et 3 ne peuvent pas être exprimés[16].

De plus, les actants peuvent dépendre syntaxiquement plus du nom prédicatif que du verbe. Prenons l'exemple de la collocation *donner un cours*. La lexie COURS a elle-même trois actants : le cours de quelqu'un (actant 1) sur quelque chose (actant 2) à quelqu'un (actant 3). La collocation *donner un cours* a elle aussi trois actants (*quelqu'un* (1) *donne un cours sur quelque chose* (2) *à quelqu'un* (3)), mais tous ne semblent pas dépendre du nom (Alonso Ramos, 1993). Alors que le premier et le troisième actant dépendent du verbe, le second actant semble lié au nom *cours* comme on peut le schématiser dans la représentation syntaxique profonde (simplifiée) de *Lulu a donné un cours de mathématiques à Léa.*

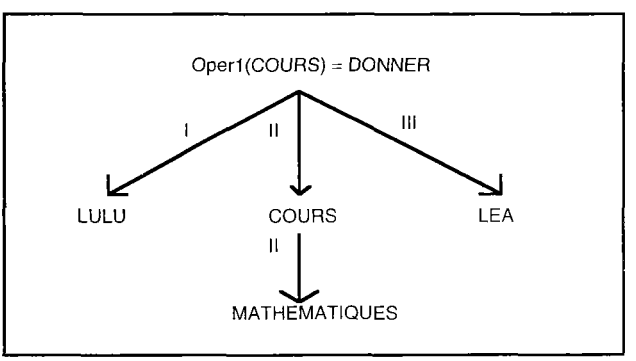


FIGURE 2 : Représentation syntaxique profonde (simplifiée) de *Lulu donne un cours de mathématiques à Léa*

16. Il serait faux de dire que cette information n'apparaît jamais dans le DEC, mais il est vrai qu'elle n'apparaît pas systématiquement. À l'article d'ENTHOUSIASME **1**, le régime comporte deux actants (l'enthousiasme de quelqu'un (1) pour quelque chose (2)). Pour la fonction lexicale **Gener**, il est indiqué que le premier actant ne se réalise pas en surface. On a ainsi :
Gener : sentiment ld'~] | $C_1 = \Lambda$

Le « déplacement » des actants 1 et 3 sur le verbe, alors que l'actant 2 reste attaché au nom semble avoir une incidence sur le comportement syntaxique de la collocation dans les transformations, comme on peut le remarquer lorsque la phrase est passivée :

Lulu a donné un cours de mathématiques à Léa.
Un cours de mathématiques a été donné à Lulu par Léa.
* Un cours a été donné de mathématiques par Léa.

Ces exemples montrent que le régime des collocations doit être extrêmement détaillé si l'on veut pouvoir manipuler ces dernières dans un véritable système informatique. Bien entendu, il apparaît possible d'effectuer quelques généralisations pour le cas des verbes opérateurs. Quelques heuristiques peuvent être utilisées comme valeurs par défaut, comme celles qui ont été proposées par Alonso Ramos (1993) pour les collocations mettant en jeu **Oper$_1$** : dans les collocations à trois actants comme *donner un cours*, le premier et le troisième actants dépendent généralement du verbe, alors que le second actant dépend du nom prédicatif.

4.2. Propriétés distributionnelles et transformationnelles des collocations

Si les collocations doivent être utilisées dans de véritables textes, elles doivent pouvoir être manipulées dans toutes sortes de contextes. Les collocations ne sont pas figées syntaxiquement dans la plupart des cas et la base et le collocatif peuvent apparaître de façon non contiguë. Par exemple, dans de nombreuses collocations de type adjectif-nom, l'adjectif peut être gradué ou apparaître en position d'attribut. On trouve ainsi :

Léa a les cheveux très blonds.
Les cheveux de Léa sont très blonds.

Ce n'est bien entendu pas toujours le cas pour les collocations nom-adjectif, comme dans l'exemple suivant où *rouge* en position d'attribut n'a pas le sens qu'il a dans *vin rouge* (l'interprétation non collocationnelle est notée par un #) :

C'est un vin très rouge.
Ce vin est rouge.

À côté de cela, il est indispensable de noter dans le lexique la position de l'adjectif par rapport au nom. *Gros* est un collocatif devant *fumeur*, mais pas en position postnominale.

Il est par ailleurs intéressant de noter que certains collocatifs productifs (c'est-à-dire, apparaissant dans de nombreuses collocations) comme *gros*, semblent imposer leurs propriétés syntaxiques à la collocation. Ainsi, *gros* quand il est un collocatif d'intensité est préposé au nom, graduable, et n'apparaît pas en position d'attribut :

Un gros fumeur.
Une grosse déprime.
Une grosse envie.

Un fumeur gros.
* Une déprime grosse.
* Une envie grosse.

Un très gros fumeur.
Une très grosse déprime.
Une très grosse envie.

Le fumeur est gros.
? Cette déprime est grosse.
* Cette envie est grosse.

Nous faisons ainsi l'hypothèse que par défaut, les propriétés distributionnelles de la collocation sont héritées du collocatif. Par souci d'économie, il apparaît ainsi plus pertinent de coder les propriétés propres au collocatif, non sur l'entrée de la base, comme cela est effectué actuellement dans le DEC (cf. 4.3.), mais sur l'entrée du collocatif.

Les propriétés « transformationnelles »[17] doivent aussi être codées dans le lexique pour les collocations verbales. Comme pour les propriétés distributionnelles, ces propriétés semblent largement dépendre du collocatif. Par exemple, le verbe support *avoir* n'accepte jamais la transformation passive, alors que tel n'est pas systématiquement le cas pour les verbes support *donner* ou *faire* :

Lulu a {faim, une grosse envie de chocolat}.
* Une {faim, grosse envie de chocolat} est eue par Lulu.

Léa a fait la vaisselle.
La vaisselle a été faite par Léa.

Léa a donné un cours.
Un cours a été donné par Léa.

Pour les collocations verbales, un nombre important de propriétés doit être codé. Parmi celles-ci, hormis la passivation, citons : la passivation réduite (a), la transformation relative (b), la transformation en *se*-passif (c), la transformation impersonnelle passive (d), la pronominalisation (e) :

(a) Le cours donné par Lulu à Léa.
(b) Le cours que Lulu a donné à Léa.
(c) Le cours se donne à l'Université de Montréal.
(d) Il a été donné le même cours à l'Université McGill.
(e) Lulu le donne à l'Université de Montréal, ce cours.

De nombreuses propriétés syntaxiques, qu'elles soient distributionnelles ou transformationnelles, semblent donc davantage héritées des collocatifs plutôt que des bases. Par souci d'économie du codage dans le lexique, il nous semble donc pertinent

17 Il nous paraît commode d'utiliser ce terme démodé pour mentionner les différents types de constructions syntaxiques quasi-synonymes que l'on peut avoir pour un verbe donné.

d'encoder ces propriétés plutôt dans l'entrée du collocatif, même si ce collocatif est sémantiquement vide, comme cela est souvent le cas pour les verbes supports. Cependant, dans le DEC, les propriétés syntaxiques, quand elles sont signalées, apparaissent codées dans l'entrée des bases.

4.3. Une proposition de codage des propriétés syntaxiques des collocations

À l'heure actuelle, le codage des propriétés syntaxiques des collocations dans le DEC s'effectue sur les articles des bases. Ainsi, la propriété [+ préposé] des collocatifs est signalée sur l'article des bases. Par exemple, à l'entrée de CHAGRIN[1]I.1, on relève :

Magn : grand, gros, **litt** noir, vif | prépos ...

Pour les collocations verbe-nom, dans la plupart des cas, l'information syntaxique est très réduite. Ainsi, dans le même article, on relève pour la FL **Oper$_1$** :

Oper$_1$: avoir, éprouver, ressentir [du ~ | C. n'a pas de dépendant]; **litt** connaître [ART ~]

Ici, il n'est pas indiqué que le passif n'est pas autorisé pour le verbe *avoir.*

Il nous paraît essentiel d'introduire des mécanismes d'héritage pour coder la syntaxe des collocations, tant pour gérer le régime des collocations que les propriétés distributionnelles et transformationnelles.

Dans le cas des collocations verbe-nom, le régime des collocations sera calculé à partir de celui de la base. Les actants de la collocation sont calculés à partir d'heuristiques telles que celles qui ont été proposées par Alonso Ramos (1993) (cf. 4.1.). Les collocations pour lesquelles le régime apparaît en contradiction avec cette heuristique sont explicitement codées dans l'entrée du collocatif.

Les propriétés syntaxiques (distribution et transformations) sont codées par défaut sur l'entrée du collocatif. La collocation hérite par défaut des propriétés du collocatif comme cela est schématisé dans la figure suivante avec l'exemple *gros fumeur.*

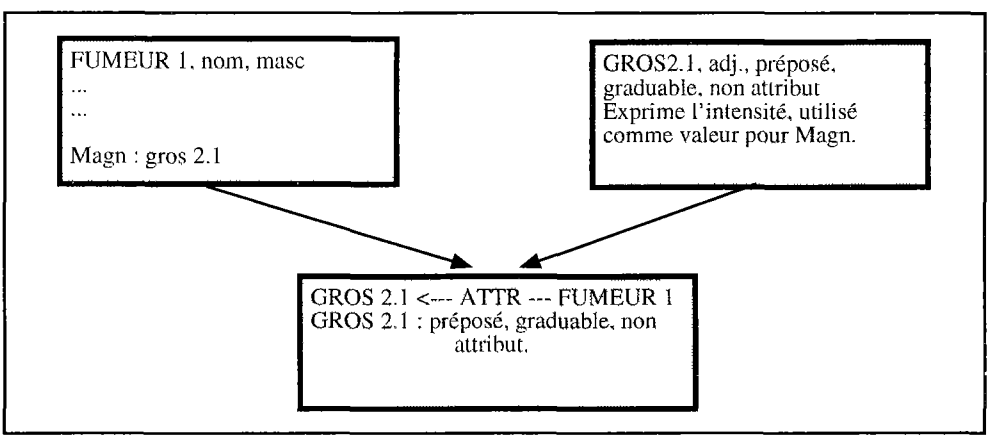


FIGURE 3 : Héritage des propriétés syntaxiques de la collocation

Cependant, les collocations n'héritent pas toujours de toutes les propriétés du collocatif. Il faut pouvoir dans ce cas prévoir un héritage « non monotone » par lequel les propriétés marquées sur la base auraient préséance sur les propriétés héritées du collocatif. Bien évidemment, ces mécanismes d'héritage ne peuvent pas être manipulés à la main. Cela implique l'utilisation d'un éditeur intégrant des mécanismes assez complexes d'héritage.

Conclusion

Dans le cadre de cette étude sur la formalisation des collocations, nous nous sommes intéressée, d'une part au mode de codage des collocations, les Fonctions Lexicales, et d'autre part, aux propriétés syntaxiques dont on devrait disposer pour les collocations. Avant que les FLs puissent être intégrées dans un système informatique, une description minutieuse des FLs (fonctions simples et fonctionnelles) et des combinaisons de celles-ci doit être effectuée. Ce travail, en cours, doit être intégré dans l'éditeur mis au point par Alain Polguère (Mel'čuk, Clas et Polguère, 1995).

Le second volet de notre étude concerne plus particulièrement le TALN, bien que l'information concernant les propriétés syntaxiques soit tout aussi indispensable dans un DEC « papier ». Il nous paraît crucial de pouvoir disposer, d'une part, d'une structure actantielle détaillée, d'autre part, de l'ensemble des propriétés syntaxiques essentielles concernant la collocation. Pour ce faire, il nous semble indispensable d'intégrer dans le lexique des mécanismes d'héritage non monotone, de façon à alléger le codage manuel concernant les collocations. Ces mécanismes d'héritage ne sont néanmoins pas facilement manipulables à la main et rendent absolument nécessaire le recours à un éditeur informatique facilitant le travail du lexicographe.

Description lexicographique des collocatifs dans un *Dictionnaire explicatif et combinatoire* : articles de dictionnaire autonomes ?[1]

Margarita ALONSO RAMOS et Suzanne MANTHA

Universidade da Coruña, Espagne et Université de Montréal, Canada

1. Présentation du problème

Dans ce travail, nous nous concentrons sur la description lexicographique des COLLOCATIONS et plus particulièrement sur celle des COLLOCATIFS dans un *Dictionnaire explicatif et combinatoire* (DEC). Notre objectif principal est d'étudier la possibilité de traiter les collocatifs comme des lexies (de plein droit), ce qui entraîne la nécessité d'avoir pour ceux-ci des articles de dictionnaire autonomes.

Nous commençons par donner quelques exemples du phénomène lexical des collocations – syntagmes spéciaux du type suivant : *OPÉRER un choix, RAVALER sa colère, MOURIR d'envie, LEVER l'ancre, prix DÉRISOIRE, café NOIR, cheveux BLONDS, peur BLEUE*, etc.

Dans tous ces syntagmes, le verbe ou l'adjectif est choisi par le locuteur en fonction du nom. Si l'on part d'autres noms, le choix des verbes ou des adjectifs pour exprimer le même sens sera différent : *PRENDRE <*OPÉRER> une décision, CONTENIR <*RAVALER> sa fureur, HISSER <*LEVER> les voiles, thé NATURE <*NOIR>, peau DORÉE <*BLONDE>, fureur NOIRE <*BLEUE>*, etc. Dans ces exemples, les noms correspondent à ce que nous appelons les « bases » des collocations et les verbes ou adjectifs à ce que nous appelons les « collocatifs »[2]. Nous nous

1. La réalisation de cette étude a été rendue possible grâce à l'aide financière accordée par le Conseil de recherches en sciences humaines du Canada (subvention n° 410-91-1844), par les Fonds FCAR (subvention n° 96-ER-0618) ainsi que par le Ministerio de Educación y Ciencia du gouvernement espagnol (bourse post-doctorale)

2. Cette correspondance (nom = base et verbe/adjectif = collocatif) n'est pas la seule possible : il existe d'autres collocations comme *coûter LA PEAU DES FESSES* et *risquer LE PAQUET* où, cette fois-ci, les verbes sont des bases et les noms, des collocatifs

intéressons à la description de ce dernier type de lexèmes – les lexèmes qui ne sont pas choisis librement.

En parlant des collocations, on insiste souvent sur l'inégalité entre les deux éléments d'une collocation : la base est sémantiquement autonome, tandis que le collocatif ne l'est pas (cf. entre autres, Hausmann, 1979 ; Heid, 1992a-b, 1994). Pour cette raison, le statut lexical des collocatifs n'est pas toujours reconnu : est-ce qu'un collocatif comme *opérer* dans *opérer un choix* est une entité lexicale suffisamment autonome pour créer un article de dictionnaire à part qui ne décrive que ce *opérer* ?

D'après notre approche (Mel'čuk, 1988, 1995a-b, 1996), ce qui caractérise spécifiquement les collocatifs, c'est leur façon spéciale d'être choisis dans le processus de la synthèse du texte. Le locuteur (humain ou machine) doit faire des choix lexicaux, c'est-à-dire qu'il doit choisir des lexies pour exprimer des sens donnés. Dans cette perspective, il faut distinguer deux types de lexies. Le premier type est constitué de lexies (la plupart, en fait) qui sont sélectionnées pour exprimer un sens donné INDÉPENDAMMENT d'autres lexies. Ainsi, si le locuteur veut exprimer le sens ‹attitude émotionnelle défavorable de X à l'égard de Y causée par le fait que X déteste Y à tel degré que X veut faire des actions agressives à Y...›, il choisit tout simplement HAINE – sans prendre en considération d'autres lexies déjà choisies. Dans ce cas, il s'agit d'un choix lexical SÉMANTIQUEMENT CONTRÔLÉ. Le deuxième type est constitué de lexies qui sont sélectionnées pour exprimer un sens donné mais qui sont sous le contrôle d'autres lexies choisies auparavant. Ainsi, pour exprimer le sens ‹intense› en parlant de la HAINE, le locuteur doit considérer la lexie à laquelle ce sens s'applique. Dans le cas de *haine*, c'est *mortelle* qui sera sélectionné pour exprimer ‹[haine] intense›. La sélection de MORTEL dans ce sens-là est doublement contrôlée : SÉMANTIQUEMENT, car on veut exprimer le sens ‹intense› et non pas le sens ‹sentir›, par exemple, et LEXICALEMENT, car le sens ‹intense› s'exprime par MORTEL auprès de HAINE mais pas auprès de RAGE, FUREUR, JOIE ou DÉSESPOIR, par exemple, où les choix lexicaux sont différents : *rage folle, fureur aveugle, vive joie, profond désespoir.*

Une collocation est donc une paire d'expressions lexicales, la base et le collocatif, telle que le choix de la base n'est contrôlé que sémantiquement, alors que le choix du collocatif est, en plus, contrôlé lexicalement – notamment, par la base. Au moment de la production du texte, la base est choisie en premier, car son sens est plus perceptible que celui du collocatif et il est exprimé de façon indépendante.

Le fait que les collocatifs soient sélectionnés en fonction de leur base affaiblit leur statut lexical et rend problématique leur description lexicographique. Les dictionnaires traditionnels n'ont pas de politique uniforme pour le traitement des collocatifs : soit qu'on crée une acception particulière pour le collocatif (p. ex., *blond* dans le *Petit Robert*), soit que le collocatif est répertorié comme un sens figuré sans marque explicite de nouvelle acception (*ravaler* [*sa colère*]), soit qu'il est consigné dans l'article de la base avec une petite description du sens de toute la collocation ([*feu*] *rouge*), ou encore qu'il est simplement consigné dans l'article de la base mais sans description (*rouge* [*de colère*]).

La création d'articles de dictionnaire pour les collocatifs présuppose qu'on leur attribue un statut de lexie de plein droit. Dans notre perspective, les notions « être une

lexie » et « avoir un article de dictionnaire » sont équivalentes. Par conséquent, une lexie est un item lexical qui a besoin d'un article de dictionnaire. Cependant, il n'est pas évident que tous les collocatifs aient besoin d'un article. Comme on le verra plus loin, les collocatifs ne sont pas tous du même type : il y en a qui se combinent avec une seule base (p. ex., *café NOIR* ; cf. section 3.1.), d'autres qui sont des syntagmes libres (p. ex., *fumer UNE CIGARETTE APRÈS L'AUTRE*). Dans ces deux cas, il n'est pas question de créer un article de dictionnaire pour les collocatifs, car ceci reviendrait à dire, dans le premier cas, que l'on attribue un statut de lexie de plein droit à un collocatif qui se combine avec une seule base et, dans le deuxième, que l'on attribue un statut de lexie à un syntagme libre.

Le problème à résoudre est de préciser le statut lexical des collocatifs de types différents et de trouver la bonne façon de les décrire lexicographiquement. Dans ce travail, nous étudierons les possibilités de description des collocatifs dans le cadre du DEC avec la rigueur et le formalisme requis.

Nous poursuivons notre discussion en procédant par les étapes suivantes :

– présentation du traitement actuel des collocations dans le DEC par l'appareil lexicographique des fonctions lexicales ;
– examen de certains traits distinctifs des collocatifs qui mesurent leur degré d'autonomie ;
– analyse des stratégies possibles de description et, pour finir, présentation d'une liste de vérification pour guider le lexicographe dans son choix de traitement à donner à un collocatif.

2. Les collocations dans le DEC : notion et description

Notre objectif principal étant de montrer comment le DEC peut décrire les collocatifs, nous devons donc commencer par caractériser, même si brièvement[3], ce type de dictionnaire. Les six propriétés de base du DEC sont les suivantes : 1) il est rédigé dans le cadre de la théorie linguistique Sens-Texte ; 2) il est orienté vers la production du texte ; 3) un article du DEC est fondé sur la définition de la lexie-vedette : sa représentation sémantique sert de base à la description de toutes ses relations paradigmatiques et syntagmatiques avec les autres lexies de la langue ; 4) il met de l'emphase sur la cooccurrence lexicale restreinte ; 5) il est formalisé, donc toutes les informations concernant le sens, la syntaxe, etc. doivent être indiquées de façon précise et explicite ; 6) il est exhaustif au niveau de la description de chaque lexie individuelle.

Une caractéristique importante du DEC est que l'unité de description lexicographique est une LEXIE, qui peut être un lexème ou un phrasème complet. Un lexème est un mot pris dans une acception déterminée, tandis qu'un phrasème complet est un syntagme non libre pris aussi dans une acception déterminée. À chaque lexie, le DEC

3 Pour une caractérisation détaillée du DEC, nous référons le lecteur aux ouvrages suivants : Mel'čuk, 1982, 1995a-b, 1996 ainsi que Mel'čuk *et al.*, 1984, 1988, 1992, 1995.

fait correspondre un seul article de dictionnaire. Dans ce sens, il se démarque radicalement des dictionnaires courants qui prennent le mot polysémique comme unité de description.

Les lexies qui sont sémantiquement liées par des ponts sémantiques (*i.e.* qui partagent des composantes sémantiques communes relativement importantes) et dont les signifiants sont identiques sont regroupées en VOCABLES. À chaque vocable correspond un superarticle qui est formé de tous les articles des lexies faisant partie de ce vocable. Comme nous le verrons plus loin à la section 3.2., il existe des vocables constitués d'une seule lexie (les « mots » monosémiques) et d'autres constitués de plusieurs (les « mots » polysémiques). Les vocables polysémiques peuvent regrouper des lexies qui ont un comportement libre, c'est-à-dire qui sont choisies d'après leur sens, et/ou d'autres lexies qui sont restreintes, c'est-à-dire qui sont contrôlées par d'autres lexies déjà choisies. En d'autres termes, un vocable polysémique peut être constitué de lexies autonomes et/ou de « lexies collocatives ». Nous appellerons les premières « contreparties libres des collocatifs ».

Les propriétés du DEC qui nous concernent le plus sont la deuxième et la quatrième. Le DEC, conformément à la théorie Sens-Texte, est orienté vers la production du texte, plutôt que vers la compréhension. Il doit donc fournir tous les moyens lexicaux pour exprimer un sens donné. Ainsi, la question à laquelle le DEC prétend répondre est « comment exprimer 'X' » plutôt que « qu'est-ce que X signifie ». En d'autres termes, le DEC est un dictionnaire adapté pour la synthèse plutôt que pour l'analyse des textes. La raison derrière cette approche est la conviction que la direction à partir du sens vers le texte (c'est-à-dire parler) est une activité linguistique par excellence, tandis que la direction inverse (comprendre) entraîne d'autres activités, qui ne sont pas purement linguistiques : le destinataire qui doit déchiffrer un énoncé doit faire appel à ses habilités logiques ainsi qu'à ses connaissances sur la situation, sur le monde, etc.

Étant donné son orientation vers la synthèse, le DEC est un dictionnaire centré sur la cooccurrence lexicale restreinte, c'est-à-dire la cooccurrence qui n'est pas déterminée sémantiquement à cent pour cent. Tous les lexèmes qui se combinent avec le lexème vedette d'un article et dont le choix est contrôlé lexicalement sont systématiquement décrits dans le DEC – au moyen des fonctions lexicales (= FL).

Avant de présenter les FL, nous rappelons ici brièvement l'approche de la théorie Sens-Texte par rapport aux collocations. Celles-ci sont comprises comme une sous-classe de phrasèmes, *grosso modo* des syntagmes non libres. Mel'čuk (1995a-b) distingue trois types de phrasèmes : 1) PHRASÈMES COMPLETS : *casser sa pipe, bas-bleu* (où le sens de l'expression n'inclut le sens d'aucun de ses constituants) ; 2) QUASI-PHRASÈMES : *donner le sein* (Mel'čuk *et al.*, 1992 : 191), *casque bleu* (où le sens de l'expression inclut les sens des deux lexèmes constituants plus un ajout sémantique imprévisible) ; 3) SEMI-PHRASÈMES qui correspondent en fait à ce que nous entendons par COLLOCATIONS (où le sens de l'expression inclut le sens d'un des lexèmes constituants et le sens de l'autre constituant est exprimé par un lexème choisi non librement).

Ces trois types de phrasèmes sont décrits dans le DEC de deux façons différentes : soit comme lexies vedettes d'articles autonomes (les phrasèmes complets et

les quasi-phrasèmes), soit comme éléments dans les articles de leurs bases (les collocations). Les deux premiers ne peuvent pas être décrits en fonction de leurs constituants, tandis que les troisièmes peuvent et doivent être décrits en fonction d'un de leurs constituants – la base – au moyen de FL dans l'article de la base.

Dans le DEC, on utilise une terminologie parallèle, où la base d'une collocation correspond au mot-clé et le collocatif, à un élément de la valeur d'une FL. Le DEC adopte cette terminologie pour mettre en relief l'aspect fonctionnel de la description des collocations.

Pour présenter le traitement des collocations dans le DEC, il nous faut introduire la notion de FL. *Grosso modo*, une FL est une fonction **f** qui associe à une lexie L – le mot-clé (= la base) de **f** – un ensemble L_i d'expressions lexicales – la valeur (= le collocatif) de **f** – qui sont choisies en fonction de L pour exprimer le sens correspondant à **f**. En d'autres mots, étant donné la collocation **A** + **B** signifiant 'A + C', alors le sens 'C' qui est exprimé par le collocatif **B** auprès de la base **A** correspond à ce qui est appelé une **FL**. Le lexème **A** qui garde intact son sens et détermine le choix de l'expression pour 'C' est le **mot-clé** de la FL. Le lexème **B** qui est choisi pour exprimer 'C' auprès de **A** est la **valeur** de la FL.

Il convient de distinguer deux sous-classes de FL : les standard et les non standard. Les premières doivent satisfaire les quatre conditions suivantes :

1. Pour toute paire de lexies L_1 et L_2, les lexies $\mathbf{f}(L_1)$ et $\mathbf{f}(L_2)$ montrent des relations (presque) identiques à ces lexies :

$$\frac{\text{'}\mathbf{f}(L_1)\text{'}}{(L_1)} \approx \frac{\text{'}\mathbf{f}(L_2)\text{'}}{(L_2)}$$

Ex. : Soit la FL **f** = **Magn** (= 'intensificateur'), L_1 = FROID et L_2 = SOURD, et $\mathbf{f}(L_1)$ = *de canard* et $\mathbf{f}(L_2)$ *comme un pot*. La relation sémantico-syntaxique entre *de canard* et FROID est la même que celle entre *comme un pot* et SOURD. Les deux expressions sont des modificateurs intensificateurs qui signifient, dans les contextes donnés, 'intense', 'très'.

2. En règle générale, $\mathbf{f}(L_1)$ et $\mathbf{f}(L_2)$ sont différents : $\mathbf{f}(L_1) \neq \mathbf{f}(L_2)$.
Ex. : *très <*grièvement> malade* vs *grièvement <*très> blessé* ; *une grippe carabinée* vs *une fièvre de cheval ; aimer à la folie* vs *haïr à mort.* Les expressions dépendent des lexies modifiées.

3. La fonction **f** a un nombre élevé de mots-clés. Le sens 'f' est très abstrait et très général et s'applique à beaucoup d'autres sens.

4. La fonction **f** a un nombre élevé d'éléments dans sa valeur (= d'expressions).

Les FL standard sont codées dans le DEC par des symboles conventionnels comme le montrent les exemples suivants :

FL mot-clé		**valeur**
Magn(*désir*)	=	*ardent*
Magn(*envie*)	=	*folle*
Magn(*interdire*)	=	*absolument*
Mult(*abeille*)	=	*essaim* [*d'abeilles*]
Func$_0$(*silence*)	=	*régner*
Oper$_1$(*baiser*)	=	*donner* [*un baiser*]
Oper$_1$(*câlin*)	=	*faire* [*un câlin*]

Lorsque les conditions 3 et 4 données ci-dessus sont violées, on a affaire à l'autre type de FL, soit les non standard. Leurs sens sont trop spécifiques, de sorte qu'ils ne s'appliquent qu'à très peu de lexies et ne possèdent qu'un infime nombre d'expressions différentes. Les FL non standard sont spécifiées en français. Ainsi, dans l'article du lexème CAFÉ, on trouverait, entre autres, les FL suivantes :

Magn	: *fort* [FL standard]
sans lait	: *noir* [FL non standard]
avec de la crème ou un peu de lait	: *crème* [FL non standard]

Les FL sont consignées dans l'article de dictionnaire du lexème qui est le mot-clé de la FL, c'est-à-dire la base de la collocation.

Pour produire correctement une collocation, il ne suffit pas, en règle générale, d'additionner les propriétés des deux lexèmes (la base et le collocatif) qui la constituent. Dans ces expressions, soit la base, soit le collocatif peut avoir des propriétés qui ne concernent qu'une collocation particulière.

Le premier cas : propriétés particulières de la base. Nous pensons aux collocations verbales où, très souvent, le nom a un comportement déviant par rapport aux déterminants. Par exemple, en espagnol, le nom *respeto* ('respect') exige ou rejette l'article selon la collocation dans laquelle il apparaît. Quand il se combine avec le verbe *tener* ('avoir'), la collocation peut signifier ('respecter [quelqu'un]') ou ('être respecté [par quelqu'un]'). Dans le premier sens, le nom ne peut pas être déterminé :

Juan tiene (**el*) *respeto por su padre*, litt. 'Juan a respect pour son père'

tandis que, dans le second, le nom doit être déterminé :

El padre tiene *(*el*) *respeto de su hijo*, litt. 'Le père a le respect de son fils'.

Dans le DEC, cette information apparaît dans le schéma de régime réduit de la collocation, où elle est présentée de la façon suivante :

Oper$_1$(*respeto*) = *tener* [~]
Oper$_2$(*respeto*) = *tener* [ART$_{déf}$ ~] | $C_1 \neq \Lambda$

La notation qui apparaît après la barre verticale indique que le premier actant du mot-clé (= la désignation de celui qui respecte) ne peut pas être vide, c'est-à-dire qu'il doit être exprimé : *El padre tiene el respeto de Juan <su respeto>* vs **El padre tiene el respeto.*

Le deuxième cas : propriétés particulières du collocatif. Considérons, par exemple, l'adjectif *fort* qui est soit préposé soit postposé avec *pluie* mais seulement postposé avec *café* :

une forte pluie ~ une pluie forte [sans changement de sens]
**un fort café ~ un café fort*

Ces deux cas ne sont pas une particularité de FORT mais seulement de la collocation concernée. Par conséquent, l'emploi seulement postposé de *fort* doit être indiqué dans la description de la collocation *café fort* :

Magn(*café*) = *fort* | postposé

Chacune de ces FL avec ses valeurs constitue une description de la collocation en question – avec la particularité que cette description est enchâssée dans l'article de la base (le mot-clé).

Si le DEC dispose d'une politique bien établie pour décrire les collocations, ce n'est pas le cas pour la description des collocatifs. Comme justification de la création d'un article de dictionnaire pour un collocatif, Mel'čuk (1995a) propose le critère de généralisation suivant : si l'élément L_i (le collocatif) de la valeur d'une FL partage les mêmes particularités de régime ou de flexion avec beaucoup de mots-clés (particularités par rapport à L_i' libre), alors, afin d'éviter la répétition d'information, L_i peut avoir son propre article – avec une description exhaustive de son régime et de toutes ses autres propriétés.

Or il y a des collocatifs qui n'ont pas de contreparties employées librement (p. ex., *bissextile*) de sorte qu'on ne peut pas compter sur l'idée d'indiquer l'information dans un autre article de dictionnaire. Il y en a d'autres comme *noir* dans *café noir* qui ont de telles contreparties, mais dont le sens est exclusif à la collocation avec le nom *café*.

Dans les trois volumes du DEC publiés, le traitement des collocatifs n'est pas uniforme. Quand les collocatifs ont des contreparties libres, on peut trouver des articles pour ceux-ci, soit parmi les autres lexies du vocable (p. ex., plusieurs lexèmes collocatifs appartenant au vocable FLAMBER), soit dans un autre vocable à part (on y voit donc un cas d'homonymie, p. ex., REMPLIR[2]).

Pour standardiser le traitement des collocatifs, il faut commencer par en ébaucher la typologie. Dans la section suivante, nous allons étudier différents traits distinctifs des collocatifs, ce qui nous aidera à déterminer le traitement lexicographique adéquat pour chaque type de collocatif.

3. Traits distinctifs des collocatifs

Étant donné la nature hétérogène des collocatifs, nous les distinguerons en nous basant sur les traits distinctifs suivants :

1. à base unique *vs* non unique ;
2. isolés *vs* ayant une contrepartie libre ;

3. décrits par une FL standard *vs* non standard ;
4. plus *vs* moins phraséologisés ;
5. plus *vs* moins autonomes.

Précisons que notre liste de traits est tout à fait provisoire. Elle n'est ni exhaustive ni définitive. Cette tentative d'établir une typologie des collocatifs au moyen de traits est nouvelle dans le cadre du DEC. Étant donné la quantité de collocations à traiter dans le lexique français et la quantité d'informations à donner pour chacune d'elles, il serait nécessaire de mener une étude beaucoup plus large que la présente. De nouveaux traits devront probablement être introduits pour permettre la distinction de tous les types de collocatifs.

Un trait à lui seul n'est pas suffisant pour décider du traitement à donner à la description d'un collocatif. Par conséquent, c'est une combinaison de traits qu'il faudrait considérer pour décider du type auquel correspond un collocatif.

Les traits ne sont pas indépendants les uns des autres. Ils n'ont pas tous la même importance et donc certaines combinaisons de traits sont plus concluantes que d'autres. Une hiérarchie dans les combinaisons de traits devrait être établie.

Notre liste doit être utilisée avec prudence. Nous la croyons quand même utile pour guider le lexicographe dans sa démarche lexicographique.

3.1. Collocatifs à base unique *vs* non unique

Certains collocatifs ne se combinent qu'avec une seule base, tandis que d'autres sont productifs, c'est-à-dire qu'ils se combinent avec un nombre élevé de bases.

Examinons d'abord des collocatifs à base unique. Ainsi, l'adjectif NOIR signifiant ‹sans lait› est un collocatif à base unique ; dans le sens donné et avec les propriétés données (postposition), il ne se combine qu'avec le lexème CAFÉ en excluant même la combinaison avec ses hyponymes :

**un expresso noir, *un déca noir, *un arabica noir.*

Un autre collocatif à base unique est l'adjectif PLATE au sens ‹sans bulles› : il a ce sens seulement en combinaison avec le nom EAU. Aucune autre boisson ne peut être qualifiée de *plate* pour exprimer le sens ‹sans bulles›.

L'adjectif BISSEXTILE est aussi un cas de cooccurrence unique mais il ne s'agit pas du même type de collocatifs que ceux vus ci-dessus. Le sens ‹qui, au lieu d'avoir 365 jours, a une journée de plus, le 366e jour étant le 29 février› ne peut être appliqué qu'au lexème ANNÉE.

Les adjectifs *noir / plate* et *bissextile* sont des collocatifs à base unique qui se distinguent par les propriétés suivantes :

– Le fait que NOIR ‹sans lait› ne se combine qu'avec CAFÉ (ou PLATE ‹sans bulles› seulement avec EAU) est tout à fait arbitraire, comme c'est le cas pour la plupart

des collocations. On pourrait dire qu'il s'agit d'UNICITÉ LEXICALE. L'adjectif NOIR signifiant ‹sans lait› est SÉMANTIQUEMENT applicable à d'autres lexèmes mais pas LEXICALEMENT : **thé noir, *tisane noire.*

– Le fait que BISSEXTILE ne se combine qu'avec ANNÉE est déterminé sémantiquement. C'est le sens du collocatif qui interdit toute autre cooccurrence. On pourrait dire qu'il s'agit d'UNICITÉ SÉMANTIQUE. L'adjectif BISSEXTILE n'est pas sémantiquement applicable à d'autres lexèmes et, par conséquent, non plus lexicalement.

Par rapport aux collocatifs à base non unique, on peut établir une échelle qui va des collocatifs qui se combinent avec un petit nombre de bases (DE FER : *volonté, santé, discipline*) jusqu'à ceux dont le nombre de bases est beaucoup plus élevé, environ une centaine (FORT : *vent, médicament, lunettes, lumière, son, goût, moutarde, café, moment,* etc.).

À un extrême de cette échelle se trouve une sorte de collocatif qui, en principe, se combine avec une ou deux bases mais qui peut aussi se combiner avec des noms propres, liés aux bases « originelles » par la relation d'instanciation. C'est le cas de SE JETER, signifiant ‹[cours d'eau] se verse dans un autre cours d'eau ou une étendue d'eau›. Les bases originelles de ce collocatif sont FLEUVE, RIVIÈRE et RUISSEAU, mais tous les noms propres désignant des cours d'eau admettent aussi la combinaison avec SE JETER : *Le Saint-Laurent <La Volga, Le Duero> se jette dans la mer.* Un autre cas proche de celui-ci est la combinaison possible de BLOND avec tous les noms qui incluent le sens ‹cheveu ou poil›. Ainsi l'adjectif BLOND, qui signifie ‹[cheveu ou poil] dont la couleur est similaire à la couleur du blé›, peut aussi se combiner avec, entre autres, MÈCHE, TIGNASSE, TRESSE ou PERRUQUE. La relation entre ces noms et les bases « originelles » est une sorte de méronymie. Nous appelons ce type de collocations, « collocations par héritage ».

À l'autre extrême de l'échelle, se trouvent des collocatifs comme FORT, GRAND, GROS, FAIRE, PRENDRE, etc. qui se combinent avec un nombre très élevé de bases. On pourrait dire que ces collocatifs jouent le rôle de « joker » : ainsi, FORT, GRAND et GROS sont les « jokers » pour l'intensificateur et FAIRE, le « joker » pour le verbe support. Dans ces cas, la première question qui se pose est de savoir, d'une part, si tous ces collocatifs sont toujours le même lexème ou pas, c'est-à-dire si, par exemple, *fort* est le même lexème dans *un café fort, une moutarde forte* ou *un son fort*. D'autre part, la productivité de ces collocatifs rend plus difficile leur description lexicographique parce qu'il n'y a pas toujours moyen de faire de regroupements sémantiques pour toutes les bases possibles et il y a des résidus inclassifiables. D'ailleurs, on doit s'attendre à de tels résidus justement parce qu'il s'agit de cooccurrence lexicale restreinte.

En somme, nous croyons que plus un collocatif est unique, moins il a de chances d'avoir un article autonome.

3.2. Collocatifs isolés *vs* ayant une contrepartie libre

Un autre trait qui aura des effets sur le traitement lexicographique à donner au col-

locatif porte sur la monosémie ou la polysémie du vocable auquel le collocatif appartient. On peut observer des collocatifs qui constituent des vocables monosémiques. C'est le cas de BISSEXTILE, RAUQUE, AQUILIN qui n'existent qu'en tant que collocatifs : nous les appellerons « collocatifs isolés ». D'autres collocatifs appartiennent à des vocables polysémiques parmi lesquels il faut distinguer les vocables dont tous les lexèmes constituants sont des collocatifs (p. ex., BLOND, CLIGNER) de ceux qui sont composés à la fois de lexèmes collocatifs et de lexèmes libres (p. ex., NOIR, PLAT, FORT, OPÉRER).

Les collocatifs isolés doivent avoir un article autonome, car il est nécessaire d'indiquer leur existence dans le lexique de la langue et on ne peut pas le faire autrement. Les collocatifs qui ont d'autres partenaires collocatifs le méritent autant. Par contre, les collocatifs ayant des contreparties libres peuvent se passer d'un article autonome si leur forme phonologique, leur partie de discours et/ou toutes autres sortes de données morphologiques, syntaxiques ou lexicales sont partagées avec les lexies libres du même vocable.

La coïncidence entre les propriétés des lexies libres et celles des collocatifs ne survient pas systématiquement. Il est fréquent de trouver certaines propriétés qui se manifestent dans l'emploi libre mais non pas dans l'emploi collocationnel et *vice versa*. Ainsi, par exemple, le verbe *opérer* employé librement permet l'aspect progressif, tandis qu'employé en collocation, cet aspect n'est pas exprimable :

a. *Jean est en train d'opérer un malade.*
b. **Jean est en train d'opérer un choix.*

Le fait que le collocatif en **b** ne présente pas la même propriété que celle de sa contrepartie libre en **a** ne prouve pas nécessairement qu'il s'agit de deux lexèmes OPÉRER différents. La perte de propriétés que subit le collocatif peut être imputée au phénomène général de la phraséologisation. Cependant le cas inverse est révélateur. Prenons l'exemple de FORT : le lexème de base FORT a comme intensificateur l'expression *comme un Turc*, tandis que le collocatif FORT, lui, ne l'admet pas (**café fort comme un Turc*, **son fort comme un Turc*, etc.). De son côté, le collocatif peut avoir son propre intensificateur : [*son*, *bruit*, *cri*,...] *fort à crever les tympans* (mais pas **café fort à crever*). Dans ces cas où le collocatif possède une caractéristique que l'emploi libre n'a pas, la création d'un article de dictionnaire pour ce type de collocatif serait plus justifiable, car ses propriétés doivent être mentionnées dans un DEC.

À la lumière de ces considérations, nous pouvons affirmer que les collocatifs isolés doivent avoir un article de dictionnaire. Dans les autres cas, nous croyons que plus l'emploi collocationnel considéré possède ses propres caractéristiques syntaxiques, lexicales, etc., plus le collocatif aura de chances de posséder un article autonome.

3.3. Collocatifs plus phraséologisés *vs* moins phraséologisés

On peut établir une échelle qui part des collocatifs très liés phraséologiquement à la base jusqu'à ceux qui le sont moins. Les collocatifs plus phraséologisés sont perçus par les locuteurs comme des éléments d'un tout, à savoir la collocation, qui sont inséparables. En d'autres termes, le sens d'un tel collocatif est très peu perceptible,

voire pas du tout, sans spécification de sa base. Ainsi, par exemple, un collocatif comme BLEU dans *un steak bleu* est très lié phraséologiquement à la base. Pour exprimer le sens 'très saignant', le français a choisi l'adjectif BLEU. La lexie de base, BLEU 'de couleur bleue', n'a pas de connotations qui établissent un pont sémantique avec BLEU 'saignant'. Le choix de BLEU pour ce sens est totalement arbitraire et idiosyncratique synchroniquement.

Les sens des collocatifs moins phraséologisés sont plus facilement perçus séparément de leur base : en fait, ces collocatifs se combinent avec un nombre très élevé de bases qui *grosso modo* sont regroupables sémantiquement. Par exemple, les verbes ÉPROUVER et SOUFFRIR, comme *éprouver de la haine* et *souffrir d'une maladie* jouent le rôle de verbe support des noms d'émotions et des noms de maladies respectivement[4].

Il y existe d'autres collocatifs encore moins phraséologisés que ÉPROUVER (avec émotion) et SOUFFRIR (avec maladie). Par exemple, le verbe CONDUIRE : il peut apparaître sans sa base (*Il conduit vite*) et il garde le même sens en emploi libre comme dans *Il conduit vite* et en combinatoire restreinte comme dans *conduire une auto*. Ici, la distance sémantique entre le collocatif et sa contrepartie libre est zéro. Ceci confère au verbe CONDUIRE un double statut : il peut tantôt être une lexie collocative et tantôt une lexie de plein droit tout dépendant du point de vue à partir duquel l'expression est produite. En effet, quelquefois *conduire une auto* est bel et bien une collocation : cette expression ne peut être considérée comme entièrement libre, car, en français, pour exprimer le sens 'faire avec l'auto ce à quoi elle est destinée', on dit *conduire*, tandis qu'en espagnol d'Amérique du Sud, on choisirait le verbe *manejar* qui correspond littéralement à 'manipuler' en français. D'autres fois, CONDUIRE est une lexie de plein droit : il s'emploie sans sa base, il possède sa propre combinatoire lexicale, ses propres dérivés, etc. Comme ces propriétés appartiennent à CONDUIRE et qu'elles sont assez nombreuses, alors il doit avoir son propre article de dictionnaire autonome.

En somme, moins le collocatif est phraséologisé, plus il a de chances d'avoir un article autonome.

3.4. Collocatifs décrits par une FL standard *vs* non standard

Le sens d'une FL standard est assez abstrait et général, comme 'très', 'causer', 'faire ce qu'on est censé faire'. Il peut être appliqué à nombre très élevé de bases et le nombre de collocations obtenues l'est aussi. Le sens d'un collocatif décrit au moyen d'une FL standard a donc besoin d'être plus spécifique et les bases avec lesquelles il se combine doivent être regroupées sémantiquement. Nous pensons à la FL **Magn**, où l'intensification n'agit pas toujours sur les mêmes composantes sémantiques. Elle dépend du sémantisme de la base. Par exemple, *fort* auprès de *médicament* intensifie la

4 Même si *éprouver* signifie 'éprouver' et *souffrir*, 'souffrir', ces expressions sont bel et bien des collocations. Dans le DEC, elles sont décrites par la FL **Oper** et, dans cette approche, toute valeur d'une FL combinée avec son mot-clé constitue nécessairement une collocation.

concentration de la substance tandis qu'auprès de *moutarde*, il intensifie le goût qui est assez prononcé. Il semble donc nécessaire de préciser les sens des collocatifs et un endroit approprié pour le faire serait la zone sémantique d'un article de dictionnaire. Il faut donc envisager la possibilité de créer un article de dictionnaire pour ces collocatifs.

À l'opposé, le sens d'une FL non standard est plus concret, donc limité à un nombre plus réduit de bases possibles et avec un nombre d'expressions moins élevé. Il n'est donc pas nécessaire de créer un article de dictionnaire, le sens d'un tel collocatif n'ayant pas besoin d'être spécifié davantage.

Si l'on veut rédiger une définition lexicographique pour un collocatif qui est décrit par une FL standard dans l'article de la base, elle sera formulée en termes assez abstraits et généraux (sauf quand le collocatif est sélectionné sémantiquement). Par exemple, la définition du collocatif FORT dans une *forte envie* [= Magn(envie)] serait tout simplement ‹intense›. Mais comme ce sens est très général, il faut contraindre les combinaisons (**un fort chagrin, *une forte haine, *une forte surprise*). Il s'agit de formuler une contrainte sémantique sur le nom modifié de façon à exclure tous les noms d'émotions avec lesquels FORT ne peut pas se combiner. Pour y arriver, il faudrait d'abord détenir une liste fermée de tous les noms d'émotions, ensuite voir lesquels peuvent se combiner avec FORT et lesquels ne le peuvent pas. En somme, la description lexicographique de collocatifs ayant un sens abstrait et se combinant avec un très grand nombre de bases représente un travail ardu voire même impossible.

Par contre, la définition d'un collocatif décrit par une FL non standard dans l'article de la base sera plus facile à formuler étant donné qu'il ne se combine qu'avec un nombre fort restreint de bases : il y a moins d'expressions à couvrir et donc moins de composantes sémantiques à considérer. Dans de tels cas, on aboutit à une définition plus spécifique. Ainsi, par exemple, la définition de BLOND dans *cheveux blonds* a la forme suivante :

> **I.1a.** adj. [X] *blond* = [Poil ou cheveu**I.1** X] dont la couleur est similaire à la couleur du blé et qui tend vers le jaune pâle.

Comme on peut l'observer, la contrainte ‹cheveux ou poil› fait en sorte que tout nom dont le sens inclut cette composante puisse se combiner avec BLOND.

Finalement, si la description du collocatif correspond à une FL standard, on a plus de chances d'avoir besoin de créer un article de dictionnaire. Aussi il est plus difficile de consigner les bases possibles d'un collocatif quand celui-ci correspond à une FL standard que quand il correspond à une FL non standard.

3.5. Collocatifs plus autonomes *vs* moins autonomes

Ce dernier trait a plus de poids que les précédents. L'autonomie d'une lexie est le trait par excellence pour juger du statut du collocatif comme une lexie à part entière et donc, pour nous permettre de décider si ce collocatif aura droit à un article de dictionnaire. Nous mesurons l'autonomie du collocatif en vertu de trois niveaux linguistiques : sémantique, syntaxique et lexical.

Une lexie est SÉMANTIQUEMENT AUTONOME quand elle a un sens donné indépendamment des lexies avec lesquelles elle se combine. Ainsi, CAFÉ**2** signifiant 'boisson...' a ce sens peu importe s'il est en combinaison avec *boire, faire, préparer, acheter, détester ou noir.* Dans des cas comme *grain <récolte> de café* ou *comptoir <terrasse, zinc, patron> d'un café*, il s'agit de deux autres lexies différentes : CAFÉ**1** signifiant 'plante...' et CAFÉ**3** signifiant 'établissement...'. On ne doit pas s'attendre à ce que les combinaisons libres de CAFÉ**2** soient possibles avec CAFÉ**1** et CAFÉ**3**. Ils sont des lexies sémantiquement autonomes et chacun possède sa propre combinatoire libre.

Par contre, *noir* signifiant 'sans lait' n'a ce sens qu'en combinaison avec *café*. Ainsi, il n'est pas sémantiquement autonome, car il nécessite l'« appui » d'une autre lexie pour exister. Cependant tous les collocatifs ne sont pas comme *noir* dans *café noir*. Il y en a comme *conduire* qui, lui, a un sens 'diriger un véhicule...' indépendamment des mots avec lesquels il se combine : *Elmuck conduit comme un fou, J'aime bien conduire sur l'autoroute, Les autos que je conduis sont toujours des Mercedes.* Le collocatif *conduire* est donc une lexie sémantiquement autonome – il n'a pas besoin de l'appui de sa base pour signifier ce qu'il signifie.

Une lexie est SYNTAXIQUEMENT AUTONOME quand elle peut admettre tout ce qui est prévu par les règles générales de la syntaxe française, par exemple, si le collocatif peut être linéairement séparé de sa base par d'autres lexies, changer de rôle syntaxique, recevoir des modificateurs, etc. Ainsi, le collocatif *blond* dans *des cheveux blonds* est plus autonome que *bleue* dans *une peur bleue*. Le premier est admis dans la position prédicative, mais pas le deuxième :

*Ses cheveux sont blonds ~ *Sa peur est bleue.*

Quant à la modification, on observe les différences suivantes :

*des cheveux très blonds ~ *une peur très bleue.*

Une lexie est LEXICALEMENT AUTONOME si elle a ses propres corrélats lexicaux, soit paradigmatiques, soit syntagmatiques. Dans un article du DEC, c'est l'appareil des FL qui décrit les corrélats de la lexie vedette. Alors, si un collocatif a ses propres synonymes, ses propres antonymes, ses propres dérivés, ses propres intensificateurs, etc., il sera plus autonome. Par exemple, le collocatif *brun* dans *tabac brun* est plus autonome que *nature* dans *thé nature*. En effet, on peut opposer *tabac brun* à *tabac blond*, mais on ne peut pas opposer *thé nature* à **thé blanc* ou **thé lacté*. Ainsi, ce collocatif *blond* a un opposé (traité par la FL **Contr**), tandis que *nature* n'en a pas. Pour l'exprimer, il faut utiliser la paraphrase *thé au lait*, aucune expression idiomatique n'existant pour exprimer ce sens.

En somme, plus le collocatif se comporte de façon autonome, plus il possède de propriétés qu'il faut décrire et, par conséquent, il tend à accéder au statut de lexie à part entière et nous penchons en faveur de lui attribuer un article de dictionnaire.

4. Les voies de solution : comparaison de différentes stratégies de description

À l'heure actuelle, nous ne sommes pas en mesure d'énoncer des critères définitifs

pour décider s'il faut ou non créer un article de dictionnaire autonome pour un collocatif. On ne peut que constater des tendances vers l'autonomie : plus un collocatif est autonome, plus nous sommes en faveur de créer un article ; mais les cas flous sont nombreux. Ce fait reflète une propriété typique des langues naturelles : elles résistent à l'encadrement dans des cases fermées. Dans cette section, nous montrons les stratégies alternatives de description possibles et les conséquences que chacune d'elles entraîne.

Les deux positions extrêmes consistent, d'une part, à créer des articles pour TOUS les collocatifs, peu importe leurs propriétés, et d'autre part, à ne créer d'articles pour AUCUN d'eux. Examinons ces deux positions à tour de rôle :

– La première position entraîne une multiplication des articles de dictionnaire. Il est vrai qu'avec les outils informatiques actuels, la taille du dictionnaire ne crée pas trop de difficultés (bien que les informaticiens cherchent toujours la façon la plus économique de stocker l'information), mais le problème est que le dictionnaire grossit d'une façon un peu artificielle : le dictionnaire compterait beaucoup d'articles pour des lexies irréelles, *i.e.* qui n'existent qu'en combinaison avec une seule base donnée. Dans notre perspective, une lexie doit avoir une définition (c'est-à-dire un sens bien délimité), un régime particulier et une propre combinatoire pour être une lexie de plein droit. Cependant, plusieurs collocatifs se comportent de façon « parasitique »[5], *i.e.* ils s'accrochent à un autre lexème du même vocable (par ex., *noir* ‘de couleur noire’), pour pouvoir exister comme collocatifs : ils ont un sens seulement dans des contextes très restreints, ils ont une flexion restreinte, ils n'ont pas leur propre combinatoire, ou toutes ces propriétés réunies.

Créer un article de dictionnaire pour tous les collocatifs serait équivalent à dissoudre tous les phrasèmes complets en lexies séparées[6]. Si l'on force l'unicité lexicale, on pourrait arriver à dire que tout lexème faisant partie d'un phrasème a un sens particulier qui n'apparaît qu'en combinaison avec les autres constituants du phrasème. Ainsi, par exemple, le phrasème complet *bas-bleu* serait séparé en deux lexèmes, l'un signifiant ‘femme’ (mettons *bas*) et l'autre signifiant ‘à prétentions littéraires’. De cette façon, *bas* aurait ce sens seulement avec *bleu* et *bleu* aurait ce sens seulement avec *bas*. Il semble évident qu'il n'y a aucun gain descriptif dans cette analyse.

– La deuxième position (= ne jamais créer d'articles pour les collocatifs) ne peut pas être adoptée de façon cohérente, car des lexies comme *conduire, tirer (un coup de fusil), fumer, souffrir, très, beaucoup*, etc., qui s'utilisent en dehors des collocations et qui doivent avoir un article de dictionnaire, seraient privées de leur article. Cette stratégie ne prend pas en considération le fait que plusieurs collocatifs peuvent être employés aussi comme des lexies libres. Si l'on adopte cette politique, il faudrait inclure

---

5 Les lexèmes constituant un phrasème complet sont aussi « parasitiques » Au moment de la production du texte, le locuteur choisit les lexèmes pour constituer un phrasème, mais il le fait sans prendre en considération leur propre sens , il emprunte la forme des lexèmes qui ont une autre vie ailleurs pour servir de blocs de construction d'une nouvelle lexie, le phrasème complet Le comportement du collocatif dans une collocation est similaire. Pour produire la collocation *café noir*, le locuteur emprunte la forme du lexème existant *noir* pour lui accrocher le sens ‘sans lait’ dans la collocation prise comme un tout

6 I Mel'čuk a discuté de cette idée dans l'article suivant · Mel'čuk, I A (1960) « O terminax "ustojčivost'" i "idiomatičnost'" » [Sur les termes « locution stable » et « locution idiomatique »] *Voprosy jazykoznanija*, n° 4, 73-80 [Il existe une traduction anglaise · FDIMT (JPRS, n° 6732)]

dans l'entrée de la base beaucoup d'informations qui ne sont pas liées à la collocation : par exemple, tous les polysèmes du vocable CONDUIRE (*Le guide nous conduit à l'intérieur du musée*) devraient être consignés sous l'entrée de *véhicule* ou de *auto*, ce qui va à l'encontre de la logique même du DEC.

Il faut donc trouver des solutions moins radicales et n'accorder un article de dictionnaire qu'à certains types de collocatifs. Pour étayer tout ce qui vient d'être dit, nous procédons à la description lexicographique du verbe CLIGNER comme dans l'expression *cligner des yeux*. Il s'agit d'un collocatif unique : il ne se combine qu'avec *yeux*. Il appartient à un vocable polysémique qui n'est constitué que de lexèmes collocatifs. Il est décrit par une FL non standard. Avant d'aller plus loin dans la description de ce collocatif, examinons d'abord comment on décrit la collocation *cligner des yeux*.

Si l'on veut que le dictionnaire serve à la production de texte, l'information qu'une base donnée sélectionne un collocatif donné doit nécessairement être consignée dans l'article de la base, comme le DEC l'a toujours fait systématiquement (cf. aussi Béjoint et Thoiron, 1987 ; Benson, 1989, 1990 ; Benson *et al.*, 1986a-b) : le locuteur part du sens de la base, car celle-ci a un sens plus perceptible et il sélectionne le lexème correspondant. Ensuite, il cherche un autre lexème qui exprime un sens donné auprès de la base. Par exemple, le locuteur qui veut savoir comment se dit 'fermer et ouvrir rapidement les yeux' part du lexème ŒIL (YEUX) et cherche un verbe décrivant ce mouvement particulier des yeux. Il devrait trouver le collocatif CLIGNER, sous l'article ŒIL, ce qui nous donne la collocation *cligner des yeux*. Par conséquent, la description de la collocation est nécessairement incluse dans l'entrée de la base. Dans le DEC, elle est décrite au moyen d'une FL non standard dans l'article de ŒIL.

Nous présentons maintenant les différentes possibilités que le DEC nous offre pour décrire le collocatif CLIGNER pris seul, à savoir : 1) inclure la description du collocatif dans l'article de la base sans attribuer d'article autonome au collocatif ; 2) créer un article de dictionnaire autonome pour le collocatif et ne présenter, dans l'article de la base, que la collocation et les propriétés qui concernent la base. Les chiffres arabes encadrés ([1],[2], ...) renvoient le lecteur à certaines explications et justifications soulevées par nos descriptions. Ces commentaires figurent immédiatement après les deux propositions suivantes :

– Proposition 1 : Nous envisageons ici la possibilité de ne pas créer d'article pour le collocatif CLIGNER et, par conséquent, nous en assurons la description entière dans l'article de la base ŒIL. On obtient le résultat suivant :

ŒIL, nom, masc [pl *yeux*].
...
I.1a. *Yeux* (*Z*) *de X* [*permettant de voir Y*] = Deux parties du visage**I.a** d'une personne X, symétriques par rapport au nez**I.1a**, chacune étant constituée d'un globe mobile ...— organe de la vue de X, qui permet**II** à X de voir Y.
...

Fonctions lexicales

Syn	: **pop** coquillards, **pop** mirettes, **pop** quinquets
A_0	: oculaire [*globe oculaire*]

...

Mouvements et positions des yeux

les paupières couvrent
le globe oculaire : [$ART_{déf}$ ~] se fermer

fermer à demi les Y.
pour mieux voir : cligner [les / des ~] ***{1er groupe ; pas de passif}*** 1, plisser ***{1er groupe; pas de passif}*** [les ~] | O. au pl et n'a pas de dépendant [*Les myopes clignent des yeux*]

F_1 = fermer et ouvrir rapidement et involontairement plusieurs fois les Y. (— $Sympt_{21}$(*fatigue, nervosité*) ou Excess(lumière)) : cligner 2 ***{1er groupe; pas de passif}***, ciller ***{1er groupe; pas de passif}***, clignoter ***{1er groupe; pas de passif}*** [les / des ~] | O. au pl et n'a pas de dépendant [*Le soleil traversant les nuages la force à cligner des yeux*]

3S_0F1 : ***clignement, cillement, clignotement [des ~] | O. au pl et n'a pas de dépendant***

4$Oper_1(S_0F_1)$: ***// faire [des N] [Les moustiques lui faisaient faire des clignements des yeux continuels]***

3$Magn^{temp}F_1$: ***sans cesse***

Gestes impliquant les yeux

F_2 = en regardant une personne Y qui regarde X, X fait un signe d'entente à Y en fermant et ouvrant rapidement une fois un O. : ⌜cligner de l'~⌝ 5 ***{1er groupe}***, ⌜faire de l'~⌝ [à N = Y] | O. au sg [*Quand il est entré dans la salle, elle a cligné de l'œil à son patron*]

S_0F_2 : ***⌜clin d'~⌝ // œillade***

– Proposition 2 : Ici, nous procédons à la création d'un article de dictionnaire pour le collocatif CLIGNER. En d'autres termes, il s'agit de distribuer les informations pertinentes à la base de la collocation dans l'article de ŒIL et les informations pertinentes au collocatif lui-même dans l'article de CLIGNER. Si l'on compare avec la proposition précédente, on y constate le retrait de toutes les propriétés qui ne concernent que CLIGNER et seule la collocation comme telle demeure consignée dans l'article de ŒIL. Ainsi, par exemple, le pluriel obligatoire de ŒIL et son impossibilité d'avoir des dépendants apparaîtront dans la description de la collocation CLIGNER DES YEUX (qui figure sous ŒIL), mais les indications morphologique (1er groupe) et syntaxique (pas de passif) trouveront leur place dans l'article du collocatif CLIGNER. Ainsi les entrées de ŒIL, de CLIGNER et du quasi-phrasème CLIGNER DE L'ŒIL auront la forme suivante :

ŒIL, nom, masc [pl *yeux*].

...

I.1a. *Yeux (Z) de X [permettant de voir Y]* = Deux parties du visage**I.a** d'une personne X, symétriques par rapport au nez**I.1a**, chacune étant constituée d'un globe mobile ...— organe de la vue de X, qui permet**II** à X de voir Y.

...

Fonctions lexicales

Syn : **pop** coquillards, **pop** mirettes, **pop** quinquets

A_0 : oculaire [*globe oculaire*]

...

Mouvements et positions des yeux

les paupières couvrent le globe oculaire : [$ART_{déf}$ ~] se fermer

fermer à demi les Y. pour mieux voir : cligner**1** [les / des ~], plisser [les ~] ! O. au pl et n'a pas de dépendant [*Les myopes clignent des yeux*]

fermer et ouvrir rapidement et involontairement plusieurs fois les Y. (— $Sympt_{21}$(*fatigue*, *nervosité*) ou Excess(lumière)) : cligner**2**, ciller, clignoter [les / des ~] ! O. au pl et n'a pas de dépendant [*Le soleil traversant les nuages la force à cligner des yeux*]

Gestes impliquant les yeux

en regardant une personne Y qui regarde X, X fait un signe d'entente à Y en fermant et ouvrant rapidement une fois un O. : ⌜cligner de l'~⌝, ⌜faire de l'~⌝ [à N = Y] ! O. au sg [*Quand il est entré dans la salle, elle a cligné de l'œil à son patron*]

CLIGNER, verbe, 1er groupe, pas de passif.

1. *X cligne Y* = X ferme les yeux**I.1a** Y à demi pour mieux voir.

Régime

X = 1	Y = 2
1. N	1. N obligatoire

1) C_2 : ART = *les, des* [6]
2) C_2 : N = *yeux* [7]
$C_1 + C_2$: *Elle cligne les <des> yeux*

Fonctions lexicales

Syn : plisser

Exemples

Il regardait longuement le tableau en clignant des yeux.

2. *X cligne Y* = X ferme et ouvre rapidement et involontairement plusieurs fois les yeux**I.1a** Y sous l'effet de la lumière ou par fatigue**I.1a** ou nervosité.

Régime
Mod1

X = 1	Y = 2
1. N	1. N obligatoire

1) C_2 : ART = *les, des*
2) C_2 : N = *paupières, yeux*
$C_1 + C_2$: *Elle cligne les <des> yeux*

Mod2

Y=1
1. N

1) C_1 : N = *paupières, yeux*
C_1 : *Les yeux de Danielle clignent*

Fonctions lexicales

Syn : ciller, clignoter
S_0 : clignement [*Les moustiques lui faisaient faire des clignements des yeux continuels*]
$Magn^{temp}$: sans cesse

Exemples

Le soleil traversant les nuages la force à cligner les yeux.

CLIGNER DE L'ŒIL, loc. verbale, 1er groupe.
X ⌜*cligne de l'œil*⌝ *à Y* = En regardant une personne Y qui regarde X, X fait un signe d'entente à Y en fermant et ouvrant rapidement une fois un œil**I.1a**.

Régime

X = 1	Y = 2
1. N	1. *à* N

Fonctions lexicales

Syn : ⌜faire de l'œil⌝ [à N]
S_0 : ⌜clin d'œil⌝, œillade

Exemples

Quand Alain est entré dans la salle, Lida a cligné de l'œil à Igor.

Nous passons maintenant aux commentaires soulevés par notre démarche :

[1] Si CLIGNER n'a pas d'article autonome, il faut ajouter une description de ses propriétés morphologiques (son groupe de conjugaison) et syntaxiques (l'interdiction du passif) du collocatif dans la description de la collocation CLIGNER DES YEUX. Les informations propres au collocatif sont saisies dans une fonte différente, « ***Berling Roman*** », et figurent entre accolades, à la suite de la valeur de la FL.

[2] CLIGNER LES/DES YEUX a comme synonyme CLIGNER DES PAUPIÈRES. Strictement parlant, cette dernière doit être recensée dans l'article PAUPIÈRE. L'inconvénient de cette démarche est que l'égalité des sens n'est plus transparente : la collocation CLIGNER LES/DES PAUPIÈRES pourtant synonyme n'est plus directement disponible à partir de CLIGNER LES/DES YEUX.

[3] Si CLIGNER n'a pas d'article autonome, les valeurs correspondant aux relations paradigmatiques et syntagmatiques de la collocation entière ne peuvent être consignées ailleurs que dans l'article de ŒIL. Ainsi, les descriptions du dérivé CLIGNEMENT et du modifieur SANS CESSE deviennent en quelque sorte des « sous-entrées » de la collocation CLIGNER DES YEUX.

[4] La description du verbe support FAIRE relève de la combinatoire lexicale restreinte des substantifs CLIGNEMENT, CLIGNOTEMENT et CILLEMENT et non pas de celle de la collocation. Ces expressions ne concernent plus le collocatif CLIGNER et encore moins la lexie ŒIL dont il est question au départ. Cette façon de faire nous force à introduire des articles gigognes qui finalement n'ont pas leur place dans l'article de la lexie ŒIL. Le DEC interdit formellement cette procédure.

[5] L'expression CLIGNER DE L'ŒIL est considérée comme un quasi-phrasème, donc avec le droit à un article autonome. On y trouve le sens ‹cligner› et le sens ‹œil› et aussi un ajout de sens majeur ‹X fait un signe d'entente à Y›. À la différence de CLIGNER1 DES YEUX et de CLIGNER2 DES YEUX où le locuteur part du sens ‹œil› pour trouver un verbe qui décrive un mouvement particulier des yeux, ici le locuteur part du sens ‹signe d'entente›, le sens ‹œil› n'occupant pas une position centrale dans le sens de l'expression. Le problème qui se pose est de voir s'il faut présenter un quasi-phrasème comme valeur d'une FL. Ce procédé est contradictoire : d'une part,

les FL ne rendent compte que des collocations et jamais des quasi-phrasèmes, mais, d'autre part, étant donné que le sens 'œil' est quand même présent dans le sens de toute l'expression, il semble nécessaire de l'introduire sous l'entrée ŒIL.

6 Par contre, les faits que ŒIL ne s'emploie qu'au pluriel et que les seuls déterminants permis sont *les* et *des* sont des propriétés qu'il faut indiquer. Ce genre de restriction n'a jamais été introduit pour les lexies de plein droit : leurs déterminants sont déduits par des règles générales de la langue. Cette nouvelle dimension de caractérisation n'est pas étonnante : les déterminants sont très souvent des victimes de la phraséologisation.

7 Il faut spécifier que le deuxième actant 'yeux' ne se réalise en surface que par le lexème YEUX et aucun autre. Cette restriction doit être introduite pour interdire des expressions impossibles comme **cligner des mirettes <des coquillards, des quinquets>*.

Suite aux différentes solutions exposées, nous proposons une liste de vérification qui indique les points importants qu'il faut considérer lors de l'élaboration des descriptions lexicographiques des collocatifs. L'idée d'une telle liste a été proposée par Hudson (1988) puis reprise par Dostie *et al.* (1992) et par Mel'čuk *et al.* (1995). Nous nous en servons à notre tour : une telle liste est utile pour signaler au lexicographe les propriétés qui doivent être prises en considération pour décider du traitement approprié des collocatifs. Signalons que cette liste est tout à fait provisoire et qu'elle doit être utilisée avec prudence.

1. L_i, le collocatif, est-il à base unique ?
Si oui, vérifiez :
- a. s'il a des contreparties libres :
  - si oui, ne pas créer d'article (*noir, plate*) ;
  - si non, en créer un (*bissextile, cligner*).
- b. s'il permet l'héritage :
  - si oui, créer un article (*tresse blonde*) ;
  - si non, ne pas en créer un (**arabica noir*).
- c. s'il est polysémique et si tous les polysèmes sont des lexies collocatives :
  - si oui, créer un article (*cligner, blond*).

Si non, vérifiez :
- a. s'il a des contreparties libres :
  - si oui, ne pas créer d'article (*endurci*) ;
  - si non, en créer un (*invétéré*).
- b. s'il permet l'héritage :
  - si oui, créer un article (*sauce forte à partir de goût fort*) ;
  - si non, ne pas en créer un (**voleur endurci*).
- c. s'il est polysémique :
  - si oui, créer un article (*fort*).
- d. si ses bases peuvent être regroupées sémantiquement :
  - si oui, créer un article.
- e. si on peut généraliser ses propriétés :
  - si oui, créer un article.

2. L_i, le collocatif, a-t-il des contreparties libres ?
Si oui, vérifiez :
a. s'il a des propriétés particulières à l'emploi collocationnel :
si oui, créer un article.

3. L_i est-il décrit par une FL standard ?

4. À quel degré L_i est-il phraséologisé ?
Si moindre degré, créer un article (*conduire*) ;
Si degré plus élevé, vérifiez :
a. s'il a des contreparties libres, il faut déterminer la distance sémantique avec la lexie de base:
plus distant, ne pas créer d'article (*steak bleu*).

5. L_i est-il autonome ?

Remerciements

Nous ne pourrions terminer cet article avant d'avoir adressé nos remerciements à M. Igor Mel'čuk qui a lu notre article et avec qui nous avons discuté plusieurs points dans le texte. Nous lui sommes reconnaissantes de tout l'intérêt qu'il a manifesté envers notre sujet et de son enthousiasme contagieux pour le travail. Nous remercions également Alain Polguère et Agnès Tutin, nos deux victimes de ce que nous pourrions appeler du harcèlement linguistique : ils nous ont apporté des exemples et des précisions au moment où c'était nécessaire. Merci aussi à Danielle Collignon pour sa patience et sa disponibilité.

Vers un nouvel outil interactif d'aide à la conception de dictionnaires électroniques spécialisés

Christophe JOUIS[a,b] et Widad MUSTAFA-ELHADI[a]

a) UFR Information, Documentation, Information Scientifique et Technique, Université Charles De Gaulle-Lille III, France
b) Centre d'Analyse et de Mathématiques Sociales, Unité Mixte CNRS, EHESS, Université Paris-Sorbonne, France

Introduction

Concevoir un dictionnaire informatisé des termes d'un domaine de connaissance spécialisé n'est pas une tâche simple pour le concepteur, notamment lorsqu'il n'a pas ou peu de connaissances sur le domaine concerné. Il doit ménager des entrevues avec des spécialistes du domaine, puis analyser ces textes d'entrevues, les glossaires déjà existants, la documentation technique, etc. Il s'agit de concevoir une base de données terminologiques structurée dont chaque entrée (un terme du domaine) contient en particulier sa définition, son contexte d'utilisation et des pointeurs (hypertextes) vers d'autres entrées du dictionnaire sous forme de liens sémantiques (synonyme, hyponyme, partie/tout, localisations spatiales, temporelles, etc.). Autrement dit, le concepteur doit se construire une représentation du domaine, sa structure conceptuelle et les relations que les concepts entretiennent entre eux à partir d'un ensemble de documents textuels.

Pour analyser ces documents textuels, nous proposons une méthode linguistique et informatique : l'exploration contextuelle. Cette méthode est fondée sur le noyau de connaissances linguistiques du concepteur qui analyse des textes sans connaissances préalables sur le domaine concerné. Ce modèle se focalise non pas sur les termes spécifiques du domaine (traditionnellement appelé « mots pleins »), mais sur les connecteurs (les mots « vides » ou mots « pivots ») entre les termes. Ces derniers constituent un ensemble d'indicateurs qui sont chargés de significations. Ils traduisent *un savoir linguistique indépendant d'un domaine de connaissance particulier.*

Ainsi, nous partons de l'hypothèse suivante : *les textes contiennent des unités linguistiques qui sont des indicateurs pertinents pour structurer des connaissances.*

Le concepteur s'appuie sur ces unités linguistiques ; il utilise une stratégie d'exploration contextuelle lorsqu'il analyse les documents.

Le système SEEK[1] est une application de l'exploration contextuelle qui se présente sous la forme d'un système interactif à base de connaissances. Le système isole *d'abord* des relations sémantiques dans des textes entre les concepts d'un domaine. Puis, *dans un deuxième temps*, c'est-à-dire lorsqu'une relation est détectée, SEEK recherche les termes (concepts), arguments de la relation.

Nous testons actuellement une deuxième version de SEEK qui a comme objectif la détection des contextes définitoires à l'aide de marqueurs linguistiques.

1. Schéma de principe de construction d'un dictionnaire électronique spécialisé

Notre objectif consiste à l'élaboration d'un outil d'aide interactif visant à la construction d'un dictionnaire informatique des termes spécifiques à un domaine de connaissance à partir d'un corpus de textes décrivant le domaine en question. La structure d'un tel dictionnaire électronique peut être vue comme une base de données terminologiques dont les entrées sont les termes du domaine. Chaque enregistrement de la base décrit un terme du domaine. La description est alors composée de liens hypertextes vers des zones textuelles du corpus et de liens étiquetés vers d'autres termes (ou enregistrements) de la base (figure 1). Plus précisément, la structure d'un enregistrement est alors formée : (i) de liens hypertextes vers les contextes définitoires du terme dans le corpus (voir § 3) et (ii) d'une liste de pointeurs vers d'autres enregistrements de la base : des relations sémantiques (figure 2).

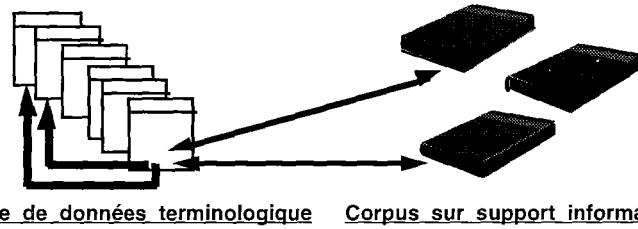


FIGURE 1 : Base de données terminologiques : liens hypertextes et liens étiquetés

Notre approche prend ses sources dans un modèle global de traitement du langage : la Grammaire Applicative et Cognitive (GAC[2]). La GAC articule plusieurs niveaux de représentations, et en particulier un niveau cognitif où l'on analyse les significations des unités linguistiques sous forme de représentations sémantico-cognitives afin de construire les représentations des connaissances associées à un texte. La GAC propose un ensemble de concepts sémantiques qui définissent un système organisé de significations. Nous distinguons les **types sémantiques** des unités linguistiques,

1. SEEK : Système Expert d'Exploration (K)contextuelle
2. Pour une description de la GAC, les auteurs renvoient par exemple à Desclés (1990), Jouis (1993).

des **relations statiques** fondamentales et des **relations évolutives** (mouvement, changement d'état, conservation d'un mouvement, itération, variation d'intensité, contraintes, causes...).

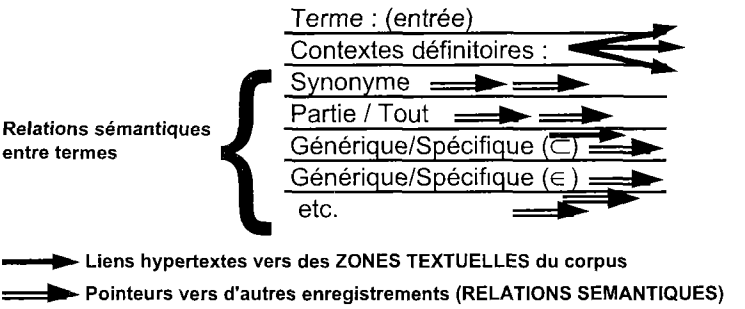


FIGURE 2 Structure d'un enregistrement.

Dans la suite, nous ne détaillerons qu'une partie des primitives de la GAC : les types sémantiques et les relations statiques car ce sont ces primitives qui sont pour le moment utilisées dans SEEK.

1.1. Les types sémantiques

Nous distinguons un certain nombre de types élémentaires permettant de classer les entités manipulées[3].

- Les entités individualisables sont celles que l'on peut désigner et montrer par pointage. On peut les compter individuellement ou les regrouper en classes. Par exemple, les entités *Jean, table, chaise, meuble, homme, enfant* sont individualisables.
- Les entités massives telles que *eau, mer, sable, vin, beurre, blé* ne sont pas des entités individualisables. Notons cependant qu'un certain nombre d'opérateurs (des classificateurs) permettent de rendre individualisable une notion massive : un *verre d'eau, un pâté de sable, un bras de mer, une bouteille de vin, un morceau de beurre.*
- Les classes distributives rassemblent des entités individuelles ayant une même propriété. Par exemple, *être-un-carré* représente une classe d'individus (ou « concept »).
- Les classes collectives se distinguent des entités individualisables parce qu'elles représentent des objets qui forment un « tout » à partir d'objets plus élémentaires. Ainsi, *foule, armée, armada, famille* sont des classes collectives.
- Le type des lieux représente des étendues ou des regroupements de positions d'une même entité (individualisable, collective ou massive) : *Paris, jardin, maison.*

3 Dans la GAC, nous pouvons construire des types plus complexes à partir des types élémentaires en utilisant des opérateurs formateurs de type : listes, n-uplets, types fonctionnels, etc. (Desclés, 1990).

1.2. Les relations statiques

Les relations statiques sont hiérarchisées et indépendantes d'un domaine particulier. Ce sont des relations binaires. Les relations statiques permettent de décrire les situations statiques du domaine, qui restent stables pendant un certain intervalle temporel où ni début, ni fin ne sont envisagés. Nous distinguons plus d'une vingtaine de relations statiques, en particulier :

- les identifications de deux entités,
- les incompatibilités entre deux entités,
- les dimensions (mesures, etc.),
- les cardinalités,
- les comparaisons de valeurs,
- les inclusions entre classes distributives (relations générique/spécifique),
- les appartenances d'une entité individuelle à une classe distributive (relations générique/spécifique),
- les relations « partie/tout » entre classes collectives,
- les localisations d'une entité par rapport à un lieu (intérieur, extérieur, frontière, fermeture, orientations, etc.).

La sémantique de chaque relation est définie par trois types de propriétés : (i) son type fonctionnel (le type sémantique des arguments de la relation) ; (ii) ses propriétés algébriques : réflexivité, symétrie, transitivité, etc. ; (iii) ses propriétés d'agencement avec les autres relations dans un même contexte. Les relations statiques s'insèrent dans un système de significations des relations de repérage entre entités[4].

En général, les systèmes de repérage de termes et de relations sémantiques entre termes sont fondés sur des approches syntaxiques, statistiques et/ou connexionnistes. Aussi, ils nécessitent pour fonctionner un dictionnaire général de la langue et parfois un lexique des termes techniques ainsi que dans certains cas une représentation sémantique préexistante spécifique au domaine de connaissances à modéliser. Dans notre problématique, c'est justement ce dernier point que nous cherchons à construire de manière semi-automatique (outil d'aide interactif).

2. Notre proposition : appliquer la méthode d'exploration contextuelle

2.1. Présentation générale de l'exploration contextuelle[5]

Pour l'analyse rapide de textes en grand nombre l'exploration contextuelle vise à construire des représentations sémantiques en se contentant d'une analyse syntaxique ap-

4 Le repérage, noté REP (ou « ε »), est un schéma général de relation : une entité X (une entité repérée) est repérée par rapport à Y (une entité repère). Le repérage se spécifie suivant les propriétés algébriques qui lui sont attribuées axiomatiquement en divers relateurs. Sur ce point, voir Desclés (1987).

5 L'exploration contextuelle a initialement été appliquée pour le traitement du temps et de l'aspect : un module de détection des valeurs sémantiques des temps de l'indicatif en français a été réalisé. Nous ne détaillerons pas ce module qui est décrit par ailleurs : Oh *et al.* (1992), Jouis (1993). D'autres applications de l'exploration contextuelle sont actuellement réalisées : résumé automatique (Leroux, Minel et Berri, 1994), analyse de la causalité (Garcia et Jackiewicz, 1995).

proximative. L'exploration contextuelle repère certains marqueurs jugés pertinents dans les phrases analysées qui deviennent des « pivots » pour établir des relations entre concepts.

D'une façon générale, un système d'exploration contextuelle se ramène à : (i) identifier des indices linguistiques pertinents (pour le problème à résoudre) ; (ii) définir un système de valeurs sémantiques ; (iii) expliciter un ensemble de règles de décision qui permettent d'associer à des cooccurrences d'unités linguistiques la valeur sémantique adéquate en fonction du contexte ; (iv) organiser l'ensemble sous forme d'un système de prise de décision sur la valeur sémantique.

L'exploration contextuelle a pour objectif de simuler une « lecture » rapide d'un texte consistant à rechercher des indices linguistiques qui permettent de construire un réseau de concepts. La « compréhension » superficielle d'un texte aboutit à la prise de décisions fondées sur le repérage des indices (marqueurs linguistiques de relations sémantiques) coprésents dans le texte. Un système d'exploration contextuelle se ramène à un ensemble de règles déclaratives qui expriment un savoir décisionnel interprétatif. Les règles (d'exploration contextuelle) se représentent sous la forme **SI** <conditions> **ALORS** <actions> ou <conclusions>. Les conditions des règles expriment la coprésence ou non d'unités linguistiques pertinentes dans le contexte. Ces indices touchent plusieurs composantes simultanément : morphologique, syntaxique, lexicale. Les conclusions de l'ensemble des règles permettent de construire progressivement des représentations sémantiques.

2.2. SEEK : une application de l'exploration contextuelle

SEEK permet, *dans un premier temps*, de rechercher des relations statiques dans des textes entre les objets d'un domaine de compétence[6]. Il produit une représentation visuelle sous forme de graphes objets/relations. Des liens hypertextes permettent de retrouver, pour chaque relation trouvée, la zone du texte ayant servi à sa construction. SEEK se présente sous la forme d'un système à base de connaissances. Il fonctionne à l'aide de règles d'exploration contextuelle, dont le but est de rechercher dans les textes des indices textuels qui permettent d'identifier une relation statique particulière, puis les arguments de la relation.

Par exemple, parmi les listes d'indices concernant la description des composants d'un objet dans un domaine (relation partie/tout ou « ingrédience ») nous avons la règle suivante, associée aux listes de marqueurs linguistiques VDecomp1, VDecomp2, VDecomp3[7] :

---

6 La base de connaissances de SEEK concernant la détection des relations statiques est composée d'une base de données de marqueurs statiques (quelques 3 300 marqueurs classés dans 240 listes) et de 220 règles d'exploration contextuelle. Cette première version de SEEK a été testée dans un contexte industriel (EDIAT/CGI, ALSTHOM/IPSé). Par ailleurs, il est soumis à une évaluation dans le cadre d'une action de recherche concertée (ARC) soutenue par l'AUPELF-UREF « Évaluation des systèmes de construction automatique de terminologie et de relations sémantiques entre termes à partir de corpus »

7 Ces indices sont utilisés dans le système SEEK (Jouis, 1993). Les listes et les règles d'exploration contextuelle concernant la relation partie-tout ont été identifiées par Agata Jackiewicz-Desbertrand *Contribution aux problèmes de l'extraction des connaissances : manifestations linguistiques et représentation informatique de la relation d'ingrédience*, Mémoire de DEA Sciences Cognitives, EHESS, 1992. Pour des raisons de place, nous ne donnerons qu'une partie du contenu des listes de marqueurs linguistiques.

SOIT	x1, x2 des unités linguistiques ; P une proposition
SI	x1 est un marqueur de la liste **VDecomp1**
	OU **VDecomp2**
	OU **VDecomp3**
	ET x2 est l'unité linguistique « **en** »
	ET x1 x2 se suivent (pas forcément immédiatement) dans la même proposition P
ALORS	Proposer la relation **partie/tout** dans P

VDecomp1 = {analyser, décomposer, démembrer, désagréger, désassembler, désintégrer, désunir, détailler, disjoindre, dissocier, réduire, séparer...} ;
VDecomp2 = {bissecter, cloisonner, compartimenter, diviser, émietter, fragmenter, graduer, morceler, parceller, partager, scinder, sectionner, subdiviser, tronçonner...} ;
VDecomp3 = {couper, découper, déchirer, fendre, trancher, casser, disloquer, rompre...}.

Repérés dans un énoncé tel que :

> (...) la zone d'accessibilité se **décompose en** deux partie : la zone d'accessibilité principale et la zone d'accessibilité secondaire.

ces indices permettent d'orienter la décision vers la relation partie/tout. Ensuite, dans cet énoncé (où une relation a été détectée), SEEK tente d'identifier les arguments (termes) de la relation. Dans une proposition où une relation est détectée, les composants d'un argument de la relation (ou terme) (figure 3) :

- n'appartiennent pas à une liste de marqueurs statiques ;
- ne sont ni des verbes, ni des adverbes, ni des pronoms, ni des prépositions, ni des « mot-phrases », ni des conjonctions, etc.

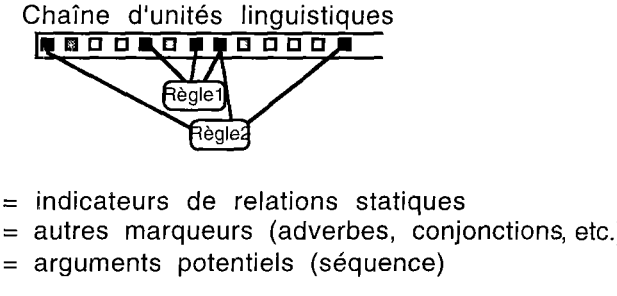


FIGURE 3 : Identification des termes par contraste.

Ainsi, grâce à la règle d'exploration contextuelle de détection d'argument ci-dessous :

SOIT	x1, x2, x3, x4 unités lexicales ; S une proposition
SI	une relation statique est détectée dans S
	ET x1 est l'unité linguistique [le] ET x2 est un terme primitif potentiel
	ET x3 est l'unité linguistique [de] ET x4 est un terme primitif potentiel
	ET x1 x2 x3 x4 se suivent immédiatement dans S
ALORS	Proposer x1+x2+x3+x4 comme argument dans S

on identifie l'argument [la zone d'accessibilité], comme argument possible pour la relation partie/tout. Finalement, on aboutit, après l'enchaînement de plusieurs règles puis validation par l'utilisateur, à une représentation sous forme de graphe, où sont mémorisées à l'aide de liens les zones textuelles qui ont permis de construire la représentation (figure 4).

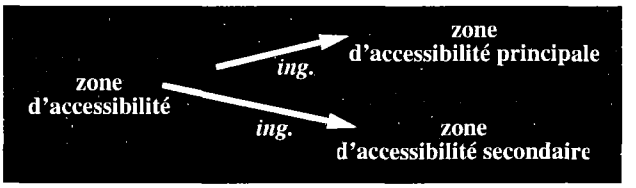


FIGURE 4. Représentation issue de SEEK

3. Extension de SEEK à la détection des contextes définitoires

Dans la perspective de la construction d'une terminologie à partir de textes de taille importante, une question se pose alors : où chercher les relations et les termes pertinents dans les textes ? Il s'agit de donner la définition de chaque concept qui est repérable dans : (i) des **zones de textes éparpillées** dans lesquelles SEEK aura repéré les relations statiques et les termes (les arguments de la relation) ; (ii) mais aussi des **zones textuelles privilégiées** associées à un terme dans lesquelles se trouvent rassemblées des informations concernant ce terme : ses **contextes définitoires.**

3.1. La définition en terminologie spécialisée

Les concepts constituent la base de toute construction terminologique. Un concept est une unité de connaissance qui comprend des propositions vérifiables concernant un élément de référence choisi, exprimé par un *terme*. La terminologie s'occupe des concepts et en conséquence des structures de la connaissance dans la mesure où celles-ci sont exprimées par le lexique des langues. Les unités lexicales ne deviennent termes que si elles sont *définies* et employées dans un contexte donné. Une théorie sur la terminologie est nécessairement liée à une structure référentielle[8] qui met en évidence les relations entre les termes et leurs référents et permet une définition correcte de l'ensemble des éléments constitutifs d'une structure référentielle.

La définition en terminologie a pour objet la délimitation du concept dans un réseau conceptuel d'un domaine donné et sert à exprimer la signification d'un terme. Ce n'est pas le terme qui fait l'objet de la définition, c'est plutôt le concept désigné par le terme. Définition, contexte, formules (mathématiques, chimiques, etc.), figures, représentations graphiques, images, etc. sont tous considérés comme constituant une information sur le concept.

A. Rey (1979 : 40) fait remarquer le lien fondamental entre le terme et sa définition. Il précise qu'ils désignent à l'origine « l'assignation d'une limite, d'une fin (*définir*) et son résultat (*terme*) ». Même si les problèmes de la définition (que ce soit en

8 Voir Dahlberg (1978, 1981)

terminologie ou en lexicologie) ont été abordés dans diverses publications[9], le champ de ces études reste largement limité aux traitements intellectuel et manuel. Peu de publications portent sur les indices de reconnaissance automatique et/ou semi-automatique des définitions.

La question que nous posons ici est de savoir comment le système doit procéder afin de repérer la triade *terme-concept-définition* lors d'une construction automatique d'une terminologie. Autrement dit, quels sont les marqueurs linguistiques susceptibles d'aiguiller le système vers les termes et les définitions ? Nous postulons que certains marqueurs serviraient à la fois pour détecter les termes et les définitions : *...se définit par..., ...se nomme..., ...c'est-à-dire....*, par exemple. Le système doit reconnaître dans une même proposition le terme et sa définition.

Nous retiendrons la distinction faite par Béjoint (1993) entre les « définitions textuelles » qu'on trouve dans les textes, manuels, articles, etc., et « les définitions dictionnairiques » qu'on trouve dans les dictionnaires. Les premières nous intéressent dans une optique de recherche de marqueurs linguistiques que SEEK va utiliser pour détecter les définitions dans les corpus. Les deuxièmes nous intéresseront en tant que « modèle ». Autrement dit, comment doit-on « aider » SEEK à compléter « idéalement » les définitions qu'il aura repérées automatiquement ?

La définition doit retenir les caractéristiques essentielles du concept et permet ainsi de le distinguer à l'intérieur d'un système conceptuel reflétant un domaine de connaissance. La littérature sur la définition « idéale » en terminologie ne manque pas. L'ISO a établi un certain nombre de règles concernant les définitions, (voir Rondeau, 1981 : 84). Mais nous savons que les règles ne sont pas toujours respectées dans les dictionnaires spécialisés. La définition par le *genre proche* et la *différence* ou la définition aristotélicienne semble être la mieux adaptée à la terminologie, car elle rend compte des relations hiérarchiques qui existent entre les concepts. Il faut signaler cependant que ce type de définition n'est pas toujours de mise ; car il y a des concepts qui s'y prêtent mal. Et c'est un des problèmes qui caractérisent la définition des concepts appartenant à des domaines techniques, ceux appartenant aux sciences exactes poseraient moins de problèmes.

Un autre problème que pose la définition par le genre et la différence est celui du choix du genre et en particulier comme nous l'avons signalé, quand il s'agit d'un domaine technique ou les genres proches s'incluent les uns les autres. Ce problème a été aussi soulevé par Béjoint (1993) qui établit une typologie de définitions génériques et montre comment leur « généricité » est douteuse.

3.2. Règles pour le repérage des termes

3.2.1. Reformulation et définition

Comme nous l'avons déjà signalé, le système doit reconnaître dans une même proposition le terme et sa définition. Nous avons établi des règles pour le repérage de termes

9 Voir Actes du colloque *La définition*, organisé par CELEX (Centre d'étude du lexique) de l'Université Paris-Nord, 1988 , *Problèmes de la définition et de la synonymie en terminologie*, Actes du Colloque international de terminologie, Université Laval, Québec mai 1982, Sager (1990 et 1982), entre autres

et les relations entre concepts. Nous exposons dans cette partie les marqueurs appropriés à la détection des contextes définitoires.

En dehors de marqueur linguistique générique tel que *définir*[10], qui détecte formellement la présence de définitions, d'autres marqueurs peuvent être retenus. Il s'agit d'éléments linguistiques qu'on trouve dans les *reformulations* (Chukwu et Thoiron, 1989 ; Thoiron et Béjoint, 1991 ; Chukwu, 1993 ; Otman, 1989). La *reformulation* peut être définie comme le processus qui permet de passer d'une forme linguistique à une autre forme porteuse de la même signification, ou de mettre en rapport des variantes d'expression que l'on souhaite considérer comme équivalentes (voir Dachelet, 1990). Pour Chukwu et Thoiron (1989 : 24) la reformulation pourrait être mise en relation avec d'autres concepts voisins tels que *paraphrase, définition, synonymie, équivalence, anaphore* et *description*. Tout en montrant la parenté entre définition et reformulation les auteurs proposent de distinguer les définitions (celles qui font référence à des documents antérieurs faisant autorité tels que les normes, les dictionnaires, les citations, etc.) des simples reformulations qui ne sont pas toujours « suffisantes » pour prétendre à des définitions à part entière (voir aussi à ce propos Chukwu, 1993 : 266).

3.2.2. Les marqueurs susceptibles de détecter des définitions

Pour établir les listes de marqueurs selon leur nature nous nous sommes largement inspirés de la classification donnée par Chukwu et Thoiron (1989), Thoiron et Béjoint (1991), Chukwu (1993), Otman (1989).

i) Les marqueurs apparaissant dans les reformulations copulatives identifiées dans Chukwu et Thoiron (1989) et Chukwu (1993) constituent une catégorie largement repérable par les marqueurs de relations statiques dans SEEK (voir la liste *supra*).

ii) Les marqueurs qui apparaissent dans les reformulations métalinguistiques appellatives (directes et inverses) identifiées dans Chukwu et Thoiron (1989). Ce sont des verbes qui appartiennent d'après les auteurs au champ sémantique de la dénomination (*appeler, nommer, désigner,* etc.) ou des expressions telles que *c'est-à-dire, autrement dit* qui signalent une explication.

iii) Les marqueurs apparaissant dans les reformulations explicatives, ou encore dans les reformulations définitoires, catégorie identifiée par Chukwu et Thoiron (1989) et Chukwu (1993), semblent être une catégorie productive pour le repérage de définitions. Chukwu et Thoiron reconnaissent à cette catégorie une fonction explicative délibérée « [...] le terme est introduit dans le discours et appelle immédiatement une définition, une explication (c'est nous qui soulignons) ou tout autre procès qui en facilite la compréhension » (Chukwu et Thoiron, 1989 : 28).

iv) Les marqueurs apparaissant dans les reformulations métalinguistiques explicatives regroupent aussi des marqueurs tels que *c'est-à-dire* ainsi que les verbes faisant ré-

10 Otman qualifie les marqueurs tels que *définir, appeler*, etc. comme étant des marqueurs « autosuffisants » pour assurer le repérage de *contextes définitoires et informatifs*. Ce terme (Otman, 1989 : 68) regroupe d'après l'auteur les notions de « contexte définitoire », « contexte encyclopédique », « contexte associatif », « contexte langagier » et « contexte métalinguistique », classification reprise par l'auteur à Boutin-Quesnel *et al.* (1985).

férence à des processus cognitifs tels que *entendre, comprendre* (Chukwu et Thoiron, 1989 : 36). Nous pouvons ajouter à cette liste aussi d'autres marqueurs que SEEK a déjà utilisé pour repérer les relations statiques *diviser, classer, distinguer, énumérer, signifier, désigner, représenter, au sens où*, etc.

De plus, nous ajoutons à ces marqueurs les listes que Otman (1989 : 71-72) a établies pour *dire, appeler, définir.*

– dire : *on dit que . , on dira que , on dit parfois que ..est, si ... on dit que c'est ., on dit que est .... si . , on dit de qu'ils sont..., nous dirons dans ce cas qu'il y a..., nous disons que . , on veut dire que. , est dit . , ...sont dits/dites , . .dits/dites, l'une dite l'autre dite... ,*
– appeler : *on appelle ., c'est ce qu'on appelle , . constitue/ent ce qu'on appelle .., dans le ...que nous appelons..., nous pourrions donc l'appeler..., ...est appelé(e)/sont appelés(es)..., l'appellation vient de ... ,*
– définir : *nous définissons ici les ...comme des ..., nous définissons que nous appelons..., est défini par ..., ...définit..., par définition...* (Listes non exhaustives).

3.3. Diversité de contextes définitoires et présence de facettes

La présence de plusieurs contextes définitoires pourrait, dans certains cas, être l'indicateur d'une facette. Autrement dit, le terme aura plus qu'une définition et par conséquent plusieurs contextes définitoires selon le point de vue ou facette sous lequel il a été envisagé. Il serait intéressant d'introduire des représentations multidimensionnelles ou à facettes du concept donné (voir Dahlberg, 1993 ; Sager, 1990 ; Mustafa-Elhadi, 1989).

3.4. Construction interactive d'une définition

La définition d'un concept est repérable dans des zones de textes éparpillées dans lesquelles SEEK aura déjà repéré les relations statiques et les termes et aussi des zones textuelles privilégiées associées à un terme dans lesquelles se trouvent rassemblées des informations concernant ce terme : ses contextes définitoires. Ces derniers sont repérables, comme pour les relations statiques, par exploration contextuelle grâce à divers indicateurs linguistiques.

À partir d'une compilation de l'ensemble des contextes définitoires il s'agit de construire ensuite une définition. Les règles d'exploration contextuelle proposent des contextes définitoires, mais il revient au spécialiste du domaine de les valider, de les intégrer et éventuellement de les compléter.

En effet, la définition terminologique doit faire apparaître les caractéristiques essentielles sans apporter des développements inutiles ou superflus. La plupart des auteurs sur la définition s'accordent sur le caractère exhaustif de celle-ci, car l'information contenue dans les définitions peut concerner divers utilisateurs potentiels qui puiseront selon leur besoin dans ce fond commun : « La définition de spécialité est une définition voulue exhaustive et dans le même temps restreinte à un seul domaine. [...] Aussi la longueur de la définition de spécialité qui est une définition nécessairement longue – ne présenterait aucun inconvénient, chaque utilisateur limitant sa recherche à la seule portion pour lui nécessaire de la définition » (Roman, 1993 : 116).

4. Conclusions

Nous avons proposé une démarche linguistique et informatique pour l'aide à la conception de bases de données terminologiques qui s'appuie sur un *savoir linguistique indépendant d'un domaine de connaissance particulier* pour repérer les définitions des concepts dans des zones de textes éparpillées. Cette démarche se focalise d'abord sur l'identification de relations statiques entre termes puis sur les termes eux-mêmes. Enfin, il s'agit de tenir compte de zones textuelles privilégiées associées à un terme dans lesquelles se trouvent rassemblées des informations concernant ce terme : ses *contextes définitoires*. Ces derniers sont repérables, comme pour les relations statiques, par exploration contextuelle grâce à divers indicateurs linguistiques.

Néanmoins, de nombreux problèmes restent en suspend : quelle serait la définition idéale ? Cette question nous conduit nécessairement à poser le problème de corpus. Quels sont les corpus « idéaux » pour une construction (semi) automatique d'une terminologie ? Est-ce que la nature des corpus (corpus descriptifs, manuels didactiques, etc.) est un paramètre déterminant quant aux performances du système ? Y a-t-il une typologie des définitions par domaine ?

La construction de dictionnaires à partir de l'analyse informatisée de corpus bruts : un outil pour le langagier

Sylvain DELISLE

Université du Québec à Trois-Rivières, Canada

1. Introduction

De nombreuses applications en ingénierie linguistique exigent la construction et la mise à jour de dictionnaires. Bien que les dictionnaires généraux soient largement disponibles, et de plus en plus sur un support adapté au traitement informatique (p. ex. CD-ROM), il en va tout autrement des dictionnaires spécialisés. Ces derniers couvrent habituellement une langue de spécialité, comme celle du monde médical ou de l'informatique, et s'adressent à un public beaucoup plus restreint. Le langagier est appelé à construire des dictionnaires spécifiques à partir d'un corpus donné, soit pour analyser un phénomène linguistique pointu, soit pour caractériser ou modéliser un (sous) langage de spécialité. Le corpus sert alors de source pour construire (ou « dériver ») le dictionnaire.

La construction d'un dictionnaire à partir d'un corpus est une tâche ardue et fastidieuse, surtout lorsque le corpus est de taille ou de complexité importantes. La plupart du temps, le langagier doit effectuer ce travail manuellement ou encore avec des moyens informatiques plutôt rudimentaires. Nous proposons une méthode qui permet de construire facilement un dictionnaire de base à l'aide d'un outil informatisé et ce, directement à partir d'un corpus composé de textes bruts ne nécessitant aucun traitement préliminaire. Cette méthode est indépendante du domaine du texte, est axée sur le verbe et ses arguments, et est basée sur l'analyse syntaxique (automatique) et sémantique (semi-automatique) de corpus bruts. Notre méthode a été implémentée et testée sur des textes anglais de nature technique. Elle est exportable à d'autres langues pour lesquelles des ressources similaires à celles décrites ci-dessous sont également disponibles.

Implémenté surtout en Prolog, le système CASSCAD[1] constitue un environnement informatique qui guide et supporte la tâche de l'usager (c.-à-d. le langagier) pendant la construction d'un dictionnaire ou pendant l'étude de phénomènes linguistiques particuliers. Les composantes du système sont organisées en trois sous-systèmes indépendants mais complémentaires : *i*) un concordancier, *ii*) un sous-système d'analyse lexicale « multi-source » et *iii*) un sous-système d'analyse syntaxique et sémantique.

Le sous-système d'analyse lexicale multi-source intègre plusieurs modules qui ont pour fonction de présenter à l'usager des informations de base sur chacun des mots trouvés dans le corpus (à partir du concordancier) afin d'en construire l'entrée à ajouter au dictionnaire spécifique en cours de construction : liste de concordances et fréquences ; fonctions grammaticales potentielles selon le dictionnaire général *The Collins* ; définition selon la base de données lexicale *WordNet*, à caractère général elle aussi ; et fonction grammaticale la plus probable selon un étiqueteur statistique. L'usager contrôle la construction de chaque nouvelle entrée lexicale à partir de ces diverses sources d'information. Quant au sous-système d'analyse syntaxique et sémantique, il permet de compléter les entrées verbales créées précédemment au cours de l'analyse lexicale. Ces deux analyseurs permettent d'identifier les patrons d'occurrence syntaxiques et les patrons d'occurrence sémantiques (thématiques) pour chaque verbe du corpus. L'analyseur sémantique compile également les fréquences d'occurrence de ces deux types de patrons. De plus, les jeux d'étiquettes thématiques sont adaptables aux exigences du langagier.

Cet article présente, à travers le système CASSCAD, l'essentiel de l'approche que nous proposons pour la construction semi-automatique d'un dictionnaire spécifique. Mais tout d'abord, voyons un exemple d'application en ingénierie linguistique pour lequel, trop souvent, le langagier est dépourvu d'outil informatique adéquat.

2. Exemple d'une approche manuelle en traduction spécialisée

Nous présentons ici un exemple d'application en ingénierie linguistique : la traduction spécialisée. Dans son ouvrage sur la traduction médicale, de la langue anglaise vers la langue française, Rouleau (1994 : 198-199) explique ainsi son analyse du terme 'traitement' :

> Rien grammaticalement ne régit l'utilisation d'une préposition particulière avec le verbe « traiter ». Il n'y aurait, selon toute apparence, aucune faute à utiliser « avec », « à » ou « par », si ce n'est que **l'usage a peut-être ses préférences.** Le seul moyen de connaître cet usage, c'est de lire des documents écrits par des spécialistes francophones et d'être attentif aux façons de dire des auteurs. Après avoir dépouillé 13 chapitres écrits par au moins autant de médecins et avoir relevé toutes les phrases où se rencontraient les mots . « traitement », « traiter », « thérapie » (corpus de plus de 300 phrases), il est possible d'affirmer que « traiter par » ou « traitement par » est la tournure la plus utilisée.

Puis Rouleau (1994 : 199-203) termine son analyse du terme 'traitement' en présentant une liste de ses cooccurrents qui peuvent être soit un nom, p. ex. « traitement

[1] CASSCAD est un acronyme construit à partir d'un réordonnancement des lettres soulignées dans « Analyse de Concordance et Analyse Syntaxique et Sémantique pour la Construction de Dictionnaires »

d'urgence », « période de traitement » ; soit un adjectif, p. ex « traitement anti-tuberculeux », « traitement préventif » ; soit un verbe, p. ex. « le traitement supprime », « prescrire le traitement » ; soit une préposition ou une locution, p. ex. « traitement par [nom d'un médicament ou voie d'administration] ». Fait remarquable, tout ce travail d'analyse de Rouleau est basé *sur des données colligées manuellement, c'est-à-dire, sans aucun outil informatique.* Dans cet article, nous présentons une approche semi-automatique qui vise justement à supporter le travail du langagier en facilitant la collecte de telles données et en permettant de construire un dictionnaire simple à partir du corpus analysé. Cette approche, dérivée de travaux en acquisition automatique de connaissances à partir de textes (Delisle, 1994), possède l'avantage d'offrir beaucoup plus qu'un simple concordancier, comme nous le verrons plus loin.

3. Approche semi-automatique : architecture et fonctions de CASSCAD

L'approche que nous proposons a pour but de rendre plus performante la construction d'un dictionnaire spécialisé ou, encore, spécifique d'un corpus particulier. Dans le cadre du présent article, nous entendons par dictionnaire une base de données fondamentales sur le vocabulaire d'un corpus – la nature exacte de ces données est présentée aux sections 3.2. et 4. Cette base de données est construite à partir du corpus analysé à l'aide du système CASSCAD. Ce système utilise plusieurs logiciels et sources d'information complémentaires. Chacun de ces éléments apporte sa contribution aux trois phases qui constituent l'approche en question.

3.1. Phase 1 – Analyse de concordance (entrée : texte/corpus brut ; sortie : statistiques diverses et liste des concordances)

Le concordancier que nous utilisons est un programme développé à l'UQTR (Boisvert, 1989) auquel nous avons apporté quelques modifications. Écrit en Pascal, ce concordancier est relativement standard et possède, entre autres, les options suivantes :

- fonctionnement en interactif ou par lots ;
- possibilité d'identifier une liste (fichier) de mots à rejeter, c.-à-d. pour lesquels on ne veut pas de concordances ;
- possibilité d'identifier une liste (fichier) de mots exclusifs, c.-à-d. les seuls pour lesquels on veut des concordances. Ces mots peuvent être identifiés à l'aide du symbole '*' (*joker*) : p. ex., ordinateur* couvrira les occurrences de 'ordinateur' et 'ordinateurs'. De plus, il est possible de demander des cooccurrences de N mots. Si un segment de la liste des mots exclusifs compte N mots, le concordancier trouvera *dans une même phrase* les cooccurrences de ces N mots (même si ces N mots ne sont pas contigus) ;
- possibilité de préciser un contexte maximum fixe (de 1 à 9 mots) avant et après chaque occurrence ou, un contexte flottant borné par le début et la fin de la phrase dans laquelle l'occurrence a été trouvée ;
- possibilité de limiter la taille des mots qui seront considérés par le concordancier (p. ex. seulement les mots ayant entre 4 et 20 caractères).

La sortie du concordancier est composée d'abord d'une série de statistiques simples telles que le nombre total de mots, de phrases et de paragraphes du texte ana-

lysé, le nombre moyen de mots par phrase et le nombre moyen de mots par paragraphe, etc. Viennent ensuite les concordances elles-mêmes qui sont ordonnées alphabétiquement. Voici un exemple obtenu à partir d'un texte anglais pour les mots 'change' (2 occurrences), 'changes' (1 occurrence), et 'chapter' (1 occurrence). Le contexte (maximum) est de 5 mots. Chaque entrée est délimitée par « ==> *mot* » *et* « > *mot [fréquence]* ».

```
==> change
later you'll learn how to change some of these defaults
should you can change this to descending order by
> change [2]
==> changes
statements if you have no changes to make to them.
> changes [1]
==> chapter
in this chapter you've learned how to produce
> chapter [1]
```

3.2. Phase 2 – Analyse lexicale multi-source (entrée : liste des concordances ; sortie : dictionnaire de base *dico_spé*)

Le programme de contrôle de l'analyse lexicale est écrit en C. Il accepte en entrée une liste de concordance sous le format décrit ci-dessus. Pour chaque index (mot) de la liste de concordance, le programme de contrôle présente à l'usager la liste des concordances de ce mot, consulte *The Collins* (Karp *et al.*, 1992) et *WordNet* (Miller, 1990)[2], et présente à l'usager l'information associée à ce mot, et, finalement, guide l'usager dans la construction d'une entrée de dictionnaire spécifique et adaptée au traitement de la phase 3 – nous qualifierons ce dictionnaire de « spécifique », appelé *dico_spé*.

L'usager a aussi la possibilité de sauter au mot suivant s'il préfère ne pas traiter un certain mot, par exemple si ce mot appartient à une catégorie grammaticale fermée (d'autant plus que l'analyseur syntaxique de la phase 3 possède son propre dictionnaire, lequel est particulièrement axé sur les catégories grammaticales fermées). Habituellement, les mots des catégories grammaticales ouvertes – nom, verbe, adjectif, adverbe – sont ceux qui présentent le plus grand intérêt pour le langagier. Soulignons également que *WordNet* est utile à cet égard puisqu'il ne porte que sur les mots de la langue anglaise appartenant aux quatre catégories grammaticales ouvertes que nous venons d'identifier.

Nous extrayons du dictionnaire *The Collins* les catégories (ou fonctions) grammaticales potentielles que peut jouer le mot. À l'aide de cette information et des concordances, l'usager peut identifier la (ou les) fonction(s) grammaticale(s) parti-

2 Il s'agit là de deux ressources du domaine public. Il existe de nombreuses autres ressources mais pas forcément aussi largement diffusées ou aussi peu coûteuses (p. ex. *The Collins COBUILD English Language Dictionary*, *The Longman Dictionary of Contemporary English*). Voir aussi De Bessé (1991) pour une liste de plusieurs ressources pour la langue française et la langue anglaise.

culière(s) de ce mot dans le corpus sous analyse. Par exemple, le mot anglais 'file' peut fonctionner en tant que nom ou en tant que verbe : c'est ce que *The Collins* nous dit. Cependant, grâce aux concordances, l'usager est à même de constater que dans le corpus sous analyse – on le suppose ici – le mot 'file' n'est utilisé que dans sa fonction de nom[3]. Ainsi, il sera possible de ne construire que le nombre minimal d'entrées du dictionnaire *dico_spé*, ce qui permet de cerner avec précision le vocabulaire d'un corpus, en plus de contribuer à réduire les problèmes d'ambiguïté lexicale pour le traitement subséquent de la phase 3. Si nécessaire, l'usager pourra même demander au programme de contrôle de soumettre une phrase particulière, tirée d'une occurrence du mot considéré, à l'étiqueteur de Brill (1992) afin de vérifier la catégorie grammaticale probable du mot en question.

La base de données lexicales *WordNet* nous offre une panoplie d'informations. Actuellement, le programme de contrôle de l'analyse lexicale extrait, par défaut, les synonymes et les hyperonymes du mot considéré. Il est cependant possible pour l'usager d'avoir accès aux autres catégories d'information de *WordNet* grâce au programme de contrôle. La mise en parallèle de l'ensemble des synonymes et des concordances d'un mot permet à l'usager de vérifier le sens de ce dernier et, si désiré, de sélectionner la définition[4] appropriée pour l'inclure dans l'entrée de dictionnaire *dico_spé*. Une fois le sens déterminé, l'accès à l'ensemble des hyperonymes permet d'identifier une catégorie taxonomique pour le mot en question. Poursuivons notre exemple avec le mot anglais 'file' qui est utilisé en tant que nom. L'examen de nos concordances et l'accès aux informations de *WordNet* nous permettent de choisir la définition du sens « data file » et la catégorie taxonomique (hyperonyme) « record ». De plus, on pourra indiquer si le mot est de spécialité ou non. Supposons que tel est le cas, alors l'usager pourra, en interagissant avec le programme de contrôle de l'analyse lexicale, participer à la création de l'entrée *dico_spé* suivante pour le mot 'file' :

```
dico_spe( file, countnoun, sg, _, ['data file'/record/157/special] ).
```

Cet exemple, dans lequel 'countnoun' signifie 'nom nombrable' et 'sg' signifie 'singulier', illustre le format Prolog utilisé par le sous-système d'analyse syntaxique (voir 3.3.). La définition des paramètres de ce format des entrées lexicales est la suivante : 1) mot, 2) catégorie grammaticale, 3) premier paramètre spécifique de la catégorie grammaticale du mot, 4) deuxième paramètre spécifique de la catégorie grammaticale du mot ('_' veut dire sans valeur), 5) liste des définitions, catégories taxonomiques, numéros de concordance[5] et indicateurs d'appartenance du mot à une langue de spécialité. Mentionnons au passage qu'il n'est pas nécessaire d'entrer toutes les formes d'un mot puisque le sous-système de la phase 3 possède un analyseur morphologique qui permet de reconnaître 'files' comme la forme pluriel du nom 'file', par exemple.

---

3 Le mot anglais 'file' possède deux fonctions grammaticales (ainsi que plusieurs sens distincts), celles de nom (p. ex. *dossier*) et de verbe (p. ex. *classer*). CLASCAO permet de distinguer ces cas.

4 Il ne s'agit pas à proprement parler d'une définition exhaustive mais plutôt d'un ou plusieurs mots qui font référence à des concepts clés permettant de distinguer les différents sens d'un mot. L'usager peut accepter ou modifier la définition fournie par *WordNet*, elle sera conservée dans le *dico_spé*.

5 Ce numéro de concordance permet d'associer une occurrence représentative à la définition du *dico_spé*. Ainsi, l'usager peut facilement retracer la provenance de ce mot dans la sortie du concordancier et, par le fait même, dans le corpus.

Si un mot est inconnu, c'est-à-dire s'il n'apparaît ni dans *The Collins* ni dans *WordNet*, il s'agit probablement d'un mot de spécialité pour lequel l'usager devra identifier de façon interactive la valeur des paramètres de son entrée lexicale dans le dictionnaire *dico_spé*. Dans ce cas, le paramètre de spécialité prendrait la valeur de 'special'. Par exemple, dans la phrase « Erythrocyte size and hemoglobinization can be estimated visually on stained films of the blood or can be calculated quantitatively from the hemoglobin, erythrocyte count, and packed cell volume », tiré de Rouleau (1994 : 279), le mot 'hemoglobinization' est inconnu du *The Collins* et de *WordNet* (mais pas 'erythrocyte' !). Dans ce cas, CASSCAD demandera à l'usager, par un menu simple, de compléter l'entrée lexicale du mot 'hemoglobinization' – notons aussi que l'étiqueteur peut faire des suggestions utiles à cet effet. Pour ce qui est du traitement de la troisième phase, le cinquième paramètre (définition, catégorie taxonomique, numéro de concordance et indicateur de spécialité) n'est pas obligatoire. L'usager peut donc le laisser indéterminé et le reconsidérer plus tard s'il le désire.

Finalement, si un mot possède plus d'un sens dans le corpus tout en appartenant à une même catégorie grammaticale, CASSCAD permet de les distinguer. Par exemple, si le mot 'file' est aussi utilisé dans le sens de meuble (classeur), l'entrée du *dico_spé* sera augmentée comme suit :

```
dico_spe( file, countnoun, sg, _,     ['data file'/record/157 special,
                                       'file cabinet'/'office furniture'
                                       /294/_] ).
```

On remarque que le numéro de concordance nous aidera alors à distinguer les deux sens du mot 'file'. Cette information sera également utile pour compléter les entrées des verbes (section 4.3.). Ici, la distinction des différents sens d'un mot doit être contrôlée par l'usager. D'autres travaux (Yarowsky, 1995 ; Chakravarthy, 1995) se sont intéressés davantage à désambiguïser automatiquement les différents sens d'un mot dans un corpus et ce, à l'aide de ressources semblables à celles que nous utilisons dans notre approche.

3.3. Phase 3 – Analyse syntaxique et sémantique (entrée : texte/corpus brut et dictionnaire *dico_spé* ; sortie : dictionnaire augmenté des entrées de verbes)

L'analyseur syntaxique, nommé DIPETT, et l'analyseur sémantique, nommé HAIKU, sont implémentés en Quintus Prolog 3.2 et en SISCtus Prolog 2.1(#9) sur des stations de travail Sun. La phase 3 commence le traitement du corpus par l'analyse syntaxique du texte original avec l'analyseur syntaxique DIPETT (*Domain-Independent Parser for English Technical Texts*). Pour chaque phrase du corpus, ce parseur produit un arbre d'analyse auquel l'usager pourra apporter des modifications simples à l'aide du module de rattachement (Delisle, 1995), si cela devait s'avérer nécessaire, par exemple, pour corriger l'attachement d'un syntagme prépositionnel[6].

6 Terry Copeck, un membre du groupe de recherche KAML de l'Université d'Ottawa, participe activement à la réalisation de ce module de rattachement

Vient ensuite l'analyse sémantique semi-automatique effectuée par le module HAIKU. L'arbre d'analyse syntaxique produit par DIPETT est maintenant décomposé par HAIKU et ce dernier détermine les relations sémantiques qui lient ses composants et ce, à trois niveaux complémentaires qui sont associés à autant d'étapes de traitement dans HAIKU[7]. D'abord, les relations entre les propositions qui forment une phrase complexe : par exemple, une proposition peut exprimer une relation de causalité par rapport à une autre proposition de la même phrase (Barker & Szpakowicz, 1995). Ensuite, les relations entre le verbe principal de chaque proposition et ses arguments : il s'agit cette fois d'une analyse au point de vue des Cas sémantiques ; et finalement, les relations entre les éléments des groupes nominaux complexes – ces derniers travaux sont en cours. Dans le présent article, nous insistons davantage sur la deuxième étape, soit l'analyse Casuelle.

3.3.1. Quelques détails sur l'analyseur syntaxique

L'analyse syntaxique nous permet, entre autres, d'accéder aux patrons syntaxiques dont nous avons besoin pour l'analyse sémantique subséquente avec HAIKU. Par opposition à d'autres stratégies de parsage plus superficielles, DIPETT effectue une analyse syntaxique détaillée et indépendante du domaine du corpus en entrée : tous les détails sur DIPETT, de même que de nombreuses références pertinentes, apparaissent dans Delisle & Szpakowicz (1991), Copeck *et al.* (1992), Delisle (1994) et Delisle & Szpakowicz (1995). En fait, DIPETT analyse un corpus brut contenu dans un simple fichier texte et dont les mots n'ont pas été lexicalement étiquetés ou annotés au préalable – l'analyse lexicale nous assure que tous les mots du texte à analyser possèdent une entrée dans le *dico_spé* utilisé par DIPETT pour ses informations grammaticales. DIPETT accepte en entrée une chaîne de caractères, une phrase ou un fragment, selon le cas, et produit en sortie un arbre d'analyse unique. Cet arbre unique n'est évidemment pas toujours parfait : le parseur utilise ses heuristiques pour produire son analyse mais, comme il n'a accès à aucune donnée sémantique, il peut construire un arbre plus ou moins correct du point de vue sémantique. C'est pourquoi la fonctionnalité du module de rattachement sera utile à cet égard.

DIPETT tente d'abord de trouver une analyse complète pour chaque phrase soumise en entrée. Lorsque cela est impossible, soit parce que la phrase est grammaticalement incorrecte ou qu'elle est extra-grammaticale par rapport à la grammaire de DIPETT ou, encore, que le temps alloué pour l'analyse d'une phrase est écoulé, l'analyseur tente alors de trouver une analyse en fragments. Il essaie de reconnaître les principales sous-structures de la phrase telles que syntagmes verbaux, syntagmes nominaux, syntagmes prépositionnels, syntagmes adverbiaux ou adjectivaux. Nous considérons qu'il est préférable d'avoir une analyse partielle que rien du tout. D'ailleurs, pour la construction du *dico_spé*, l'analyse fragmentaire permet de répondre à nos objectifs initiaux sans perte importante, car ce sont les structures prédicat-arguments[8] qui importent, et nous les obtenons par cette analyse par fragments. DIPETT constitue un environnement d'analyse syntaxique qui témoigne de l'im-

7 Les étapes 1 et 3 de HAIKU sont la contribution de Ken Barker du Département d'informatique de l'Université d'Ottawa

8 La pertinence des structures prédicat-arguments pour le traitement informatisé du texte semble avoir effectué un retour en force. Voir à ce sujet Marcus *et al.* (1994) et Grishman (1994)

portance accordée à l'aspect ingénierie du langage dans nos travaux. Des tests avec la version la plus récente du parseur (v3.0) nous donnent les résultats suivants : jusqu'à 95 % des phrases d'un corpus (anglais, technique) sont analysées : 60 % d'analyses complètes et 35 % d'analyses par fragments.

3.3.2. Quelques détails sur l'analyseur sémantique

Nous traitons ici de la partie principale de l'analyse sémantique, c.-à-d. l'analyse Casuelle semi-automatique et interactive – les fondements de cette analyse sont présentés dans Delisle (1994) et *Delisle et al.* (à paraître). Les Cas (Fillmore, 1968 ; Somers, 1987), représentent les relations sémantiques entre le verbe principal d'une proposition et ses arguments syntaxiques, c.-à-d. le sujet, l'objet, les syntagmes prépositionnels et les adverbes. Les relations nommées par les Cas correspondent à des rôles dans l'action associée au verbe. Par exemple, le Cas Agent identifie l'instigateur de l'action. Les Cas se retrouvent dans la syntaxe comme des structures prédicat-arguments dans lesquelles chaque Cas est dénoté par un marqueur et réalisé par un syntagme. Ainsi, dans la phrase « Maxime a réparé sa voiture avec ses nouveaux outils », le Cas Agent est associé à « Maxime », le Cas Objet est associé à « sa voiture » et le Cas Instrument est associé à « ses nouveaux outils ». Nous avons opté pour une analyse sémantique basée sur les Cas pour deux raisons majeures. Premièrement, l'analyse Casuelle permet d'établir un lien explicite entre la syntaxe et la sémantique ; ceci est essentiel dans une approche basée sur la syntaxe. Deuxièmement, l'analyse Casuelle s'effectue en des termes relativement simples et intuitifs qui en rendent les concepts accessibles à l'usager ; il s'agit là d'un point important dans le contexte d'une approche semi-automatique orientée vers le langagier.

Nous avons construit un système de Cas général et indépendant de tout domaine particulier. C'est ce système de Cas qui, par défaut, est utilisé par HAIKU. Il comporte 28 Cas regroupés en 5 catégories (les abréviations des Cas apparaissent après le '/') : 1) PARTICIPANT : Agent/agt, Beneficiary/benf, Experiencer/expr, Instrument/inst, Object/obj, Recipient/recp ; 2) CAUSALITY : Cause/caus, Effect/eff, Opposition/opp, Purpose/purp ; 3) TIME : Frequency/freq, Time_at/tat, Time_from/tfrm, Time_to/tto, Time_through/ttru ; 4) SPACE : Direction/dir, Location_at/lat, Location_from/lfrm, Location_to/lto, Location_through/ltru, Orientation/ornt ; 5) QUALITY : Accompaniment/acmp, Content/cont, Exclusion/excl, Manner/man, Material/matr, Measure/meas, Order/ord La justification et la définition de ces Cas sont présentées dans Barker et al. (1993). Ce sont les abréviations de ces Cas qui serviront à la construction des patrons sémantiques.

4. La construction des entrées complémentaires pour les verbes

L'analyseur Casuel de HAIKU accepte comme entrée un arbre d'analyse produit par DIPETT et y associe, semi-automatiquement, les patrons Casuels qui représentent le mieux le sens de la phrase[9]. Les Cas sont réalisés dans la syntaxe de deux façons : 1) de façon lexicale (c.-à-d. par un marqueur explicite dans la syntaxe de surface), par

9 Lorsque la phrase contient plusieurs propositions, HAIKU la découpe en une suite de propositions qui sont analysées les unes à la suite des autres

exemple lorsqu'une préposition introduit un syntagme prépositionnel, et 2) de façon *positionnelle*, par un marqueur implicite associé au sujet (psubj), à l'objet direct (pobj) ou à l'objet indirect (piobj). Tout comme pour les Cas ci-dessus, ce sont les symboles associés à ces marqueurs qui serviront à la construction des patrons. Par exemple, le patron syntaxique (PSY) `psubj-pobj-at` est associé à une proposition dans laquelle le verbe principal possède un sujet, un objet direct et un syntagme prépositionnel introduit par la préposition 'at'. De même, le patron sémantique (PSÉ) `agt-obj-lto` peut être associé à une proposition dont le PSY est `psubj-pobj-at` et dans laquelle le verbe principal possède un sujet qui tient le rôle agent, un objet direct qui tient le rôle objet et un syntagme prépositionnel qui tient le rôle de location_to (destination)[10].

L'analyseur Casuel effectue un type d'apprentissage automatisé qui possède les trois principales caractéristiques de l'apprentissage basé sur les occurrences (ou *instance-based learning, voir Aha et al.*, 1991) : *i*) c'est un apprentissage supervisé, c.-à-d. contrôlé par l'usager ; *ii*) c'est un apprentissage incrémentiel ; et *iii*) c'est également un apprentissage basé sur les similarités entre le nouveau patron à identifier et ceux déjà assimilés. Pour ce faire, HAIKU utilise quatre dictionnaires simples (voir 4.1. à 4.4.) qui peuvent être vides au début de l'analyse d'un corpus. Dans ces circonstances, la contribution de l'usager sera plus importante initialement et s'allègera à mesure que HAIKU garnira ses dictionnaires de façon incrémentielle. Ce sont ces quatre dictionnaires construits par HAIKU qui viendront compléter le *dico_spé* initial résultant des phases 1 et 2.

Tous les dictionnaires de HAIKU sont continuellement mis à jour pendant l'analyse du corpus. Chaque proposition se voit attribuer un PSÉ unique à la suite de l'intervention de l'usager, soit qu'il approuve la suggestion du système, soit qu'il la modifie. Pour faire une suggestion à l'usager, l'analyseur Casuel fouille ses dictionnaires dans le but de trouver un PSÉ qui correspond le mieux au PSY de la proposition considérée. Pour ce faire, on utilise un algorithme de contrôle de l'analyse Casuelle (Delisle *et al.*, à paraître) couplé à un algorithme simple de filtrage (Delisle *et al.*, 1993) qui permet de trouver le ou les meilleurs PSÉ candidats en fonction du PSY – l'analyseur Casuel utilise aussi la phrase exemple conservée dans le dictionnaire `cmpDict` afin d'illustrer la situation à l'usager et ainsi simplifier sa décision.

Si le ou les PSÉ suggéré(s) par le HAIKU ne semblent pas acceptables à l'usager, ce dernier est alors appelé à intervenir en identifiant lui-même le PSÉ approprié. HAIKU affiche des informations complémentaires comme la liste des Cas associés aux marqueurs de Cas de la proposition analysée et la liste des Cas manipulés par le système. Notons que l'usager peut ajouter de façon dynamique ses propres Cas à l'ensemble des 28 Cas prédéfinis dans le système à tout moment au cours de son interaction avec HAIKU. Ce dernier permet également de sauver automatiquement la liste de Cas de l'usager afin d'en faciliter la réutilisation.

Voyons maintenant la structure de ces quatre dictionnaires. Pour illustrer nos propos, nous utiliserons le petit corpus suivant à titre d'exemple : « Bob printed the

10 L'ordre n'importe pas dans les patrons, seule leur interprétation sémantique est importante. Ainsi, `psubj-pobj-at-by` est équivalent à `psubj-pobj-by-at`, et, de façon similaire `agt-obj-lat-tat` est équivalent à `agt-obj-tat-lat`.

new data file. Beth and Tom will print their letters. Their boss could not print the production report on the new laser printer. The new computer caused a power failure yesterday. We know that your boss would not delete all your data. These new employees have deleted my letters from my disk. »

4.1. mDict *(meaning dictionary)*

Le dictionnaire de sens (`mDict`) contient des entrées pour les mots individuels : verbes, prépositions ou adverbes. Pour les deux dernières catégories, le `mDict` contient la liste fixe des Cas qui peuvent être marqués par ces mots (voir Barker *et al.*, 1993). Pour les verbes, une entrée contient : 1) la liste des PSY trouvés dans le corpus, ainsi que le nombre d'occurrences de chaque PSY ; et 2) la liste des Cas qui ont été associés à chaque marqueur de Cas, ainsi que le nombre d'occurrences de chaque association. Voici le contenu intégral en Prolog des entrées des verbes du `mDict` après l'analyse syntaxique et l'analyse Casuelle du petit corpus ci-dessus (le `mDict` était vide au départ) :

```
mDict(cause, ['psubj-pobj-adv':1],
       [[adv,[tat:1]],[pobj,[obj:1]],[psubj,[agt:1]]]).
mDict(delete, ['psubj-pobj':1,'psubj-pobj-from':1],
       [[from,[lfrm:1]],[pobj,[obj:2]],[psubj,[agt:2]]]).
mDict(know, ['psubj-pobj':1],
       [[pobj,[obj:1]],[psubj,[agt:1]]]).
mDict(print, ['psubj-pobj':2,'psubj-pobj-on':1],
       [[on,[lto:1]],[pobj,[obj:3]],[psubj,[agt:3]]]).
```

4.2. cmpDict *(Case-marker pattern dictionary)*

Le dictionnaire des patrons de marqueurs de Cas (`cmpDict`) contient une entrée pour chaque PSY. Chaque entrée associe au PSY la liste des PSÉ qui ont été attribués à ce PSY au cours de l'analyse du corpus, de même que le nombre d'occurrences de chacun des PSÉ. De plus, chaque PSÉ est illustré par une phrase exemple, tirée du corpus analysé, que l'usager aura considérée comme représentative du PSÉ en question. Voici le contenu intégral en Prolog des entrées des verbes du `cmpDict` après l'analyse syntaxique et l'analyse Casuelle du petit corpus ci-dessus (le `cmpDict` était vide au départ) :

```
cmpDict('psubj-pobj', [['agt-obj':4,
      '''[bob,printed,the,new,data,file,.]''']]).
cmpDict('psubj-pobj-adv', [['agt-obj-tat':1,
      '''[the,new,computer,caused,a,power,failure,
      yesterday,.]''']]).
cmpDict('psubj-pobj-from', [['agt-obj-lfrm':1,
      '''[these,new,employees,have,deleted,my,letters,from,
      my,disk,.]''' ]]).
cmpDict('psubj-pobj-on', [['agt-obj-lto':1,
      '''[their,boss,could,not,print,the,production,
      report,on,the,new,laser,printer,.]''']]).
```

4.3. cpDict *(Case pattern dictionary)*

Le dictionnaire des PSÉ (`cpDict`) contient une entrée pour chaque PSÉ rencontré dans le corpus et lui associe la liste des verbes qui se sont vus attribuer un tel PSÉ. Voici le contenu intégral en Prolog des entrées des verbes du `cpDict` après l'analyse syntaxique et l'analyse Casuelle du petit corpus ci-dessus (le `cpDict` était vide au départ) :

```
cpDict('agt-obj',[delete,know,print]).
cpDict('agt-obj-lfrm',[delete]).
cpDict('agt-obj-lto',[print]).
cpDict('agt-obj-tat',[cause]).
```

Notons qu'il est possible de distinguer plus finement les verbes en associant à ceux-ci le numéro de concordance introduit à la section 3.2. Par exemple, supposons que les deux PSÉ associés au verbe 'delete' correspondent à autant de sens et que nous désirions les démarquer. Les deux premières entrées du `cpDict` pourraient alors être comme suit :

```
cpDict('agt-obj',[delete/489,know,print]).
cpDict('agt-obj-lfrm',[delete/502]).
```

4.4. ccvpIndex

Le dernier dictionnaire sert en fait de structure indexée afin de faciliter l'accès aux résultats produits par la phase 3 et conservés dans un fichier de sortie qui est indépendant des quatre dictionnaires dont il est question ici. Dans ce fichier de sortie, on retrouve deux structures pour chaque unité[11] (`units` dans le `ccvpIndex`) du corpus analysée par le système : l'arbre d'analyse syntaxique et la structure de Cas de HAIKU. Ces deux structures sont co-indexées grâce à un simple numéro d'identification unique (# dans le `ccvpIndex`). Le `ccvpIndex` met donc en association toutes les occurrences distinctes de PSY, PSÉ, verbe, et autres détails sur le contexte d'occurrence de ces patrons, c'est-à-dire le numéro de l'unité dans laquelle un PSY, un PSÉ et un verbe particuliers ont été rencontrés dans le corpus ; la sous-catégorisation de surface[12] du verbe (`sr_types` dans le `ccvpIndex`) telle que trouvée dans le corpus ; et le temps du verbe dans cette occurrence. Voici le contenu intégral en Prolog des entrées des verbes du `ccvpIndex` après l'analyse syntaxique et l'analyse Casuelle du petit corpus ci-dessus (le `ccvpIndex` était vide au départ) :

```
ccvpIndex('psubj-pobj', 'agt-obj', delete,
      units([[#(5), sr_types('np-np'),
                  tense([would_conditional_present_simple])]])).
ccvpIndex('psubj-pobj', 'agt-obj', know,
```

11 Nous appelons unité chaque segment considéré pour analyse. Un segment peut correspondre à une phrase complète ou à un syntagme isolé (ou une phrase incomplète).

12. 'np' (noun phrase) veut dire syntagme nominal ; 'nom_cl' (nominal clause) veut dire proposition nominale ; 'adv' signifie groupe adverbial ; et 'on', 'from' désignent un syntagme prépositionnel introduit, respectivement, par la préposition 'on' ou la préposition 'from'.

```
        units([[#(5), sr_types('np-nom_cl'),
                      tense([infinitive])]])).
ccvpIndex('psubj-pobj', 'agt-obj', print,
        units([[#(1), sr_types('np-np'),
                      tense([past_simple])],
               [#(2), sr_types('np-np'),
                      tense([future_simple])]])).
ccvpIndex('psubj-pobj-adv', 'agt-obj-tat', cause,
        units([[#(4), sr_types('np-np-adv'),
                      tense([past_simple])]])).
ccvpIndex('psubj-pobj-from', 'agt-obj-lfrm', delete,
        units([[#(6), sr_types('np-np-from'),
                      tense([present_perfect_simple])]])).
ccvpIndex('psubj-pobj-on', 'agt-obj-lto', print,
        units([[#(3), sr_types('np-np-on'),
                      tense([could_conditional_present_simple])]])).
```

5. Aperçu de quelques travaux connexes

L'extraction automatique ou semi-automatique de connaissances, d'informations ou de données, à partir de textes de tous genres est, depuis la fin des années 80, un domaine de recherche très actif en informatique linguistique. Citons, à titre d'exemple, les travaux récents d'Agarwal (1994), d'Appelt *et al.* (1993), de Gomez *et al.* (1994), d'Ogonowski *et al.* (1994) et de Delisle (1994).

Parmi ces travaux, plusieurs portent sur la construction automatique de dictionnaires et de lexiques avec des objectifs similaires à ceux de l'approche que nous avons décrite dans le présent article. Mentionnons, entre autres, Cardie (1993), qui propose une approche permettant d'acquérir à partir d'un corpus les fonctions grammaticales et les sens des mots appartenant à une catégorie ouverte ; Grishman *et al.* (1994a) et Sanfilippo (1994), qui décrivent certains aspects de la problématique de la construction d'un grand lexique pour des fins de traitement informatique de textes, ainsi que certaines solutions qu'ils proposent ; Riloff (1993) et Soderland *et al.* (1995), qui présentent chacun un système conçu pour la construction automatique d'un dictionnaire spécifique d'un domaine donné, et ce, dans le contexte d'une application en extraction d'information ; et Sanfilippo & Poznanski (1992), qui suggèrent une approche au problème de la mise en correspondance des différents sens d'un mot lorsque ceux-ci proviennent de différents dictionnaires informatisés.

Il existe aussi de nombreux travaux portant davantage, quoique non exclusivement, sur les entrées des verbes de ces dictionnaires. Soulignons, entre autres, Framis (1994), Grishman & Sterling (1994) et Manning (1993), qui présentent des approches à l'identification automatique des restrictions (ou contraintes) de sélection à partir de l'analyse d'un corpus ; Myaeng *et al.* (1994) et Pugeault *et al.* (1994), qui s'intéressent particulièrement à l'extraction des structures prédicat-arguments à partir des textes ; et Basili *et al.* (1992) et Sekine *et al.* (1992), qui proposent des méthodes pour acquérir automatiquement à partir d'un corpus des collocations de nature sémantique.

6. Conclusion

Le contenu des dictionnaires construits par le système CASSCAD peut grandement aider le langagier (p. ex. terminologue ou traducteur) ou l'ingénieur de la connaissance dans la construction d'un dictionnaire spécialisé (ou spécifique) et, de façon plus particulière, dans l'étude des verbes d'un corpus pour en préciser les propriétés syntaxiques et sémantiques. Ainsi, le *dico_spé* nous dit :

- quels mots apparaissent dans un corpus et quelles informations s'y rattachent (catégorie grammaticale, définition, catégorie taxonomique, indicateur de spécialité, etc.) ;

et les quatre dictionnaires construits par HAIKU nous disent, en plus, pour les verbes :

- quels sont leurs patrons syntaxiques et leurs fréquences d'occurrence respectives ;
- quels sont leurs sous-catégorisations de surface ;
- quels sont leurs patrons sémantiques et leurs fréquences d'occurrence respectives ;
- quels verbes ont des patrons (syntaxiques ou sémantiques) identiques ou similaires ;
- dans quelles phrases du corpus apparaissent un verbe, un PSY ou un PSÉ particuliers ?
- dans quelles phrases du corpus apparaît le verbe V avec le PSY ou le PSÉ P ?
- dans quelles phrases du corpus apparaissent ensemble le PSY P1 et le PSÉ P2 ?

Il nous semble qu'un système comme CASSCAD pourrait être d'un grand secours pour le langagier qui souhaite construire une classification de verbes (voir Dixon, 1991 ; Levin, 1993). Dans le futur, nous prévoyons améliorer le traitement des mots composés afin de permettre à l'usager de les considérer comme des unités linguistiques lorsque désiré : les éléments sont en place dans le concordancier, l'analyseur lexical et l'analyseur syntaxique, mais il nous reste à les intégrer de façon cohérente. De plus, il serait avantageux de pouvoir utiliser les catégories taxonomiques tirées de *WordNet* (ou créées par l'usager) afin de rendre plus spécifiques les collocations sémantiques et les contraintes de sélection des verbes accumulées par HAIKU. Nous planifions également une expérimentation sur de gros corpus afin d'évaluer notre approche sur une plus grande échelle : cela permettrait de répondre à des questions comme « quel est la proportion des verbes identifiés lors de l'analyse lexicale qui se voient associer des PSY ou PSÉ une fois la phase 3 complétée ? ». Une autre question intéressante est celle de la généralisation de notre approche : est-elle utile à la construction d'un dictionnaire à caractère général ?

Remerciements

Je remercie tous ceux qui ont contribué aux travaux mentionnés dans cet article : d'abord, René Boisvert pour avoir réalisé le programme de concordance ; ensuite, Georges Diop Rogandji pour avoir implémenté le module d'analyse lexicale multi-source ; puis, tous les membres du groupe de recherche KAML du Département d'informatique de l'Université d'Ottawa qui, au fil des années, ont testé DIPETT et HAIKU sans pitié aucune, en plus d'apporter des idées qui ont contribué à mes recherches. Je remercie aussi le CRSNG (Conseil de Recherches en Sciences Naturelles et Génie du Canada) de son support financier. Finalement, je remercie Maurice Rouleau pour avoir relu cet article.

Réseau notionnel, intelligence artificielle et équivalence en terminologie multilingue : essai de modélisation

Marc VAN CAMPENHOUDT

Centre de recherche TERMISTI, Institut supérieur de traducteurs et interprètes, Bruxelles, Belgique

1. Introduction

L'approche notionnelle constitue l'un des fondements de la terminologie. Depuis plusieurs années, des équipes de recherche ont développé des logiciels permettant de naviguer au travers des réseaux notionnels et ainsi de mieux appréhender la notion au sein de son microdomaine. Des gestionnaires comme *MC4* (Université de Clermont-Ferrand), *Termisti* (ISTI, Bruxelles) ou *Code* (Université d'Ottawa) constituent autant de pas successifs vers la construction de bases de connaissances et vers l'intelligence artificielle.

Cet article a pour principal objectif de montrer que l'exploitation logique des réseaux notionnels au sein de bases de connaissances terminologiques (B.C.T.) multilingues devrait aussi permettre de gérer divers problèmes d'équivalence. Il se fonde sur un corpus d'exemples extraits de *De la quille à la pomme de mât* (Paasch, 1901), un vaste dictionnaire nautique trilingue dont l'organisation notionnelle exemplaire a conduit à l'ébauche du modèle théorique ici exposé[1]. De par sa tâche d'expert maritime, son auteur, le capitaine Heinrich Paasch, a été inévitablement confronté au non-isomorphisme[2] des langues. Il est très aisé d'affirmer que tel ou tel dictionnaire est fondé sur une approche notionnelle. Rares sont pourtant, à nos yeux, les auteurs de terminographies multilingues qui vont jusqu'au bout de cette logique et distinguent réellement chacune des notions propres à chacune des langues envisagées.

1. La relation hyponymique retiendra plus particulièrement notre attention La place des autres relations notionnelles dans le modèle a déjà été décrite dans la thèse que nous avons consacrée à ce dictionnaire (Van Campenhoudt, 1994) et dont cet article est issu
2. À la suite de Lyons (1970 : 45), nous parlerons de (*non-*)*isomorphisme* entre les langues et de *chevauchement culturel*.

2. Découpage notionnel et confrontation des langues

2.1. La tradition viennoise face à l'équivalence

La linguistique a depuis longtemps montré que toutes les langues n'approchent pas la réalité de la même manière et que de nombreux problèmes se posent lors de l'établissement d'équivalences. Eugen Wüster, le chef de file de l'école viennoise, avait assurément pris conscience du fait que les systèmes de notions varient d'une langue à l'autre. En divers passages de son œuvre[3], il a rappelé cet état de fait et regretté que de nombreux terminographes réalisent des œuvres dans lesquelles le système notionnel est conditionné par une langue particulière, ce qui débouche inévitablement sur des impossibilités de traduction.

Face aux problèmes d'équivalence soulevés par la divergence notionnelle entre les langues, Wüster (1971 : 44-45) proposait pour solution d'adopter un système notionnel commun, normalisé[4] au niveau international. Son principal héritier, Helmut Felber (1987 : 131) ne semble pas échapper à la confusion qui ferait de la terminologie une discipline foncièrement normative, habilitée à déterminer une fois pour toutes ce qui existe et ce qui n'existe pas, soumettant toutes les langues de l'humanité au *diktat* conceptuel de quelques langues européennes. Aujourd'hui encore, Felber (1994 : 165) propose de procéder à une unification notionnelle en cas de non-isomorphisme. Il est pourtant paradoxal que dans le même temps il présente comme normal le fait qu'un même objet puisse être conceptualisé de manière différente selon les disciplines envisagées[5].

2.2. Un réseau notionnel interlinguistique (R.N.I.)

Dans un article intitulé *Terminological Equivalence and Translation*, Reiner Arntz (1993 : 6-7) se fonde sur le problème de la divergence dans la manière dont les langues désignent les couleurs pour montrer qu'il convient avant toute chose de décrire les systèmes notionnels propres à chaque langue. Pour Arntz, l'approche descriptive constitue le fondement d'une terminologie multilingue orientée vers la traduction. Elle permet de comparer les systèmes notionnels de chaque langue pour découvrir toutes les divergences à prendre en compte lors de l'établissement des équivalences. Il préfère toutefois ne pas recourir à la normalisation dans une perspective de traduction et propose de résoudre les difficultés éventuelles par des procédés linguistiques tels l'emprunt, la néologie et la paraphrase.

Cette perspective est intéressante, car elle consiste à rendre compatibles les réseaux notionnels de chaque langue plutôt que de les standardiser internationalement. Dans une terminographie multilingue, chaque langue doit pouvoir servir indistinc-

3 Lire notamment Wuster (1971 : 36ssq. et 44-45 ; 1968 : 2.19 ; 1981 : 66 et 71). Ce constat est également présent chez Felber (1987 : 128ssq.).

4 Dans le même article, Wuster (1971 : 40-41) va même jusqu'à parler d'*épuration*, mot sans ambiguïté quant à la nature de la tâche de normalisation.

5 Felber (1994 : 169) propose une intéressante modélisation de cette variation notionnelle en fonction des disciplines. Il est intéressant de noter que l'auteur ne tient pas compte du cas où la différence de conceptualisation est marquée par un terme différent. Il est vrai qu'un tel cas s'apparenterait étrangement à celui d'une inacceptable différence de découpage notionnel entre les langues.

tement de langue source ou de langue cible. La seule manière de satisfaire à cette exigence sans verser dans la normalisation semble bien être de fusionner les réseaux notionnels de chacune des langues considérées de manière à rendre compte de toutes leurs particularités. Pour établir ce réseau notionnel commun, que nous nommerons dorénavant **réseau notionnel interlinguistique** ou **R.N.I.**, le terminologue doit nécessairement partir de l'observation des désignations de chaque langue pour identifier les concepts qu'elle véhicule (sémasiologie). La recherche des équivalents (onomasiologie) s'effectue ensuite, mais elle doit, autant que possible, être respectueuse des faits décrits.

Dans une telle perspective, l'activité de normalisation n'est pas une condition nécessaire à l'établissement de l'équivalence. Arntz (*ibid.*) décrit d'ailleurs la normalisation terminologique comme une activité parallèle, quand bien même elle est également précédée d'une phase descriptive. Contrairement à ce qu'affirme Felber (1987 : 152), l'approche descriptive n'est donc pas qu'« une phase préliminaire qui prépare le travail terminologique normatif » ; elle peut aussi constituer le fondement d'une démarche d'établissement de l'équivalence.

S'il est arrivé à Wüster (1981 : 79) de parler de « *système de notions international* », il ne semble pas avoir voulu désigner par ces mots la démarche du R.N.I. décrite ci-dessus, mais plutôt les systèmes notionnels unifiés internationalement qui existent comme tels dans quelques domaines et qui ne requièrent donc pas de normalisation. Toutefois, l'introduction du *Dictionnaire multilingue de la machine-outil* montre que, confronté à la réalité des langues, Wüster (1968 : 2.19) a adopté une démarche plus descriptive que normative[6].

2.3. Principe d'équivalence notionnelle et découpage conceptuel

Dans un article fort intéressant, Bernard Levrat et Gérard Sabah (1990 : 93) rappellent que dans divers réseaux sémantiques, un lien d'équivalence permet de représenter les relations de synonymie. Ils montrent que « lors de la gestion automatique du réseau, ce lien peut être utile pour mettre en évidence des polysémies potentielles : si A est synonyme de B et si A est synonyme de C alors que B n'est pas synonyme de C, c'est que probablement A possède deux sens qui devraient être différenciés par deux nœuds du réseau. »

Les réseaux, qu'ils soient notionnels ou sémantiques, sont bâtis sur une perspective conceptuelle. Cette citation montre que dans un réseau sémantique, la synonymie est basée sur l'équivalence entre deux concepts, comme l'est la traduction dans un réseau notionnel interlinguistique. Tout semble donc nous autoriser à transposer la loi qui vient d'être énoncée pour l'adapter à la distinction des notions (ou concepts) en terminologie traductionnelle. L'énoncé qui suit explique comment identifier les termes qui renvoient à plusieurs notions et qui devront vraisemblablement faire l'objet d'un dégroupement hyponymique au sein du R.N.I. :

6 À ce sujet, lire également Arntz (1993 : 6-7).

Si A de L_1 est équivalent à α de L_2 et si A de L_1 est équivalent à ß de L_2 alors que α de L_2 n'est pas synonyme de ß de L_2, c'est que probablement A de L_1 possède deux sens qui devraient être différenciés par deux nœuds du réseau.

	L_1		L_2
notion 1	A	=	α
notion 2 :	A	=	ß

Ce principe, que nous dénommerons **principe d'équivalence notionnelle (P.E.N.)**, est scrupuleusement respecté dans notre corpus de référence. Grâce au dégroupement homonymique, le terminographe a veillé à ce qu'à chaque notion identifiée corresponde un terme adéquat. Ces dégroupements homonymiques peuvent être dus à une ou plusieurs langues :

Watch. The act of vigilance.	**Veille** Action de veiller.	**Wache ; Wachen.**
Watch. The divisions of time by day and night on board a ship, when a certain portion of a vessel's crew are on duty	**Quart.** Division du temps tant le jour que la nuit à bord d'un navire, pendant laquelle une certaine partie de l'équipage est de service sur le pont	**Wache.** Die Zeiteintheilung bei Tag und Nacht an Bord eines Schiffes, an der ein gewisser Theil der Bemannung Dienst auf Deck hat.
Watch. The men employed to form a watch ; for instance, the half of the crew.	**Bordée.** Nom donné à la partie d'un équipage formant le quart.	**Wache.** Benennung fur die Leute, welche eine Wache bilden (zu einer Wache gehören) (Paasch 1901 : 576)
Pilotage. The skill or knowledge of a pilot respecting coasts, rivers, channels, currents, etc.	**Pilotage.** La connaissance d'un pilote des côtes fleuves, courants, etc.	**Lootsenkunde.** Die Kenntniss eines Lootsen in Betreff der Kusten, Flüsse, Stromungen, des Fahrwassers u s w
[.]	[. .]	[ ]
Pilotage. The money paid for the services of a pilot.	**Droits de pilotage.** Contributions perçues pour les services rendus par les pilotes	**Lootsengeld.** Das, fur die Dienste eines Lootsen gezahlte Geld.
[...]	[.]	[.]
Pilot-office. The building or the rooms in a sea-port, in which the Pilot-master and assistants conduct the business in connection with pilotage.	**Pilotage.** Bureaux de l'Administration du Pilotage dans un port, où l'inspecteur du pilotage et ses assistants dirigent les affaires se rapportant au pilotage des navires.	**Lootsenwesen.** Gebäude, in welchem sich die Büreaus einer Lootsenbehörde befinden und woselbst alle dieses Fach betreffenden Angelegenheiten erledigt werden. (Paasch 1901 512)

Bien entendu, l'application stricte du principe d'équivalence notionnelle implique que la présence d'un synonyme dans l'une des langues concernées suffise à justifier le principe de dégroupement, conformément à la loi d'établissement des nœuds du réseau monolingue (Levrat et Sabah *op.cit.*). Par exemple, dans le passage suivant :

Breakwater. A structure of timber; iron or steel plates, say from one to four feet in height according to the size of the vessel, fitted across forecastle-decks (notably of large steamers) to break the force of any sea shipped over the bows.	**Brise-lame.** Construction en bois, en fer ou en acier, ayant une hauteur de un à quatre pieds selon la grandeur du bâtiment, fixée en travers d'un pont de gaillard (notamment sur les grands steamers) pour briser les lames ou pour diminuer la force de celles-ci lorsqu'elles s'élèvent sur l'avant du navire.	**Brechwasser.** Ein Gefüge von Planken, eisernen oder stählernen Platten, je nach der Grosse des Schiffes, ein bis vier Fuss hoch, welches quer über ein Backdeck (besonders bei grossen Dampfern) angebracht ist, um die Gewalt der über den Bug stürzenden Wellen zu brechen (Paasch 1901 43)
Breakwater. A stone-wall built up from the bottom of the sea, at the entrance of a bight, etc., to form a harbour, or to shelter one	**Brise-lames.** Sorte de digue (ou mur de pierres) érigée sur le fond de la mer en avant d'un port et qui s'élève jusqu'au-dessus des eaux, pour amortir la violence des vagues, et protéger le port	**Wellenbrecher; Brechwasser.** Eine am Eingange einer Bucht u.s.w vom Grunde der See aufgebaute, deichähnliche Mauer, an welcher sich die Gewalt der Wellen bricht (Paasch 1901 : 424)

De nombreux cas de dégroupements homonymiques, tel le dernier cité, apparaissent d'autant plus justifiés que les notions concernées relèvent de sous-domaines différents et ne sont donc pas liées. Très souvent d'ailleurs, la prise en compte des liens notionnels corrobore la nécessité de distinguer plusieurs notions en vertu du P.E.N. Ainsi, il suffit de s'apercevoir que le terme peut être classé dans deux arborescences espèce-genre différentes pour se rendre compte qu'il recouvre vraisemblablement deux notions différentes.

On notera toutefois que des notions se distinguent parfois sur la base du seul réseau notionnel, sans qu'intervienne le P.E.N. : elles sont désignées par des termes homonymes dans chaque langue, mais recouvrent des réalités distinctes, liées par une relation fonctionnelle[7] qui ne s'exprime pas aisément.

Course. The direction, over sea, from one point of land to another.	**Route.** Chemin à parcourir par voie de mer, de l'un point de terre à un autre	**Kurs; Curs.** Die Richtung über See, von einer Landspitze zu einer anderen
Course. The direction in which a vessel sails by compass.	**Route.** La direction qu'un navire suit d'après la boussole	**Kurs; Curs.** Der Kompasstrich, auf dem ein Schiff segelt, um einen bestimmten Ort zu erreichen. (Paasch 1901 · 443)

7. Sur les relations fonctionnelles, lire Lerat (1990).

2.4. Vers une multiplication du nombre de notions ?

Une terminographie multilingue conçue sur la base du réseau notionnel d'une seule langue ne fonctionne correctement que lorsque ladite langue sert de langue source. Le rôle du principe d'équivalence notionnelle est précisément de répondre à une des exigences fondamentales du R.N.I. : que chaque langue puisse indifféremment servir de langue source ou de langue cible. Ce principe a toutefois pour corollaire inévitable un net accroissement du taux d'homonymie pour les langues qui possèdent les notions de plus grande extension. Tel est le cas chez Paasch, puisque comme l'attestent les extraits déjà cités, de nombreux termes homonymes sont présents dans *De la quille à la pomme de mât*.

À travers l'étude de ce dictionnaire, nous avons tenté d'isoler les principes théoriques qui expliquent comment et pourquoi le recours à l'homonymie permet, autant que l'emprunt, la néologie ou la périphrase, de résoudre des problèmes d'équivalence partielle. Nous avons ainsi été amené à découvrir que contrairement à nos prévisions, l'inévitable accroissement du nombre de notions au sein du R.N.I. était souvent restreint par un étrange mécanisme régulateur dont nous nous proposons d'analyser le fonctionnement.

3. Relation d'hyponymie, homonymie et équivalence

L'idée que les relations qui lient les notions d'un même domaine ou sous-domaine forment un réseau porteur d'informations est fort proche de celle qui a conduit à l'élaboration des réseaux sémantiques. La comparaison peut aller beaucoup plus loin, puisque les relations qui entrent en jeu dans les réseaux notionnels et dans les réseaux sémantiques sont de nature voisine. Les cogniticiens, qui ont joué un rôle fondamental dans l'établissement des premiers réseaux sémantiques, ont mis en valeur le rôle fondamental de la relation hyponymique espèce-genre (ci-après, relation TY). À la suite des travaux de Quillian (1967), on pensait copier ainsi un processus cérébral de stockage lexical fondé sur le principe d'héritage des propriétés au sein d'arborescences fondées sur la relation hyponymique[8]. Or, ce type de relation occupe une place prépondérante dans la macrostructure du dictionnaire de Paasch et c'est l'exploitation de quelques principes liés à l'hyponymie qui permet d'y résoudre divers problèmes d'équivalences.

3.1. Trois réseaux notionnels à confronter

Pour montrer la manière dont fonctionne le R.N.I., nous allons isoler, à titre d'exemple, une petite partie du réseau notionnel du sous-domaine de la voiture (Paasch, 1901 : 338-352). Ce domaine se révèle particulièrement intéressant dans la mesure où, pour dénommer des réalités identiques, l'anglais, le français et l'allemand ont adopté des systèmes de désignation fort proches et fondés sur l'hyponymie. Toutefois, diverses divergences de point de vue posent des problèmes d'isomorphisme entre ces trois langues.

8 Les terminoticiens s'intéressent beaucoup aux travaux des cogniticiens, notamment à ceux qui ont abouti à la création de *Wordnet* (Miller, 1990).

En anglais, comme en français et en allemand, le système de désignation est fortement motivé, puisque les voiles sont nommées en fonction de leur emplacement. Le tableau n° 1 montre ainsi qu'en français, les voiles carrées se nomment de bas en haut *basses voiles, huniers, perroquets* et *cacatois.* On distingue le mât sur lequel elles se situent en joignant à leur nom (ci-après *N*) les adjectifs *petit N* (situé sur le mât de misaine), *grand N* (sur le grand mât), *grand N avant* (sur le grand mât avant), *grand N central* (sur le grand mât central) ou *grand N arrière* (sur le grand mât arrière). Pour le mât d'artimon, les désignations sont particulières (de bas en haut : *perroquet de fougue, perruche, cacatois de perruche* et *contre-cacatois de perruche*). À l'époque considérée, les huniers et les perroquets se subdivisent le plus souvent en deux voiles superposées ; celle du dessous est dite *fixe* et celle du dessus est dite *volante.*

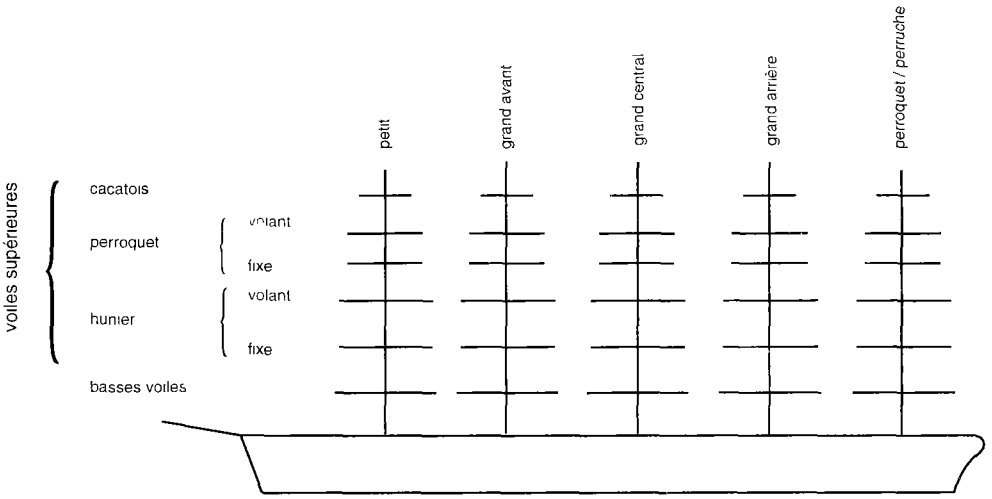


TABLEAU 1

La typologie des voiles carrées regroupe quelque 90 notions dans notre corpus (Paasch, 1901 : 338-342). Ce nombre étant beaucoup trop important, nous avons choisi de restreindre l'objet de notre démonstration aux seules voiles dénommées *cacatois* (*Royal*, en anglais et en allemand)[9]. Un décompte très précis permet de dénombrer 8 notions se rapportant aux cacatois dans le corpus. Mais ces 8 notions appartiennent à un R.N.I. trilingue et constituent le résultat de la confrontation des réseaux notionnels anglais, français et allemand. En effet, si l'on se fonde sur les légendes des illustrations et les systèmes de désignation propres à ces trois langues, on obtient des arborescences distinctes, comportant chacune un nombre différent de notions qui se rapportent pourtant toutes aux mêmes réalités matérielles. Les arborescences anglaise et allemande englobent chacune 6 notions (tableaux 2 et 3), alors que l'arborescence française en comporte 7 (tableau 4). La confrontation des langues et la prise en compte des faits de chevauchement montre que l'on aboutit au total à 10 notions différentes[10].

9 Notre choix s'est porté sur ces voiles, car elles ne se subdivisent d'ordinaire pas en cacatois fixe et cacatois volant, ce qui a le mérite de simplifier le propos.

10 Soit 1. *Royal* = *cacatois* = *Royal*, 2 *fore-royal* = *Vor-Royal*, 3. *main-royal* = *Gross-Royal*, 4. *middle-royal* = *grand cacatois central* = *Mittel-Royal*, 5 *mizen-royal* = *Kreuz-Royal*, 6 *jigger-royal* = *Jigger-Royal*, 7 *grand cacatois*, 8 *grand cacatois avant*, 9 *grand cacatois arrière*, 10. *cacatois de perruche*

Pourtant, pour rendre compte de la même réalité et permettre une traduction qui fonctionne quelle que soit la langue source et la langue cible, Paasch bâtit un réseau notionnel unique (R.N.I.) de 8 notions. Nous nous attacherons à découvrir dans les pages qui suivent comment une telle réduction peut se justifier.

[1] Royal	Cacatois	Royal; Oberbramsegel	
[2] Fore-royal	Petit cacatois	Vor-Royal	
[3] Main-royal	Grand cacatois	Gross-Royal	
[4] Main-royal	Grand cacatois avant	Gross-Royal	4MC 4MB 5MB[11]
[5] Middle-royal	Grand cacatois central	Mittel-Royal	5MB
[6] Mizen-royal	Grand cacatois arrière	Kreuz-Royal	4MC 4MB 5MB
[7] Mizen-royal	Cacatois de perruche	Kreuz-Royal	3MC
[8] Jigger-royal	Cacatois de perruche	Jigger-Royal; Besahn-Royal	4MC

(Paasch 1901 341)

——— . relation hyponymique TY

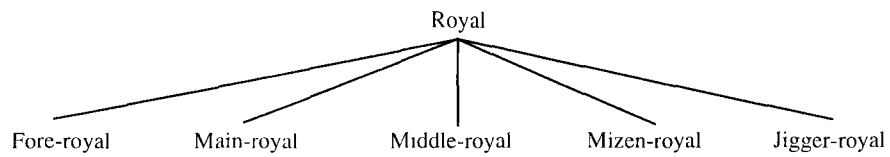


TABLEAU 2

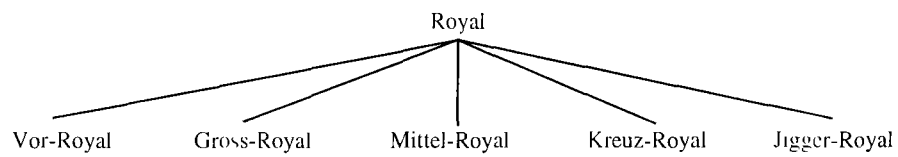


TABLEAU 3

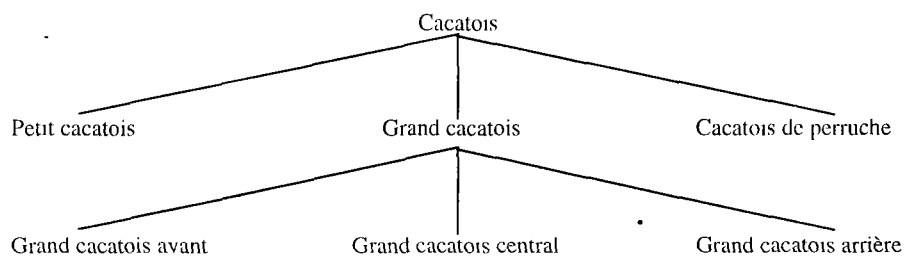


TABLEAU 4

11 3MC = trois-mâts carré, 3MB = trois-mâts barque, 4MC = quatre-mâts carré, 4MB = quatre-mâts barque, 5MB = cinq-mâts barque Les gréements carrés se distinguent des gréements de barque par la présence de voiles carrees sur le dernier mât

3.2. Hypothèse de la notion « zéro »

Comme on le constate dans les arborescences 2, 3 et 4, les réseaux notionnels allemand et anglais sont identiques, mais diffèrent de celui du français. Lorsqu'on fusionne ces trois arborescences dans un R.N.I. trilingue, on s'aperçoit que telle notion de telle langue ne possède pas de correspondant dans une autre langue. Ainsi, le cacatois situé tout à l'arrière d'un trois-mâts carré (3MC) ou d'un quatre-mâts carré (4MC) se nomme toujours *cacatois de perruche* en français. Par contre, les locuteurs anglophones et germanophones distinguent le cacatois de perruche d'un 3MC (*mizen-royal = Kreuz-Royal*) de celui d'un 4MC (*jigger-royal = Jigger-Royal*). La notion française *cacatois de perruche* possède donc une acception plus large et ne possède pas d'équivalents dans les deux autres langues. Inversement, les notions *mizen-royal = Kreuz-Royal* et *jigger-royal = Jigger-Royal* sont plus restreintes et ne possèdent pas d'équivalent en français. L'arborescence n° 5 confirme cette différence, qui nous conduit à distinguer trois notions au sein du R.N.I. : la notion française, perçue comme hyperonyme, et les deux notions « anglo-allemandes », perçues comme hyponymes.

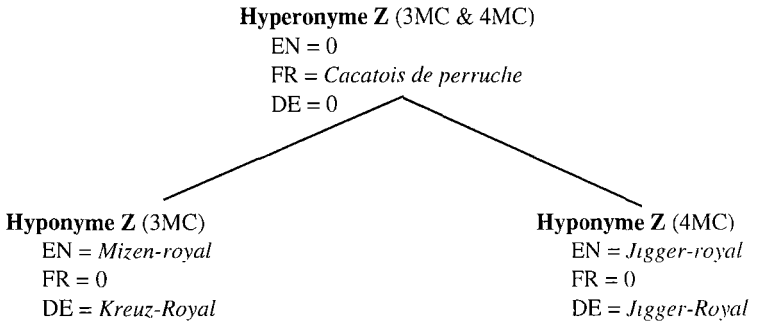


TABLEAU 5 · Mise en commun des trois réseaux notionnels au sein du R.N.I.

Nous proposons de désigner sous le nom de **notion « zéro »** (ci-après abrégée **notion Z**) toute notion du R.N.I. qui apparaît comme non prise en compte dans une langue précise lors de la comparaison des réseaux notionnels propres à chaque idiome.

3.3. Désignation par « hyperonomase »

Si l'on observe les équivalences proposées dans le corpus de référence, on s'aperçoit que le terminographe fournit un équivalent à la notion Z hyponyme en ayant recours à son hyperonyme immédiat[12]. Ainsi, pour désigner en français les notions hyponymes *mizen-royal = Kreuz-Royal et jigger-royal = Jigger-Royal,* il réutilise le terme hyperonyme *cacatois de perruche,* qui, en français, renvoie à ce type de voile quel que soit le nombre de mâts.

12. Selon un principe de substitution très fréquemment attesté dans les textes spécialisés (« *le hunier* » utilisé pour « *le grand hunier fixe* »). Malheureusement, ce procédé fonctionne mal dans les phrases négatives (l'énoncé « *ce n'est pas le grand hunier fixe* » ne peut pas toujours être remplacé par « *ce n'est pas un hunier* »)

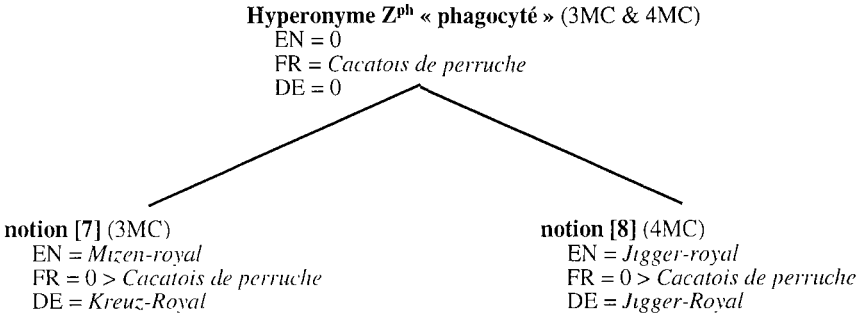


TABLEAU 6 : R N I. adapté aux besoins de la traduction

Nous risquons le terme **hyperonomase**[13] pour rendre compte du processus qui consiste à désigner, dans une langue déterminée, une notion Z hyponyme à l'aide de son hyperonyme. Dès à présent, on perçoit que c'est l'hyperonomase qui engendre l'homonymie et que les notions Z désignées par hyperonomase ont toujours une extension plus restreinte que celle de leur hyperonyme, dont elles sont évidemment homonymes.

3.4. « Phagocytose » de l'hyperonyme Z

L'explication de l'équivalence n'est assurément pas aussi simple, car si l'hyperonomase permet de désigner les notions [7] et [8] dans chacune des langues, elle ne rend pas compte de la disparition de la notion Z hyperonyme *cacatois de perruche* dans le dictionnaire de Paasch. En effet, l'hyperonomase fournit un équivalent français aux deux notions hyponymes, mais point d'équivalents anglais et allemand à la notion hyperonyme. Dans le dictionnaire, les équivalents sont bel et bien prévus pour les notions [7] et [8], mais non pour la notion Z hyperonyme. En effet, la notion *cacatois de perruche* constitue au sein du R.N.I. un générique inutile, qui peut être aisément supprimé. Un cacatois de perruche se situe nécessairement à bord d'un 3MC (notion [7]) ou d'un 4MC (notion [8]).

Il semble bien que dans certains cas l'hyperonomase entraîne la disparition pure et simple de la notion Z hyperonyme. Elle est littéralement « phagocytée » par les notions Z hyponymes dès lors que l'hyperonomase rend inutile sa traduction dans les autres langues (on parlera ci-après de **notion hyperonyme Z phagocytée** ou **Z^{ph}**). Nous sommes persuadé que le dictionnaire de Paasch, conçu pour la traduction, obéit à ce principe de la **phagocytose**, lequel mérite assurément d'être approfondi d'un point de vue théorique.

Cette disparition pose inévitablement un problème de traduction : s'il est aisé de traduire les termes anglais *mizen-royal* et *jigger-royal* vers le français en usant de l'hyperonomase, force est de reconnaître que l'inverse n'est pas exact. Si, dans le

13 Ce néologisme est, certes, critiquable, mais permet d'éviter de lourdes circonlocutions. En récupérant *onomase* pour lui adjoindre *hyper-*, nous complétons la famille *hyponyme, hyponymie, hyperonyme* (proposée par Lyons, 1970 : 347) tout en suivant – du moins en synchronie – le modèle de la famille *paronyme, paronymie, paronomase*

cadre d'une relation générique (TY), l'hyperonyme peut toujours désigner l'hyponyme (car il l'englobe), inversement, l'hyponyme ne permet pas de désigner l'hyperonyme (car il est plus restreint). Confronté au terme français *cacatois de perruche* utilisé comme générique, le traducteur hésitera entre les deux hyponymes anglais (regroupés dans le dictionnaire de Paasch) et sans doute les coordonnera-t-il dans sa traduction.

3.5. Approche théorique de l'hyperonomase et de la phagocytose

3.5.1. Le chevauchement culturel et le référent commun

Le caractère spécialisé du domaine abordé peut donner une complexité apparente aux deux phénomènes qui viennent d'être présentés : l'hyperonomase et la phagocytose. Il ne s'agit pourtant, *a priori*, que de cas où l'absence d'isomorphisme se traduit par une inclusion de la notion d'une langue dans la notion d'une autre langue.

On constate, en effet, que dans tous les cas où la phagocytose est envisageable, le référent des notions hyponymes peut être désigné par le terme hyperonyme. Dans tous les cas de notion Z^{ph} rencontrés, il apparaît que l'extension de Z^{ph} correspond parfaitement à l'addition des extensions de chacun de ses hyponymes. On dira que l'hyperonyme Z^{ph} est capable de désigner les objets conceptualisés comme co-hyponymes dans d'autres langues : il désigne les mêmes référents. Lyons (1970 : 349-350) a déjà évoqué à sa manière le problème de la notion Z^{ph} en montrant bien que dans le cadre d'une relation hiérarchique, le choix d'utiliser le terme hyperonyme pour désigner l'hyponyme permet de résoudre ce qu'il dénomme *le non-isomorphisme des langues.* Remarquons toutefois que Lyons concluait à l'absence de règle sémantique et au règne de l'intuition, constat que nous allons tenter de dépasser dans les pages qui suivent.

3.5.2. Le R.N.I. pour contexte

Il convient de rappeler que les notions Z et Z^{ph} n'existent que dans le cadre du R.N.I., c.-à-d. dans le cadre d'une confrontation des langues. À notre connaissance, le principe du recours à l'hyperonyme n'a jamais été établi en termes d'adaptation du R.N.I. aux besoins de la traduction. En accordant une si grande importance à la relation espèce-genre, vue comme foncièrement hiérarchique, Wüster avait assurément l'intuition de ce principe de l'hyperonomase. Toutefois, il n'a pas cherché à l'expliquer et ne l'a guère exploité[14], dans la mesure où il acceptait difficilement l'homonymie entre l'hyperonyme et l'hyponyme, perçue comme un sommet de l'ambiguïté plutôt que comme le résultat inévitable de la confrontation des langues.

> Certains termes sont tellement ambigus qu'ils désignent à la fois une notion et l'un des spécifiques de cette notion (*homonymes verticaux*). En terminologie, on distingue ces deux notions en ajoutant le chiffre romain [1], en exposant, après

14 Dans le *Dictionnaire multilingue de la machine-outil,* Wüster (1968) utilise divers symboles qui permettent d'annoncer les cas d'équivalence partielle, mais non de les résoudre. Il ne recourt qu'exceptionnellement aux dégroupements homonymiques.

> le terme lorsqu'on veut parler du sens large d'un point de vue logique. Dans l'autre cas, on utilise le chiffre [1], en exposant, après le terme [...]. (Wüster, 1981 : 88)

Felber (1987 : 153) montre lui-même que ces « homonymes verticaux », qu'il nomme *homonymes polysèmes*, entretiennent bien une relation hyponymique, voire une relation partie-tout. Il ne semble toutefois pas établir de lien entre cette perspective de passage de la polysémie à l'homonymie et la comparaison des réseaux notionnels de chaque langue.

3.5.3. Hyponyme = hyperonyme + actualisation d'un caractère virtuel

Dans la théorie viennoise, les notions se composent d'un ensemble de **caractères.** Ces caractères, qui sont des propriétés des objets conceptualisés, permettent de différencier ou de rapprocher les notions[15]. Normalement, les genres sont distingués des espèces selon un même **type de caractère**, c.-à-d. en fonction de caractères fondés sur un même **critère de subdivision**[16] (p. ex. le nombre de mâts, dans la subdivision des voiliers en trois-mâts, quatre-mâts, cinq-mâts, etc.).

Tous les co-hyponymes d'un même hyperonyme possèdent inévitablement un certain nombre de caractères en commun, lesquels correspondent exactement aux caractères de leur hyperonyme. Tel est par exemple le cas pour les types de grands cacatois. On constate clairement dans le tableau qui suit que les trois hyponymes *grand cacatois avant, grand cacatois central* et *grand cacatois arrière* possèdent les mêmes caractères que leur hyperonyme *grand cacatois*, dont ils se différencient par un caractère au moins[17]. Rien n'interdit toutefois de dire que l'hyperonyme possède également ces caractères de manière virtuelle. Comment expliquer autrement que le terme *grand cacatois* puisse servir à désigner chacun des trois hyponymes ? L'idée d'une prise en compte de caractères virtuels paraît d'autant plus envisageable que les terminologies dénomment fréquemment les hyponymes par des syntagmes qui adjoignent un caractère lexicalisé derrière le terme hyperonyme.

grand cacatois	'cacatois'[18]	'sur un grand mât'	0
grand cacatois avant	'cacatois'	'sur un grand mât'	'avant'
grand cacatois central	'cacatois'	'sur un grand mât'	'central'
grand cacatois arrière	'cacatois'	'sur un grand mât'	'arrière'

TABLEAU 7

Nous nommerons **caractères virtuels** les propriétés d'un objet qui, dans une langue donnée, ne sont pas conceptualisées pour délimiter la notion alors qu'elles le sont

15 « Caractère : Représentation mentale d'une propriété d'un objet (2.1) et qui sert à en délimiter la notion (3.1). » (ISO 1087, 1990 : 2)

16 « Type de caractère : Toute catégorie de caractère utilisée comme critère dans l'établissement d'un système de notions génériques » (ISO 1087, 1990 : 2). La norme ISO 704 (1987 : 4) parle de *critère de subdivision*, terme qui nous paraît plus transparent et que nous adopterons dans la suite de l'étude.

17 « [...] le concept spécifique a les caractères du concept générique plus un au moins. Au fur et à mesure qu'on monte vers du plus général, on est en présence de concepts dits plus 'abstraits' » (Lerat, 1990 : 81).

18. Par convention, les caractères sont représentés entre guillemets simples.

dans d'autres. Une telle approche implique, pour la rigueur du propos, que l'on reconsidère la définition de la notion proposée par l'ISO 1087 (1990 : 1) : dans un contexte multilingue, la notion doit, en effet, être définie comme la conceptualisation d'un ou de plusieurs objets à partir de certaines de leurs propriétés (caractères), identifiées comme pertinentes dans une langue donnée.

3.5.4. R.N.I. et instabilité notionnelle

En terminologie, les notions sont réputées stables. Ce qui est vrai tant qu'on demeure dans une perspective monolingue tend pourtant à devenir très relatif dans une perspective multilingue. En effet, dans le cadre d'une recherche d'équivalences au sein du R.N.I., le terme n'apparaît plus comme monosémique et, selon son sens, se traduira de telle ou telle manière. Ainsi, confronté à l'anglais et à l'allemand, le sens du terme français *cacatois de perruche* se met à varier (la notion se dédouble). Le tableau 8 mentionne les caractères considérés comme pertinents dans le système notionnel de chaque langue pour distinguer les cacatois de perruche. On y observe une correspondance exacte des caractères de la notion anglaise *mizen-royal* avec ceux de la notion allemande *Kreuz-Royal,* d'une part, et des caractères de la notion anglaise *Jigger-royal* avec ceux de la notion allemande *Jigger-Royal,* d'autre part.

FR *cacatois de perruche*[0]		'cacatois'	'dernier mât'	0
EN . *mizen-royal*	.	'cacatois'	'dernier mât'	'à bord d'un 3MC'
EN · *jigger-royal*	.	'cacatois'	'dernier mât'	'à bord d'un 4MC'
DE : *Kreuz-Royal*	:	'cacatois'	'dernier mât'	'à bord d'un 3MC'
DE . *Jigger-Royal*	:	'cacatois'	'dernier mât'	'à bord d'un 4MC'

TABLEAU 8

On remarquera également que les caractères « à bord d'un 3MC » et « à bord d'un 4MC » ne sont pas nécessaires pour décrire la notion française *cacatois de perruche.* Toutefois, ces mêmes caractères « à bord d'un 3MC » et « à bord d'un 4MC », deviennent pertinents dans le cadre d'une traduction vers l'anglais ou l'allemand. En effet, dans le cadre du R.N.I., ces caractères doivent être pris en considération de manière à trouver une équivalence en vertu du principe d'équivalence notionnelle ; c.-à-d. que pour arriver à désigner le même objet, il apparaît indispensable de le conceptualiser de la même manière.

FR : *cacatois de perruche*[1]	:	'cacatois'	'dernier mât'	0 > 'à bord d'un 3MC'
EN . *mizen-royal*		'cacatois'	'dernier mât'	'à bord d'un 3MC'
DE : *Kreuz-Royal*		'cacatois'	'dernier mât'	'à bord d'un 3MC'
FR . *cacatois de perruche*[2]	:	'cacatois'	'dernier mât'	0 > 'à bord d'un 4MC'
EN . *jigger-royal*	.	'cacatois'	'dernier mât'	'à bord d'un 4MC'
DE . *Jigger-Royal*	·	'cacatois'	'dernier mât'	'à bord d'un 4MC'

TABLEAU 9

En théorie, tout hyperonyme, même éloigné, est apte à désigner de manière univoque le même objet que son lointain hyponyme : il suffit que la notion hyperonyme ne se distingue de la notion hyponyme que par des caractères virtuels. Cruse (1986 : 155) fait toutefois remarquer que le recours à l'hyperonyme engendre toujours une sous-spécification. Ceci explique sans doute que Paasch utilise toujours l'hyperonyme immédiat, lequel est d'ailleurs perçu comme le plus utile par Pierre Lerat (1988 : 20).

3.5.5. Pourquoi la phagocytose ?

Dans le corpus de référence, seules les deux notions hyponymes subsistent après phagocytose de la notion Z^{ph} *cacatois de perruche*[0]. Par le mécanisme de l'hyperonomase, les caractères virtuels de l'hyperonyme sont activés au niveau hyponymique, de sorte qu'il se révèle apte à désigner chaque hyponyme. Dans la mesure où tous les caractères virtuels de l'hyperonyme Z^{ph} se trouvent ainsi activés sous toutes leurs valeurs possibles au niveau des hyponymes, ledit hyperonyme Z^{ph} ne désigne plus aucun objet qui ne soit concrètement représenté par ses hyponymes. La notion hyperonyme Z^{ph} devient donc superflue au sein du R.N.I. d'un dictionnaire de traduction et peut être phagocytée.

[7] Mizen-royal	Cacatois de perruche	Kreuz-Royal	3MC
[8] Jigger-royal	Cacatois de perruche	Jigger-Royal; Besahn-Royal	4MC

(Paasch 1901 : 341)

Dans le cadre d'une entreprise visant à permettre la communication entre des locuteurs de langues différentes, il paraît plus utile de traduire deux notions spécifiques dans la langue qui ne les désignait pas que de rendre compte d'une notion générique que ladite langue était la seule à prévoir. S'agissant de désigner des objets, l'extension du générique correspond toujours au total des extensions des notions spécifiques. Dès lors que ces dernières sont dénommées dans chacune des langues, le générique ne constitue plus qu'une abstraction de peu d'utilité. La phagocytose paraît ainsi s'imposer d'elle-même dans le cas de la distinction entre les notions *mizen-royal* = *cacatois de perruche*[1] = *Kreuz-Royal* (3MC) et *jigger-royal* = *cacatois de perruche*[2] = *Jigger-Royal* (4MC), lesquelles rendent inutile toute référence à une notion recouvrant en même temps le cacatois de perruche d'un 3MC et celui d'un 4MC. En d'autres termes, au sein du R.N.I., l'hyperonyme Z^{ph} ne désigne rien que ne désignent déjà ses hyponymes.

L'actualisation du caractère virtuel au niveau hyponymique entraîne en traduction une modification implicite du sens de l'hyperonyme, mais il ne s'agit jamais que d'un artifice terminographique visant à établir l'équivalence. Que l'on parle du cacatois de perruche d'un 3MC ou de celui d'un 4MC, on le désigne toujours par le terme *cacatois de perruche*. Jamais il n'est demandé au locuteur francophone de revoir la manière dont il appréhende le réel au nom d'une quelconque normalisation.

Par ailleurs, si les notions sont classées en vertu du lien TY – et tel est le cas chez Paasch –, les homonymes nés d'une phagocytose se retrouvent normalement regroupés dans le dictionnaire. Le traducteur francophone peut ainsi découvrir que dans le cadre d'un contexte anglais ou allemand qui établit une nette distinction entre

mizen-royal et *jigger-royal,* entre *Kreuz-Royal* et *Jigger-Royal* (cf. 3.4.), il convient de spécifier davantage la portée du terme *cacatois de perruche* en y adjoignant un complément déterminatif (*de trois-mâts carré, de quatre-mâts carré*). Réciproquement, un traducteur anglais ou allemand découvrira que la traduction du générique français *cacatois de perruche* appelle une interprétation du contexte pour décider du caractère virtuel activé.

3.5.6. Existe-t-il des hyperonomases sans phagocytose ?

Dans les exemples produits jusqu'à présent, l'hyperonomase s'accompagne toujours d'une phagocytose de l'hyperonyme Z. Il est assurément des cas où l'hyperonyme n'est pas une notion Z et n'est donc pas « phagocytable ». Toutefois, ces cas sont rares dans *De la quille à la pomme de mât*, sauf exceptions propres au chapitre des *Termes généraux*. C'est ainsi que la possibilité d'opérer une distinction entre les notions *observatoire astronomique* et *observatoire météorologique* n'exclut pas la nécessité de devoir éventuellement faire référence à la notion générique *observatoire,* dans le cas d'un établissement qui réunirait les deux fonctions, voire davantage.

Observatory. Any place from where a view may be observed.	**Point d'observation**. Un endroit quelconque duquel on jouit d'une vue	**Warte.** Ein erhabener Ort von wo man eine freie Aussicht hat
Observatory. A building fitted with installations and instruments necessary for making astronomical, etc observations.	**Observatoire.** Etablissement pourvu des installations et des instruments nécessaires pour les observations astronomiques, météorologiques, etc.	**Warte.** Ein für astronomische, meteorologische u s.w. Beobachtungen eingerichtetes, und mit den hierzu erforderlichen Instrumenten ausgestattetes Institut.
Observatory. (astronomical)	**Observatoire** (astronomique)	**Sternwarte.**
Meteorological-observatory.	**Observatoire** (météorologique)	**Wetterwarte.**

(Paasch 1901 : 505)

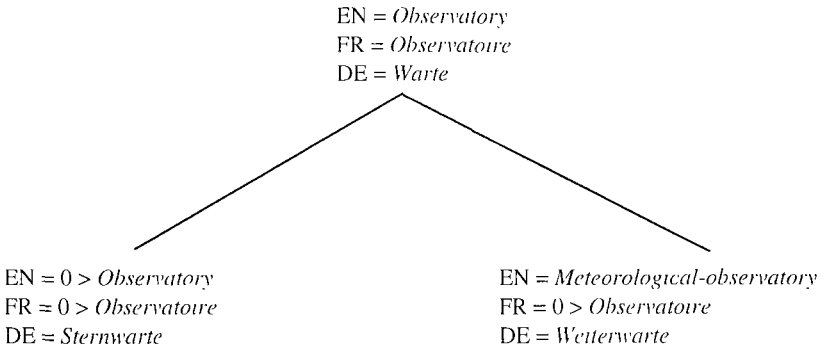


TABLEAU 10

3.5.7. La relation TY, condition de l'hyperonomase et de la phagocytose

Comme on le constate, l'arborescence n° 10 ne peut rendre compte de l'équivalence *observatory* = *point d'observation* = *Warte,* car la notion ainsi désignée n'appartient pas à la même arborescence TY. Il nous semble important de remarquer que si l'hyperonomase répond au principe d'équivalence notionnelle, elle n'en est qu'une forme d'accomplissement très particulière, liée à une relation hyponymique observable au sein du R.N.I.

Il ne peut être question de parler d'hyperonomase et encore moins de phagocytose pour les cas cités en 2.3 (*watch, pilotage, route*), car aucune des trois langues ne possède une notion qui serait l'hyperonyme de notions propres aux autres langues. Les caractères communs permettent tout au plus de déterminer un lien notionnel indéterminé qui fonde une relation fonctionnelle. Par exemple, la bordée (*watch*) « est responsable de » la veille (*watch*) « pendant » le quart (*watch*) ; ou encore, le bureau de pilotage (*pilotage*) « est le centre des activités de » pilotage (*pilotage*).

3.6. Équivalence partielle et hyponyme virtuel Z^v

La confrontation des notions peut aboutir à observer des cas où deux langues ne possèdent pas de désignation pour une réalité particulière, conçue comme incluse dans des notions plus larges mais dont l'extension varie d'une langue à l'autre. L'étude du dictionnaire de Paasch nous conduit à poser l'hypothèse de l'existence dans le R.N.I. de **notions zéros virtuelles (Zv)**. Il s'agit de notions Z hyponymes qui, bien qu'elles ne soient propres à aucune langue, doivent être désignées par hyperonomase pour résoudre le problème d'équivalence posé par de tels cas.

3.6.1. Un cas de notion Z^v

Le phénomène de la notion Z^v s'observe dans l'arborescence des types de cacatois. Les notions équivalentes *mizen-royal* = *Kreuz-Royal* correspondent à deux notions hyponymes en français : *grand cacatois arrière* (à bord d'un 4M ou d'un 5M) et *cacatois de perruche* (à bord d'un 3M). Conformément au principe de l'hyperonomase et de la phagocytose, le dictionnaire ne retient donc que les deux notions hyponymes du R.N.I. : d'une part, *Mizen-royal*$^{[1]}$ = *Grand cacatois arrière* = *Kreuz-Royal*$^{[1]}$ et, d'autre-part, *Mizen-royal*$^{[2]}$ = *Cacatois de perruche* = *Kreuz-Royal*$^{[2]}$.

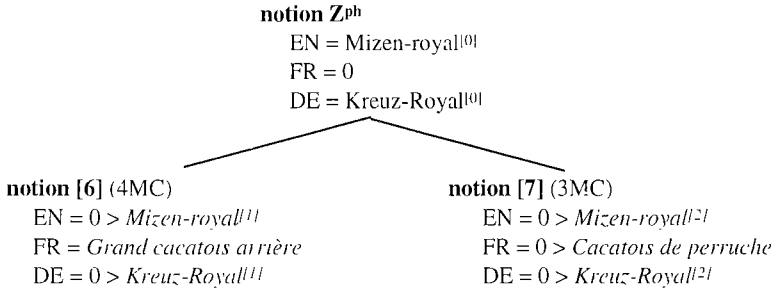


TABLEAU 11

Si l'on considère à présent l'ensemble des désignations des cacatois au sein du R.N.I., on s'aperçoit que le second hyponyme (notion [7]) correspond à une notion que nous avons déjà décrite comme résultant d'une autre hyperonomase accompagnée de phagocytose, celle décrite dans le tableau 6. Dans la mesure où tout ceci doit paraître bien abstrait, nous avons essayé de recréer, dans le tableau 11 une vue d'ensemble du R.N.I. avant phagocytose. La partie gauche de l'arborescence correspond au tableau 11 ; la partie droite, au tableau 6. Pour clarifier les notions, nous avons représenté les objets conceptualisés (voiles) par chacune d'entre elles. Il apparaît clairement que la notion [7] conceptualise exactement le même objet, quand bien même elle peut être appréhendée au départ d'hyperonymes distincts, mais possédant des caractères parfaitement compatibles : la partie gauche de l'arborescence distingue deux types de cacatois en fonction de l'emplacement du mât ; la partie droite, en fonction du nombre de mâts. Ceci explique que dans le R.N.I. de *De la quille à la pomme de mât*, il ne s'agit que d'une seule et même notion, celle que nous avons identifiée par le chiffre [7].

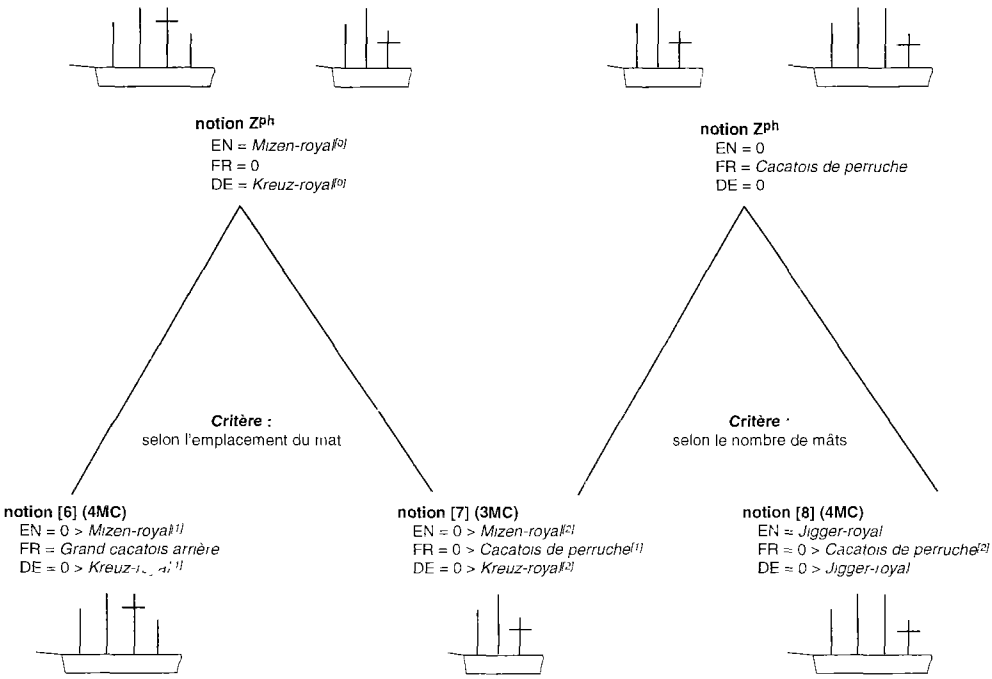


TABLEAU 12

[6] Mizen-royal[1]	Grand cacatois arrière	Kreuz-Royal[1]	4MC, 4MB, 5MB
[7] Mizen-royal[2]	Cacatois de perruche[1]	Kreuz-Royal[2]	3MC
[8] Jigger-royal	Cacatois de perruche[2]	Jigger-Royal; Besahn-Royal	4MC

(Paasch 1901 : 341)

Il s'agit d'un cas patent de notion Z^{v}. En effet, la reconstitution du R.N.I. montre que la notion [7] n'existe dans aucune langue : elle est tout à la fois hyponyme de la

notion Z^{ph} *Mizen-royal*$^{[0]}$ = *Kreuz-Royal*$^{[0]}$ (tableau 11) et de la notion Z^{ph} *cacatois de perruche*$^{[0]}$ (cf. tableau 6). La notion interlinguistique [7], présente dans le dictionnaire, est clairement une notion qui n'existe ni en anglais, ni en français, ni en allemand. Aucune de ces trois langues ne possède une notion aussi restreinte, qui ne renverrait qu'au seul cacatois du dernier mât d'un trois-mâts carré (3MC). Si le terminographe a créé cette notion virtuelle « de toute pièce », c'est bien pour permettre la traduction la plus juste, compte tenu de tous les référents envisageables.

L'illustration et la dénomination des référents permettent d'ailleurs d'aboutir empiriquement à une solution rigoureusement identique. Dans le tableau 13, inspiré du problème du découpage des couleurs proposé par Lyons (1970 : 46-47)[19], chaque case correspond à chacun des objets (cacatois) désignés dans chaque langue par un terme différent ; en d'autres termes, chaque case représente l'extension de la notion dénommée par ce terme. La confrontation des découpages confirme bien que dans le cadre d'un dictionnaire trilingue, il faut envisager trois notions différentes au sein du R.N.I. pour arriver à désigner les trois référents envisageables.

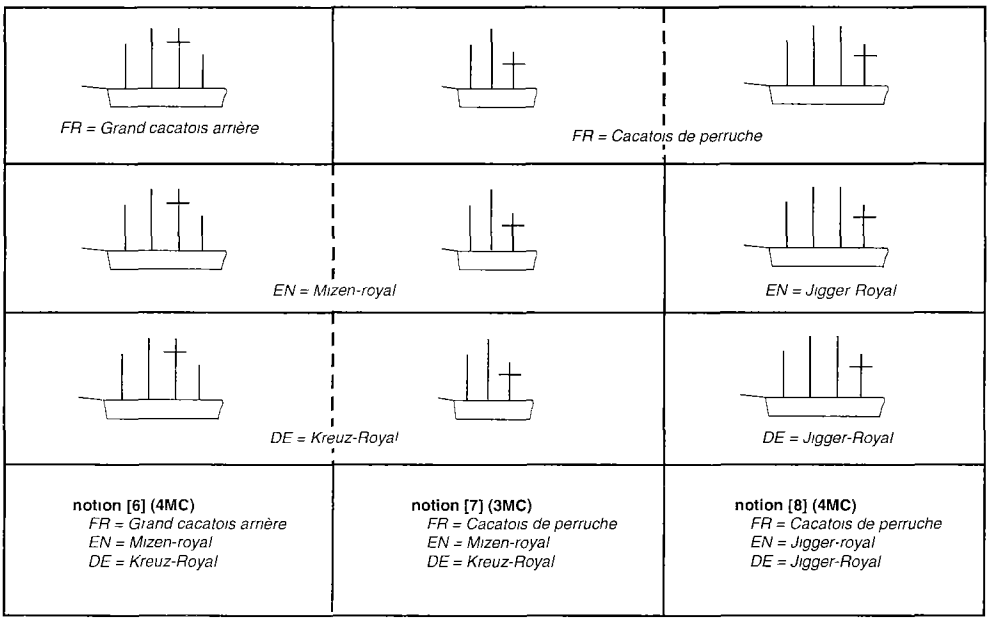


TABLEAU 13

3.6.2. Un cas complexe et exemplaire

Le modèle des notions Z^{ph} et Z^{v} permet d'expliquer la manière dont de nombreux problèmes d'équivalence particulièrement complexes ont été résolus au sein du corpus de référence. On sera, par exemple, intrigué d'y découvrir quatre entrées *diablotin*, alors

19 Nous avons déjà analysé ailleurs le lien entre le problème des couleurs et le P.E.N. (Van Campenhoudt, 1991 et 1994 70-71)

que les marins français considèrent que ce terme désigne une seule et même voile triangulaire (voile d'étai), toujours située devant le dernier mât (le mât d'artimon).

[1] Mizen-topmast-staysail	Diablotin	Kreuz-Stengestagsegel	(3MC)
[2] Mizen-topmast-staysail	Diablotin	Besahn-Stengestagsegel	(3MB, BAR, 3MG)
[3] Jigger-topmast-staysail	Diablotin	[Kreuz-Stengestagsegel][20]	(4MC)
[4] Jigger-topmast-staysail	Diablotin	Besahn-Stengestagsegel	(4MB, 5MB)

(Paasch 1901 · 343)

Un examen approfondi montre que l'on distingue en allemand deux types de diablotin selon que cette voile se situe devant un mât d'artimon qui ne comporte que des voiles axiales (syntagme formé avec *Besahn*) ou qui comporte également des voiles carrées (syntagme formé avec *Kreuz*). En anglais, on se fonde sur le nombre de mâts pour distinguer les diablotins d'un 3M (syntagme formé avec *mizen*) et ceux d'un 4M (syntagme formé avec *jigger*). En français, on considère qu'il s'agit à chaque fois d'une seule et même voile. Le tableau 14 (arborescence) rend compte de la situation de ces notions au sein du R.N.I. dès lors que l'on prend en compte les caractères distinctifs propres à chacune des langues considérées.

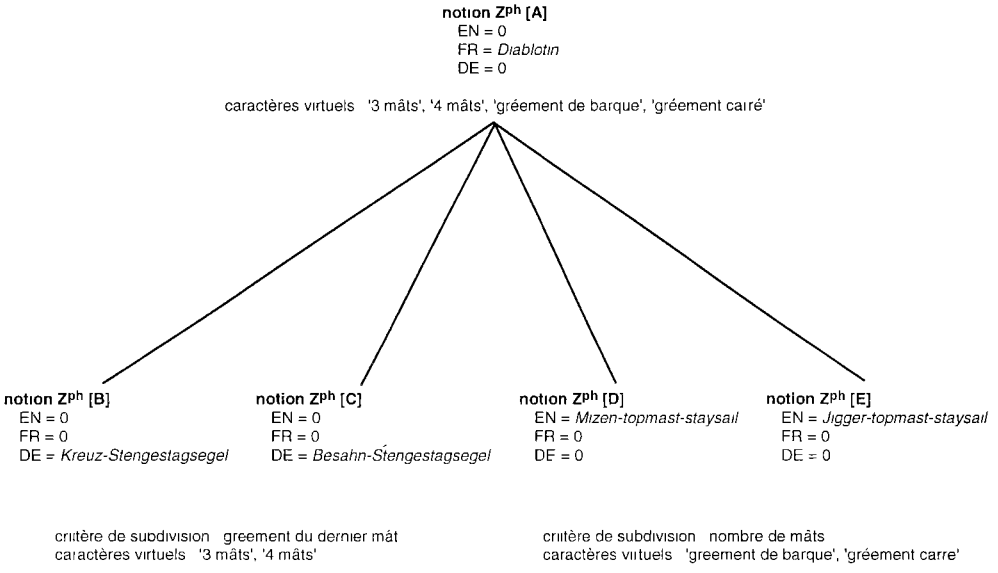


TABLEAU 14

La solution du dictionnaire correspond à la prise en compte de notions virtuelles Z^v qui préservent l'intégrité référentielle de chaque terme tout en permettant la traduction dans un R.N.I. trilingue. Cette solution est représentée sous forme d'arborescence dans le tableau 15. On y découvre que l'extension très large de la notion *diablotin* en

20 Par souci de simplifier l'exposé, nous reproduisons exceptionnellement le terme allemand tel qu'il apparaît dans la quatrième édition du dictionnaire (1908), après stabilisation du système de désignation

français fait de celle-ci un hyperonyme Z^{ph}, tant vis-à-vis des notions anglaises que vis-à-vis des notions allemandes. Le découpage hyponymique en fonction de la disposition des voiles (en allemand) ou du nombre de mâts (en anglais) intervient au niveau immédiatement subordonné. On considérera donc que dans le R.N.I., il existe quatre hyponymes de *diablotin*[0] : *Kreuz-Stengestagsegel*[0], *Besahn-Stengestagsegel*[0], *Mizen-topmast-staysail*[0] et *Jigger-topmast-staysail*[0].

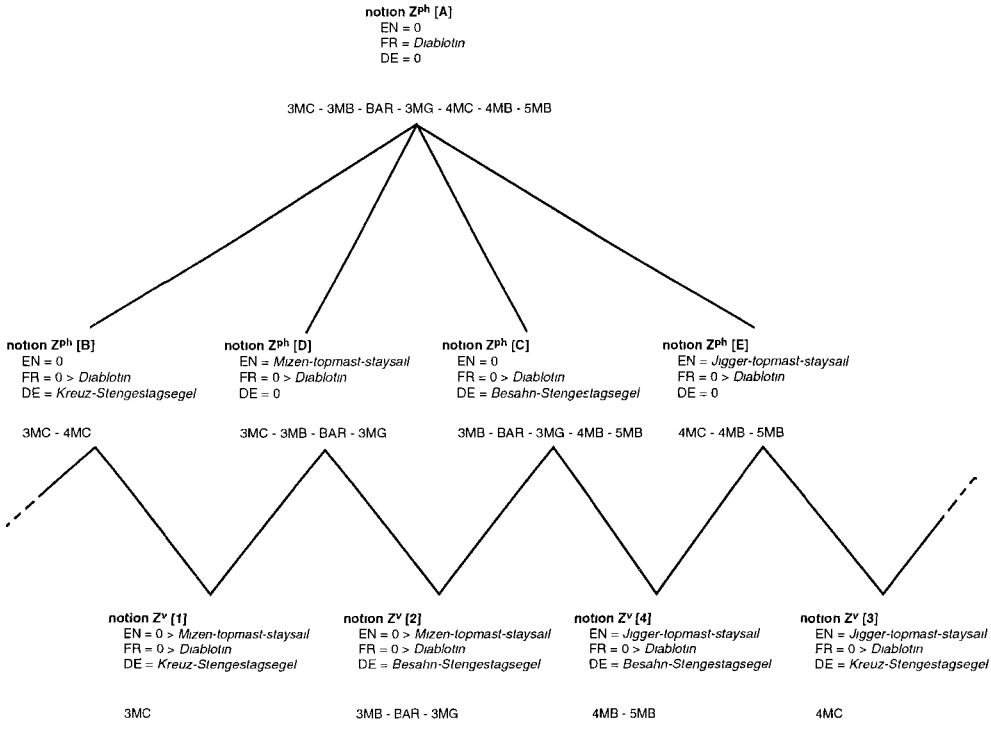


TABLEAU 15

Paradoxalement, toutes ces notions co-hyponymes constituent des notions Z^{ph}, à l'instar de *diablotin*[0]. En effet, elles ne possèdent aucun équivalent dans les deux autres langues. Toutefois, l'activation des caractères virtuels – qui correspond, plus simplement, à la prise en considération des référents – permet de dégager quatre notions Z^{v}, propres à aucune langue, mais aptes à permettre une traduction dans les six sens envisageables dès lors qu'on les désigne au moyen de leurs hyperonymes respectifs. On observe dans l'arborescence n° 15 que les quatre notions Z^{v} correspondent parfaitement aux notions mentionnées et illustrées dans le dictionnaire.

3.7. Approche théorique de la notion virtuelle

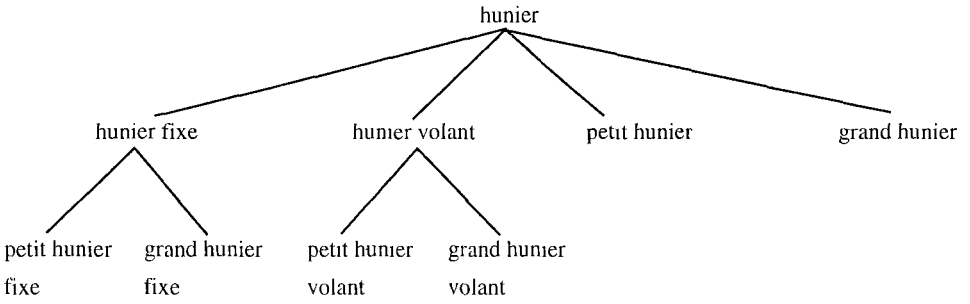
Ce mode de résolution suppose que dans un réseau multilingue une notion hyponyme puisse être subordonnée à deux hyperonymes Z^{ph} différents. En fait, même dans un réseau monolingue, une notion hyponyme peut dépendre de plusieurs hyperonymes dès lors qu'elle conserve en les combinant les caractères différenciateurs desdits hypero-

nymes et qu'elle actualise leurs caractères virtuels. Une telle notion ne possédera pas de caractère différenciateur propre.

Force est de constater que dans les cas rencontrés, la notion virtuelle Z^v combine toujours des caractères propres au système hyponymique de chacune des langues prises en compte. Les caractères combinés constituent ce que la théorie viennoise nomme des **caractères indépendants**[21].

Les caractères sont dits **dépendants** lorsqu'ils doivent nécessairement intervenir à des niveaux hyponymiques différents de la hiérarchie arborescente (Felber 1987 : 100). Par exemple, dans la distinction des types de navires, le caractère « muni d'une chaudière » précède nécessairement des caractères comme « muni de roues à aubes » ou « muni d'une hélice », auxquels il est supérieur. En effet, les notions *vapeur à roues* et *bateau à vapeur à hélice* sont hiérarchiquement subordonnées à la notion (*bateau à*) *vapeur.*

Par contre, les caractères **indépendants** « peuvent se suivre à différents niveaux d'une série verticale de notions et être combinés arbitrairement » (Felber, 1987 : 101). En d'autres termes, dans une même arborescence TY, des caractères indépendants peuvent servir à distinguer des co-hyponymes sur la base de critères de subdivision différents. Par exemple, les caractères « fixe » ou « volant », d'une part, et « situé sur le mât d'artimon » ou « situé sur le grand mât », d'autre part, permettent de distinguer quatre types particuliers de la notion *hunier : hunier fixe* et *hunier volant* d'une part, *petit hunier* et *grand hunier* d'autre part. Fondés sur des critères différents (la mobilité et l'emplacement), ces caractères sont indépendants dans la mesure où ils peuvent se combiner à un niveau inférieur pour distinguer les notions *petit hunier volant, petit hunier fixe, grand hunier volant* et *grand hunier fixe.*



(d'après Paasch 1901 . 338-339)

TABLEAU 16

Les notions hyponymes situées au croisement de deux typologies peuvent avoir autant d'hyperonymes qu'elles actualisent de critères propres à chacun d'eux. Ainsi, la notion *petit hunier volant* possède deux hyperonymes (*petit hunier* et *hunier volant*) et constitue le point de liaison de deux arborescences fondées chacune sur un critère de subdivision différent : la mobilité et le mât.

21 La distinction entre caractères indépendants et caractères dépendants n'est malheureusement plus prise en compte dans les dernières normes ISO 704 (1987) et ISO 1087 (1990)

3.8. Vers une application informatique

Notre approche théorique montre que l'équivalence possède un fondement relativement logique lorsqu'elle est obtenue au sein de la relation TY. Ce constat nous conduit à penser qu'une B.C.T. multilingue devrait être à même de déceler, voire de traiter, les notions Z au sein du R.N.I.

3.8.1. Du réseau à l'équivalence

Dans la pratique, le R.N.I. d'un glossaire multilingue se doit d'être immédiatement utile pour le traducteur qui souhaite connaître l'équivalent idoine. Cette perspective est celle qui est logiquement suivie dans un dictionnaire conçu et présenté sur papier, tel que *De la quille à la pomme de mât.*

Idéalement, une B.C.T. multilingue devrait proposer un réseau par langue. Le R.N.I. ne serait constitué que dans un second temps par une comparaison des notions de chaque langue. Une exploitation logique de chaque réseau, fondée notamment sur les caractères et la relation TY, devrait permettre d'isoler les cas de non-isomorphisme et de proposer des équivalences acceptables. Jusqu'à cette date, aucun logiciel gestionnaire de données terminologiques n'a réellement été développé dans cette perspective[22]. Or, la construction d'un réseau notionnel interlinguistique peut se révéler complexe et risque d'être remise en cause dès qu'il sera décidé d'y intégrer une nouvelle langue.

Un logiciel de terminologie « intelligemment assistée par ordinateur » devrait, en réalité, être à même de formuler diverses propositions face aux impossibilités de traduction. Ainsi, lors d'une phase d'évaluation qui suivrait l'élaboration des réseaux de chaque langue, il pourrait émettre diverses propositions comme :

> Proposition 1 : « Le terme français *cacatois de perruche* n'a pas d'équivalent en anglais, voulez-vous connaître les notions hyponymes en anglais ? »
> Proposition 2 : « Les hyponymes anglais de *cacatois de perruche* sont *mizen-royal* et *jigger-royal* et n'ont pas d'équivalents en français. Voulez-vous utiliser l'hyperonyme *cacatois de perruche* pour désigner ces hyponymes en français ? »

3.8.2. Implications définitoires

Toutefois, des précautions s'imposent : il ne saurait être question de permettre au logiciel d'altérer l'information initiale. Il s'agit plutôt d'exploiter celle-ci, de l'interpréter à la manière d'un véritable système expert chargé de seconder le terminologue.

L'idée d'une hyperonomase et d'une phagocytose assistées demeure, bien sûr, une hypothèse qui doit se vérifier à l'épreuve des faits. Elle pose des problèmes qui méritent d'être étudiés avec beaucoup de précautions, notamment le report de la définition de l'hyperonyme phagocyté au niveau de l'hyponyme. La conséquence logique de l'hyperonomase serait que dans le R.N.I., l'extension des hyperonymes re-

portés au niveau subordonné (p. ex. *cacatois de perruche*) soit plus restreinte que celle qu'ils auraient dans un ouvrage monolingue[23].

4. Synthèse : R.N.I. et approche notionnelle

4.1. Une approche contrastive du découpage notionnel

Il est apparu en 2.3. et 2.4. que le principe d'équivalence notionnelle conduit à une multiplication des homonymes au sein du R.N.I. La plupart des notions citées en guise d'exemples sont classées par Paasch dans le chapitre des *Termes généraux.* Elles ne sont guère spécialisées et pourraient fort bien être traitées de la même manière dans un dictionnaire de traduction consacré à la langue générale. Par exemple, pour déterminer l'équivalent anglais du mot *coque*, il faut nécessairement préciser à quel concept on entend faire référence : à l'enveloppe d'un fruit ou d'un œuf (= *shell*), à la carapace d'un mollusque (= *cockle*), à la carène d'un navire (= *hull*), etc.

En fait, il semble bien que le principe du dégroupement homonymique n'est qu'un avatar de la distinction entre homonymie et polysémie dans la tradition lexicographique. Tant que la perspective demeure monolingue, le lexicographe qui adopte une perspective homonymique ne dispose que de peu de critères pour décider s'il y a lieu ou non d'attribuer plusieurs entrées à un même signifiant. Par exemple, *pomme* reçoit quatre entrées dans le *D.F.C.* (1966) et six dans le *Lexis* (1987). Par contre, les dictionnaires fondés sur l'approche polysémique, comme les *Petit* et *Grand Robert,* utilisent le critère de l'étymologie pour décider du nombre d'entrées : ainsi, ils distinguent deux entrées *bière*, parce que le néerlandais *bier* et le francique *bera* ont connu des évolutions qui conduisent à attribuer des signifiants identiques à deux variétés bien distinctes de contenants.

Comparaison n'est certes pas raison, mais ce critère étymologique équivaut en quelque sorte, *mutatis mutandis,* à se servir de systèmes conceptuels propres à des langues étrangères pour présider au dégroupement homonymique. Traduites par exemple en anglais, les deux entrées *bière* requièrent des traductions différentes : *beer* et *coffin,* ce qui justifie l'existence de deux notions différentes au sein du R.N.I.

Les théoriciens de la lexicographie ont déjà abondamment disserté sur le caractère arbitraire du dégroupement homonymique basé sur une approche purement sémantique des notions. Il est bon de souligner que le P.E.N. fonde la norme non point sur des décisions arbitraires, mais sur une approche contrastive et une prise en compte de l'usage. Finalement, s'il y a un écart entre deux approches du sens, celui-ci sépare moins la terminographie et la lexicographie en soi que l'approche monolingue et l'approche multilingue. Cette dernière est *nécessairement* conceptuelle. Dès lors, s'il importe de faire référence à Wüster, ce n'est pas au nom d'une adhésion inconditionnelle à ses théories – vieillies sous plus d'un aspect –, mais parce que sa pensée fait écho à l'impérieuse nécessité, pour la traduction des langues de spécialité, de veiller à délimiter clairement le champ de l'équivalence.

23 Paasch a veillé à arranger ses définitions en fonction des hyperonomases et phagocytoses qu'il a réalisées et des dégroupements homonymiques qui en découlent

4.2. Notion zéro et relation d'hyponymie

Au fondement de l'hypothèse développée dans cet article, se situe la notion zéro. Par-delà son appellation nouvelle, nous pensons que ce concept doit faire figure d'évidence aux yeux de tout terminologue attentif à la prise en compte des liens notionnels. Lyons (1970 : 348) constatait déjà que « les vocabulaires des langues naturelles ont tendance à présenter beaucoup de cases vides, d'asymétries et d'indéterminations » à la différence de ce qui se produit dans les taxinomies scientifiques. Cruse (1986 : 145ssq.) a longuement montré que, dans une perspective monolingue, la prise en compte de la relation espèce-genre conduisait à observer des « vides » (*gaps*) à divers niveaux de superordination de l'arborescence TY.

Par la confrontation des réseaux de différentes langues dans le cadre du R.N.I., notre étude confirme cette hypothèse et montre que le cas du vide notionnel peut également concerner le bas de l'arborescence. De ce point de vue, il faut admettre que la physionomie du réseau des relations hyponymiques observées en terminologie nautique demeure proche de celle observée dans la langue générale. On peut penser qu'il en va de même dans les nombreux domaines de spécialité qui possèdent une longue histoire et qui, au contraire des taxinomies visées par Lyons (*ibid.*), ne font l'objet d'aucune harmonisation interlinguistique.

Le concept de *notion zéro* ne s'applique pas seulement au cas du vide notionnel classique. En effet, le dépouillement de *De la quille à la pomme de mât* atteste l'existence de notions virtuelles Z^v, qui n'existent dans aucune langue mais qui sont nécessaires à l'établissement de l'équivalence dans le R.N.I. À notre connaissance, ce phénomène n'a jamais été décrit de la sorte[24]. Pourtant les notions Z^v correspondent à des cas de chevauchement culturel et doivent être impérativement prises en compte si l'on veut bâtir une B.C.T. rigoureuse, apte à fournir des équivalents fiables. On peut penser, en effet, que l'oubli des notions Z^v explique un bon nombre d'insuffisances des dictionnaires lorsqu'il s'agit de résoudre des problèmes d'équivalence partielle.

4.3. L'équivalence partielle revisitée

Les principes d'exploitation des notions Z ont été dégagés dans cet article à partir de faits terminographiques concrets observés au sein de systèmes clos. Ils semblent permettre une description plus fine des problèmes de chevauchement culturel que ne le permet la classique distinction entre « supériorité » et « intersection » présentée par Felber (1987 : 129). Cette représentation paraît, en effet, peu adéquate pour rendre compte du rapport étroit entre l'équivalence et la relation TY, qui, l'une comme l'autre, sont identifiées à l'aide des caractères des notions concernées. Certes, Felber utilise les caractères pour expliquer l'équivalence, mais il n'approfondit pas la liaison entre lesdits caractères et la relation TY et néglige ainsi le rôle fréquent de la relation hyponymique dans l'établissement d'une équivalence.

24 La norme ISO R 1087 (1969 : 8) précise qu'« une notion peut résulter de la combinaison d'autres notions, même sans égard pour la réalité », mais cet énoncé vise plutôt des découvertes scientifiques annoncées et non encore vérifiées.

Dans toute arborescence fondée sur la relation espèce-genre TY, il y a une intersection partielle entre les compréhensions des co-hyponymes. Cette intersection rassemble tous les caractères communs aux co-hyponymes. Toutefois, il ne faut pas nécessairement assimiler le phénomène de l'intersection de deux notions à celui de l'équivalence partielle, parfois nommée *intersection partielle.* Ce n'est pas parce que deux notions possèdent plusieurs caractères en commun que l'on peut parler d'équivalence partielle : qui songerait à évoquer une équivalence partielle entre les mots *sail* et *drap de lit* du fait qu'ils désignent des notions qui partagent les caractères « tissu », « blanc » et « couture » ?

Le fait même de parler d'*intersection* sans faire référence à la relation hyponymique apparaît donc comme gênant. Tant dans le cas du phénomène dit de la *supériorité* que dans celui dit de l'*intersection*, les caractères de la notion hyperonyme correspondent à l'intersection en compréhension des caractères des différents co-hyponymes. Dans un cas comme dans l'autre, la notion hyperonyme représente donc une possibilité de dénomination du subordonné par recours au principe de l'hyperonomase. Notre analyse de diverses équivalences (*mizen-royal, cacatois de perruche*) atteste d'ailleurs que l'hyperonomase fonctionne aussi bien dans le cas dit de la *supériorité* que dans celui de l'*intersection.*

Assimiler l'intersection partielle à l'équivalence partielle peut même conduire à négliger le cas des notions virtuelles. Ainsi, on pourrait être tenté de dire que la notion Z^{v} *mizen-royal* = *cacatois de perruche* = *Kreuz-Royal* constitue un cas « d'intersection partielle » entre le français d'une part et l'anglais et l'allemand, d'autre part. Or, cette notion est clairement un cas de conjonction[25] en compréhension (c.-à-d. du point de vue des caractères concernés). Parler ici d'intersection, équivaut à considérer la notion en extension (c.-à-d. du point de vue des référents concernés) et à traiter de l'équivalence partielle en mélangeant deux approches définitoires fondamentalement différentes.

Les concepts théoriques de la notion zéro – qu'elle soit hyponyme ou hyperonyme, réelle ou virtuelle – de l'hyperonomase et de la phagocytose nous paraissent plus précis et plus adéquats. Ils permettent de rendre compte de l'assise de la traduction proposée en vertu du principe d'équivalence notionnelle défini au début de cet article. Ils sont certes plus difficiles à comprendre, mais leur rigueur nous paraît à la mesure des exigences du modèle notionnel.

4.4. Le R.N.I. face à l'approche viennoise

Le principe du R.N.I. trouve sa justification dans une approche terminologique qui prend en compte le terme, la notion et l'objet. Sous cet aspect, il demeure compatible avec le modèle triangulaire proposé par l'École de Vienne et contribue à insister sur le rôle prédominant des caractères dans la distinction des notions. Toutefois, la prise en compte des caractères dans le R.N.I. se double de l'observation des différences dans

25 Dans la tradition viennoise, la détermination, la conjonction et la disjonction sont les types de rapports de combinaison qui peuvent unir trois notions dans le cadre d'une relation logique TY (Felber, 1987 104-105) Notre modèle conduit à remettre en cause cette approche (Van Campenhoudt, 1994 103ssq)

la manière dont chaque langue les appréhende, alors que dans le modèle viennois les caractères émanent d'objets matériels ou immatériels sur lesquels les langues n'ont censément aucune prise.

La perspective du R.N.I. est sous-tendue par une approche descriptive qui combine les démarches sémasiologique et onomasiologique. La première permet de dresser un inventaire des notions à partir des termes utilisés dans les diverses langues envisagées et de confronter les caractères activés. La seconde consiste à dénommer toutes les notions répertoriées dans le R.N.I., mais en préservant, si possible, l'intégrité référentielle dans chacune des langues. En effet, des mécanismes comme l'hyperonomase, la phagocytose et la notion virtuelle permettent de respecter la manière dont la réalité est conçue et dénommée dans chaque langue.

4.5. Un modèle pour le terminographe ?

Il convient cependant d'observer que notre modèle est conçu à partir d'une terminographie particulière, orientée vers des objets essentiellement concrets. S'il s'applique fort bien à des réalités tangibles, aux frontières aisément identifiables, il ne pourrait prétendre rendre compte des équivalences entre les notions juridiques caractéristiques de la *Common Law* et celles propres au Code Napoléon.

Vue sous cet angle, cette modélisation doit avant tout être appréhendée comme un outil théorique permettant de mieux comprendre le mécanisme de l'équivalence et de l'analyser ponctuellement. Elle constitue également une piste de recherche pour l'élaboration d'un système informatique capable de mener rapidement un très grand nombre de raisonnements. Il est évident que la pratique de ce type de description n'est que d'un faible rendement pour le terminographe qui conçoit tout un dictionnaire et qui peut arriver au même résultat grâce à un travail soigné et respectueux du principe d'équivalence notionnelle.

Les mots-clés métalinguistiques comme outil d'interrogation structurante des dictionnaires anciens

Russon Wooldridge et Isabelle Leroy-Turcan

Université de Toronto, Canada et Université de Lyon III, France

Introduction

Les dictionnaires anciens mettent en œuvre une pluralité de systèmes de structuration textuelle, tant pour la macrostructure que pour la microstructure. Dans le domaine de la lexicographie française générale, le cas le plus marqué à cet égard est sans doute le *Thresor de la langue françoyse* (= *TLF*) de Jean Nicot (1606), combinaison de dictionnaire monolingue, bilingue et multilingue, de dictionnaire de langue, dictionnaire encyclopédique et dictionnaire étymologique. Dans celui plus spécialisé de l'étymologie, le premier grand répertoire français, le *Dictionnaire étymologique, ou Origines de la langue françoise* (= *DEOLF*) de Gilles Ménage (1694), associe lui aussi les deux genres du dictionnaire étymologique et du dictionnaire général de langue au service d'une perspective ouverte, à visées linguistiques et encyclopédiques. Les articles individuels de ces deux ouvrages emploient différents modèles de contenu et d'articulation selon l'objet particulier de la description ou de l'analyse. Les modèles utilisés ne sont souvent qu'imparfaitement réalisés, ce qui crée un certain flou structurel, flou renforcé chez le lecteur par le caractère imprévisible des structures[1].

Aussi la base informatique qui a été réalisée pour le *TLF* de Nicot – celle du *DEOLF* de Ménage est en cours[2] – ne contient-elle, comme métatexte, que des jalons indiquant la localisation, les vedettes, la typographie, la langue des unités textuelles et les alinéas. Pour donner accès aux champs informationnels – catégorie grammaticale, définition, marque d'usage, exemple, citation, source, étymologie, etc. – sans dénaturer le texte original et sans empiéter sur les compétences de chaque lecteur en le

1. Sur la récurrence déficiente, cf Wooldridge, 1977 , sur la récurrence parfois prévisible grâce à une interprétation stylistique du texte dictionnairique, cf Leroy-Turcan, 1994a et 1994b

2 Ce qui a été fait correspond à un corpus thématique (les végétaux) et à l'échantillon Ga, I, J, K, Ra-Re

conditionnant dans des pistes d'orientation ou dans des interprétations particulières, il a été élaboré des listes de mots-clés métalinguistiques, lesquels réunissent sous une forme lemmatique toutes les occurrences textuelles d'un marqueur de champ informationnel. Ainsi, par exemple, le lemme FEMININ permet de retrouver dans la base Nicot tous les contextes où le lexicographe a indiqué – par « feminin », « f. », « fem. », « foem. » ou « foemin. » – le genre d'un nom ou d'un adjectif féminin[3].

Le projet d'informatisation des huit éditions complètes du *Dictionnaire de l'Académie française* (1694-1935) – projet annoncé à l'Institut de France en novembre 1994 (Wooldridge, 1994 ; cf. Leroy-Turcan et Wooldridge, 1995) – rencontre le même type de flou structurel que dans Nicot et Ménage, dans une mesure moindre mais bien réelle. Le texte du *Dictionnaire de l'Académie,* notamment celui de la première édition de 1694, quoique présentant une microstructure apparemment plus simple et plus récurrente que celles de Nicot et de Ménage, délimite souvent mal la frontière entre langue et métalangue, mot et référent, définition et marque d'emploi, locution, collocation et exemple. Ajouter un métatexte hautement structuré risquerait ainsi de dénaturer le texte en lui imposant une perspective moderne, donc anachronique[4].

De fait s'opposent deux orientations radicalement différentes de balisage du texte informatisé selon les relations choisies : 1) via l'analyse du spécialiste qui propose une interprétation du fonctionnement du texte, ce qui se matérialise sous la forme d'un encodage complexe (cf. le balisage fin tel qu'il a été proposé pour Ménage et dont la mise en œuvre est complexe – Leroy-Turcan, 1994b) ; 2) via des grilles de lecture destinées à compléter le balisage minimal des marques de localisation et de typographie (cf. *supra* et *infra* les éléments retenus pour le balisage minimal) : les listes de mots-clés métalinguistiques.

Le but de notre communication est de donner une première mesure, fondée sur des échantillons informatisés[5], de l'efficacité des mots-clés métalinguistiques, en relation avec la position microstructurelle et les marqueurs typographiques, comme outil d'interrogation structurante du *Dictionnaire de l'Académie* et, par extension, des dictionnaires anciens en général. La Base Académie Échantillon comprend, pour chaque édition, les tranches ÂME, DOUAIRE à DOUZIL, GAGNER, GRAS, GROS, LOIN à LOISIR, QUE, QUEUE, TIGE à TINTOUIN, VOLER. Un jalonnage indique l'édition, la vedette, l'alinéa, le caractère d'imprimerie et la page-colonne. À l'exception de la vedette, l'identification des champs balisés se fonde objectivement sur des critères formels systématiques. La communication démontre que le balisage explicite a été adopté comme méthode de recherche des vedettes plutôt que l'interrogation de la base à partir de marqueurs typographiques (capitales et position), alors que la localisation des champs informationnels se fonde sur le caractère d'imprimerie et les mots-clés métalinguistiques.

Un concept important pour cette méthode d'interrogation est celui de la « requête floue » (Wooldridge, 1993). En gros, le flou signifie que plutôt que de dépenser un effort énorme pour obtenir 100 % de ce qu'on cherche et rien de plus, on fait

3 Pour une discussion du concept de mot-clé métalinguistique, cf. Wooldridge, 1988 et à paraître

4 Cf. par exemple le cas des sous-vedettes non marquées dans Acad 1694 mais repérables en tant que telles grâce à Acad 1718.

5 Les différentes bases ont été créées et interrogées avec le logiciel WordCruncher

mieux et on obtient pratiquement les mêmes résultats en se contentant, avec beaucoup moins d'effort, d'une fourchette de 95 % à 105 % du total théorique, quitte à rejeter le 5 % de bruit. La requête floue convient particulièrement bien comme modèle d'interrogation d'un texte à flou structurel. Si nous introduisons la notion de mot-clé métalinguistique, c'est notamment en raison du flou macrostructurel des vedettes et sous-vedettes dans Acad 1694, flou qui touche aussi la microstructure.

1. Vedettes et sous-vedettes

Le système de classement des unités de la nomenclature employé dans la première édition (1694) est différent de celui des autres (1718-1935). Dans la première édition, les mots sont regroupés en familles étymologiques, lien caduque dans Acad2-8 ; dans la macrostructure principale, les mots de base des différentes familles sont rangés par ordre alphabétique, tandis que les autres membres de chaque famille sont organisés suivant des principes de dépendance dérivationnelle[6].

Les propriétés formelles des unités de la macrostructure alphabétique sont les grandes capitales et la position initiale d'alinéa. À celles-ci s'ajoutent, dans certaines éditions, le gras (Acad6-8) et un interligne précédent plus grand (Acad8). Bien que les grandes capitales s'emploient aussi dans les renvois de la première édition (« TIMPAN. *Voy* TYMPAN. ») et que la position initiale d'alinéa puisse être occupée par des objets de toutes sortes, les deux ensemble suffisent en général à identifier toutes les vedettes et seulement des vedettes. Les rares exceptions sont à considérer comme des accidents[7].

Le niveau secondaire de la macrostructure, celui des sous-vedettes, est plus problématique. Afin de rendre possible une comparaison du contenu de la macrostructure de la première édition avec les macrostructures des autres, il est nécessaire d'attribuer un jalon de vedette aux items subsidiaires d'Acad1 susceptibles d'appartenir à la nomenclature alphabétique d'Acad2-8 (ex. TIMIDITÉ, INTIMIDER, TIMIDEMENT ET TIMORÉ). Ces sous-vedettes ont deux propriétés formelles : petites capitales et position initiale d'alinéa ; cependant ces deux propriétés sont souvent partagées par ce qui dans l'ensemble des éditions (Acad1-8), comme dans la tradition lexicographique générale, doit être considéré comme des sous-adresses fonctionnant au niveau de la microstructure. Les difficultés posées dans Acad1 par la distinction des sous-vedettes et des sous-adresses sont nombreuses[8].

---

6 Ainsi, pour les mots en TIM , la nomenclature principale donne TIMBALE, TIMBRE, TIMIDE, TIMON ET TIMPAN, alors que SOUS TIMIDE on trouve TIMIDE, TIMIDITÉ, INTIMIDER, TIMIDEMENT et TIMORÉ. (Est laissée de côté dans cette courte communication la question mineure des formes participiales, par ex. INTIMIDÉ.) À partir de la 2e édition, tous les mots sont donnés dans une seule macrostructure strictement alphabétique, ce qui a pour conséquence de séparer INTIMIDER, TIMIDE-TIMIDEMENT-TIMIDITÉ et TIMORÉ

7 Ex. QUI VIVE, dans l'alinéa « QUI VIVE. Voy VIVRE. » placé dans l'article QUI, fonctionne de la même façon que QUICONQUE OU QUIDAM, comme sous-vedette de QUI.

8. Nous nous bornerons ici à donner une idée du problème en examinant les quelques premiers débuts d'alinéa en capitales de l'article LONG « LONG, LONGUE [..] SE FORLONGER [..] LOIN. [ ] AU LOIN [...] LOIN À LOIN, DE LOIN À LOIN. [...] LOINTAIN, AINE. [...] ELOIGNER. [..] ». Après application de la règle que la fin de la première unité d'un début d'alinéa en capitales est marquée par la première virgule à moins que celle-ci soit précédée d'un point, il nous reste : LONG. SE FORLONGER, LOIN, AU LOIN, LOIN À LOIN, LOINTAIN et ÉLOIGNER. Une comparaison avec Acad2-8 et la tradition lexicographique générale nous enseigne que LONG, FORLONGER (SE), LOIN, LOINTAIN et ÉLOIGNER sont des (sous-) vedettes et que AU LOIN et LOIN À LOIN sont des sous-adresses de LOIN.

Comme les critères formels objectifs sont insuffisants pour permettre un jalonnage automatique des sous-vedettes d'Acad1[9], une procédure raisonnable consiste à ajouter systématiquement un jalon de vedette à l'endroit de la première unité de chaque début d'alinéa en capitales, puis, dans une post-édition manuelle interprétative, à éliminer les jalons qui correspondent à une sous-adresse.

2. Caractères d'imprimerie et champs informationnels

Les deux principaux caractères d'imprimerie utilisés dans le *Dictionnaire de l'Académie* sont le romain et l'italique. Ils ont chacun des fonctions sémiotiques différentes, tout comme les capitales et les minuscules. Le gras ajouté aux vedettes d'Acad6-8 augmente la consultabilité du texte mais il est sémiotiquement redondant. Dans le système sémiotique de base, le romain minuscule (caractère non marqué) sert au niveau textuel fondamental du discours métalinguistique du lexicographe, lequel contient catégorie grammaticale, marque d'usage, filiation sémantique, définition et les copules articulatrices des différentes unités linguistiques et métalinguistiques de la microstructure ; les capitales romaines, l'italique et le gras (caractères marqués) s'emploient pour les autonymes – c.-à-d. les unités de l'objet de description, la langue : mots, expressions idiomatiques, cooccurrents, exemples, synonymes, etc. Mais peut-on se servir du caractère d'imprimerie, en rapport avec la position (position absolue ou relative d'un item dans la microstructure), pour rechercher les champs informationnels ?

Comme on l'a vu, le romain minuscule est utilisé pour plusieurs champs informationnels : la catégorie grammaticale est normalement signalée immédiatement après la vedette (« DOUBLE. adj. de tout genre. »), exceptionnellement ailleurs (« Il est aussi subst. ») ; les marques d'usage et de qualification sémantique tendent à être non initiales dans les alinéas discursifs des premières éditions (« On dit fig. et fam. [...] *une cervelle, une teste bien timbrée, mal timbrée* » Acad2-5 s.v. TIMBRER), initiales dans les dernières (« Fig. et fam., *Une cervelle, une tête timbrée* » Acad6-8). L'italique s'emploie systématiquement dans les exemples, de façon occasionnelle à l'endroit des cooccurrents et des synonymes : « GAGNER, se joint quelquefois avec la préposition *Sur* » (Acad2 ; Acad1 « *sur* ») *vs.* « SANS DOUTE, [...] se joint quelquefois avec Que » (*id.* ; Acad1 « *que* ») ; « DOUBLON. [...] On dit aussi, *Pistole* » (Acad6 ; Acad5 « [...] que nous appelons Pistole ») vs. « *Ne... que* peut, dans certains cas, être considéré comme entièrement synonyme de l'adverbe Seulement » (*id.* s.v. QUE = Acad7-8).

Le gras seul suffit pour trouver toutes les vedettes et co-vedettes d'Acad6-8 (325 séquences dans la Base Échantillon = 100 % des (co-)vedettes). Les grandes capitales romaines (387) sont utilisées pour des vedettes (Acad1-5) et des co-vedettes (Acad2-5) dans 374 cas (96,64 %) et pour des renvois dans 13 cas (Acad1, 3,36 %). Les petites capitales romaines (790) sont hautement polysémiques : elles servent régulièrement pour les co-vedettes d'Acad1 (5 occurrences = 0,63 %), les sous-vedettes et les sous-adresses d'Acad1-8 (726 = 91,90 %) et les renvois d'Acad4-8 (55 = 6,96 %) ;

9 Ce que confirme le cas de certains mots en italique qui fonctionneront ensuite comme adresse (cf. pour l'article LOUP, Leroy-Turcan et Wooldridge, 1995)

leur statut de caractère marqué expliquerait quatre occurrences idiosyncratiques, ou irrégulières : un synonyme (« On dit aussi DUPLICATA » Acad8 s.v. DOUBLE), un cooccurrent (« Il se joint quelquefois avec la préposition SUR » Acad8 s.v. GAGNER) et un élément d'exemple (« *âme rachetée par le sang de* JÉSUS-CHRIST » Acad6-7 s.v. ÂME ; cf. Acad5 « [...] *JÉSUS-CHRIST* », Acad2-4 « [..] *Jésus-Christ* »).

Selon la logique générale du dictionnaire, les définitions (métalangue) sont imprimées en romain, les cooccurrents, synonymes et antonymes (langue) en italique. Lorsque, comme c'est souvent le cas des adjectifs et des adverbes, la définition est un mot plutôt qu'une périphrase, la distinction entre définition et synonyme est gommée ; en conséquence, l'emploi des différents caractères peut devenir aléatoire : « Il signifie aussi, *Espais*, et est opposé à delié, delicat » (Acad1-7 s.v. GROS), au lieu de « Il signifie aussi, Espais, et est opposé à *delié, delicat* » (cf. « DOUBLE [...] Il est opposé à *simple* » Acad1).

Il devient nécessaire alors d'avoir recours aux mots-clés métalinguistiques, tels que SIGNIFIE, SE JOINT AVEC, ON DIT AUSSI, OPPOSÉ À, etc., pour la recherche des définitions, cooccurrents, synonymes et antonymes.

3. Mots-clés métalinguistiques dans académie

La relative régularité, à travers les huit éditions, de l'emploi du caractère d'imprimerie s'observe aussi dans la terminologie du métalangage dictionnairique. Les noms sont normalement donnés comme noms masculins ou féminins, les verbes comme verbes transitifs ou intransitifs. Les formules de présentation des expressions lexicalisées, des définitions, de l'articulation sémantique et des niveaux d'usage restent les mêmes. En l'absence de l'étymologie et de la prononciation, qui n'est donnée que dans des cas exceptionnels, le nombre des champs informationnels est relativement petit. Pour la recherche des informations, quelques termes métalinguistiques sont caractéristiques par leur efficacité.

La Liste de mots-clés est un index alphabétique qui contient les adresses dans la base des occurrences des mots-clés métalinguistiques. Les items de la Liste sont des lemmes regroupant des formes variantes textuelles ; par exemple, le lemme FEMININ donne accès aux formes textuelles « f. », « fem. », « fém. », « fémin. » et « féminin ». La fréquence dans la Base Échantillon du mot-clé brut FEMININ est 204. 201 (98,53 %) des occurrences indiquent le genre de l'unité lexicale sujet d'énoncé ; dans presque tous les cas, le mot est précédé soit du mot-clé SUBSTANTIF (196), soit du mot-clé ADJECTIF (2). Un examen des six autres cas (6 sur les 8 occurrences de la forme « féminin ») révèle que trois d'entre eux concernent des signes linguistiques autonymes (« GROSSE, au féminin » Acad6-7 ; « Au féminin » Acad8 s.v. GROS), tandis que les trois autres se réfèrent à une propriété sémantique au niveau de la métamétalangue (« On appelle en termes de Grammaire, *Noms douteux*, Ceux que les uns mettent au masculin, et d'autres au féminin. » Acad5-7). Le mot-clé FEMININ qualifiant une unité lexicale peut alors être corrigé pour en réduire le nombre d'occurrences à 201. Il est clair cependant que le mot-clé explicite FEMININ, tel qu'il vient d'être défini, ne donne pas accès à toutes les formes féminines de la nomenclature : le féminin de l'adjectif est normalement signalé par la forme elle-même et non pas par une étiquette (« DOUX, DOUCE. adj. »), tandis que les noms à genre double

sont marqués négativement par une absence d'étiquette de genre. Par exemple, les 150 occurrences du mot-clé SUBSTANTIF suivi ni de MASCULIN, ni de FEMININ renferment 30 concernant le féminin :

> « TIGRE, TIGRESSE. s. » (Acad1-8)
> « DOUILLET, est aussi substantif, dans la seconde acception. *Faire le douillet. C'est un douillet, une douillette.* » (Acad6-8)

Les occurrences du mot-clé FEMININ peuvent alors être augmentées par l'ajout des adresses des formes féminines de la nomenclature non étiquetées.

Dans le cas du genre, le manque d'une étiquette explicite n'exclut pas, comme nous venons de le voir, la recherche objective des items pertinents : TIGRESSE est donné comme féminin en vertu et de sa position comme seconde de deux co-vedettes et de l'indication « s. » ; DOUILLETTE est donné comme nom féminin dans l'exemple « *C'est une douillette.* ». Pour ce qui est du niveau d'usage et de la filiation sémantique, on doit se fier aux étiquettes explicites, sans lesquelles on est amené à faire une interprétation subjective du texte.

On peut adopter une méthode légèrement différente pour des termes comme « aussi ». Dans presque toutes ses occurrences en romain (770 sur 786 = 97,96 %), « aussi » est métalinguistique ; cette copule polysémique s'emploie dans des informations sur la catégorie grammaticale, le sens, la synonymie et la syntaxe. Pour la définition du mot-clé AUSSI, on a alors le choix entre la règle simple, floue mais efficace « 'aussi' précédé d'un jalon de caractère romain » (*f* 786), la liste globale plus précise des occurrences métalinguistiques (*f* 770) et la création de plusieurs mots-clés AUSSI correspondant à chaque type d'information particulier (catégorie grammaticale, sens, etc.).

L'usage familier est marqué comme tel dans le texte au moyen des termes « familier », « fam. », « famil. », « familière », « familières », « familiers » ou « familièrement ». Le mot-clé FAMILIER renvoie, dans la Base Échantillon, aux 319 occurrences de ces variantes. Il est important de distinguer la subjectivité de la décision du lexicographe de qualifier un item de familier – plutôt que, par exemple, de populaire[10] ou de bas[11] – de l'objectivité de la recherche des étiquettes textuelles.

Pour rechercher les occurrences d'usage figuré, on peut choisir soit de se limiter au mot-clé FIGURÉ (formes textuelles « fig. », « figur. », « figuré », « figurées », « figurém. », « figurément » – *f* 517), soit d'y inclure ANALOGIE (« par analogie », « par une sorte d'analogie » – *f* 13) et/ou PROVERBIAL (« prov. », « proverb. », « proverbe », « proverbiale », « proverbialem. », « proverbialement » – *f* 275). On peut remarquer que dans 112 de ses occurrences PROVERBIAL se combine avec FAMILIER (ex. « On dit prov. et fig. *Joüer à quitte ou à double*, pour dire, Hazarder tout pour se tirer d'une affaire. » Acad1-5 s.v. DOUBLE ; cf. « [...] figurément et familièrement [...] » Acad6-7, « Voyez QUITTE » Acad8).

10. Le mot-clé POPULAIRE – « pop. », « popul », « populaire », « populairement » – a une fréquence de 41.
11. Le mot-clé BAS – « bas », « bass. », « bassement » – a une fréquence de 19.

Si l'identification de la catégorie grammaticale, du genre et des marques d'usage est facile, celle d'autres champs informationnels, tels que la définition et l'exemple d'emploi, peut être complexe et nécessiter une interprétation subjective. Comme nous l'avons vu dans la section précédente, une condition préalable pour la définition est qu'elle soit en romain, pour l'exemple qu'il soit en italique. L'emploi de mots-clés métalinguistiques à l'endroit de ces deux types d'informations (SIGNIFIE, SE PREND POUR, COMME...) est occasionnel ; aucune combinaison de caractère d'imprimerie et de mots-clés ne permet de rechercher tous les cas de définitions/exemples et uniquement les définitions/exemples. Une considération préalable à l'application de marques formelles (ou à celle de jalons dans un dictionnaire moderne dont les champs informationnels ont été systématiquement balisés) est la définition de ce qui constitue une définition ou un exemple.

La « définition » peut fonctionner en métalangue de contenu ou en métalangue de signe (Rey-Debove, 1971) ; elle peut traiter le mot au niveau du lexique ou du discours :

Métalangue de contenu : « AME. s. f. Ce qui est le principe de la vie dans les choses vivantes. » (Acad1)
Métalangue de signe : « DOUBLE [...] se dit aussi des choses plus fortes, de plus grande vertu que les autres de mesme nature. » (*id.*)
Lexique : « ÂME [...] se dit aussi figurément de Ce qui est le principal fondement d'une chose, qui la maintient. *La discipline militaire est l'âme d'une armée. La bonne foi est l'âme du commerce.* » (Acad8)
Discours : « Fig., *Donner de l'âme à un ouvrage, mettre de l'âme dans un ouvrage,* Exprimer vivement ce qu'on y représente, y mettre beaucoup de feu, de sentiment. » (*id.* s.v. ÂME)

Les copules explicites reliant occasionnellement l'unité lexicale (sujet) à la définition (prédicat) comprennent « signifie » (*f* 414)[12], « pour dire » (*f* 801) et « se prend pour » (*f* 53).

« QUEUE, Signifie aussi, La derniere partie, les derniers rangs de quelque Corps, de quelque Compagnie » (Acad3)
« On dit, *Manger gras, faire gras*, pour dire, Manger de la viande les jours que l'on devroit manger maigre. » (Acad4 s.v. GRAS)
« Il se prend plus particulierement, et d'une maniere absoluë, pour Façon d'agir douce, et éloignée de toute sorte de violence. » (Acad1 s.v. DOUCEUR)

Une autre marque occasionnelle de la définition est l'explicitation du statut d'espèce (hyponyme) de l'unité lexicale par opposition au genre (hyperonyme) du terme nucléaire de la définition. Ainsi, « espece/espèce » (*f* 78) et « sorte » (*f* 86) qualifiant, par exemple, DOUBLON et LOIR de types de monnaie et d'animal respectivement :

« DOUBLON. s. m. Espece de monnoye d'Espagne, qui est d'or, et que nous appellons *Pistole.* » (Acad1 et cf. Acad2-5) ; cf. « DOUBLON. s. m. Monnaie d'or espagnole qui a différentes valeurs. » (Acad6 et cf. Acad7-8)
« LOIR. s. m. Sorte de petit animal semblable à un Rat qui vit dans le creux des arbres et qui dort durant tout l'hyver, à ce que disent les Naturalistes. »

12. Cf. « sign. » (= « signifie » 9, « signification » 1) 10, « pour/peut signifier » 17, « signifioit/signifiait » 5, « phrases [...] qui signifient » 1, « signification(s) » 29.

(Acad1 et cf. Acad2-4) ; cf. « LOIR. s. m. Petit animal semblable à un rat, qui vit dans les creux des arbres, et qui dort durant tout l'hiver. » (Acad5 et cf. Acad6-8)

Pour ce qui est des exemples, il n'y a aucun moyen absolu de déterminer, dans le texte du dictionnaire, la frontière entre unités lexicales et exemples, entre syntagmes lexicalisés et syntagmes libres. Dans un alinéa qui contient plusieurs séquences en italique, les items lexicalisés précèdent normalement les items libres. Dans la plupart des cas, un syntagme lexicalisé est suivi d'un traitement sémantique, alors qu'un exemple libre est donné en position finale. Dans le premier extrait suivant, la première séquence en italique est une unité lexicale suivie d'une définition, la seconde une série de trois exemples ; dans le deuxième extrait, l'unique séquence en italique est une unité lexicale suivie d'une définition ; dans le troisième extrait, les multiples séquences en italique sont à considérer comme collocations ou phrases exemplificatrices même si plusieurs d'entre elles sont suivies d'une définition du mot en usage.

« On dit, *Filer doux,* pour dire, Demeurer dans la retenüe, dans la soumission à l'égard de quelqu'un que l'on craint, souffrir patiemment une injure. *C'est un homme avec qui il faut filer doux. je le feray bien filer doux. quand il s'entendit menacer, il fila doux.* » (Acad1 s.v. DOUX)

« On dit prov. *Aller doucement en besogne.* Et tantost il signifie, Sagement, meurement, sans rien précipiter ; tantost il signifie, Laschement, mollement. » (*id.* s.v. DOUCEMENT)

« DOUCEMENT. adv. d'Une maniere douce. *Dormir doucement. il faut marcher doucement dans la chambre d'un malade. heurtez doucement à la porte,* c'est à dire avec le moins de bruit que l'on peut. *Allez-y plus doucement. il faut traiter doucement les vaincus. reprendre quelqu'un doucement de ses fautes. je luy fis doucement la guerre de ce que,* etc. *quand on a souffert de grandes douleurs, et que l'on ne souffre plus, on se trouve bien doucement. on peut vivre doucement la campagne pour peu de chose. ce cheval galoppe fort doucement. cette affaire veut estre traitée, veut estre maniée doucement,* c'est à dire delicatement. *Il faut s'y prendre doucement. on craignoit qu'il n'arrivast quelque desordre dans l'assemblée : mais toutes choses s'y passerent fort doucement,* c'est à dire fort paisiblement. *C'est une chose qu'il faut faire doucement ;* c'est à dire, sourdement, sans faire esclat. » *(ibid.)*

Mais ce qui est valable pour l'Académie ne l'est pas forcément pour d'autres dictionnaires : ainsi les mots-clés métalinguistiques FEMININ et FAMILIER ne sont pas opératoires dans le cas du dictionnaire de Ménage qui n'aborde que rarement la synchronie.

4. Les variations fonctionnelles des mots métalinguistiques selon les genres de dictionnaires

Un même mot métalinguistique peut fonctionner différemment dans un dictionnaire de synchronie et dans un dictionnaire historique à dominante étymologique. C'est, par exemple, le cas de FEMININ dans Académie opposée à Nicot et à Ménage. Nous exa-

minerons, pour la marque de l'usage ancien, celui du mot métalinguistique ANCIEN.

Sous le lemme ANCIEN sont regroupées toutes les modalités de marques d'une graphie, d'un mot, d'une collocation ou d'un emploi qualifiés d'anciens (éventuellement par rapport à un usage en cours) ; sont donc compris sous ce lemme toutes les formes se rattachant à la base *ancien-* et les termes exprimant le même sémantisme comme *vieux* et *vieillir*, et leurs formes fléchies, ou les adverbes *autrefois, jadis*, sans négliger toutes les marques temporelles de passé dans les verbes qui peuvent être eux-mêmes métalinguistiques (comme *signifier, appeler* ou *dire*) ou éléments de définition (comme *valoir* s.v. DOUBLE : « Espece de monnoye qui valoit deux deniers » Acad2-5).

Le résultat des interrogations des trois bases – c.-à-d. le texte intégral de Nicot et des échantillons de Ménage et d'Académie (cf. *supra*) – donné sous forme de tableau (ci-dessous) nécessite quelques commentaires en raison des difficultés d'appréciation liées à la nature même de chaque dictionnaire.

	Nicot 1606	Ménage 1694	Acad 1694-1935
ancien-	234	24	15
vieux	3	11	11
[il] vieillit / a vieilli	0	0	14
autrefois	0	4	15
jadis	19	1	0

Principales marques d'usage ancien

Ancien- dans Ménage. Sur 90 occurrences d'*ancien-*, 49 ne sont pas du tout pertinentes, 12 concernant un discours socio-culturel, 7 étant dans des citations et 30 appartenant à la bibliographie ; les occurrences restantes se répartissent entre l'étymologie (sur 10 occurrences, 2 étymons = « ancien mot » ; 4 renvois à d'autres langues dont 3 séquences « de l'ancien » ; un emploi « d'ancienne origine »), des références à l'ancien français (3 occurrences = « mot ancien ») et l'usage ancien (10 emplois d'*anciennement* tous combinés à des marques d'imparfait et 14 d'*ancien*), sans compter les doubles emplois dans un même article. La diversité des occurrences d'*ancien-* rend nécessaire la définition des différentes conditions de l'environnement du mot métalinguistique réparti dans des sous-catégories de séquences métalinguistiques levant toute ambiguïté.

Vieux dans Ménage. Sur les 21 occurrences de *vieux* dans Ménage, seulement 11 sont pertinentes pour l'identification d'un usage ancien ; 4 emplois qualifient des références bibliographiques, un emploi qualifie un nom de poète, un autre un proverbe, 3 se trouvent dans des citations, 5 emplois de la forme du féminin n'appartiennent pas à la métalangue, un emploi concerne l'étymologie ; la proportion importante de rebut nous conduit à proposer des modalités de structuration ou de modélisation de l'environnement du mot métalinguistique susceptible, dans ce cas, de devenir plutôt une séquence métalinguistique qui inclut les éléments textuels permettant une interrogation plus fine. De fait, l'interrogation par la séquence « Vieux mot », en début d'article, ou « , vieux mot », en groupe apposé, donne le résultat des 11 occurrences pertinentes, s.v. GABAN, GABER, GALLER, GAUSSER, JOUCARITE, JUS, ISNEL, RAIN, RAMON, RAMPONNER ET RESE.

Des problèmes analogues s'observent dans le discours fortement étymologique de Nicot, la proportion des remarques d'usage restant dominante (les 234 occurrences d'*ancien-* sont à trier).

On peut faire le même genre d'analyse pour les séquences « on dit », « on disoit », « on a dit », qui n'ont pas le même fonctionnement dans Nicot, Ménage et Académie.

Conclusion

Bien que les dictionnaires modernes ne soient jamais entièrement systématiques, ils le sont relativement ; quand on les informatise par rétroconversion, on balise systématiquement leurs champs informationnels à un degré plus ou moins détaillé. Les dictionnaires anciens sont, dans une mesure variable, moins systématiques que les dictionnaires modernes. Pour ne pas les enfermer dans une interprétation univoque, on doit éviter un balisage systématique des champs informationnels. En revanche, on peut, dans la majorité des cas, obtenir un taux de succès très satisfaisant dans la recherche des champs informationnels au moyen des indicateurs que sont le caractère d'imprimerie et les mots-clés métalinguistiques. Dans l'utilisation des jalons de caractère et la définition des mots-clés, il faut réfléchir au rendement de la recherche floue par opposition à une post-édition ardue : la seule interrogation de ce genre de base par les mots métalinguistiques ne saurait produire des statistiques utilisables de façon automatique ou manuelle pour la fréquence ou le repérage des champs informationnels ; même une définition rigoureuse des différentes modalités d'environnement du mot métalinguistique exige les compétences linguistique, dictionnairique et pragmatique du lecteur/utilisateur de la base.

Représentation de la polysémie dans un dictionnaire électronique

Michel MATHIEU-COLAS

Laboratoire de Linguistique Informatique, Université Paris XIII - CNRS - INaLF, Villetaneuse, France

• *Abstract* •

The traditional treatment of polysemy causes many problems of representation : the coexistence of several meanings within the same entry generates extremely complex structures, which are difficult for a computer to exploit. In developing electronic dictionaries, our suggestion would be to generalize and to systematize the splitting up (dégroupement) of the entries : it is better for each use to be considered as a full « word », i.e to receive an independent address and a specific description (morphological, syntactic and semantic information, translations, etc.). The model allows as many descriptions as there are meanings. This does not necessarily lead to dispersion of information : relations between the different uses (branchings, shifts in meaning ..) can be reintroduced into appropriate fields, which allows a more flexible representation of polysemic connections.

Étant admis que la polysémie constitue une donnée fondamentale des langues naturelles et l'une des principales difficultés pour le traitement automatique, nous nous interrogerons ici plus particulièrement sur les modalités de *représentation* de cette pluralité dans le cadre des dictionnaires électroniques. Après un bref rappel de la conception lexicographique classique, nous plaiderons en faveur d'une nouvelle disposition des entrées de dictionnaire et tenterons de répondre aux objections que pourrait soulever le modèle proposé.

Précisons que ces réflexions s'inscrivent dans le cadre des recherches que nous menons, avec Gaston Gross, au Laboratoire de Linguistique Informatique de Villetaneuse (LLI). Si certaines de nos propositions peuvent paraître évidentes en terminologie, elles le sont peut-être moins du point de vue linguistique et lexicographique, à en juger par la diversité des approches[1].

1. Pour une autre approche de la polysémie, voir ici même la communication de Pierrette Bouillon.

1. La conception classique

Quelques remarques suffiront à illustrer le traitement traditionnel de la polysémie. La lexicographie classique repose sur une conception fondamentalement **unitaire** du mot : tous les emplois sont regroupés au sein d'un même article, la multiplicité étant prise en charge par différents systèmes de hiérarchisation et de classement des sens. Seuls sont exclus de ce système les véritables homonymes, fondés sur des étymons différents (les trois mots *baie* qui coexistent en français), et quelques familles anciennement éclatées, à l'instar de *voler* (*to fly / to steal*) ou de *grève.*

Il en résulte, pour les termes polysémiques, de nombreux problèmes de représentation : disposition linéaire ou arborescente, ordre logique ou historique, opposition entre langue générale et langues de spécialité, etc. Quels que soient les choix effectués, la coexistence de plusieurs emplois au sein d'un même article se traduit par des structures d'une grande complexité et au surplus très différentes d'un dictionnaire à l'autre.

Il est vrai que ce modèle unitaire a connu, récemment, quelques aménagements. L'exemple le plus familier en est sans doute le *Dictionnaire du français contemporain* (*DFC*, Larousse, 1971), où de nombreux termes polysémiques sont dégroupés, ce qui revient à les traiter comme de simples homographes ; des lexèmes comme *bureau, cher* ou *commander* se trouvent ainsi décomposés en deux ou trois entrées, ce qui a pour effet de faciliter la description synchronique des emplois et de permettre un traitement plus rigoureux des séries dérivationnelles.

Mais il ne s'agit là que d'une solution de compromis, car l'on s'arrête à mi-chemin : au sein de chaque entrée peut subsister une multiplicité d'emplois qui reproduit, au second degré, le modèle traditionnel. Si *bureau*, désormais, a droit à deux adresses, chacune d'elles n'en est pas moins subdivisée en trois ou quatre descriptions :

> **1. bureau** n.m **1°** Table, munie ou non de tiroirs, dont on se sert pour écrire [...] – **2°** Pièce où est installée cette table [...] – **3°** Mobilier de cette pièce [...]
> **2. bureau** n.m. **1°** Établissement public où sont installés des services administratifs [...] – **2°** Caisse d'un théâtre [...] – **3°** Ensemble des employés ou des fonctionnaires qui travaillent dans une administration [...] – **4°** Membres d'une assemblée, d'une association, élus pour diriger les travaux [...]

Les difficultés liées à la polysémie demeurent ici entières : le problème est déplacé, il n'est pas résolu.

2. Les mérites du dégroupement

Nous proposons en conséquence de systématiser et de généraliser le principe du **dégroupement** : même lorsqu'il s'agit de langue générale, chaque emploi gagne à être considéré comme un « mot » à part entière, ce qui revient à lui attribuer une adresse autonome et une description spécifique (*bureau* donnerait ainsi lieu, pour reprendre l'exemple précédent, à sept entrées distinctes). Pratiquant cette technique depuis quelques années, dans le cadre des travaux du LLI, nous sommes de plus en plus conscients des avantages qu'elle offre.

Rappelons que notre conception des dictionnaires électroniques s'inspire largement de la pratique des bases de données : chaque entrée constitue un « enregistrement », cependant que la description se trouve répartie en une série de « champs » (rubriques) clairement définis, correspondant aux différents paramètres de l'information lexicographique. Voici, à titre d'exemple, l'ébauche de description d'une unité monosémique (*taille-crayon*) :

MOT :	taille-crayon
CAT. GRAM. :	nm
STRUCTURE :	v00 [*verbe* + *nom*]
FLEXION :	00 ; 01
VARIANTES :	taille-crayons
TRAITS :	inc [inanimé concret]
CLASSE :	instrument
DOMAINE .	écrit., dess.
ANGLAIS :	pencil sharpener
ALLEMAND :	Bleistiftspitzer

On trouve ici représentées des données morphologiques (structure formelle, flexions, variantes graphiques), des informations sémantiques (domaines) et/ou syntaxiques (les traits et, plus précisément, ce que Gaston Gross et moi appelons les « classes d'objets[2] »), ainsi que des traductions. Les mêmes informations peuvent être visualisées dans d'autres formats structurellement équivalents, notamment sous forme linéaire, les rubriques étant délimitées par des séparateurs et des identificateurs :

taille-crayon /G:nm /M:v00 /F:00; 01 /V:taille-crayons /T:inc /C:instr. /D:écrit.,dess. / AN:...

ou sous forme de tableau, les lignes et les colonnes correspondant respectivement aux entrées lexicales et aux champs de description :

MOT	G:	M:	F:	V:	T:	C:	D:
TAILLE-CRAYON	nm	v00	00,01	taille-crayons	inc	instr.	écrit ,dess
TAILLEUR-PANTALON	nm	nn00	01-01		inc	vêt fém	habill
TALK-SHOW	nm	d62	01		évé.	émission	télév
TALKIE-WALKIE	nm	d62	01-01		inc	appar.	radiocomm

2. G. Gross, 1994 , M. Mathieu-Colas, 1994, pp 162-173

Ces exemples simplifiés ne rendent pas compte, naturellement, du nombre réel des paramètres qui articulent nos analyses. Ainsi, pour les mots prédicatifs, nous indiquons la structure argumentale (y compris les traits et les classes qui spécifient les arguments : voir *infra*), à quoi s'ajoute, pour les noms abstraits, l'indication des verbes supports (*voyage* se construit avec « faire », *ordre* est introduit par « donner »). S'agissant des informations sémantiques, nous mentionnons, quand il y a lieu, les synonymes, les antonymes, les relations méronymiques (relations partie-tout : Otman, 1995 : chap. 6), de même que nous notons, sur le plan pragmatique, les registres temporels, régionaux ou sociaux. La liste n'est pas close, et d'autres informations pourraient ici trouver leur place : indication des dérivés, « fonctions lexicales » (I. Mel'cuk), indices de fréquence, etc.

Quel que puisse être dans le détail le choix des rubriques, on retiendra surtout, pour la présente analyse, l'importance de la *fiche* en tant que principe organisateur de l'information lexicographique (ce qui nous rapproche de son utilisation en terminologie : voir Lerat, 1990). Ce mode de structuration comporte deux aspects complémentaires : d'une part, il implique que les données lexicales puissent être décomposées en paramètres discrets formalisables (fondant ainsi la possibilité de procéder à des extractions et des traitements automatiques) ; d'autre part, il signifie que chaque entrée du dictionnaire correspond à un emploi strictement défini. Il en résulte qu'on est conduit, en cas de polysémie, à développer **autant de descriptions qu'il y a de sens différents** – soit par exemple, pour le mot *crapaud* :

MOT	TRAIT	CLASSE	DOMAINE	REGISTRE	*ANGLAIS*
CRAPAUD #1	ani	batracien	zool		*toad*
CRAPAUD #2	hum	qualif		fam. (gamin)	*brat*
CRAPAUD #3	hum	qualif		fam. (pers. laide)	
CRAPAUD #4	ina	mal. anim.	vétér.		*greasy heel*
CRAPAUD #5	inc	instr. mus.	mus.		*baby grand*
CRAPAUD #6	inc	siège	ameubl.		*tub easy-chair*
CRAPAUD #7	inc	dispos.	ch. de f.		*sleeper clip*
CRAPAUD #8	inc	dispos.	pyrotechn.		*jumping cracker*
CRAPAUD #9	inc	défaut	joaill.		*flaw*
CRAPAUD #10	inc	support	topogr.		
etc.					

Le dégroupement ainsi conçu a le mérite de la simplicité, tant du point de vue linguistique (clarté et lisibilité) que sur le plan informatique (facilité de traitement pour la machine). Plus particulièrement, chaque paramètre de la description est susceptible, par ce moyen, de gagner en précision.

a) Cela vaut déjà, d'une certaine manière, pour les informations morphologiques : chaque emploi est susceptible d'avoir son propre genre (*un espace/une espace*), son type de conjugaison (*saillait/saillissait*), ses variantes graphiques (*porte-aiguille[s]* en couture, mais non en chirurgie), sa mise au féminin :

« *VENDEUR, EUSE* n. 1. Personne dont la profession est de vendre, en partic. dans un magasin.
2. DR. Personne qui fait une acte de vente. (En ce sens, le fém. est *venderesse*.) » (PETIT LAROUSSE)

Le souci d'unité conduit ici au paradoxe (la parenthèse finale contredit la formulation de l'entrée), alors que le dégroupement que nous proposons permet de décrire plus simplement chacun des deux emplois :

vendeur #1	/D:comm.	/F:6B	(= fém. *vendeuse*)
vendeur #2	/D:dr.	/F:68	(= fém. *venderesse*)

Le même problème peut se poser pour la mise au pluriel, comme l'illustre l'entrée *œil, yeux* du *TLF*, contredite par une remarque livrée en fin d'article : « Dans les sens techn., le plur. de *œil* est *œils* : *les œils d'une voile.* » Mais rien ne permet de savoir, dans le détail, à quels emplois précis s'applique cette remarque. Le dégroupement, au contraire, rend à chacun son dû :

œil #1	/D:lg	/F:06	(= plur. *yeux*)
œil #2	/D:arm.	/F:01	(= plur. *œils*)
œil #3	/D:bourrell.	/F:01	
œil #4	/D:hort.	/F:06	
œil #5	/D:impr.	/F:01	
œil #6	/D:jeux (go)	/F:06	

b) Les avantages du dégroupement sont plus sensibles encore, naturellement, pour la composante proprement sémantique de la description. La notation, pour chaque sens, de toutes les informations pertinentes (classes, domaines, marques d'usage, etc.) assure une représentation plus fine de la polysémie et facilite, en conséquence, les procédures de levée d'ambiguïtés.

Un seul exemple, situé aux confins de la sémantique et de la syntaxe, suffira à illustrer notre propos : il s'agit de la construction des termes « prédicatifs » (verbes, adjectifs, noms abstraits), que nous évoquions précédemment. Soit le verbe *conduire* et les deux phrases suivantes :

Ce sentier conduit à la mer
Pierre conduit un poids lourd

Chacun des deux emplois se caractérise par une distribution spécifique : sujet <voie de communication> et complément locatif pour le premier, sujet humain et objet <véhicule> pour le second. On remarquera en particulier l'intérêt des « classes d'objets » pour la définition des arguments (si l'on omet de préciser que le sujet de la première phrase est une « voie », on s'expose à générer des phrases non acceptables : **ce crayon/*ce bébé/*cette surprise conduit à la mer*). Or de telles informations sont plus aisées à représenter si chaque emploi bénéficie d'une description différenciée :

conduire #3	/N0: <voie>	/N1: loc	(*Ce sentier conduit à la mer*)
conduire #12	/N0: hum	/N1: <véhic>	(*Pierre conduit un poids lourd*)

Il en irait de même pour un adjectif comme *juste* :

juste #1	/N0: hum	(*Cet examinateur est juste*)
juste #5	/N0: <instrum. de mesure>	(*Cette balance est juste*)
juste #6	/N0: <instrum. de musique>	(*Ce piano n'est plus très juste*)
juste #8	/N0: <vêtement>	(*Cette veste est un peu juste*)

Le dégroupement, ici et là, est source de clarté.

c) Les mêmes principes inspirent notre traitement des mots composés et des locutions (plusieurs dizaines de milliers d'unités sont en voie de description). Dans les dictionnaires que nous élaborons, chaque unité complexe possède sa propre entrée et fait l'objet d'un traitement distinct (par exemple *carte bleue, carte orange, carte grise, carte verte, carte de séjour, carte de travail, carte d'électeur, carte de crédit, carte de visite, carte d'identité, carte à jouer, repas à la carte, jouer cartes sur table, donner carte blanche à* <Nhum>, etc.), ce qui permet d'élaborer des descriptions plus précises.

d) Enfin, *last but not least*, le traitement multilingue tire directement avantage du dégroupement : chaque emploi faisant l'objet d'un description spécifique, il suffit d'indiquer, pour chaque entrée, la ou les traductions appropriées (revoir *supra* l'exemple du mot *crapaud*).

3. Objections et réponses

Le dégroupement systématique auquel nous proposons de recourir est toutefois susceptible de soulever un certain nombre d'objections. Nous nous limiterons ici à deux critiques majeures.

3.1. Continuité ou discontinuité

La première concerne notre conception *discontinue* de la polysémie : nous traitons les mots comme des ensembles d'emplois discrets et clairement différenciés. Or on sait que d'autres travaux privilégient plutôt une représentation *continuiste* du sens (voir, par exemple, Kayser, 1987 et Fuchs, 1988). S'opposant au point de vue « homonymique » fondé sur une pluralité de significations disjointes, B. Levrat (1993) plaide pour une « optique polysémique » et postule une « signification unique », un « sémantisme de base » qui s'enrichit sous l'influence du contexte pour donner naissance à un ensemble de valeurs apparentées.

Cette question, trop fondamentale pour que nous puissions l'évoquer en quelques lignes, mériterait à elle seule de plus amples développements. Disons seulement, pour notre propos, que de telles recherches offrent un grand intérêt du point de vue de la constitution d'une théorie du sens, mais semblent très complexes à mettre en œuvre dans la pratique et difficiles à appliquer à une grande échelle (quand il s'agit de décrire plusieurs dizaines de milliers de mots). En outre, même du point de vue théorique, nous rejoindrions volontiers les analyses de R. Martin, pour qui les figures de *surdétermination*, d'*indétermination* et de *neutralisation* décrites par C. Fuchs « ne

peuvent se définir qu'à partir de sens ou d'acceptions préalablement distingués » (Martin, 1994 : 92). Même les métaphores ou les métonymies les plus audacieuses, les créations poétiques les plus libres ne contredisent pas l'existence d'emplois plus stables organisés de façon discrète ; au contraire, ils les présupposent, ils prennent appui sur eux pour mieux produire leurs effets par un jeu subtil d'allusions, de détournement et de reconstruction du sens. Ce sont les emplois stables et lexicalisés qui constituent, avant toute chose, l'objet des dictionnaires : dans cette perspective, le traitement lexicographique de la polysémie s'accommode mieux, nous semble-t-il, d'une représentation discontinue.

3.2. Articulation des emplois

La deuxième objection que nous voudrions examiner met en cause, par-delà la discontinuité, l'émiettement et l'éclatement des descriptions, et la perte d'information qui pourrait en résulter. Même si, dans un dictionnaire traditionnel, les divers sens d'un mot sont clairement différenciés (hypothèse discontinuiste), ils demeurent **articulés** les uns aux autres dans l'unité du mot (matérialisée par la cohésion de l'article) ; à l'inverse, dans notre système, le lien semble rompu : on perçoit bien ce qui distingue les emplois, on ne voit plus ce qui les relie... Cette objection, à notre sens, n'invalide pas le dégroupement, mais conduit au contraire à un enrichissement et à un approfondissement du modèle.

Qu'on nous permette de commencer par un exemple artificiel. Supposons qu'un mot se trouve ainsi décrit :

MOT	I.	Sens A.
	II.	*Par anal.* sens B.
	III. 1.	*Par méton.* sens C.
	2.	*Spécialt,* sens D.

L'application du dégroupement, tel que nous l'avons défini, conduit à un « aplatissement » de la description, donc à un appauvrissement :

mot #1 sens A
mot #2 sens B
mot #3 sens C
mot #4 sens D

Rien n'empêche cependant de réintroduire ici l'information relationnelle, de façon structurée et explicite, en consignant dans un champ supplémentaire les éventuelles dérivations sémantiques :

ENTRÉES	EMPLOIS	DÉRIVATION SÉMANTIQUE
mot #1	sens A	
mot #2	sens B	< 1 (analogie)
mot #3	sens C	< 1 (méton.)
mot #4	sens D	< 3 (spécial.)

Les indications fonctionnent ici comme un système de « pointeurs » permettant de relier, de proche en proche, les entrées associées : l'emploi n° 4 dérive, par spécialisation de sens, de l'emploi n° 3, lequel procède, par métonymie, de l'emploi n°1... On peut aussi, si on le souhaite, réintégrer dans la description l'arborescence sous-jacente au modèle classique – ces rapports hiérarchiques qui s'expriment traditionnellement sous forme de lettres ou de chiffres (I.A.1.a...) :

ENTRÉES	EMPLOIS	DÉRIVATION SÉMANTIQUE	ARBRE
mot #1	sens A		I
mot #2	sens B	< 1 (analogie)	II
mot #3	sens C	< 1 (méton.)	III.1
mot #4	sens D	< 3 (spécial.)	III.2

On voit ainsi qu'il est possible, dans le cadre d'une présentation dégroupée, de représenter toute l'information relative aux liens logiques ou historiques qui articulent les emplois.

Cela pourtant ne suffit point. Nous voudrions montrer que le format proposé ne permet pas seulement une simple « récupération » d'informations déjà présentes par ailleurs, mais qu'il apporte en outre une amélioration. Il n'est, pour s'en convaincre, que d'observer l'état présent de la lexicographie : d'un dictionnaire à l'autre, on note une assez grande stabilité dans les emplois décrits, mais une extrême disparité dans leur disposition (variations affectant l'ordre et la hiérarchie des acceptions). Voici comment trois dictionnaires – *Petit Robert, Petit Larousse illustré, Dictionnaire du français contemporain* – présentent le mot *colle* :

PR	*PLI*	*DFC*
1. Substance	1 Substance	1. Substance
2. *Arg scol* Interrogation	2 *Arg scol*	2.1. *Fam.* Question embarrassante
Cour Question difficile	a) Interrogation	2.2. *Arg. scol.* Interrogation
Punition	b) Punition	3. *Arg scol* Punition
	3. *Fam.* Question embarrassante	

Les mêmes emplois sont présentés, mais la structuration diffère. Au risque d'employer une image paradoxale, on pourrait dire qu'ici, les feuilles de l'arbre sont constantes : seule change l'arborescence.

Une telle observation est parfaitement compréhensible du point de vue linguistique : les emplois, en effet, reflètent directement la pratique de la langue, ils constituent, pourrait-on dire, sa réalité première – alors que l'organisation lexicographique, relevant du niveau métalinguistique, est une structure au second degré, sujette comme telle à interprétation et à révision. D'où l'intérêt qu'il peut y avoir à dissocier, dans la présentation, les deux niveaux de structuration.

Supposons que l'on veuille modifier, d'une édition à l'autre, la disposition des emplois, par exemple remplacer l'arborescence du *DFC* par celle du *Petit Larousse.* Cela supposerait, dans le format traditionnel, une refonte complète de l'article. Dans un dictionnaire électronique tel que nous le concevons, il suffit de modifier l'information relationnelle dans le champ approprié :

		DFC	*PLI*
colle #1	substance	1	1
colle #2	question embarrassante	2.1	3
colle #3	interrogation scolaire	2.2	2a
colle #4	punition	3	2b

Une telle souplesse de traitement est susceptible de faciliter la maintenance des dictionnaires.

On peut aussi, si on le souhaite, juxtaposer plusieurs dispositions dans des champs différents – l'un reflétant l'ordre historique, l'autre figurant l'arborescence logique, etc. – ce qui est évidemment impossible dans le schéma classique ; ou encore, dans le cadre d'une procédure de désambiguïsation automatique, attribuer aux emplois un ordre séquentiel correspondant à un algorithmique de décision (Martin[3], 1994) ; ou même choisir de n'imposer aucune hiérarchie (comme le fait le *GDEL* pour le mot *colle*). Les emplois constituant la partie stable de l'édifice, les liens qui les unissent peuvent être définis (ajoutés, modifiés, supprimés) avec toute la liberté souhaitable.

Nous conclurons par un point de méthode. Pour un phénomène aussi complexe que la polysémie, il s'avère plus opératoire, du point de vue linguistique comme du point de vue informatique, de commencer par rendre compte de la **diversité** des éléments (dégroupement maximal) avant de pouvoir décrire, avec plus de précision, les **liens** qui les unissent. Disons-le autrement : étant donné une structure à la fois une et multiple – comme l'est la polysémie –, il est techniquement plus simple de partir du multiple pour y introduire l'unité que d'effectuer l'opération inverse. Je ferais volontiers mienne la devise épistémologique du philosophe Jacques Maritain : *distinguer pour unir*[4].

3. R. Martin, 1994, pp. 101 et suiv. Une lecture attentive de l'article *remettre* dans le *TLF* permet à l'auteur d'identifier une quarantaine d'emplois distincts et de les réorganiser systématiquement dans la perspective d'un traitement automatique (construction d'un algorithme permettant la sélection des sens par un automate).
4. Voir le volume paru en 1932 sous le titre *Distinguer pour unir ou Les degrés du savoir* (Bibliothèque française de philosophie, Paris, Desclée de Brouwer)

Une base de données lexicale multilingue interactive

Catherine WALTHER et Éric WEHRLI

Laboratoire d'analyse et de technologie du langage (LATL), Département de linguistique, Université de Genève, Suisse

1. Introduction

Les applications liées au traitement automatique du langage (TAL) exigent d'une manière générale de très gros lexiques, de l'ordre de plusieurs dizaines, voire de plusieurs centaines de milliers d'entrées. Or, l'établissement d'une base de données lexicale est une entreprise de longue haleine, exigeant des moyens importants. On comprend mieux, dès lors, que deux préoccupations majeures dans l'établissement des lexiques au cours de ces dernières années aient été, d'une part, l'utilisation de dictionnaires informatisés et, d'autre part, la réutilisabilité. La première de ces tendances vise à extraire des dictionnaires conventionnels disponibles sur support informatique les données propres à l'élaboration de lexiques utilisables pour le TAL (cf. Atkins et Zampoli (1994) ; Boguraev et Briscoe (1988)). Quant à la seconde tendance, elle tend à définir le contenu des lexiques de façon suffisamment générale (c'est-à-dire non spécifique à une application particulière) de façon à faciliter l'utilisation des lexiques pour d'autres applications. Idéalement, on souhaiterait disposer d'un lexique susceptible d'être utilisé, moyennant quelques inévitables ajouts, pour toute une série d'applications pour une langue ou un groupe de langues données.

C'est dans ce contexte, et avec ces objectifs que le LATL a entrepris, il y a quelques années, le développement d'une base de données lexicale multilingue. Limitée dans un premier temps au français et à l'anglais écrits, cette base de données a été augmentée à plusieurs reprises, d'une part par l'ajout de l'allemand, d'autre part par celui des données nécessaires au traitement de la langue orale. L'approche adoptée s'est voulue résolument pragmatique, et les ajouts ont été effectués en fonction des besoins des applications. Initialement développée pour servir des applications dans le domaine de l'analyse syntaxique et de la traduction interactive, cette base de données est maintenant utilisée pour des projets aussi divers que l'étiquetage (*tagging*), la synthèse vocale à partir de textes (*text-to-speech*), l'établissement de concordances lemmatisées et, dans un avenir proche, l'analyse de la parole.

Dans cet article, nous décrivons la structure de cette base de données lexicale

multilingue interactive (LMI), qui comprend à ce jour des dictionnaires monolingues pour le français, l'anglais et l'allemand, les dictionnaires bilingues (de transfert) correspondants, ainsi que des dictionnaires d'utilisateur et des dictionnaires spécialisés, construits sur la même architecture. Les dictionnaires monolingues et bilingues ont été adaptés de manière semi-automatique (interactive) à partir de plusieurs dictionnaires informatisés, dont l'*Oxford Advanced Learner Dictionary of Current English,* le *Micro Robert-Collins,* le *Grand Robert,* et les bases de donnés CELEX, Brulex et BDlex.

2. Structure des dictionnaires monolingues

Les dictionnaires monolingues français, anglais et allemand sont tous organisés sur le modèle de la spécification morphologique complète, ce qui signifie qu'il s'agit de dictionnaires de mots et non de morphèmes (cf. Jackendoff (1975) et Wehrli (1985)). Les diverses formes morphologiques associées à un lexème particulier correspondent à des entrées distinctes mais liées les unes aux autres, les relations morphologiques étant exprimées sous la forme d'un ensemble de relations entre entrées lexicales indépendantes. Ce mode d'organisation s'éloigne radicalement des modèles plus classiques utilisés en linguistique informatique, dans lesquels la morphologie constitue habituellement une composante indépendante entre le dictionnaire et l'analyseur ou le générateur.

Comme toutes les variantes morphologiques sont présentes en tant que telles dans les dictionnaires monolingues du système, il n'est pas nécessaire de décomposer les mots en analyse, ou de recomposer des formes dérivées en génération.

À l'usage, cette base de données s'est avérée particulièrement efficace (la recherche lexicale se réduit à une recherche dans la base de données), fiable (le système ne peut produire que des formes existantes) et souple (LMI peut être utilisée pour pratiquement n'importe quel type d'application). De plus, le fait que les dictionnaires monolingues contiennent toutes les formes fléchies permet d'associer à chaque variante morphologique d'un mot les informations qui lui sont propres, comme la représentation phonétique, la structure syllabique ou encore sa fréquence.

Il convient pourtant de noter que les dictionnaires monolingues ne se réduisent pas à une simple liste de mots dans laquelle les traits syntaxiques et sémantiques seraient répétés pour tous les membres d'un paradigme morphologique. Dans le but d'éviter ce genre de redondance, nous distinguons deux types d'entrées, soit les formes orthographiques de surface (ou *mots*), et des formes d'un niveau plus abstrait qui correspondent aux lectures des mots (ou *lexèmes*). Les expressions idiomatiques (ou *idiomes*) constituent le troisième type d'entrée des dictionnaires monolingues et bilingues dans la base de données LMI.

2.1. Information liée aux mots

Les entrées de type *mot* contiennent l'information typiquement associée aux formes orthographiques de surface, comme les traits grammaticaux et d'accord (catégorie lexicale, nombre, genre, cas, temps, etc.), la représentation phonétique du mot (y com-

pris la structure syllabique, les accents, pour le français la consonne latente à réaliser dans les contextes de liaison, etc.), et la fréquence d'occurrence. À titre d'exemple, les informations associées avec une forme verbale (*mangeront*) ou adjectivale (*grand*) sont illustrées partiellement en (1) et (2).

(1) **mangeront**
verbe
indicatif, futur, 3ème personne, pluriel

(2) **grand**
masculin, singulier
/grɑ̃/
consonne latente = /d/
...

2.2. Information associée aux lexèmes

L'information associée aux *lexèmes* est de nature syntaxique (structure argumentale, traits sélectionnels, etc.) et sémantique (rôles thématiques des arguments, traits sémantiques, propriétés quantificationnelles, etc.). Elle comprend également une indication de la fréquence d'occurence du lexème (toutes variantes morphologiques confondues). L'exemple (3) ci-dessous donne une illustration des informations associées au verbe *dire* :

(3) **dire**
[thème$_1$, but$_2$] : [__ NP$_1$ PP$_2$], [CP$_1$ PP$_2$]
[thème$_1$] : [__ NP$_1$]
...

Relevons que dans (3), les différentes lectures syntaxiques du verbe *dire* sont représentées sous la forme de structures thématiques, par exemple [thème$_1$, but$_2$], où *thème* est le premier argument et *but* le second, avec l'indication de la réalisation syntaxique de ces arguments. Ainsi, dans notre exemple, le premier argument (thème) peut-il être réaliser soit comme un syntagme nominal, soit comme une proposition. Dans les deux cas, le deuxième argument est réalisé sous la forme d'un syntagme prépositionnel.

2.3. Information liée aux expressions

Nous considérons comme expressions idiomatiques des expressions à mots multiples dont la forme est figée ou semi-figée et/ou dont la sémantique n'est pas complètement compositionnelle comme *casser sa pipe* ou *rendre hommage*, que nous distinguons des mots composés comme *chauve-souris* ou *pomme de terre* (cf. Habert et Jacquemin (1993) ; M. Gross (1986) ; G. Gross (1990), pour un exposé des problèmes liés aux mots composés, Abeillé et Schabes (1989) ; Wasow *et al.* (1994), pour les expressions idiomatiques).

Une entrée de type *idiome* contient le lexème support de l'expression (le verbe

ou le substantif de base), un terme secondaire, la liste des constituants de l'expression, et une liste des contraintes associées à l'expression entière ou à ses constituants (passivisation, plurification, modification adjectivale, référents des possessifs, etc.). Le terme secondaire sert de clé de recherche secondaire, ce qui permet d'optimiser la recherche d'une expression. Cela est particulièrement utile dans le cas d'expressions basées sur un verbe support très fréquent, comme dans les expressions *casser sa pipe, donner sa langue au chat,* ou *faire le zouave*)[1]. Accessoirement, le terme secondaire permet d'établir facilement la liste de toutes les expressions contenant un terme donné. Par exemple, pour le mot *tombe, creuser sa tombe avec ses dents, être muet comme une tombe, avoir un pied dans la tombe,* etc. Un exemple d'entrée est donné en (4) :

(4) *casser sa pipe*
verbe support = casser
terme secondaire = pipe
[casser] [POSS pipe]
[- passif], [POSS = sujet], ...

4. Rôle et place de la morphologie

La morphologie joue un double rôle, à la fois dynamique et statique, dans la base de données LMI. C'est tout d'abord une interface dynamique invoquée lors de l'insertion de nouveaux termes dans la base de données. En effet, pour garantir la cohérence dans les différents dictionnaires monolingues, les nouveaux termes ne peuvent être insérés que par l'intermédiaire de leur forme de référence (infinitif pour les verbes, singulier pour les substantifs, masculin singulier pour les adjectifs). Sur la base de spécifications initiales (catégorie, type de conjugaison ou de déclinaison) fournies au système par l'utilisateur, l'interface morphologique génère automatiquement ou de manière interactive un paradigme complet, qui doit être validé par l'utilisateur avant son insertion. Ce mode de vérification garantit la complétude et l'exactitude des entrées de LMI, à savoir, que toutes les formes fléchies d'un paradigme, et rien qu'elles, sont insérées dans la base de données. À côté du rôle dynamique que nous venons de discuter, la morphologie joue un rôle statique ou relationnel dans LMI, où elle prend la forme de liens entre entrées lexicales complètes. Ces liens expriment des relations flexionnelles, homographiques/homonymiques, ainsi que certaines relations d'ordre dérivationnel.

2.4.1. Relations flexionnelles

Comme l'illustre la figure 1, un paradigme flexionnel entier est relié à la liste des lexèmes qu'il réalise ; inversement, une liste de lexèmes est reliée à l'ensemble du paradigme de mots orthographiques. Ce double lien assure la correspondance entre mots et lexèmes aussi bien en analyse qu'en génération.

1 Pour une discussion plus détaillée du problème de l'identification des expressions idiomatiques dans l'analyseur syntaxique du LATL, voir Campone et Wehrli (1996)

Mots	**Lexèmes**
go	
goes	go 1 (John has gone home)
going	go 2 (he is going to buy it)
gone	go 3 (he is going next)
went	...

FIGURE 1 : Relations morphologiques flexionnelles dans LMI.

2.4.2. Homographie/Homonymie

Lorsque des homographes sont associés à des paradigmes différents, comme c'est le cas pour l'exemple anglais illustré dans la figure 2, ces paradigmes sont enregistrés séparément dans le système.

Mots	**Lexèmes**
depression	[MEDICINE] (to suffer from depression)
depression	[METEOROLOGY] (a deep depression)
depressions	[ECONOMICS]

FIGURE 2 Homographie.

Cette façon de faire permet d'éviter certains pièges bien connus. Par exemple, considérons le cas de la traduction de la phrase (5), extraite d'un traité de médecine.

(5) Les dépressions sont plus fréquentes en automne qu'en été.

Comme il existe une entrée *dépression* portant le marqueur contextuel **MEDICINE** associée à un paradigme qui n'inclut aucune forme du pluriel, le système de génération pourra correctemement produire la phrase (6a) plutôt que la forme incorrecte (6b).

(6) a. Depression is more frequent in Falls than in Summer.
b. *The depressions are more frequent in Falls than in Summer.

Si la phrase (5) provenait d'un rapport météorologique plutôt que d'un traité de médecine, c'est la traduction (6b) et non (6a) qui serait alors appropriée.

2.4.3. Relations dérivationnelles

Certaines relations d'ordre dérivationnel sont également représentées dans la base de données sous la forme de liens entre entrées lexicales. C'est le cas, notamment, de certains dérivés nominaux. Ainsi, des substantifs anglais comme *elaboration* ou *destruction* sont reliés aux verbes dont ils sont dérivés. Cette relation permet de récupérer facilement de l'information pertinente (typiquement, de l'information thématique pour les substantifs dérivés de verbes).

Lexèmes

destruction (N, ...)
destroy (V, [Agent, Thème], ...)

2.5. Traitement des mots composés

Dans les dictionnaires monolingues les mots composés courants (*rendez-vous, papillon de nuit, little by little*) apparaissent comme des entrées indépendantes des mots dont ils sont constitués. Rappelons qu'en général, la sémantique de ces mots composés n'est pas compositionnelle, ce qui signifie que leur sens ne peut être complètement dérivé du sens de leurs parties (p. ex. *zoot suit* « costume de zazou »).

En ce qui concerne l'allemand, langue dans laquelle le processus de composition de mots est particulièrement productif, plusieurs stratégies sont mises en œuvre en fonction de la nature du mot composé. Les verbes à particules séparables (comme *abfahren* « partir en véhicule » ou *weggehen* « partir à pied ») sont associés au verbe support (*fahren* et *gehen* respectivement), et chacune des combinaisons correspond à un lexème particulier (ou à une liste de lexèmes particuliers). Les substantifs composés lexicalisés, comme *Bahnhof* « gare », quant à eux, sont insérés directement dans le dictionnaire. Les mots composés qui ne seraient pas trouvés dans le dictionnaire font l'objet d'une analyse morphologique qui se termine avec succès lorsque toutes les parties du mot composé sont présentes dans le dictionnaire (*Anziehungskraft* « force de gravitation ») ou lorsqu'un préfixe est validé comme un préfixe séparable (*weg* dans *weggegangen* « parti »).

3. Les dictionnaires bilingues

Les dictionnaires bilingues de la base de données LMI font un usage crucial de la distinction entre formes orthographiques et lexèmes, puisque seuls ces derniers servent de base au transfert lexical (cf. Wehrli (1985), entre autres).

Il est aisé de montrer que le transfert lexical ne peut s'effectuer au niveau des formes orthographiques de surface. En effet, ce niveau de repésentation lexicale encode des traits morphologiques tels que le cas, le genre, le nombre, le temps, etc., qui sont généralement propres à une langue et sont susceptibles de varier en fonction du contexte syntaxique dans lequel ils apparaissent. Ainsi, dans une langue aux cas morphologiquement réalisés comme l'allemand, la forme exacte d'un élément cible de type nominal ou adjectival ne peut être établie avant que certaines propriétés syntaxiques de ce constituant n'aient été déterminées, et en particulier sa fonction grammaticale. Cette dernière dépend de multiples facteurs liés aux propriétés lexicales du verbe gouverneur et aux éventuelles transformations grammaticales appliquées lors de la dérivation (transformation passive, etc.). Pour ces raisons, il est indispensable que le transfert lexical soit exprimé au niveau plus abstrait des lexèmes. Dans le processus de traduction, la sélection de la forme morphologiquement appropriée d'un lexème intervient à un stade relativement tardif de la dérivation de la phrase cible, lorsque tous les traits syntaxiques pertinents ont été identifiés.

Les dictionnaires bilingues spécifient l'ensemble des relations possibles entre les lexèmes de la langue source et ceux de la langue cible. Comme les relations entre unités lexicales de deux langues sont très fréquemment de type 1 à n, un dictionnaire bilingue doit également contenir des informations susceptibles d'aider la composante de transfert à sélectionner la traduction la plus appropriée. De telles informations doivent permettre de lever des ambiguïtés du type illustré dans les exemples suivants :

(7)

a. français : *temps*	anglais : *time, weather, ...*
b. anglais : *time*	français : *temps, fois, ...*
c. français : *mur*	allemand : *Mauer, Wand, ...*
d. allemand : *stimmen*	français : *voter, accorder* (musique)

Les lexèmes des langues source et cible sont reliés de façon complètement réversible, puisque chaque entrée dans le dictionnaire bilingue spécifie un lexème source et un lexème cible. Lorsqu'un lexème source peut correspondre à plus d'un lexème cible, le dictionnaire contient autant d'entrées que de correspondances, comme le montre la figure 4.

français	**anglais**
avoir (V, [__ NP])	have (V, [__ NP])
avoir besoin (V, [__ PP])	need (V, [__ NP])
avocat (N)	avocado (N)
avocat (N)	lawyer (N)
casser sa pipe (V)	kick the bucket (V)

FIGURE 4 : Exemples de correspondances bilingues.

Dans le but de résoudre des incompatibilités argumentales ou d'aider à la sélection de l'élément cible le plus approprié, il est nécessaire d'ajouter aux entrées de la figure 4 des informations supplémentaires, en particulier sur les correspondances d'arguments, le contexte d'utilisation ou le sous-langage pertinent, des descripteurs, ainsi que des informations statistiques. Quelques-uns de ces éléments sont décrits dans les sections ci-dessous.

3.1. Transfert d'arguments

Pour tous les éléments lexicaux susceptibles de sélectionner des arguments, comme les verbes, certains adjectifs et substantifs, il est nécessaire de spécifier dans le dictionnaire bilingue la façon dont les arguments du prédicat source correspondent aux arguments du prédicat cible. Cela permet de gérer les cas de non-correspondance comme celui illustré en (8) :

(8) a. Cet homme [vous] fournira [tous les renseignements dont vous avez besoin].
b. This man will provide [you] [with all the information you need].

Le verbe *fournir* dans la phrase (8a) sélectionne un objet direct et un objet indirect. Toutefois, dans la traduction de cette phrase, donnée en (8b), l'objet direct du

verbe *provide* correspond à l'objet indirect de *fournir,* et le complément prépositionnel [*with all the information you need*] à l'objet indirect du verbe français.

Les cas de non-correspondance ne sont pas limités à la réalisation syntaxique des arguments internes, comme pour l'exemple précédent. C'est un fait bien connu qu'un argument externe dans une construction source peut correspondre à un argument interne dans la construction cible. Par exemple, l'argument externe du verbe français *manquer* correspond à l'argument interne de *miss* en anglais ou de *vermissen* en allemand, comme les phrases (9) le montrent :

(9) a. Héloïse manquait à Abélard.
b. Abélard missed Héloïse.
c. Abélard vermisste Héloïse.

Si les correspondances argumentales sont spécifiées dans le dictionnaire, la gestion de tels cas devient relativement simple. Dans la mesure où l'analyseur reconnaît la structure argumentale de la construction source, la composante de transfert utilise l'information de correspondance argumentale pour déterminer comment chaque argument doit être réalisé dans la langue cible.

3.2. Descripteurs

Les descripteurs sont des descriptions brèves (synonymes, définitions, paraphrases, etc.) qui permettent de distinguer les différentes lectures d'un terme source particulier (dans les cas d'homographie et de polysémie). Pour illustrer ce point, admettons que le dictionnaire bilingue français-anglais contienne les deux correspondances données dans la figure 4 pour le mot *avocat*. La première de ces correspondances (*avocat / avocado*) pourrait avoir pour descripteur **Fruit**, et la seconde (*avocat / lawyer*) le descripteur **Homme** ou **Homme de loi**. Ces descripteurs sont utilisés comme matériel de désambiguïsation par le système de traduction automatique interactif. Ainsi, pour le syntagme nominal *un avocat* dans la phrase (10a), l'utilisateur (non nécessairement anglophone) pourra sélectionner la bonne lecture sur la base d'un menu de dialogue du type (10b) :

(10) a. Jean a besoin d'un avocat.
b. avocat 1. fruit
2. homme de loi

3.3. Contextes et sous-langages

L'information contextuelle est une marque associée à des correspondances qui sont spécifiques à un sous-langage donné. Par exemple, le mot français *compositeur*, qui se traduit généralement en anglais par *composer*, devient *typesetter* dans le sous-langage de la typographie. Par conséquent, la correspondance entre *compositeur* et *typesetter* porte le trait contextuel **typography**. De même, la correspondance entre *accord* et *chord* porte le trait **music**, comme on peut le voir dans la figure 5.

français	**anglais**	**contexte**
compositeur	composer	standard
compositeur	typesetter	typography
accord	chord	music

4. LMI en chiffres

Pour conclure ce bref exposé sur la base de données LMI, donnons quelques chiffres qui montreront tout à la fois l'ampleur du travail déjà accompli et de celui qui reste à faire !

Les dictionnaires monolingues comptent approximativement 20 000 lexèmes pour l'allemand et pour le français, 45 000 pour l'anglais, ce qui correspond à plus de 160 000 formes orthographiques en français et en allemand, et environ 85 000 en anglais.

Pour les dictionnaires bilingues, nous disposons de près de 20 000 correspondances entre le français et l'anglais, plus de 10 000 entre l'anglais et l'allemand, alors que le dictionnaire français-allemand est en préparation.

Enfin, en ce qui concerne les expressions idiomatiques, elles n'ont pour l'instant été prises en considération que pour le français, et notre lexique en compte un peu plus de 2 000.

Acquisition semi-automatique du lexique

Evelyne VIEGAS et Sergei NIRENBURG

Computing Research Laboratory, New Mexico State University, Las Cruces, USA

• *Abstract* •

In this paper, we present the process of lexical acquisition as we defined it to build ***Spanlex****, a Spanish lexicon, for* ***Mikrokosmos****, a semantics-oriented machine translation (MT) system between Spanish and English In this paper, we discuss the types of information which must be included in a computational lexicon for this and similar applications such as, for instance, text generation We also present the acquisition methodology we developed which supports team acquisition*[1].

1. Introduction

Mikrokosmos est un système de traduction automatique entre l'espagnol et l'anglais, de textes journalistiques appartenant en priorité au sous-domaine terminologique d'opérations d'achat, de vente ou de fusion entre compagnies. Le système Mikrokosmos est basé sur la sémantique (Nirenburg *et al.*, 1994) et adopte une approche interlangue pour la traduction. La représentation interlangue est appelée une (TMR) *Text Meaning Representation* ou représentation de sens du texte. Nous ne pouvons développer, dans cet article, le processus complet de traduction ; nous nous contenterons de décrire très partiellement la représentation interlangue TMR, qui est essentiellement constituée des TMR provenant du lexique, où elles apparaissent à l'état non-saturé[2]. Les lexèmes acquis proviennent d'un corpus de textes journalistiques qui présente un vocabulaire spécifique, mais aussi du langage courant, de par les produits dont il est question (voitures, pharmacies,. .). Ainsi, dans notre tâche d'acquisition

1. Toute notre gratitude à Victor Raskin, pour sa participation très active dans la mise en place du protocole d'acquisition. Nous remercions également tous nos lexicographes-acquéreurs pour leur acquisition effective, en particulier Oscar Cossio, Margarita Gonzales, Jeff Longwell, Maya, Javier Ochoa

2. Par « état non-saturé », nous entendons non-instancié Par exemple, le mot *manger* donne des indications dans sa TMR sur ses contraintes de sélection : ANIMAL pour l'agent et MANGEABLE pour le thème ; ce n'est qu'au niveau de la TMR complète, lorsque toutes les TMRs venant d'autres lexèmes ont interagi, que l'on a une TMR saturée ; ainsi, dans *Jean aime manger chaud*, ANIMAL sera contraint à HUMAIN

lexicale, nous travaillons en collaboration étroite avec des lexicologues et des terminologues.

Notre approche théorique se situe dans une perspective de linguistique computationnelle qui privilégie la notion d'organisation lexicale, au sein d'une sémantique semi-compositionnelle. En d'autres termes, nous tirons avantage des techniques de l'intelligence artificielle et de la sémantique linguistique.

Notre but est d'acquérir un lexique d'environ 40 000 sens de mots. Cela rend nécessaire, si ce n'est inévitable, de semi-automatiser la tâche d'acquisition pour les tâches « rebutantes », comme, par exemple, vérifier l'orthographe d'un lexème de façon à ce que l'acquéreur puisse se concentrer sur des tâches plus productives et intéressantes (informations syntagmatiques, paradigmatiques, stylistiques ou pragmatiques).

Cela suggère en priorité d'avoir accès à des dictionnaires informatisés, à des corpus informatisés et surtout à des interfaces permettant une manipulation aisée de ces ressources. Nos interfaces ont été élaborées en accord avec l'utilisateur et continuent d'évoluer en fonction des besoins et des problèmes rencontrés lors de leurs utilisations.

Dans cette première étape de Mikrokosmos, nous avons essentiellement mis l'accent sur l'acquisition de l'information syntactico-sémantique, effectuée par le biais de dépendances sémantico-syntaxiques.

Nous présentons à l'acquéreur une série de moules sémantico-syntaxiques prédéfinis, le guidant dans la phase d'acquisition. Lexicographes et terminologues utilisent le même outil d'acquisition, puisqu'il leur est possible de spécifier différents types d'informations (le domaine d'application, le lieu d'emploi, les collocations, etc.).

Dans ce qui suit, nous présentons tout d'abord le type d'information que l'on trouve dans les lexiques computationnels et la façon dont cette information est structurée et organisée. Puis, nous motivons et présentons le type d'information que nous codons dans nos propres lexiques, en donnant l'exemple de **Spanlex**, le lexique espagnol acquis dans le cadre du projet Mikrokosmos. Nous passons ensuite à la tâche même d'acquisition et présentons les différentes étapes intervenant dans ce processus, en montrant qu'il est possible d'acquérir un lexique de haute qualité et à grande échelle.

2. Quels types de lexiques pour quels types d'application : la question est-elle vraiment pertinente ?

Les principaux lexiques computationnels que l'on trouve sur le réseau ou qui sont en cours de développement à l'heure actuelle, renferment essentiellement deux types d'information : essentiellement syntaxique, comme *Comlex*, (Macleod et Grishman, 1994) et/ou sémantique, par exemple *Acquilex*, (Sanfillippo, 1992). Un autre type de différences entre ces deux types de lexiques, se situe au niveau de la représentation, qui adopte soit une hiérarchie lexicale, où les feuilles de la hiérarchie sont des lexèmes ; ou qui adopte une hiérarchie conceptuelle, où les lexèmes sont reliés entre eux via les concepts. C'est cette dernière approche que nous adoptons dans Mikrokosmos.

En ce qui concerne l'organisation globale du lexique et quels lexèmes vont donner lieu à une entrée dans le lexique computationnel, il est à noter que les lexiques qui mettent l'accent sur la syntaxe adoptent une approche par énumération de sens basée sur des sous-catégorisations différentes : *vouloir* et *vouloir que...*, reçoivent des entrées différentes, comme nous l'expliquons dans le paragraphe suivant.

2.1. Motivation de la description de nos lexiques

Le type d'information contenu dans un dictionnaire dépend très souvent du **type d'application** pour lequel il est utilisé. Par exemple, pour faire de la traduction multilingue, des dictionnaires multilingues, où l'on juxtapose les lexèmes des différentes langues, peuvent être suffisants : *manger/comer/eat*, respectivement en français, espagnol et anglais. Pour faire de la génération, de l'information sur l'ordre des mots, par exemple la place de l'adjectif dans un groupe nominal, *une maison bleue* et non *une bleue maison*, est nécessaire ; de même, il est nécessaire de coder les collocations dans un dictionnaire, *a heavy smoker* versus *un grand fumeur*, respectivement en anglais et français. Il est clair que l'information collocationnelle n'est pas indispensable en phase d'analyse.

Cependant, l'acquisition de dictionnaires à grande échelle est un travail coûteux, c'est pourquoi il est préférable d'acquérir un lexique qui soit réutilisable pour d'autres domaines, d'autres applications.

Cela nous amène à nous concentrer maintenant sur le type d'**organisation et** de **structuration du lexique**. Il est bien établi, à l'heure actuelle, en sémantique lexicale computationnelle, qu'une simple énumération des mots par ordre alphabétique est computationnellement chère (Boguraev et Pustejovsky, 1990). Par ailleurs, il est également reconnu qu'une structuration du sens des mots basée sur des sous-catégorisations différentes est également chère, computationnellement parlant, et surtout inadéquate, en ce qu'elle empêche de capter le noyau sémantique de l'item lexical, par exemple *oublier* qui sous-catégorise pour un groupe nominal dans *J'ai oublié les clés de ma voiture* ou un xcomp dans *j'ai oublié de prendre mes clés* ou encore un comp dans *j'ai oublié que tu avais mes clés,* (Viegas et Nirenburg, 1995). Les lexiques que nous construisons intègrent les résultats les plus avancés, qu'ils proviennent de la sémantique lexicale (distinction de sens de mots), de la linguistique computationnelle (dépendances sémantico-syntaxiques), ou de l'intelligence artificielle (en termes de techniques, comme les hiérarchies à héritage simple ou multiple). La construction de ce type de lexiques implique de faire appel à des experts en lexicographie, terminologie, linguistique computationnelle, et représentation des connaissances.

Dans ce qui suit, nous donnons une description documentée de l'information contenue dans nos lexiques. Le lexique est composé de super-entrées (suivant la convention adoptée par Meyer *et al.*, 1990). Chaque super-entrée consiste en une liste d'entrées représentant des sens de mots différents, et ce indépendamment de la catégorie syntaxique du mot. Chaque sens de mot est identifié par un unique identificateur, ou un lexème (suivant la terminologie de (Mel'čuk *et al.*, 1984). À l'intérieur d'une entrée on peut trouver des noms, verbes (en anglais, *walk-N* versus *walk-V*), ou des homonymes. Les critères utilisés pour décider ou non de la création d'une

nouvelle entrée ou d'une sous-entrée sont les suivants (Onyshkevych et Nirenburg, 1994) :

1 Candidats potentiels à une super-entrée :

- les noms composés représentés par un seul mot ou séparé par un tiret (rouge-gorge), sont des candidats potentiels ;
- les noms propres, composés ou non, sont stockés dans une base de connaissances, autre que celle du lexique de base ; ils sont néanmoins codés en utilisant le même format ;
- les idiomes sont indexés sur la tête, par exemple, *battre* dans *battre le fer (tant qu'il est chaud).*

2 Critères de sélection pour un nouveau lexème :

Nous renvoyons à (Meyer *et al.*, 1990) pour la justification des critères linguistiques et lexicographiques qui permettent de décider la création ou non d'une nouvelle sous-entrée.

Concrètement, en phase d'acquisition, la façon dont est structurée notre ontologie, joue un rôle primordial dans la façon d'acquérir un lexème. Si nous reprenons l'exemple de *livre*, qui est ambigu entre l'ASPECT PHYSIQUE et l'ASPECT DOCUMENT, la décision de créer une entrée ou deux est « suggérée » au lexicographe, dans la mesure où l'ontologie prévoit les deux types pour *livre*, comme le montre la figure ci-dessous, (Kavi, en préparation).

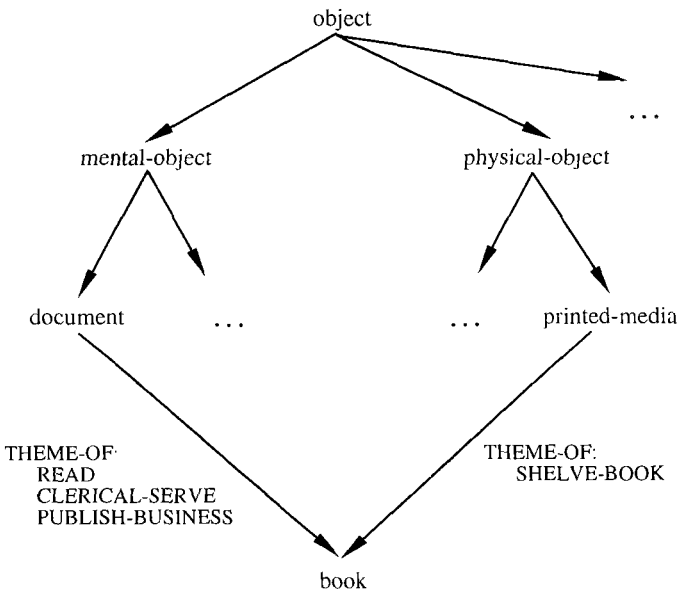


FIGURE 1 : Ontologie pour livre.

2.2. Les différentes zones d'un lexème

L'information contenue dans un lexème est répartie selon des **zones** correspondant à divers niveaux d'information lexicale, que nous décrivons ci-dessous (Meyer *et al.*, 1990).

1 **CAT**égorie syntaxique : *Nom, Verbe, Pronom*
2 **MORPH**phologie : pour les formes irrégulières et les changements de stems *mouse* versus *mice* en anglais.
3 **COMMENTS** : cette zone est subdivisée en plusieurs zones donnant de l'information administrative (date d'entrée du système pour l'acquisition du lexème, date de modification, nom du lexicographe), une définition du mot, des exemples dans la langue source, le domaine d'application du mot (langue, région).
4 **ORTH**ographe : pour les abréviations ou variantes *United States of America versus USA.*
5 **PHON**onologie
6 **SYN**tactic-**STRUC**ture : donne des indications sur les dépendances syntaxiques au niveau de la proposition ou de la phrase ; cette zone renferme essentiellement de l'information sous-catégorielle.
7 **SEM**antic-**STRUC**ture : la partie sémantique lexicale du mot, donnant sa TMR non saturée, ou représentation du sens.
8 **LEX**ical-**REL**ations, donne de l'information collocationelle.
9 **LEX**ical-**RULES** : répertorie l'ensemble des règles qui s'appliquent à ce lexème.
10 **STYL**istique : donne de l'information sur les facteurs stylistiques, tels que le degré de familiarité, de formalité. Cette zone contient aussi des sous-zones contenant des fonctions « déclencheuses » pour l'analyse (pour traiter la coréférence) ou pour la génération (donnant la préférence au niveau de l'ordre des mots, par exemple.)

2.3. Une base de connaissances multi-propos

De la description précédente, par rapport au type d'information codée dans le lexique, il découle que nos lexiques sont multi-propos, en étant conformes aux trois points ci-dessous :

a **multi-lingues** : ils acceptent plusieurs langues naturelles aussi différentes que l'espagnol et le japonais,

b **multi-media** : ils renferment de l'information linguistique, pour le traitement du langage naturel, de l'information phonologique, essentiellement pour un traitement de reconnaissance de la parole, et enfin de l'information pour la vision, pour une reconnaissance visuelle, par le biais de l'ontologie.

c **multi-usages** : ils peuvent être utilisés en analyse, et génération monolingue ou multilingue ; en traduction (semi)-automatique ; pour la reconnaissance/production de la parole.

3. Le processus d'acquisition : les différentes tâches

Nous consacrons les paragraphes suivants au processus même d'acquisition de la connaissance syntactico-sémantique, tel que nous l'avons développé pour **Spanlex**. Le schéma ci-dessous représente tous les modules de travail et de ressources nécessaires à l'acquisition du lexique décrit ci-dessus. Dans ce schéma figurent également les outils de test/évaluation des entrées lexicales, ainsi que l'analyseur sémantique qui utilise dynamiquement toutes les ressources statiques (lexique, ontologie).

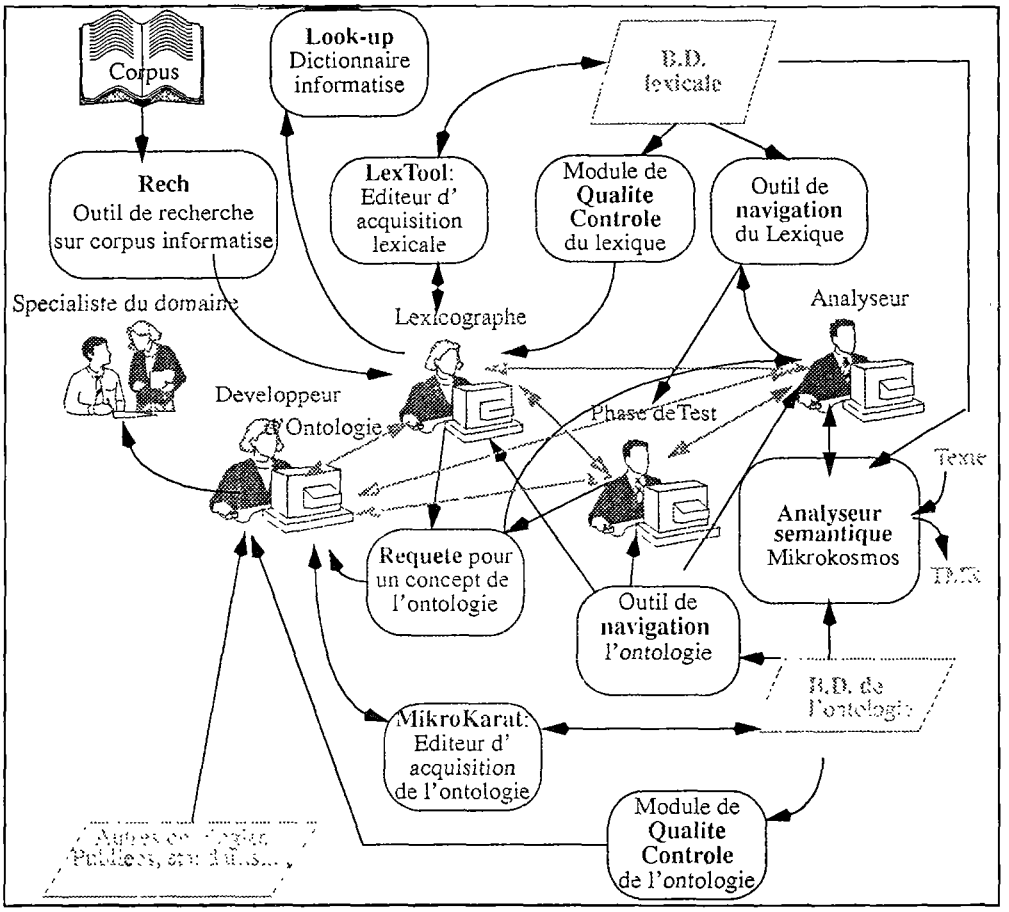


FIGURE 2 . Processus d'acquisition : modules de travail et ressources.

Il est important de noter le nombre d'interactions entre les « développeurs » de lexiques et les « développeurs » d'ontologies. En effet, l'ontologie n'est pas figée une fois pour toutes, elle évolue en fonction des lexèmes rencontrés dans différentes langues. Pour cela, nous avons développé des outils de communication où des requêtes aux « développeurs » d'ontologie sont formulées périodiquement.

4. Quelques chiffres explicatifs

Nous nous concentrons maintenant sur le processus d'acquisition tel que nous l'avons développé pour Spanlex, le lexique espagnol que nous sommes en train de construire. Notre tâche consiste à acquérir de 30 000 à 40 000 sens de mots espagnols. Le type d'information que nous codons dans le lexique, à savoir essentiellement lexico-sémantique, ne peut se faire si ce n'est de façon semi-automatique. Pour cela, nous avons conçu et implémenté des outils d'aide à l'acquisition proprement dite (voir figures en Annexe 1), de façon à faciliter la tâche aux lexicographes, ainsi que des outils d'évaluation de l'acquisition, de façon à pallier les inconvénients dus à une intervention humaine, pour une application computationnelle.

Nous décrivons ci-dessous le développement du processus sur une période de douze mois, de novembre 1994, date de l'initialisation de l'acquisition lexicale jusqu'à décembre 1995. Nous montrons comment il est possible effectivement de réaliser une acquisition lexicale à grande échelle, en incorporant essentiellement de l'information sémantique.

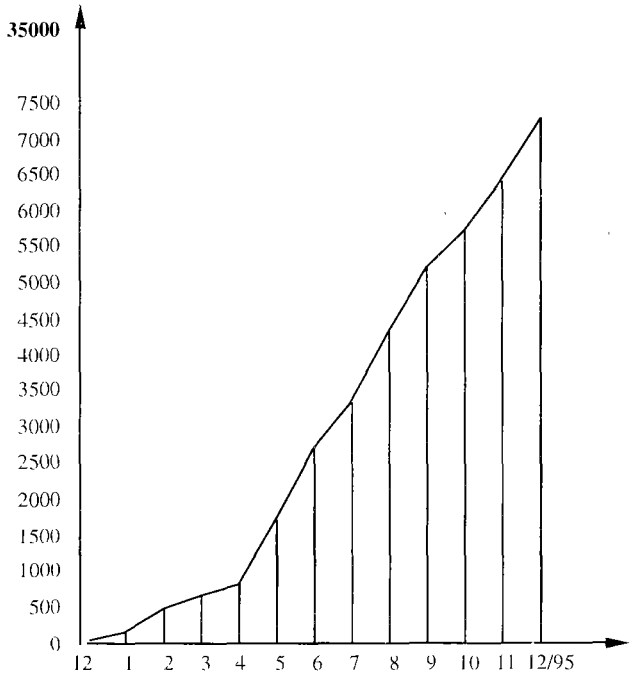


FIGURE 3 : Acquisition semi-automatique du lexique

La figure 3 montre l'évolution du nombre d'entrées sur les douze mois. Les chiffres représentés ici, indiquent le nombre de sens de mots ou de lexèmes acquis semi-automatiquement. Le premier travail a consisté en une analyse linguistique de façon à déterminer les formats pré-définis renfermant de l'information sous-catégorielle et les rôles thématiques associés, comme le format Trs-Ag-Th qui spécifie que le verbe est transitif et sous-catégorise pour un sujet et un objet qui ont les rôles respectivement

d'agent et de thème, par exemple le verbe *manger*. Ces formats, résultats d'une étude linguistique, ont été mis en place de façon à guider le lexicographe dans sa tâche d'acquisition. Nous avons également conçu une interface d'acquisition, qui a été élaborée en suivant le processus d'acquisition tel que les lexicographes le perçoivent et non pas en suivant la structure de la base de données. En plus de l'interface d'acquisition elle-même, nous avons mis à la disposition de nos lexicographes des dictionnaires monolingues espagnols et bilingues, espagnol-anglais. Par ailleurs, un outil de recherche de mot en contexte, dans un corpus de textes espagnols en ligne a été également créé (figure en Annexe 2). Ces outils sont décrits plus amplement dans (Viegas, 1995). Ce type d'acquisition requiert un important effort d'apprentissage de la part des lexicographes, et est aussi coûteux. Cependant, si l'on veut constituer un lexique syntactico-sémantique de base, cet effort est inévitable.

Nous avons développé en parallèle un programme produisant entièrement automatiquement les entrées, attestées par les dictionnaires et/ou corpus, des nouvelles formes dérivationnelles de lexèmes acquis semi-automatiquement, et ce à l'aide de règles morpho-sémantiques.

Voici un exemple partiel, des formes générées automatiquement, avec une règle sémantique associée, pour le verbe espagnol *comprar* (acheter) :

comprar, v, LR1event
comprador, n, LR2social-role_relation1a
compra, n, LR2event10
compra, n, LR2theme_of_event10
compradero, adj, LR3feasibility_attribute2a
comprable, adj, LR3feasibility_attribute1
comprado, adj, LR3event_telic
compradizo, adj, LR3feasibility_attribute5a
comprador, adj, LR3social_role_relation1a
malcomprar, v, LRneg_affect1, LR1event
malcomprado, adj, LR3event_telic
recomprar, v, LRrepetition1, LR1event
recompra, n, LR2event10
recompra, n, LR2theme_of_event10
recomprado, adj, LR3event_telic

Par exemple, *comprable, adj, LR3feasibility_attribute1*, est dérivé morphologiquement de *comprar*, et ajoute à la sémantique de *comprar* la caractéristique d'être possible ou non.

Les nouvelles entrées générées, sont ensuite testées par les lexicographes, à l'aide des mêmes outils qui testent les entrées acquises semi-automatiquement ; nous allons pouvoir ainsi multiplier la taille de notre lexique[3], par 5 ou 6 (figure 4), et ce automatiquement (Viegas et Gonzales, 1995).

3. Nous avons commencé la génération morpho-sémantique à partir des 1 500 verbes déjà acquis, et le nombre moyen de formes lexicales générées par verbe est de 35.

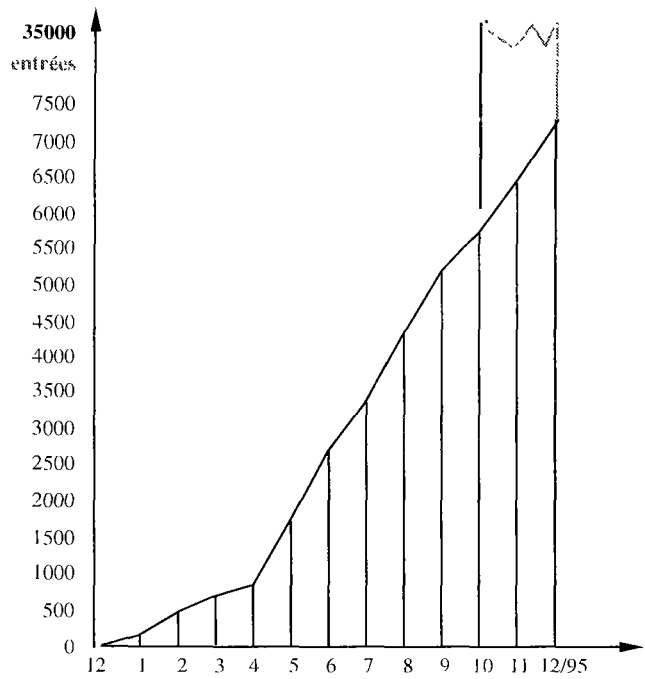


FIGURE 4 Développement du lexique à partir de règles morphologiques dérivationnelles.

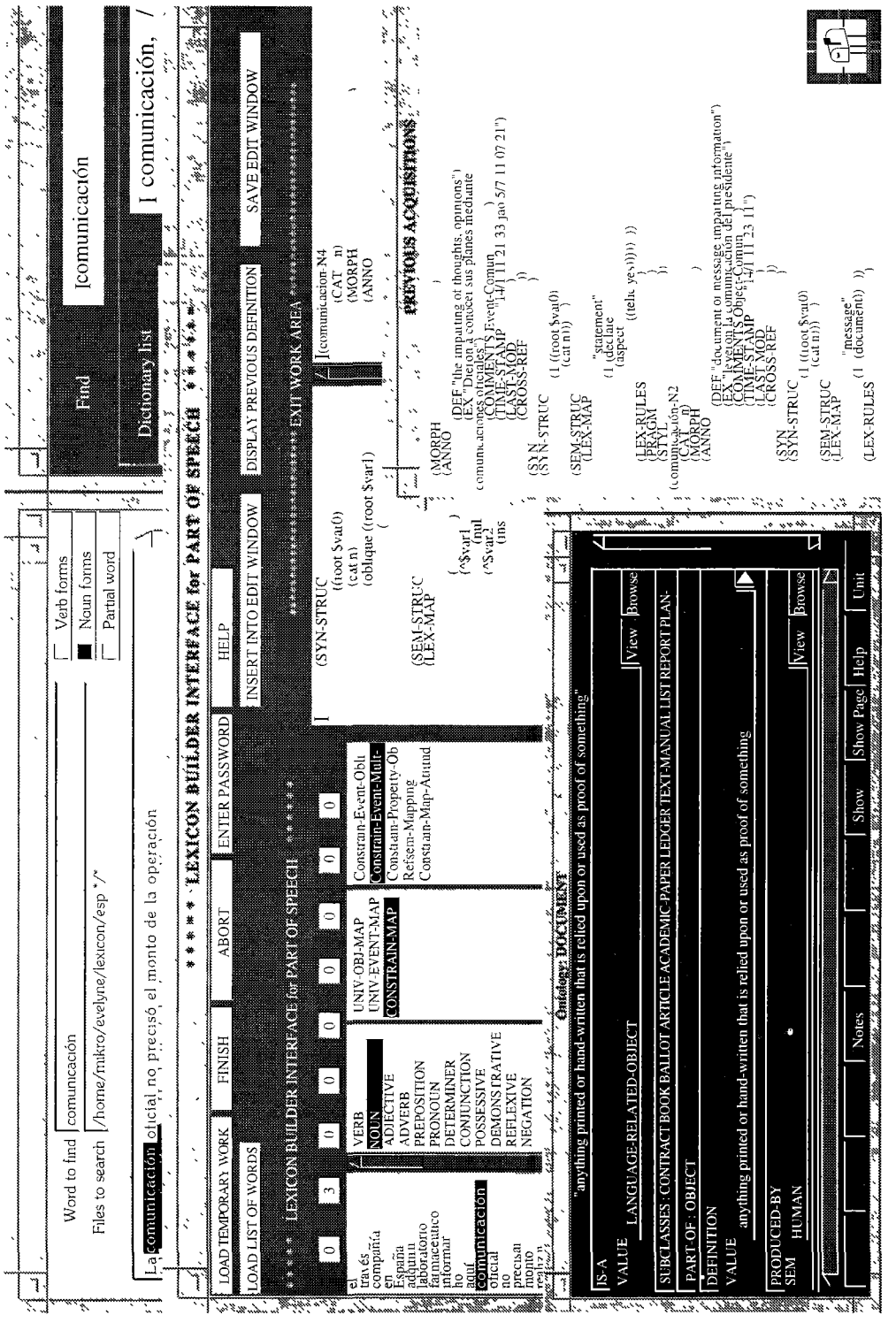
Dans ce processus de génération automatique, le but est de pouvoir générer automatiquement la sémantique des dérivations morphologiques des lexèmes déjà acquis, afin de les présenter à l'acquéreur pour vérification, tout en allégeant la tâche de l'acquisition manuelle, et tout en enrichissant notre lexique[4].

5. Conclusion

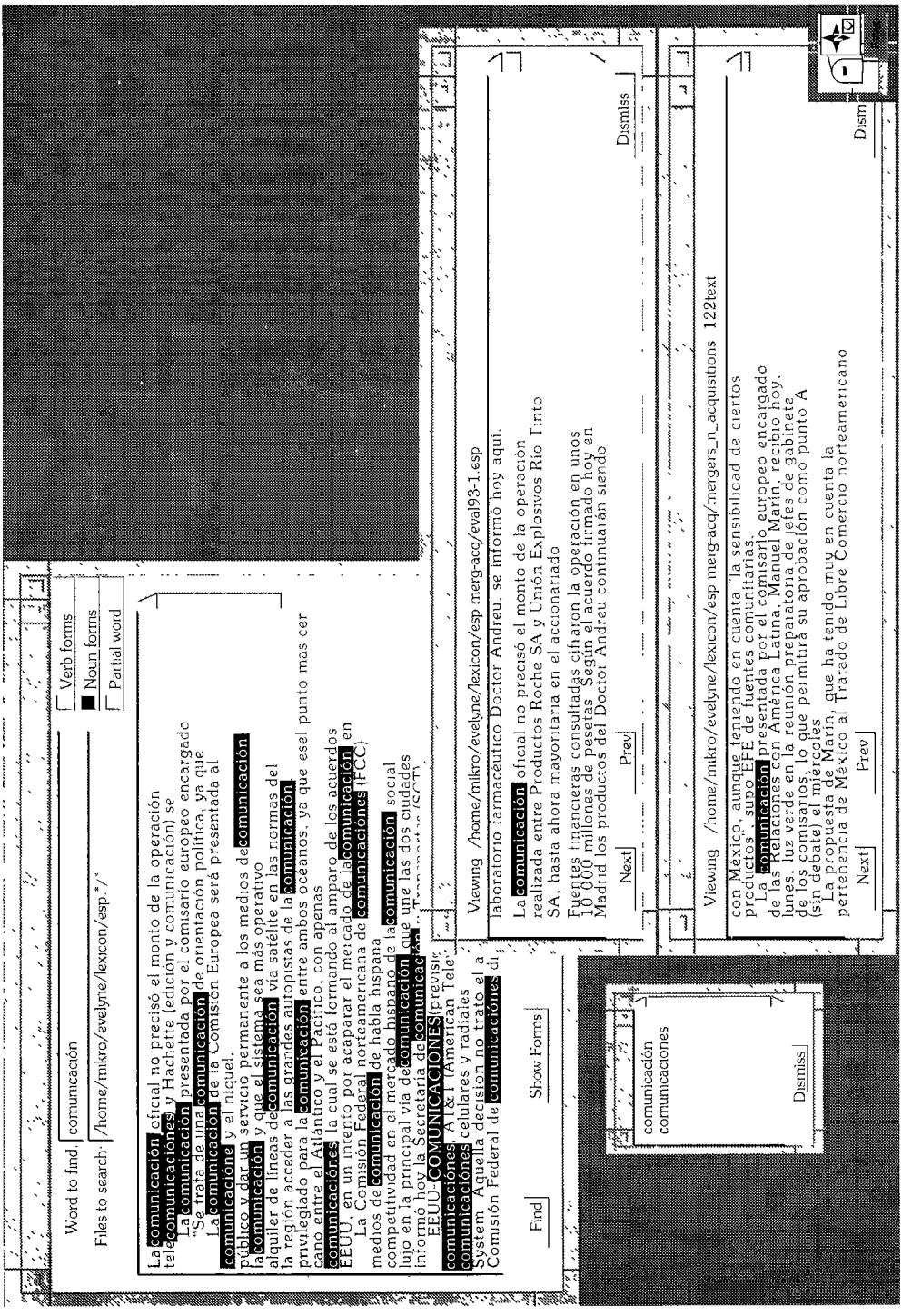
Dans cet article, nous avons présenté les ressources nécessaires pour réaliser l'acquisition semi-automatique de lexiques riches en information sémantique. Nous avons également mis l'accent sur, l'importance d'une intervention humaine d'une part, et d'autre part sur l'importance d'avoir des outils pour guider l'acquisition, et des outils pour évaluer les résultats de l'acquisition, de façon à élaborer des lexiques de haute qualité et utilisables pour un traitement (semi-)automatique du langage naturel. Il est aussi primordial, d'aller vers la construction de bases de connaissances lexicales qui puissent être utilisées par différentes applications ou dans différents domaines.

4. Viegas et Nirenburg (1996) expliquent en détail le processus de génération morpho-sémantique.

Annexe 1 : Outils d'acquisition



Annexe 2 : Recherche d'occurrences dans le corpus



Pistes de description sémantique : le cas de Biolex, dictionnaire des bio-industries

François Gaudin et Myriam Bouveret

URA CNRS 1164[1], Université de Rouen et Praxiling, Université de Montpellier, France

Notre texte sera centré sur un dictionnaire des bio-industries qui nous servira de corpus et de fil conducteur. Ce dictionnaire a été conçu et dirigé par Louis Guespin, jusqu'à sa disparition, au sein de l'URA CNRS 1164, équipe rouennaise de recherche en sociolinguistique. Sa réalisation a été rendue possible grâce aux aides de l'ANVAR et du Ministère de la Recherche.

Ce dictionnaire a été développé sur logiciel Termex. Il recense une nomenclature de 1 200 termes dont il propose une description française en 19 rubriques et des traductions en anglais et en allemand. Mais, au-delà de ces caractéristiques techniques modestes, nous voudrions livrer quelques-unes des pistes que nous avons suivies dans un effort qui visait à décrire un vocabulaire tout en conciliant recherche fondamentale et recherche appliquée ; c'est-à-dire en cherchant à mettre à l'épreuve certaines options théoriques. Rappelons que ces options théoriques, sur lesquelles nous ne reviendrons pas ici, ont fait l'objet de quelques publications collectives rassemblant travaux et réflexions autour de la socioterminologie[2].

Nous nous intéresserons ici principalement aux rubriques que nous avons choisies pour décrire les relations lexicales et prédicatives en mettant en lumière les choix qui ont guidé notre travail. Parmi ces choix, nous relèverons le refus du recours aux définitions, une préférence accordée aux contextes, une prévention à l'égard de la notion de domaine, ainsi que la tentative de distinguer les niveaux de la signification et du concept.

1 Unité de recherche associée au Centre national de la recherche scientifique, n° 1164 . « Sociolinguistique, usage et devenir de la langue ».
2 Voir Gaudin et Assal, 1991 ; Gambier et Gaudin, 1993 ; Gardin *et al.*, 1994 ; Gaudin, 1995

Les relations lexicales

Concernant la description des relations lexicales, Biolex offre 5 rubriques, inégalement productives : hyperonyme, hyponyme, isonyme, antonyme, synonyme.

Les relations d'**hyperonymie** et d'**hyponymie** se sont avérées très utiles, ce qui n'a rien pour surprendre puisqu'elles sont au cœur de tout discours lexicographique. La relation de **synonymie** autorise le regroupement de relations diverses : du syntagme à sa réduction, de la forme en usage à la forme officielle, de l'emprunt à la forme autochtone. Relevons également quelques cas de synonymies commerciale et historique. En revanche, la relation d'**antonymie** est très peu utilisée, même élargie aux contraires complémentaires et réciproques. Outre le caractère propre de cette relation, peu productive, cela s'explique par le fait que la relation d'antonymie se trouve en concurrence, dans notre description, avec celle d'**isonymie.**

Cette relation lexicale unit ce que l'ISO appelle des *notions coordonnées*, celles-ci étant définies comme des notions qui, dans un système hiérarchique, se situent « au même niveau qu'une ou plusieurs autres notions » (ISO, 1990 : 2). Sans entrer dans un débat terminologique, disons que nous appelons « relation d'isonymie » toute relation unissant deux unités mises en concurrence, le plus souvent de même niveau, sans que l'on puisse poser une hiérarchie valable selon tous les points de vue[3].

Cette relation est fructueuse, dans la mesure où elle permet de recenser le plus directement les unités auxquelles le terme étudié s'oppose le plus directement dans un paradigme discursif. Dans notre corpus, les termes *néphélométrie* (nephelê, nuage) et *turbidimétrie* (turbidus, trouble) sont isonymes et ne pourraient être englobés que par une paraphrase du type « technique de mesure de la turbidité » et pas par une unité lexicale attestée. L'analyse dégage là un archisémème sans archilexème correspondant.

Après le premier repérage, au sein d'une catégorie, que permet la relation d'hyperonymie, les renseignements paradigmatiques aident la construction du signifié. Hyponymes, isonymes, antonymes, synonymes permettent de situer la valeur linguistique du terme, voire de sa base lexicale lorsqu'une famille dérivationnelle existe. C'est là un postulat de travail.

Les relations lexicales peuvent également permettre de procéder à une levée d'homonymie quand l'histoire entérine des distinctions lexicales nouvelles. Notre corpus nous a ainsi conduits à distinguer deux entrées *fermentation* :

– *fermentation* 1 : fermentation anaérobie, sens historiquement premier ;
– *fermentation* 2 : fermentation aérobie, sens utilisé en milieu industriel.

Un tel éclatement de la description pourrait paraître excessif. À tout coup, il serait illégitime dans une dictionnaire de langue. Mais ici, l'histoire a vu évoluer les pratiques des bio-industriels et leurs discours, discours dans lesquels le terme *fermenta-*

3. Pour un développement illustré d'exemples tirés de notre corpus, cf Assal *et al*., 1992.

tion a pris deux significations différentes : le sens « aérobie », lié aux fermentations industrielles, est venu s'installer à côté de celui d'« anaérobie ».

L'évolution des techniques a suscité la circulation concurrente des deux emplois. Il s'est d'abord agi de spécifications locales, liées à des textes ; mais l'apparition régulière de tels contextes a fini par modifier le sens en discours de ces deux types d'emplois. Ensuite, cette distinction, liée à la pratique technologique, s'est stabilisée en différence de signification.

Cette signification s'est lexicalisée et, aujourd'hui, l'utilisation d'un syntagme comme *fermentation continue* est licite, mais sa compréhension suppose le recours à *fermentation* 2, car cette sorte de fermentation a toujours lieu en présence d'oxygène. En termes de sémantique structurale, on peut diagnostiquer ici une différence de sèmes inhérents.

Dans cet exemple, la divergence lexicale est attestée également dans le paradigme dérivationnel puisque l'on trouve le nom d'instrument *fermenteur* qui sert à désigner des appareils liés à la fermentation aérobie (fermentation 2). Au sein de cette sphère d'activité, on a bien deux signifiés qui s'opposent. Il y a eu retombée de la distinction des deux signifiés sur la famille dérivationnelle. On voit là une illustration de la pertinence des paradigmes dérivationnels, tels que les a mis en lumière Louis Guilbert (1975 : 173 et suiv.), et dont Pierre Lerat a montré toute la pertinence pour les langues spécialisées (Lerat, 1995).

La distinction des signifiés peut être vue comme une lexicalisation d'une différence conceptuelle, lorsqu'un continuum existe entre deux signifiés. Ainsi, dans le cas du terme *viscosité*, la propriété dénotée a d'abord été spécifiée dans des discours étudiant les propriétés des fluides, en restant encore rattachée au trait « caractère visqueux, épais ». Ensuite, au XIX^{e}, elle s'est isolée pour désigner une propriété mesurable. C'est alors, quand on put isoler une grandeur quantifiable de façon mathématique par des appareils lentement perfectionnés, que les hommes de l'art usèrent de termes nouveaux et propres en recourant à une dérivation basée sur le radical *viscos-* et n'intéressant que le sens physique de *viscosité* ; on vit alors apparaître des termes tels que *viscosimètre* en 1831, *viscosimétrie* en 1933, etc.

C'est dans cette logique de la retombée d'une distinction conceptuelle sur le système de la langue que nous avons été conduits à distinguer sous deux entrées le sens de *viscosité*[1] lié au mot visqueux, « gluant », et synonyme d'*adhésivité*, rendu par *Klebrigkeit* en allemand et celui de la *viscosité*[2], conçue comme propriété de certains fluides que mesure le viscosimètre. L'existence d'un nom d'appareil de mesure vient en quelque sorte attester au plan dérivationnel de la spécificité conceptuelle du second terme.

D'ailleurs, la distinction que nous avons opérée au plan paradigmatique se retrouve au plan syntagmatique, puisque typiquement la *viscosité*[2] mesurable est celle des fluides, et non la viscosité du « chapeau de certains bolets », « de la main d'œuvre » ou « du marché des capitaux », comme nous le suggère le *Nouveau Petit Robert*, pour ne rien dire de la redoutable « viscosité mentale » qu'il mentionne.

Le lien entre *viscosité*[2] et *fluide* est une relation que nous dirons prédicative.

Les relations prédicatives

Biolex présente quatre rubriques descriptives consacrées aux relations prédicatives, ce sont les rubriques « action typique », « objet typique », « agent typique » et « application typique ». Les rubriques prédicatives que nous avons utilisées sont inspirées des fiches proposées par Pierre Lerat pour les travaux du Centre de terminologie et de néologie dont il a été le fondateur. Son idée initiale était d'explorer le rendement des propriétés de :

> prédicat privilégié (« action typique ») et d'argument logique contigu (« objet connexe ») dans une relation prédicative de premier ordre. Dans la rubrique « objet », il s'agit de repérer les notions qui sont associées le plus spontanément par les experts (la cause, le prévenu, le tribunal, autour de l'action « procès », pour le juge, par exemple) (Lerat, 1988 : 18).

C'était là permettre d'enregistrer des faits syntagmatiques significatifs et non plus seulement des faits paradigmatiques.

Nous avons conservé ces rubriques d'action et d'objet, dénommées « action typique » et « objet typique », en les élargissant, sous une terminologie commune, aux relations d'« agent typique » et d'« application typique ».

Ces quatre rubriques permettent l'enregistrement de renseignements d'ordre syntagmatique assez variés. De tels éléments ne tombent pas du ciel : ils sont repérés dans des discours et l'on va recenser les plus fréquents, les plus centraux – bref, les plus **typiques**. Après le temps de la sélection, vient le temps de la combinaison.

Ces rubriques permettent donc de recenser les cooccurrents les plus fréquents, et donc les éléments de discours les plus centraux, les plus courants. Si les premières rubriques permettaient de cerner la signification, les relations prédicatives permettent de mettre en lumière le sens le plus fréquemment produit dans une formation discursive.

Pourquoi parler de « formation discursive » ? Pour insister sur le fait qu'il s'agit de rendre compte d'un ensemble de discours tenus au sein d'une sphère d'activité, cet ensemble n'étant pas réductible à des découpages institutionnels (cf. Gaudin, 1993). En fait, comme le souligne Michel Foucault, « on ne peut pas établir de relation biunivoque entre les disciplines instituées et les formations discursives » (1969 : 233). C'est là un des intérêts que présente la description d'un vocabulaire comme celui que rassemblent les bio-industries. Mais revenons aux relations prédicatives.

En se plaçant sur le plan syntagmatique, on est passé de la signification stable au sens produit qui va permettre la construction du concept.

Prenons les arômes en exemple. Tel qu'utilisé en bio-industries, *arôme* ne présente qu'une famille dérivationnelle. Nous en concluons à l'existence d'un seul signifié. Mais ce signifié unique permet de désigner plusieurs concepts. En effet, les arômes peuvent être distingués selon leurs actions typiques, *aromatisation* ou *fumaison*, l'objet typique restant identique : *denrées alimentaires*. Ici deux rubriques seulement sont intéressantes car l'agent typique n'est autre que le vocable étudié et l'application typique se situe à un niveau peu pertinent puisque cette rubrique renvoie

aux industries alimentaires. On est dans un cas de figure où l'application typique rejoint le domaine.

Pareillement, pour la chymosine les deux rubriques renseignées – l'action typique, *coagulation*, et l'objet typique, *lait* – permettent d'en construire le sens utile : la chymosine est l'enzyme permettant la coagulation du lait.

L'utilisation de ces rubriques n'est pas toujours simple. Tout d'abord, l'**action typique** n'est pas nécessairement unique ; ainsi, l'agitateur-disperseur sert, comme son nom l'indique, à plusieurs actions. Mais la morphologie est trompeuse, ou réductrice, car les actions recensées sont les suivantes : *mélange, dissolution, mise en suspension, homogénéisation, aération.* Il y a là une pratique implicitée sous le terme qui ne peut se réduire à *agiter* et *disperser*. Les éléments typiques permettent donc de donner des éléments de construction du concept et de les distinguer du niveau de la motivation morphologique.

De même, le *disperseur* sert à *homogénéiser, émulsifier* et *disperser*, ce dernier verbe ne suffisant pas à lui seul pour rendre compte des utilisations de l'appareil.

La difficulté peut tenir à une lacune de métalangage. Ainsi, l'action typique de l'*antimousse* est de « détruire les bactéries ou à empêcher leur développement ». Ici manque un verbe hyperonyme permettant de neutraliser l'opposition entre « détruire » et « empêcher le développement ».

Il peut exister une corrélation entre la variété des actions typiques et des applications typiques. Ainsi, la *gomme xanthane* est utilisée dans les :

{app. typ} industries chimiques, comme
{act. typ} émulsifiant, lubrifiant

et dans les :

{app. typ} industries alimentaires, comme
{act. typ} additif.

La pluralité qui caractérise certaines actions typiques se retrouve pour les **objets typiques**. Il en est ainsi lorsque les utilisations d'un même microorganisme dans différents secteurs obligent à introduire une polysémie au sein d'une entrée unique. On est alors conduit à tourner le dos à la démarche homonymique qui caractérise souvent les travaux terminologiques. C'est ainsi que pour la bactérie lactique nous avons dû décrire quatre actions et objets typiques :

{act. typ} 1. acidification
{obj. typ} 1. fromages

{act. typ} 2. protéolyse
{obj. typ} 2. précurseurs d'arômes

{act. typ} 3. production de polysaccharides
{obj. typ} 3. laits fermentés

{act. typ} 4. fermentation malo-lactique
{obj. typ} 4. vins et cidres

Et une présentation souple comme celle-ci permet de présenter la diversité des applications en évitant la lourdeur et l'arbitraire qui conduit à multiplier les homonymes. En fait, *bactérie lactique* renvoie à une catégorie qui inclut toute une taxonomie de bactéries lactiques différentes. [Dans un dictionnaire plus ciblé, les bactéries feraient l'objet d'un traitement homonymique. La description sémantique se fait en fonction de l'univers de discours considéré.]

La notion d'**agent typique** pose des difficultés de description liées à la notion même d'agent, « être qui accomplit l'action exprimée par le verbe » (Dubois *et al.*, 1994). En fait, l'agent peut être un animé ou un inanimé, un agent direct ou indirect. L'action n'est que rarement accomplie de façon directe : si la *calmoduline* est bien, simple protéine, l'être qui accomplit l'action de *fixer les ions calcium*, cette rubrique pose souvent des difficultés quasi ontologiques, et de diverses natures, notamment dans la description d'appareils et de techniques.

Ainsi l'*électrophorèse en gel de polyacrylamide* utilise le pouvoir de séparation de ce gel. Notre grille de description nous a conduit à poser comme agent typique le « pouvoir sélectif du gel de polyacrylamide ». Pareillement, l'*électrodialyse* décrite comme suit :

{act. typ} désionisation
{obj. typ} l'eau
{ag. typ} un champ électrique

a donc comme agent typique champ électrique. La notion d'agent se trouve ici pourvue d'une extension large et assez imprévisible, mais c'est bien l'histoire des techniques qui le veut ainsi.

Quant à l'*électroélution*, qui sert à la récupération de protéines contenues dans un gel, elle utilise l'« effet du champ électrique dans la cuve d'électroélution ». La frontière entre accomplir l'action, la provoquer ou la permettre est bien mince. Mais les scrupules du descripteur sont peu de choses et tiennent, ici, principalement à un cadre grammatical trop simple. Il faut donc faire jouer les catégories pour y faire entrer le réel.

La notion d'**application typique** nous est apparue utile en cours de travail afin de situer le niveau auquel intervenait l'action décrite. Son choix est lié à une attitude initiale de prévention contre celle de domaine. C'est le fait de ne pas recourir à la description en termes de domaines et sous-domaines qui nous a conduits à ajouter cette rubrique, plus indicatrice de la fonction que du domaine proprement dit.

Cette rubrique s'est avérée utile par sa souplesse dans la mesure où elle permettait de viser le niveau pertinent, qu'il s'agisse d'une sous-application très locale ou au contraire d'un terme intéressant l'ensemble d'un domaine.

Elle s'est avérée également précieuse notamment quand les utilisations sont diverses : par exemple, la dextrane sert aussi bien dans la fabrication de résines que

dans la préparation de plasma. Il y a là des utilisations de ce polysaccharide très différentes et bien plus précises et limitées que chimie, santé ou hématologie. Et nous évitions encore l'artifice qui nous eût conduits à distinguer deux dextranes, là où il existe simplement deux conceptualisations liées à des pratiques différentes.

Ces quatre rubriques nous permettent donc de dégager des réalisations discursives différentes qui, selon nous, tiennent à des concepts différents. On en trouve illustration dans des divergences de traduction pour un signifié unique en français. Ainsi, *arôme* possède deux actions typiques : la fumaison et l'aromatisation. Et l'on retrouve cette distinction dans les traductions :

arôme de fumée = *smoke flavour*, et
arôme de transformation = *transformation aroma*.

On voit ici que les signifiés sont bien culturels et que les unités de traduction nécessitent des distinctions proches permettant de mettre en évidence le niveau conceptuel. Autre exemple patent, les divers *agitateurs* recueillis dans notre corpus appartiennent à une seule et même classe lexicale, classe qui se scinde en deux au contact de l'anglais pour lequel est pertinente l'opposition entre *stirrer* et *shaker*, selon le mode d'agitation. Mais, bien entendu, le niveau proprement conceptuel se situe à un niveau plus fin que le niveau lexical, celui dont rendent compte les catégories prédicatives.

Les relations que nous venons de décrire ont pour but de faciliter le passage de la description du signifié au concept par un élargissement d'une description purement oppositive, celle des paradigmes, à une description positive, celle des relations syntagmatiques attestées. Les relations prédicatives permettent d'aboutir à un noyau prédictif, de décrire des phrases de base élémentaires, telles que :

> « La bioabsorption permet le stockage de métaux toxiques par des micro-organismes pour la dépollution », ou
> « La chambre d'électroporation permet la fragilisation par un faible courant électrique de membranes cellulaires pour la fusion cellulaire ».

C'est assez élémentaire, certes, mais cela ne se devine pas.

Pour une approche contextuelle

Nos relations prédicatives jouent en fait le rôle de définitions minimales, par l'exploitation des renseignements contenus dans les énoncés dépouillés. On s'est ainsi rapproché de la « définition naturelle », tout en écartant les risques liés aux discours métalinguistiques. Mais le lecteur a besoin de renseignements plus nombreux afin d'esquisser la construction de la notion. Ces renseignements, il les obtient à la lecture des contextes qui constituent le troisième volet de notre description.

Les avantages des contextes sont nombreux et connus. On l'a vu, ils autorisent une description en catégories plus ouvertes. Abordant ce travail en sociolinguistes, nous souhaitions éviter le caractère artificiel de la position métalinguistique, et nos contextes sont des énoncés spontanés. Par ailleurs, la variété des énoncés recueillis

nous permettait de faire place à la variation. C'est, là encore, un argument de type sociolinguistique, et non normatif. Une option reste certes une option ; mais nous pensons que, même en se penchant sur des communautés professionnelles apparemment homogènes, il ne faut pas oublier que la variation est un facteur majeur, présent dans toute communauté de parole. Enfin, on sait l'intérêt des contextes pour permettre de rendre compte de la variété des points de vue exprimés sur la notion, le référent, le processus, etc. Ce sont là des questions qui renvoient à la problématique de la catégorisation.

De façon plus modeste, nous voudrions souligner des contextes en ce qu'ils permettent souvent une construction contrastive des concepts. Pour ce faire, ils jouent fréquemment sur la relation d'isonymie. Ainsi le contexte proposé sous *chémostat* présente ensemble le *chémostat* et le *turbidostat* :

> *chémostat*
> {ctxl} [...] **Dans un turbidostat**, la vitesse à laquelle les cellules quittent le réacteur (mesurée par la turbidité ou l'opacité du flux sortant) commande l'introduction des substances nutritives. **Dans un chémostat**, la concentration d'une substance nutritive critique, présente dans le flux entrant, est calculée de façon à contrôler la vitesse de la réaction, en limitant la prolifération des microorganismes : c'est le contrôle par ajustement préalable.

Pareillement, les contextes de *colorimètre* permettent d'opposer ce terme à *spectrophotomètre* et *photomètre* :

> *colorimètre*
> {ctxl} À la différence du spectrophotomètre, de bande passante réduite, dans le colorimètre, la sélection de la longueur d'onde se fait par filtres colorés interchangeables, et non par monochromateur [...].

La relation d'isonymie, utilisée ici, autorise des stratégies cognitives économiques et efficaces pour construire des catégories. En effet, elle permet de partir de ce qui est le plus proche au lieu de recourir au genre prochain. Et, même si cela est préférable, il n'est pas nécessairement besoin que la notion voisine soit beaucoup plus familière, car en fait, on peut arriver à construire deux notions de façon conjointe. Nous avons déjà rencontré cette stratégie dans des dictionnaires de sciences (cf. Gaudin, 1992).

Les éléments qui précèdent s'en tiennent à des problèmes de description. Il sont peu reliés aux travaux, à caractère plus théorique, évoqués en introduction. Se pose en fait à nous la question de la possibilité d'un travail terminographique qui s'inspire de positions socioterminologiques. Cette possibilité est attestée pour les travaux relatifs aux politiques linguistiques : en témoigne l'effort méthodologique développé principalement au Québec et, depuis peu, en France.

En revanche, de récents débats l'ont illustré, la sociolinguistique a été de peu de secours à la lexicographie, concernant « le contexte social et situationnel dans lequel chacun des mots peut apparaître. La sociolinguistique n'a pas de solution à proposer aux lexicographes » sur ce point, note Vachon-L'Heureux (1995 : 2). En fait, les praticiens sont un peu réduits à, par exemple, proposer l'exclusion de la marque populaire, auquel cas « la suite privilégiée serait soutenu, neutre, familier et très familier » (Vachon-L'Heureux, 1995 : 3)

Que le problème ait un niveau de pertinence, cela n'est pas douteux : le mouvement vers une normalisation des produits imposera vraisemblablement des ouvrages « lexicographiquement corrects ». Mais, faute d'assises théoriques, la description des variétés sociolectales est restée rudimentaire eu égard au nombre des études sociolinguistiques menées depuis vingt ans. Les descriptions des variations diachroniques et diatopiques se sont affinées alors que force est de constater que les progrès ont été fort minces concernant les faits diastratiques.

Cette lexicographie particulière qu'est la terminographie pour les agnostiques du concept peut-elle tirer profit de l'application de principes méthodologiques issus de la sociolinguistique ? L'exemple de la lexicographie englobante, ou générale, montre que les faits sociolinguistiques ne se laissent pas décrire à peu de frais. En revanche, les technolectes ou sociolectes professionnels, ou langues spécialisées, permettent d'étudier des communautés linguistiques aux effectifs limités. Cela permet, notamment, de travailler finement les questions d'individuation sociolinguistique, mais aussi d'étudier la circulation des vocables, en recourant à des typologies fonctionnelles assez fines.

Mais plus profondément, il reste que le but d'un dictionnaire est de satisfaire des usagers. Les changements des habitudes dans ce domaine sont difficiles et il n'est pas sûr que les insuffisances mises en évidence par la réflexion soient ressenties par les utilisateurs. Mais la période de formidable mutation technologique que connaît l'art lexicologique devrait favoriser les propositions théoriquement stimulantes.

Le lexique génératif : Une alternative au traitement de la polysémie
Le cas des adjectifs qui dénotent un état mental

Pierrette BOUILLON

ISSCO (Institut pour les Études Sémantiques et Cognitives), Université de Genève, Suisse

1. Introduction

Récemment, les travaux en sémantique lexicale ont connu des développements intéressants : motivées par des soucis de cohérence du lexique, des théories nouvelles commencent enfin à voir le jour et à exploiter des descriptions sémantiques plus expressives, ainsi que des méthodes de composition plus puissantes (voir notamment Pustejovsky, 1995 et Briscoe et Copestake, 1993).

Cet article exploite l'une de ces théories : celle du Lexique Génératif (*Generative Lexicon*, GL) (Pustejovsky, 1991 et 1995), en l'étendant au traitement de la polysémie des adjectifs français qui dénotent un état mental. L'article commence par cerner les problèmes posés par ce type d'adjectifs (section 2). Il oppose ensuite GL à l'approche lexicographique traditionnelle (section 3). Enfin, il conclut en montrant l'adéquation de GL pour le traitement des adjectifs d'état mental (sections 4 et 5).

2. La polysémie des adjectifs d'émotion et orientés-agent

Les adjectifs d'état mental qui dénotent un état émotionnel (exemples (1)) et une compétence ou un comportement (« orientés-agent », selon la dénomination de Ernst, 1984) (2) présentent un comportement polysémique remarquable, bien mis en évidence dans la littérature (notamment dans Lehrer, 1990 et Croft, 1984).

(1) *triste, irrité, déprimé, gai,* etc.

(2) *intelligent, habile, ingénieux, stupide, doué, grossier, poli, impoli,* etc.

a. Tout d'abord, la plupart d'entre eux peuvent modifier des noms de types sémantiques « humain », « objet » et « événement », sans que ce comportement ne soit généralisable à tous les adjectifs de la classe : « *doué* » et « *irrité* », par exemple, constituent des exceptions à cet égard (4).

(3)	(humain)	a. *un homme triste*	a. *un homme ingénieux*
	(objet)	b. *un livre triste*	b. *un livre ingénieux*
	(événement)	c. *un examen triste*	c. *un examen ingénieux*
(4)	(humain)	a. *un homme irrité*	a. *un homme doué*
	(objet)	b. **un livre irrité*	b. **un livre doué*
	(événement)	c. **un examen irrité*	c. **un examen doué*

b. Ensuite, ils présentent des sens différents en fonction du type sémantique modifié : quand ils sélectionnent un nom de type « humain », ils dénotent normalement l'état dans lequel se trouve l'individu (5) ;

(5) *un homme triste* → qui est dans l'état de tristesse
un homme intelligent → qui est dans l'état d'intelligence

quand, au contraire, ils modifient un événement ou un objet, ils présentent, soit un sens « causatif » (6b) (dans ce cas, l'objet ou l'événement est la cause de l'état), soit « explicatif » (6c) et (7c) (dans ce cas, l'objet ou l'événement dénote la manifestation de l'état). Certains peuvent combiner les deux sens, comme « *triste* » en (6b,c). Enfin, dans des contextes marqués, le sens « causatif » est aussi possible avec des noms de type « humain » (8).

(6) *un livre/voyage triste* →
a. *qui est dans l'état de tristesse
b. qui cause la tristesse de quelqu'un
c. où se manifeste la tristesse de quelqu'un

(7) *un livre/voyage intelligent* →
a. *qui est dans l'état d'intelligence
b. *qui cause l'intelligence de quelqu'un
c. où se manifeste l'intelligence de quelqu'un

(8) *un homme triste à voir* →
a. *qui est dans l'état de tristesse
b. qui cause la tristesse de celui qui le voit

3. Approches monomorphiques et semi-polymorphiques

Pour représenter la polysémie de ces adjectifs, deux approches opposées peuvent être envisagées : l'approche monomorphique ou celle, semi-polymorphique, du « Lexique Génératif » de Pustejovsky.

a. L'approche traditionnelle, monomorphique, consiste à créer dans le lexique une entrée différente pour chaque type sémantique dénoté par le mot ; ainsi, elle créera au moins deux entrées différentes pour chaque adjectif de (1) et (2), comme en (9) pour « *triste* » :

(9) a. *triste* (+ humain) : qui est dans un état de tristesse
b. *triste* (+ humain/objet/événement) : qui cause un état de tristesse
c. *triste* (+ objet/événement) : où se manifeste un état de tristesse

Cette approche pose évidemment un certain nombre de problèmes. Tout d'abord, elle postule qu'il est possible d'énumérer toutes les interprétations des mots, ce qui ne rend pas compte de la création des sens en contexte. Certains adjectifs changent en effet de sens en fonction du nom modifié, comme « *rapide* » qui a un sens différent dans « *une dactylographe rapide* » (qui exécute avec promptitude), « *un pas rapide* » (qui s'accomplit rapidement), « *un cheval rapide* » (qui se meut rapidement), « *un moyen rapide* » (qui conduit vite au but escompté) ou « *une piste rapide* » (qui permet de hautes performances). Comment, dans ce cas, énumérer tous les sens du mot ? Ensuite, elle ne rend pas compte du fait que certaines ambiguïtés sont régulières, comme celles des adjectifs d'état mental. En troisième lieu, elle considère les entrées comme indépendantes l'une de l'autre et n'explicite pas le lien qui existe entre certaines d'entre elles, comme par exemple entre les deux sens de « *triste* », l'état de tristesse et sa cause. Enfin, elle crée une ambiguïté, difficile à résoudre dans la suite du traitement. Si on crée, par exemple, deux entrées pour « *cuire* », pour distinguer les deux interprétations de (10), il sera très difficile de faire dans la suite la bonne sélection puisque les deux sens sont très proches et se chevauchent.

(10) a. *Jean cuit les pommes de terre* (changement d'état, sans création)
b. *Jean nous a cuit un cake* (création)

b. L'approche semi-polymorphique de Pustejovsky (Pustejovsky, 1991 et 1995) s'oppose à une énumération de types sémantiques. Elle propose de donner à chaque mot une riche représentation, qui constitue une réserve de types, en indiquant pour chacun d'eux les différentes extensions de sens possibles. Pour se faire, elle rompt avec l'approche monomorphique de deux manières : tout d'abord, elle incorpore les différents types sémantiques dénotés par le mot dans une méta-entrée, qui indique les relations qu'ils entretiennent entre eux. Ensuite, elle définit des mécanismes génératifs capables d'opérer sur la méta-entrée, pour changer éventuellement sa dénotation en contexte. La suite de ce chapitre se concentre sur la méta-entrée et les mécanismes génératifs, puis la représentation de la polysémie nominale en GL.

3.1. La méta-entrée dans le GL

Dans GL, tous les mots sont représentés dans le lexique avec une structure complexe. Celle-ci peut combiner jusqu'à trois niveaux de représentation, connectés entre eux dans un réseau hiérarchique (Pustejovsky, 1991 et 1995).

a. La structure argumentale définit le nombre et le type d'arguments du mot. « *Construire* », par exemple recevra trois arguments (11) : un sujet de type « animé » (« arg1 ») et un objet « artefact » («arg2 »), ainsi qu'un troisième argument (« D-arg1 ») de type « matériel » qui diffère des deux autres puisqu'il ne doit pas nécessairement être exprimé au niveau syntaxique (11) (« *default argument* » D-ARG) ;

(11) argstr = arg1 = animé
arg2 = artefact
D-arg1 = matériel

b. La structure événementielle spécifie le type d'événement d'un mot ou d'une phrase (état, procès ou transition), ainsi que sa structure interne ; « *construire* », par exemple, sera défini comme une transition, composée de deux sous-événements : un procès initial (l'acte de construire) (« E1 ») et un état résultatif (12) (celui de l'objet construit) (« E2 »). La tête (« *head* ») indique le sous-événement le plus important dans la structure, « E 1» dans le cas de « *construire* » ; « E2 » dans celui de « *mourir* ».

(12) evenstr = E1 = procès
E2 = état
head = E1/E2[1]

c. La structure des qualia décrit les caractéristiques sémantiques du mot, en indiquant comment les événements et les arguments sont sémantiquement liés. Pour ce faire, elle utilise quatre rôles (13), définis comme suit : formel (quelle est sa catégorie sémantique ?), constitutif (quels sont ses éléments constitutifs ?), agentif (comment il est créé ?) et télique (quelle est sa fonction ?)

(13) qualia = const = ...
form = ...
telic = ...
agentif = ...

« *Couteau* » (14), par exemple, dénote un artefact. Il a un télique : sa fonction est de « *couper* », ainsi qu'un agentif : il est créé par un être humain.

(14) couteau
argstr = arg1 = x :artefact
qualia = artefact-lcp
form = x
telic = couper(e1,x,y)
agentif = créer(e2,z,x)

Dans la méta-entrée, il est important de distinguer les types de base d'un item (qui définissent son « paradigme lexical conceptuel » LCP), des autres types qui ne sont accessibles que par des opérations génératives. « *Couteau* », par exemple, en (14), dénote un « artefact », mais d'autres types, comme les événements « e1 » et « e2 », pourront être accessibles par une des opérations génératives suivantes (Pustejovsky, 1995 : chap. 7).

a. La coercition de types permet de remplacer le type sémantique d'un item lexical par celui qui est requis par le prédicat, pour autant que ce dernier soit présent dans la qualia de l'item. « *Commencer* », par exemple, sélectionne un objet de type « événement ». Dans « *cette usine vient de commencer les couteaux de luxe* », « *couteau* », de type artefact, peut prendre le type « événement » requis par « *commencer* », puisque sa structure des qualia contient des événements : « *cette usine vient de commencer les couteaux de luxe* » signifie entre autres « *cette usine vient de commencer à fabriquer des couteaux de luxe* » (voir Pustejovky et Bouillon, 1995, pour plus de détails).

1 La barre oblique «/» indique une disjonction

b. Le liage sélectif permet à un item lexical d'opérer sur un type qui ne se trouve pas dans le paradigme lexical conceptuel de l'item qu'il modifie. Dans « *une voiture rapide* », par exemple, « *rapide* » sous-modifie l'événement défini dans le télique du nom, mais sans changer sa dénotation, comme en (15) : « *une voiture rapide* » signifie « *une voiture qui roule rapidement* » (Pustejovky et Bouillon, 1995).

(15) lambda x [... Telic(x) = lambda e [rouler(x)(e) et rapide(e)]

c. La cocomposition est une fonction bilatérale qui permet à un argument de changer la sémantique du prédicat, en composition. Elle n'intervient pas dans le traitement des adjectifs d'état mental et nous ne nous y attarderons pas ici.

3.3. Traitement de la polysémie nominale en GL

GL distingue au moins deux niveaux polysémiques différents.

a. Les noms non-polysémiques dénotent un type simple, comme « *couteau* » en (14). Ces derniers ont un seul argument (« arg1 »), qui correspond directement au formel (« x ») (16).

(16) alpha
argstr = arg1 = x:t
qualia = t-lcp
form = x
const = ...
telic = ...
agentif = ...

b. Les noms polysémiques font référence à plusieurs types en même temps. Ils dénotent un type complexe (« pointé ») qui correspond au produit cartésien des types définis dans la structure argumentale (noté t1.t2) (17). Dans ce cas, le mot a deux arguments (« arg1 » et « arg2 »), dont la relation est définie au niveau du formel (ici relation P). Il dénote trois types de base différents, qui peuvent être projetés indépendamment : le type pointé « t1.t2 » et les deux sous-types « t1 » et « t2 » (17).

(17) alpha
argstr = arg1 = x:t1
arg2 = y:t2
qualia = t1.t2-lcp
form = P(x,y)
const = ...
telic = ...
agentif = ...

Le « *livre* » appartient à cette catégorie (18). Il dénote le type complexe « contenu-imprimé » (« info.obj-phys ») qui est le produit des deux types « info » et « obj-phys ». Il présente donc deux arguments, dont le formel indique la relation : « y » contient « x ». Sa fonction est d'être lu et il est créé par l'acte d'écriture. À partir de cette méta-entrée, il sera possible de projeter indépendamment les trois types de base : le type pointé (« contenu-imprimé ») et les deux sous-types « info » et « obj-phys » (19a,b,c).

Les deux types « e1 » et « e2 » (événement) pourront aussi être accessibles par des mécanismes génératifs, comme la coercition ou le liage sélectif (19d,e) (Pustejovsky et Bouillon, 1995).

(18) livre
argstr = arg1 = x:info
arg2 = y:obj-phys
qualia = contenu-imprimé_lcp
form = contenir(y,x)
const = ...
telic = lire(e1,w,x.y)
agentif = écrire(e2,v,x.y)

(19) a. *je lis un livre de Proust* (« contenu-imprimé »)
b. *je ne m'intéresse pas à la couverture, mais au livre* (« info »)
c. *je ne m'intéresse pas au contenu, mais au livre* (« obj-phys »)
d. *je commence un nouveau livre* (« événement »)
e. *c'est un long livre* (« événement »)

La thèse que nous voudrions défendre dans le chapitre suivant est que les adjectifs d'état mental dénotent, eux aussi, des types complexes. Ces derniers expliquent le comportement polysémique esquissé dans la section 2.

4. Traitement des adjectifs d'émotion et orientés-agent dans GL

4.1. Proposition générale

Nous proposons le traitement global suivant :

a) distinguer deux types d'adjectifs d'état mental par leur structure de qualia[2] : les adjectifs causatifs (d'émotion) (exemples (1)) et orientés-agent (2) ;
b) représenter la polysémie de ces adjectifs au niveau de leur structure des qualia, par des types complexes ;
c) expliquer les différences de sélection à l'intérieur des classes par la notion de projection et de tête.

Les deux premiers points seront l'objet de la section 4.1.1, le troisième de 4.2.2., 4.2.3 examinera ensuite plus en détail le comportement des adjectifs causatifs.

4.1.1. Deux sortes d'adjectifs, avec type complexe

Les adjectifs de (1) et (2) recevront respectivement les représentations (20) et (21). Ceux qui peuvent appartenir aux deux classes (voir section 2) seront codés avec la structure (22) qui n'est pas spécifiée quand au second rôle : télique ou agentif.

2 Ces deux types d'adjectifs correspondent aux classes 3 et 1 de Picabia, 1978

(20) adj_caus
eventstr = E1 = e1:état
E2 = e2:AGENTIVE_INTELLECTUAL-évén
head = e1/e2/e1,e2
argstr = arg1 = x:humain
D-arg1 = e2
qualia = e1.e2_lcp
form = Adj(e1,x)
agentive = P(e2,x,...)

(21) adj_agent-orienté
eventstr = E1 = e1:état
E2 = e2:AGENTIVE_INTELLECTUAL-évén
head = e1/e2/e1,e2
argstr = arg1 = x:humain
D-arg1 = e2
qualia = e1.e2_lcp
form = Adj(e1,x)
telic = P(e2,x,...)

(22) adj_causatif-agent-orienté
eventstr = E1 = e1:état
E2 = e2:..._EXPERIENCER-évén/..._INTELLECTUAL-évén
head = e1/e2/e1,e2
argstr = arg1 = x:humain
D-arg1 = e2
qualia = e1.e2_lcp
form = Adj(e1,x)
agentif/telic = P(e2,x,...)

Ces structures indiquent :

a. que ces adjectifs exigent deux arguments : un de type « humain » (« arg1 ») et, un second (« D-arg1 »), facultatif au niveau syntaxique, événementiel, de type « perceptif » actif ou statif[3] pour les adjectifs causatifs et qui dénote un « acte intellectuel » pour les adjectifs orientés-agent (voir Croft, 1984 : 21, pour une proposition similaire) ;

b. qu'ils font référence à deux types sémantiques différents (« e1.e2 »), dont la structure des qualia indique la relation (tout comme pour « *livre* ») :

> i. un état résultatif (« e1 ») (encodé dans le formel) (état de tristesse, d'intelligence, d'ingéniosité, etc.) ;
> ii. un événement (« e2 »). Encodé dans le rôle agentif (20), il dénote alors la cause de l'état émotionnel et doit être de type perceptif, statif ou actif : comme Croft (1990), nous considérons en effet qu'il y a deux processus impliqués dans un état émotionnel : l'expérienceur porte son attention, intentionnellement ou non, sur un stimulus, qui suscite ensuite un état mental ; dans le rôle télique au contraire (21), l'événement dénote la manifestation de l'état.

3 Sur cette distinction, voir Lehrer, 1990.

En d'autres termes, (20) signifie que quelqu'un (« x ») est dans un certain état (« e1 ») à cause d'un événement (« e2 »); (21) que l'individu (« x ») est dans un état (« e1 ») dont l'événement (« e2 ») est la manifestation.

4.1.2. La notion de tête (« head »)

Tous les adjectifs ne vont cependant pas pouvoir projeter indépendamment les deux types qu'ils dénotent, en fonction de leur tête (voir chapitre 3.1.), qui va jouer le rôle de filtre au niveau de l'interface syntaxe/sémantique (Pustejovsky, 1995 : chap. 6). En (20) et (21), il apparaît que trois types de configuration sont possibles, qui déterminent autant de projections différentes.

a) Tête sur « e1 » (« head = e1 ») : l'adjectif est projeté via le formel et il en résulte un sens « statif » (exemples (23) et (24)) ; dans ce cas, l'adjectif a un seul argument obligatoire « *x* » de type « humain ». Un second, de type « événementiel », est cependant possible s'il fait référence à l'agentif (23) ou au télique (24). La différence au niveau du type des événements suffit à expliquer des comportements spécifiques au niveau de la complémentation des adjectifs (1) et (2) :

> i. d'une part, les adjectifs causatifs partagent, avec les autres prédicats « émotifs » (« *regretter* », etc.), la particularité de pouvoir apparaître avec un complément subjonctif ou infinitif, en distribution complémentaire (23) (« obviative phenomenon ») (Kiparsky et Kiparsky, 1979 : 169 et Rochette, 1988 : 251) : dans ce cas, l'infinitive sature directement le prédicat « P(e2,x,...) » de l'agentif, tandis que la complétive au subjonctif est interprétée comme l'objet (« y ») d'un événement perceptif sous-entendu (en (23b), je suis triste suite à « *la perception de ton départ* » « Exp(e,je,que-tu partes) »).
>
> (23) a. *je suis triste de partir*
> b. *je suis triste que tu partes*
> c. **je suis triste que je parte*
>
> ii. D'autre part, la plupart des adjectifs orientés-agents permettent, comme second argument, une infinitive introduite par « *à* » (A-VP) (24b) ou « *de* » (24a) (DE-VP), le A-VP saturant directement « e2 », tandis que le DE-VP étant considéré comme l'objet (« y ») de l'acte intellectuel (en (24b), « je suis ingénieux de prendre l'initiative de partir »).
>
> (24) a. *je suis ingénieux de partir*
> b. *je suis ingénieux à jouer aux échecs*
> c. **je suis ingénieux que tu partes*
> d. **je suis ingénieux que je parte*

b) Tête sur « e2 » (« head = e2 ») : l'adjectif est projeté via l'agentif ou le télique et il en résulte un sens « causatif » ou « explicatif » (25) ; dans ce cas, l'argument du nom est nécessairement de type événementiel (25a) ou bien est interprété comme tel au niveau sémantique (25b,c).

(25) a. *une lecture triste* → une lecture qui cause la tristesse du lecteur

b. *un livre triste* → un livre dont la lecture cause la tristesse du lecteur
c. *un examen intelligent* → un examen dont la conception témoigne de l'intelligence de la personne qui l'a conçu

c) Pas de spécification de la tête (« head = e1, e2 ») : l'adjectif est projeté via le formel ou le télique/agentif, ce qui explique son ambiguïté.

La suite de cet article donne quelques exemples pour chacune de ces configurations.

4.1.3. De quelques exemples

Voyons d'abord plus en détail le cas de l'adjectif « *triste* », dans son sens causatif uniquement. Comme il n'est pas spécifié quant à la tête (26), il peut être projeté via le formel ou l'agentif.

(26) triste

eventstr =	E1 = e1:état
	E2 = e2:STATIF_ACTIF_EXPERIENCER-évén
	head = e1, e2
argstr =	arg1 = x:humain
	D-arg1 = e2
qualia =	exp-causative_lcp
	form = triste(e1,x)
	agentive = P(e2,x,...)

a. Quand le formel (« e1 ») reçoit la tête, la phrase exprime l'état de tristesse d'un individu, comme en (27a). Dans cet exemple, la cause de l'état n'est pas spécifiée : Jean est triste pour une raison non déterminée. Elle peut cependant être rendue explicite, comme en (27b,c) où le complément sature l'agentif.

(27) a. *Jean est triste*
b. *Jean est triste de lire ce livre*
c. *Jean est triste que je lise ce livre*

b. Quand l'agentif (« e2 ») reçoit la tête, la phrase exprime la cause de l'état de tristesse d'un individu, comme en (28). Dans ce cas, le nom peut être :

i. un nom événementiel, comme « *lecture* » en (28a) ;
ii. d'autres types de noms, dont la structure des qualia indique une relation directe avec un événement de perception, comme « *livre* », en (28b) (dont la qualia contient les événements « *lire* » et « *écrire* » (18)). Dans ce cas, la perception est intentionnelle et contrôlée ;
iii. d'autres types de noms, dont le type permet la mise en relation avec un événement perceptif statif générique, comme « *percevoir par ses sens* » (28c). Ici, au contraire, la perception n'est ni intentionnelle, ni contrôlée.

(28) a. *une lecture triste*
b. *un livre triste* → à lire ou à écrire
c. *un sapin triste* → à percevoir pas ses sens

En cas d'ambiguïté entre les deux lectures (stative et causative), la syntaxe permet de distinguer les deux sens : en (29a), il s'agit de la place de l'adjectif et, en (29b), du choix de la préposition « *à* » (et non « *de* », comme en (27b)).

(29) a. *de tristes enfants* → à percevoir par ses sens
b. *un homme triste à voir*

Enfin, (30a,b) est impossible, puisque deux événements essayent de saturer le rôle agentif (« *lecture du livre* » et « *voir le film* » en (30a)). (30c), au contraire, est possible puisque « *enfant* » et « *mère* » saturent des variables différentes (« x » et « y »), comme explicité en (31).

(30) a. **un livre triste à cause du film qu'il a vu*
b. **de tristes enfants de la mort de leur mère*
c. *des enfants tristes de la mort de leur mère*

(31) qualia = exp-causative_lcp
form = triste(e1,enfants)
agentive = Exp(e2,enfants,mort_de_leur-mère)

D'autres adjectifs causatifs vont être spécifiés quant à la tête, ce qui explique un comportement différent de « *triste* » : les participes présents et la plupart des adjectifs en « *-able* » et « *-ible* », qui dérivent d'un verbe psychologique, auront la tête sur l'agentif ; les participes passés l'auront sur le formel. Cette restriction explique pourquoi certains adjectifs causatifs (1) ne peuvent modifier des noms de type « contenu-imprimé » (32), ainsi que le sens exclusivement causatif des adjectifs en « *-ant* » et « *-able* » (33).

(32) **un livre ennuyé*

(33) a. *un homme ennuyant* → qui cause de l'ennui
b. *un livre agréable* → qui cause du plaisir

5. Conclusion

Dans cet article, nous avons étendu la théorie du Lexique Génératif au traitement des adjectifs d'état mental français. Nous avons montré ainsi l'élégance de l'approche semi-polymorphique, pour :

a. éviter la prolifération des entrées : les différents sens de (1) et (2) sont intégrés dans une même méta-entrée ;
b. expliquer les liens entre les différents sens : la structure des qualia rend explicites les liens entre les différents types sémantiques de (1) et (2) – en (1), l'événement cause l'état ; en (2), il en est la manifestation ;
c. faire découler le comportement polysémique de (1) et (2) de leur représentation sémantique : les différences explicitées dans le chapitre (2) s'expliquent par la structure des qualia et la manière de la projeter ;
d. distinguer deux sortes d'adjectifs, en fonction de leur dénotation : type simple ou complexe. Cette distinction reprend l'opposition traditionnelle entre adjectifs

« statiques » et « dynamiques » (dénotant des qualités sujettes au contrôle du possesseur) (Quirk *et al.*, 1994 : 434, par exemple), pour la caractériser et la représenter sémantiquement : un adjectif est dynamique s'il fait référence soit à ce qui crée l'état ou à sa manifestation ultérieure.

Remerciements

Nous remercions vivement James Pustejovsky et Laurence Danlos pour leurs commentaires sur une version antérieure de cet article.

Quand l'informatique tutoie le dictionnaire des difficultés de la langue française...

Daniel Blampain

Institut supérieur des traducteurs et interprètes (ISTI), Bruxelles, Belgique

Quand on consacre une partie importante de son temps à analyser les difficultés de la langue française et à les présenter au plus large public sous la forme d'un dictionnaire (Hanse et Blampain, 1994), s'il est bien une expression qui, appliquée à la langue française, fascine, c'est *se jouer des difficultés*, expression où le sens ludique du verbe l'emporte sur le sens d'effort et de peine inscrit dans la mémoire sémantique du substantif.

Jouant constamment des deux points de vue, celui du linguiste et celui de l'usager, je m'efforce d'isoler et de décrire les difficultés afin de répondre au mieux à l'attente de l'usager... qui se trouve en difficulté.

Les difficultés sont nombreuses. Un choix pertinent de celles-ci, un traitement qui conjugue l'analyse rigoureuse et la communication agréable déterminent la qualité et l'accueil de l'ouvrage. Difficultés liées à l'oral ou à l'écrit, difficultés phonétiques et orthographiques, grammaticales et lexicales, difficultés liées aux registres de langue ou aux français régionaux qui sont les joies de la francophonie, toutes sont traitées au sein d'articles qui occupent des espaces variables, proportionnels à la complexité des cas posés et qui déclinent tout sur le mode d'exemples multiples, au point que le critère de qualité est l'utilisation satisfaisante de l'ouvrage aussi bien par des francophones que par des allophones.

Et quand l'informatique, au service de l'usager de la langue française, propose son aide en la matière... qu'en est-il de ce trait d'union placé entre deux outils ? Qu'en est-il de cette familiarité naissante entre le dictionnaire des difficultés et ce que, dans le cadre des industries de la langue, on appelle de manière austère les « correcteurs grammaticaux » ? Qu'en est-il de cette dictionnairique qui tente d'unir le lexique à la syntaxe ? À une époque où plus de 70 % du travail porte sur l'écrit, sur le traitement et l'échange de messages d'information, où, dans les secteurs spécialisés de la presse et de l'édition, il revient de plus en plus à chacun d'assumer la correction de son texte,

le métier de correcteur disparaissant en raison du coût de la révision (25 % de la saisie), l'utilisateur de l'informatique accueille avec beaucoup d'intérêt les outils d'assistance mis à sa disposition pour assurer la qualité de ses textes.

Je voudrais d'emblée limiter mon propos au correcteur dit de la 3e génération, disponible sur ordinateur personnel courant. On désigne par là le correcteur qui repose sur un analyseur syntaxique, à la différence des correcteurs locaux qui travaillent sur des listes de mots et des accords grammaticaux localisés. Ma réflexion et mon expérimentation, qui ne se confond pas avec les tests traditionnels portant sur la vitesse d'exécution ou sur l'étendue du dictionnaire, ont principalement porté sur le *Correcteur 101* (logiciel Machina Sapiens - Copyright 1992-1994. Montréal), reconnu par l'ensemble des revues spécialisées pour ses performances exceptionnelles. Il ne s'agira pas non plus de comparer l'imprimé et l'informatisé. Les partages sont évidents. Le dictionnaire des difficultés traite de la langue orale, décrit la langue sémantiquement et diachroniquement, réduit l'arbitraire par l'ampleur de l'analyse et par la présentation de la diversité des usages. Il est en outre un dictionnaire d'auteur, c'est-à-dire qui a son style.

Notre objectif est plutôt de cerner la problématique déployée par le correcteur grammatical dans son évolution, sur le terrain commun des difficultés grammaticales, tant il est vrai que le « grammairien sous Windows », comme on l'a parfois appelé, peut nous aider à améliorer les algorithmes de reconnaissance syntaxique et la description de la combinatoire. Dans le domaine des industries de la langue, on a aujourd'hui compris qu'il est préférable de penser en termes de petits progrès répondant à des besoins précis plutôt qu'en termes de compétence linguistique générale. De son côté, le linguiste a bien perçu l'intérêt qu'il y a à mener le plus loin possible la description des phénomènes linguistiques de manière à rendre celle-ci opératoire pour la machine.

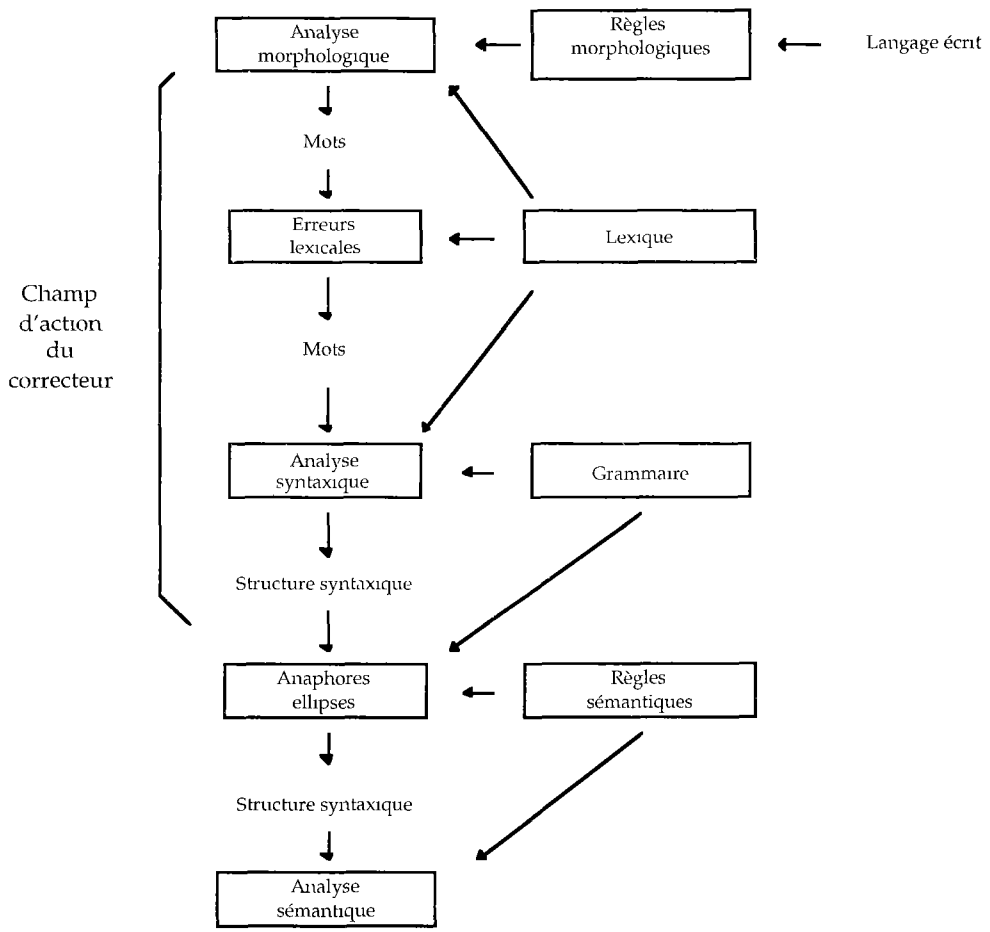
Si je me réfère au 101 (Nouvelle version 2.0), les caractéristiques suivantes sont à prendre en considération. Né de l'intention de développer un outil d'apprentissage de la grammaire par ordinateur, il se présente aujourd'hui comme un outil destiné aux « professionnels de la langue » disposant d'un ordinateur compatible IBM de type 386 ou 486. Indépendant des traitements de texte, il fonctionne sous Windows et reconnaît le format d'un nombre important de documents issus des traitements de texte les plus connus. Il occupe 2 Mo sur disque rigide. Le dictionnaire issu du *Petit Robert*, du *Petit Larousse* et du *Multidictionnaire* compte 66 000 entrées ou lemmes, soit plus de 500 000 formes déclinées. La grammaire repose sur l'actualisation de 3 500 « règles ». Plus de 1 000 règles ont été ajoutées dans la deuxième version. Deux modes de correction sont proposés : autonome et interactif, ce dernier permettant de mener l'analyse phrase par phrase. Par un lien hypertexte on accède à l'explication de l'erreur et de la correction suggérée.

Nous adopterons successivement le point de vue du linguiste et de l'usager, conformément à l'attitude définie plus haut. Le linguiste s'intéressera au dictionnaire et à la syntaxe.

Le dictionnaire

La mission du correcteur, qui met en œuvre étiqueteur et lemmatiseur, demeure avant

tout orthographique. À ce titre, il occulte le joyeux flottement de l'orthographe française, qui du reste n'apparaît que lorsqu'on ouvre côte à côte plusieurs dictionnaires. Le problème de l'« orthographe d'usage », c'est-à-dire qui concerne les mots en eux-mêmes se trouve d'autant mieux résolu qu'il est permis d'ajouter des mots, de créer ses propres dictionnaires, hiérarchisés s'il s'agit d'une mise en réseau. L'autre orthographe, appelée traditionnellement « orthographe de règle », a été l'objet de toutes les préoccupations au cours de ces dernières années. Il s'agit de maîtriser le problème du transfert des catégories morphologiques (genre, nombre, personne) associées à des classes grammaticales (nom, pronom) sur d'autres classes grammaticales telles que le déterminant, l'adjectif, le verbe. Les mécanismes de l'accord se répartissent en trois types : à l'intérieur du syntagme nominal, dans le cadre de la phrase et au-delà des limites de la phrase. Nous sommes aujourd'hui au niveau 2.



Pour en arriver là, il a fallu bien sûr répondre à l'impératif d'occuper un minimum d'espace sur le disque tout en visant l'exhaustivité du dictionnaire. On subdivise donc les dictionnaires, on utilise des codes indiquant classes et catégories grammaticales pour éviter de placer toutes les formes dérivées dans le dictionnaire lui-

même, le programme de correction se chargeant de les retrouver. Si l'homonymie se trouve aujourd'hui signalée dans un grand nombre de cas, la phrase « *Un chèque sans provision, c'est rare* » est déclarée tout aussi correcte qu'« *Un cheik sans provision, c'est rare* ». Le syntagme « *les sports divers* » est tout aussi acceptable que « *les sports d'hiver* ». Dès que le sens régit l'orthographe de l'unité syntagmatique, le travail du correcteur s'arrête. « *Une chêne de TV* » sera d'abord rectifié par rapport au genre de « chêne », masculin. Mais faut-il prendre en considération le fait lui-même ou sa relativité par rapport à la fréquence d'une telle proposition ?

On l'a compris, le *dictionnaire* reste la base de l'analyseur. La procédure d'étiquetage des mots dans la phrase correspond d'abord à la consultation du dictionnaire qui permet d'associer aux mots de la phrase analysée des informations contenues dans le lexique. Si un mot est inconnu, l'analyse se bloque. La technique de l'intégration dans le dictionnaire qui est à mettre en œuvre répondra aux demandes les plus fondamentales de l'analyseur. On précisera donc si le mot ajouté est un nom, un verbe, un adjectif en indiquant le genre et le pluriel du nom, la conjugaison du verbe ou le féminin et le pluriel de l'adjectif, c'est-à-dire que les critères morphologiques sont exclusivement d'ordre flexionnel. Cette fenêtre permet par un système de notes d'indiquer une définition, une variante orthographique, un paronyme, un homonyme, une abréviation...

Trois remarques sont à faire. La classe des affixes dérivationnels, c'est-à-dire la morphologie lexicale ou « syntaxe interne » des mots, n'est pas prise en considération, alors que l'on pourrait tirer beaucoup de l'étude de la suffixation par rapport au genre ou aux changements de classe grammaticale pour l'analyse de la trame formelle et plus largement pour explorer les voies de la néologisation. On observera en outre que les listes de mots de liaison sont considérées comme closes. Enfin, la constitution de lexiques du français régional ou d'anglicismes est aisée à condition de se limiter à l'unité lexicale simple, les syntagmes n'ayant pas encore le droit d'entrée. Ceci constitue un problème majeur par rapport aux langues de spécialité.

La précision de l'identification dans la phrase est ainsi liée à la précision du dictionnaire. On doit se prononcer sur le fonctionnement syntaxique du mot intégré, sans passer par la sémantique. De là des options du type suivant pour le nom : « *Aquaculturiste* » - *Cette personne est* — ; *Cette personne est un* — ; *Je veux de l'*—. Les différents types de transitivité d'un verbe à intégrer sont posés selon les schémas de fonctionnement : *Elle* — *bien*; *Elle* — *quelqu'un; Elle* — *à quelqu'un; Elle* — *de quelqu'un; Elle* — *heureuse* (verbe attributif).

On est évidemment loin des descriptions des relations lexique-syntaxe établies par M. Gross ou I. Mel'čuk. Non seulement la sémantique n'est pas convoquée ici, mais la complexité des constructions verbales transitives indirectes n'est pas encore prise en considération. « *Les enfants jouent tennis* » est admis comme « *Les enfants jouent leur avenir* »; « *Il prend soin à lui* » est admis comme « *Il prend tout à son frère* ». L'ensemble des restrictions combinatoires est loin d'être en jeu. Le grammaticalement correct correspond au grammaticalement conforme aux règles choisies, qui ne constituent pas à elles seules l'*acceptable* tel que l'entend le locuteur dont le français est la langue maternelle.

Du dictionnaire à la syntaxe

Le découpage syntaxique se fait donc sur la base de l'unité mot, mot orthographique identifié par rapport à sa classe grammaticale, à ses catégories et à sa fonction, c'est-à-dire à sa définition en termes relationnels selon des désignations conformes au *Bon Usage* de Grévisse-Goosse.

L'unité mot apparaît donc comme l'unité formellement saisissable et analysable par rapport à l'unité phrase. C'est dire que la procédure distributionnelle qui consiste à déterminer la classe grammaticale par le type de fonction remplie n'est pas d'application ici. C'est dire aussi que deux voies sont encore à explorer : celle de l'organisation de l'unité morphématique, qui mettrait en évidence le pouvoir de néologisation de la langue, et celle de l'unité syntagmatique qui, à mon avis, permettrait d'aller plus loin dans la maîtrise des phrases complexes, tant il est vrai que le modèle syntaxique ne génère pas des phrases mais des suites de syntagmes.

Puisqu'aucune homologie directe n'existe entre le niveau morpho-syntaxique et le niveau « sémantico-référentiel », pour reprendre les termes de C. Hagège, on exploite ici fondamentalement la solidarité morphologique : un verbe conjugué ne peut se concevoir sans une marque de temps et de personne ; tout déterminant du nom sera au masculin ou au féminin, au singulier ou au pluriel ; l'ordre d'apparition des morphèmes obéira à une... fixité précise : *grand* + *e* et *non* **e* + *grand*, *mang* + *era* et *non* **era* + *mang*.

Mais le système grammatical actualisé par l'analyseur n'est pas sans poser quelques problèmes. Dans le cas du 101, il est précisé que le logiciel est en constante amélioration et que certaines constructions n'ont pas encore été analysées.

1. On sait que la fonction d'un élément n'est pas directement déterminée par sa classe grammaticale. Là se trouve toute la différence entre la physiologie et l'anatomie, disait Todorov. De là, toutes les anomalies entre le dictionnaire et l'analyse syntaxique.

Mais on sait aussi combien il est difficile d'articuler les facteurs qui président à la fonction : le critère positionnel (p. ex. l'adj. épithète), le critère morphologique (accord avec l'élément régisseur ou la forme du mot : les pronoms *je*, *me*, *moi* indiquant eux-mêmes la fonction) ou encore les critères de classe grammaticale.

2. Une ponctuation correcte et courante est nécessaire au bon fonctionnement de l'analyseur. Le point détermine chaque étape de l'analyse et limite celle-ci (unité phrase). Le problème de l'anaphore (ou référence à un élément présenté antérieurement) n'est pas résolu, à l'intérieur de la phrase et *a fortiori* dans la relation transphrastique. « *Pierre connaît ma maison, mais pas le tien* » sera accepté. Particulièrement fréquente dans le cas des pronoms, elle ne pose plus de problèmes que dans le cas du pronom relatif. L'explication se trouve dans la proximité immédiate de l'antécédent. Toutefois, l'antécédent complexe remet le sémantique à l'avant-plan et nécessite le recours à des options.

Il est vrai que l'on retrouve ici la vieille question : l'anaphore, qui joue un rôle important dans les accords fait-elle partie des phénomènes syntaxiques ou séman-

tiques ? Dans le même sens, d'autres analyses... ne se font pas. Le problème de la coréférentialité en est un exemple. « *Je veux que je fasse ceci* » est-elle une phrase grammaticale ? Aucune erreur ne sera détectée par l'analyseur. Il s'agit pourtant là de progrès réalisables. Certes rares seront les locuteurs qui avanceront ce type de phrase. Le cas de l'ordre des pronoms personnels compléments représente un autre exemple d'amélioration souhaitable et faisable. « **On lui la donne* » est déclaré de structure inconnue, mais *lui* est analysé comme pronom sujet. Or *on* est toujours sujet. Les problèmes d'analyse ne sont pas dans ce cas liés à l'orthographe. L'articulation du grammaticalement correct et de l'orthographiquement correct doit encore être approfondie : « *Lui être intelligent beaucoup* », aucune erreur détectée.

3. La phrase complexe.

Le degré de complexité de la phrase, lié à la subordination et l'insertion, a toujours limité les performances des analyseurs. De nets progrès ont été réalisés. Une phrase extraite du programme de ce colloque (« *On s'interrogera enfin sur l'avenir du dictionnaire, sous quelque forme que ce soit, en s'intéressant à l'apparition de nouveaux outils d'aide à la compréhension (et non à la traduction) avec recours à des interfaces dans la langue maternelle de l'utilisateur* ») peut amener le message pudique suivant : « Je manque de temps ». Rien ne change, si l'on attend. L'analyse fonctionne si l'on supprime le complément entre virgules « *sous quelque forme que ce soit* » et le syntagme entre parenthèses « *et non à la traduction* ». Toutefois dans les cas complexes, des analyses partielles et des options peuvent être proposées. Il apparaît d'ailleurs que les langues de spécialité, où les possibilités combinatoires sont plus restreintes dans les syntagmes et où l'acte illocutoire détermine les structures phrastiques, offrent un terrain privilégié à l'expérimentation d'analyseurs. La condition d'amélioration initiale est de pouvoir intégrer les syntagmes générés par les domaines.

L'utilisateur

S'il est vrai que dans ce genre de logiciel, l'interface est très importante, il faut indépendamment de cette convivialité prendre en considération les caractéristiques de l'habitus (au sens de Bourdieu) que le développement d'un correcteur grammatical peut installer chez l'usager.

1. Le logiciel n'intervient pas dans l'opération d'encodage, à la différence du dictionnaire classique. Son utilisation suppose la systématisation de la production du texte brut, tel qu'on peut le voir circuler sur Internet aujourd'hui.

2. Le logiciel ne corrige pas tout et n'analyse pas tout. Ses limites sont parfois avouées, parfois dissimulées sous la convivialité de l'interface (*Je manque de temps*). Les indices de lisibilité, qui accompagnent certains correcteurs, font partie de ces écrans qui mettent en avant une pseudo-pragmatique occultant l'analyse réelle. Négligeant la complexité grammaticale et sémantique, ils se réfèrent à la longueur moyenne des phrases et du texte et proposent des résultats aberrants sur la base de l'indice de lisibilité de Flesch, sévèrement critiqué il y a déjà 25 ans par G. De Landsheere de l'Université de Liège.

En réalité, le maintien de 4 ou 5 % d'erreurs doit-il être analysé dans l'absolu du

linguiste ou selon le type de document, voire selon le type de destinataire ? Plus de 50 % des erreurs corrigées ne sont jamais faites par des étudiants universitaires francophones, mais 68 % d'entre eux ne corrigent pas **tout ceux*, **bien que je cours* (40 %), **les corrections qu'effectuent ce logiciel* (40 %), erreurs que l'ordinateur corrige. Il est peut-être moins important de soumettre les phrases compliquées que de procéder au préalable à un étalonnage des erreurs dans la pratique sociale ! Et de ce point de vue, on ne peut ignorer la chance que peut représenter un correcteur grammatical pour défendre le français à l'étranger, en réduisant sa difficulté. Sur 804 mots d'un texte composé par des étudiants allophones universitaires, 78 types de fautes ont été recensés (soit près de 10 % des mots concernés). Un correcteur grammatical en corrige 40, présente 15 options, ignore 13 erreurs et se bloque dans 10 cas.

3. Ce type de logiciel fait appel à une compétence plus que minimale de l'usager, qui doit parfois choisir parmi deux solutions et qui, surtout, doit faire face – chose nouvelle - à l'invention de la faute, appelée aussi « fausse détection », parfois supérieure à la vraie détection.

4. L'écriture se trouve ramenée à un outil technique et efficace. La règle se trouve rappelée dans sa simplicité scolaire, dans les limites imposées par l'écran, c'est-à-dire bien sûr dans son arbitraire. Est-ce une étape dans la réduction de la sacralisation de l'orthographe ?

Conclusion

L'évolution des études linguistiques s'est toujours faite par séparation des champs de recherche : synchronie-diachronie avec Saussure, forme-sens avec Sapir, Bloomfield, Harris, Chomsky. La formalisation d'une grammaire orthographique de la langue représente-t-elle une étape isolable dans le traitement de la langue par la machine ? L'articulation lexique-syntaxe ainsi réétudiée module-t-elle une nouvelle approche dictionnairique ?

Un nouveau type d'écriture est-il à envisager ? Il serait conforme à une langue grammaticalement analysable par la machine. La diversité de la langue serait d'autant plus réduite que le correcteur par ses dictionnaires spécifiques signalerait les anglicismes, québécismes, belgicismes...

Le dépassement de la vérification locale pour embrasser la phrase représente un grand progrès et une étape importante dans l'établissement d'une grammaire cohérente qui rend compte de l'aspect mécanique du comportement syntaxique. Mais ni l'addition de règles, ni l'ajout de mégas ne suffisent. Plus les descriptions seront exhaustives, plus elles seront opératoires, plus elles permettront de faire coïncider le grammaticalement possible avec le grammaticalement acceptable. La frontière sémantique sera probablement redéfinie grâce aux recherches importantes dans la gestion du documentaire. Si l'orthographe pouvait ainsi se retrouver complètement désacralisée, le champ serait libre pour se jouer des difficultés autres qu'orthographiques et autrement plus passionnantes, pour tutoyer sans réserves le dictionnaire des difficultés.

Choix de grammaire et organisation du lexique*

Philippe BARBAUD

Université du Québec à Montréal, Canada

1. Les faits

Trois séries de faits au moins permettent de soutenir, sur le plan empirique, que les entrées lexicales dénominatives ne sont pas toutes des atomes syntaxiques qui « projettent » leur catégorie au niveau maximal de leur structure syntagmatique symbolisée par SN ou N″ ou encore X^{max}. Ce phénomène caractérise un phénomène d'« EXOCENTRICITÉ CATÉGORIELLE » et il s'avère particulièrement productif en composition lexicale grossièrement définie comme un processus de formation des mots composés. Nous montrerons dans ce travail qu'une approche syntaxique de ce phénomène sur le plan théorique fournit une base d'implantation motivée dans certaines grammaires d'unification.

1.1. Faux-endocentriques et composés déverbaux

On mentionnera d'abord le cas de certains composés nominaux que l'on qualifiera de « faux-endocentriques », c'est-à-dire des SN dont le genre n'est pas déductible de leur pivot nominal :

(1) a. *UN* (*mille-pattes, mille-feuilles, rouge-gorge, deux-temps, trois-pièces, quatre roues motrices, tête-à-queue, du porte-à-porte*, etc.)
b. *UNE* (*deux-chevaux, six-cylindres*, etc.)

Dans cet échantillon de faux-endocentriques, le SN est d'un genre différent du ou des lexèmes nominaux qu'il contient.

* Ce travail a bénéficié du concours financier de l'AGENCE FRANCOPHONE POUR L'ENSEIGNEMENT SUPÉRIEUR ET LA RECHERCHE (AUPELF-UREF) et du CONSEIL DE RECHERCHE EN SCIENCES HUMAINES (CRSH) du Canada. Nous adressons nos remerciements à Fernande Dupuis, Benoît Habert et Christian Jacquemin pour leurs commentaires relatifs à certaines sections figurant dans une version préliminaire de cet article.

On mentionnera ensuite le cas des Noms Composés à Pivot Verbal (NCPV) du type *tire-bouchon* qui sont toujours de genre masculin (Surridge, 1985) :

(2) *UN* (*porte-parole, baise-main, brise-glace, casse-tête, gagne-pain*, etc.)

Dans cet échantillon, le genre masculin du SN ne provient ni du verbe ni du lexème nominal qui tient lieu de « COD pro-forma » du composé déverbal (Barbaud, 1994 : 9).

On mentionnera enfin le cas des dérivés nominaux du type [*mise* Prep N] dont le pivot est un participe passé et non pas un nom commun dérivé par conversion, comme nous le soutenons dans Barbaud (à paraître) :

(3) *UNE* (*mise en cause, à feu, en quarantaine, aux fers, sous vide*, etc.)

Dans ce paradigme particulier, il n'est pas évident que le nœud SN soit une extension catégorielle du V_{pp} *mise*.

À cette dernière observation il est loisible d'ajouter celle qui concerne les SN composés dérivés de verbes à l'infinitif et exceptionnellement à temps fini :

(4) *LE* (*laissez-passer, rendez-vous, faitout, laisser-faire, laisser-aller*, etc.)

Voilà un état de fait très succintement décrit à partir duquel, néanmoins, il est loisible d'évaluer l'adéquation de certains modèles de grammaire ou de certains programmes de Traitement Automatique des Langues Naturelles (TALN).

Précisons toutefois que, bien que nous ne disposions d'aucunes données quantifiées pour appuyer notre position, nous croyons que le phénomène de l'exocentricité catégorielle est loin d'être marginal en français et dans l'ensemble des langues romanes en raison de leur morphologie particulière. Si le phénomène de l'exocentricité catégorielle semble effectivement absent des corpus de « langues spécialisées », au sens de Lerat (1995), il est en revanche dûment consigné dans les dictionnaires d'usage élaborés à partir de la langue courante. De ce fait, il est presque toujours associé, comme on peut s'en rendre compte, à un phénomène d'exocentricité sémantique ou d'interprétation « idiomatique ». Autrement dit, on note une très forte corrélation entre l'exocentricité formelle et la création d'entrées lexicales par composition, ce qui permet d'envisager le problème sous l'angle de la formation d[illegible] mots et, par voie de conséquence, de la terminologie.

1.2. Deux modèles de grammaire en compétition

De manière générale, on peut définir l'exocentricité catégorielle comme une rupture ou une disjonction entre les propriétés formelles du nœud syntagmatique dominant et celles du terme lexical dominé qui lui sert de pivot (« tête » = "head"), peu importe la nature et le nombre des nœuds intermédiaires. Si l'on se réfère aux représentations arborescentes des grammaires génératives, l'exocentricité s'illustre grossièrement au moyen du graphe donné en (6), lequel s'oppose radicalement au graphe endocentrique donné en (5) :

(5) G1 : X'' ... X° ...

(6) G2 : Y'' ... X° ...

Une grammaire G2 qui est compatible avec le graphe donné en (6) est sous-tendue par des propositions théoriques et descriptives aussi substantielles que divergentes par rapport à une grammaire G1 qui n'admet que des descriptions conformes au graphe donné en (5). Mais une telle perspective déborde évidemment du cadre limité de cet article.

Revenons à nos graphes pour y apporter une précision importante. Bien que le graphe exocentrique donné en (6) fournisse une représentation adéquate des cas patents de rupture catégorielle illustrés en (2-4), il s'avère incapable de rendre compte du cas des faux-exocentriques illustrés en (1) puisqu'à proprement parler, il n'existe aucune différence catégorielle entre SN = X'' et son pivot X° = N. La rupture catégorielle n'est pas le fait d'une disjonction entre X et Y'' mais plutôt le fait d'une disjonction entre certains traits de X par rapport à ceux de X''. Ce type d'exocentricité se laisse donc mieux caractériser par le graphe suivant, lequel fait état d'une rupture entre les traits α de X° et les traits γ de X''. Aussi le graphe (7) n'est-il possible, malgré les apparences, que dans une grammaire de type G2.

(7)

X_{γ}^{max}

...

... X_{α} ...

Par conséquent, on constate que la manipulation de traits acquiert autant d'importance vis-à-vis de la structure que la manipulation transformationnelle de nœuds atomiques ou syntagmatiques, ce qui favorise au départ un choix en faveur des grammaires dites « d'unification » par opposition aux grammaires de projection.

Quel que soit le moyen choisi pour assurer l'unification des deux niveaux de la représentation catégorielle d'un syntagme nominal (règles de projection, règles d'unification, conventions de percolation, mécanisme d'héritage ou de transfert de traits, etc.), on saisit d'emblée la difficulté majeure que doit affronter n'importe quel programme sensible aux traits morphologiques ou aux catégories de niveau atomique symbolisées par X°.

1.3. Problèmes de parsage

Un parseur ou un analyseur dont l'ingénierie de base consiste à catégoriser des suites de lexèmes pour les regrouper en syntagmes cohérents de catégorie SN se heurte inévitablement à un problème majeur d'analyse interne puisque non seulement le genre du DET peut ne pas correspondre au genre du N mais aussi la catégorie du SN peut ne pas correspondre à la catégorie du lexème-pivot, ce que nous illustrons par les dif-

férentes analyses données en (8) :

(8) a. $[_{SN}\ DET_{<+masc>}\ [...\ N_{<+fém>}...]]$
b. $[_{SN}\ DET_{<+fém>}\ [...\ N_{<+masc>}...]]$
c. $[_{SN}\ DET\ [_{SV}...\ V\ ...]]$

Tout algorithme strictement conforme à une grammaire de type G1 non seulement exclut radicalement l'exocentricité sur le plan théorique mais échouera aussi sur l'unification de ces trois descriptions qui seront déclarées mal formées bien que le même programme, peut-on présumer, fournisse un parsage et une catégorisation rigoureusement exacts des données. Par contre, dans une grammaire de type G2 qui n'exclut pas l'exocentricité, il est loisible en théorie de faire correspondre les graphes (6) et (7) aux trois descriptions données en (8) à la condition de modifier substantiellement l'ingénierie de base des modèles inspirés de G1 en ce qui concerne l'interface entre le lexique et la syntaxe. Il s'ensuit, du point de vue applicatif, que les cas d'exocentricité catégorielle n'ont rien d'insignifiant en TALN, en acquisition et en mise à jour terminologique parce qu'ils sont inévitablement source de bruit. Quant à la théorie linguistique, elle ignore purement et simplement le problème. Il n'en demeure pas moins que ni la catégorie N de ces composés ni leur sens dénominatif ne sont directement prédictibles sur la base des lexèmes qui les composent mais qui sont pourtant reconnus individuellement par n'importe quel parseur le moindrement performant. Non seulement le programme doit-il alors neutraliser la reconnaissance des lexèmes constitutifs pour éviter tout blocage d'une lecture contradictoire de leurs traits mais encore force-t-il à recourir à la lecture experte à des fins de repérage et d'encodage dans le lexique.

1.4. La solution de l'atomicité

La lecture experte est évidemment coûteuse mais c'est le moyen le plus efficace pour contourner ce problème d'application sans le résoudre au plan théorique pour autant. Le but de la lecture experte est donc d'encoder ce type d'expressions en tant qu'entrées lexicales particulières, ce qui est correct quant au résultat mais implique le recours à un dispositif de CONVERSION passablement arbitraire et aux incidences théoriques peu désirables. En effet, toute opération de conversion implique que la constituance syntagmatique verbale des NCPV par exemple et des expressions apparentées se voit réduite à une constituance atomique nominale afin de pouvoir assurer l'insertion de l'entrée lexicale dans le domaine du SN et non pas dans un domaine différent catégoriellement. Bref, l'ingénierie de conversion d'une grammaire de type G1 doit permettre d'assigner la catégorie N à des formes préalablement reconnues comme étant de catégorie non-N selon un schéma de règles donné en (9) et ce, pour finalement contourner les analyses données en (8) et plus particulièrement (8c)[1] :

1 Voir J Thiele (1987) pour une illustration typique de cette pratique La nominalisation par conversion se distingue de la nominalisation par affixation du type illustré en (1), cette dernière ayant fait l'objet d'un grand nombre d'études, comme on sait

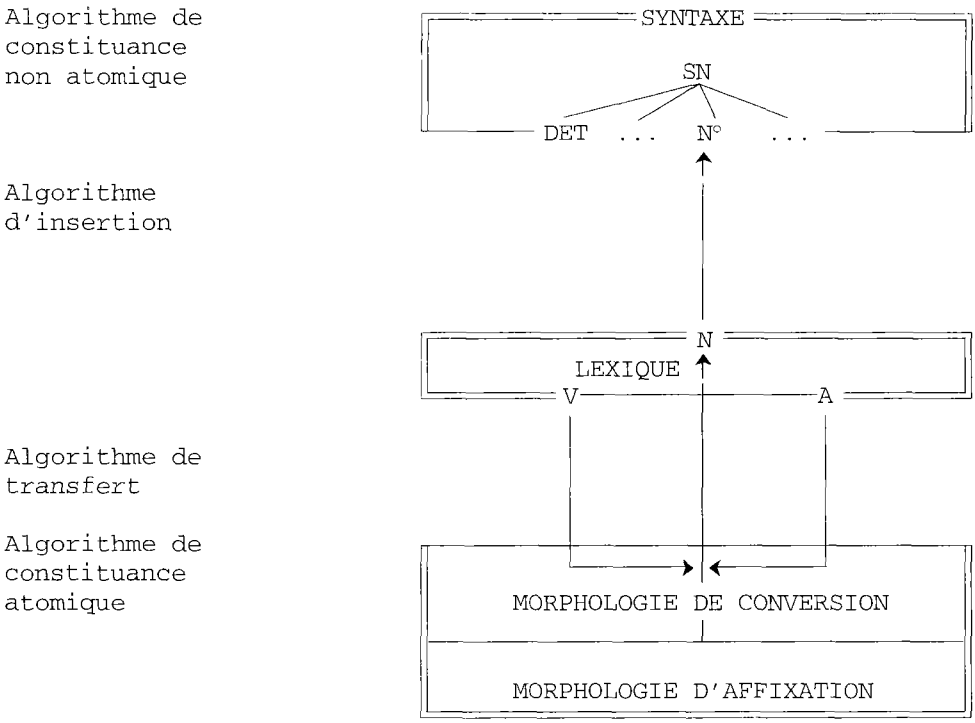
(1) a *La destruction (de Rome par Alaric)*
b *L'invasion (de la Gaule par Attila)*

Dans ce type de nominalisation, il est loisible d'assimiler le suffixe à un prédicat lexical auquel on assigne une variable argumentale – le morphème radical – en position de spécifieur, *i e* celle qui précède le prédicat, ce qui permet d'établir un rapport formel avec la structure syntaxique. Pour une discussion plus approfondie de la conversion, voir Barbaud (1994).

(9) a. A → N (*blanc, possible, important,* etc.)
b. V_{inf} → N (*sourire, déjeuner, lancer, être, avoirs,* etc.)
c. $V_{ppé}$ → N (*attaché, émigré, blessé, préposé,* etc.)
d. V_{ppt} → N (*dirigeant, protestant, militant, dégivrant,* etc.)

Convertir en atome nominal toute expression reconnue par un parseur robuste comme un constituant non-N est une opération catégorielle qu'il faut quand même accomplir manuellement afin de respecter l'architecture de la grammaire G1, architecture que l'on peut représenter au moyen du schéma suivant :

(10) **Fonctionnement d'une grammaire de type G1**



Dans une telle grammaire, l'endocentricité obligatoire de la constituance non atomique en syntaxe oblige à respecter la condition d'atomicité des entrées lexicales afin de permettre au mécanisme de l'insertion lexicale de s'appliquer. C'est donc à la morphologie que revient la tâche de réduire les composés nominaux déverbaux sous forme d'atomes N au moyen de règles quelconques de conversion. On obtient alors un modèle essentiellement « projectif », c'est-à-dire orienté de la morphologie vers le lexique et du lexique vers la syntaxe. Il fonctionne nécessairement « du bas vers le haut » (*bottom to top*). Dans ce type de grammaire, le lexique se conçoit entre autres comme une base de données MONOSTRATIFIÉE dont chaque élément discret est uniformément encodé sous une seule forme disponible, la forme atomique des X° de la grammaire.

1.5. L'opération d'insertion lexicale

On peut ici préciser que c'est l'opération d'INSERTION lexicale telle que Chomsky l'a conçue dans les années 60 qui force à adopter la contrainte de l'atomicité des entrées lexicales d'où découle la conception actuellement dominante de l'endocentricité projective des structures syntagmatiques. Mais l'insertion lexicale est une opération pseudo-grammaticale qui n'a aucun fondement cognitif. Le cerveau ne fonctionne pas à sens unique, comme Changeux (1983) le dit explicitement. L'opération d'insertion lexicale si familière aux générativistes est en réalité une servitude des grammaires transformationnelles primitives des années 50 conçues en fonction des limites inhérentes aux machines de Turing. Cette conception est toujours prédominante. Il n'est pas dit que cet héritage ne doive pas être remis en cause de nos jours et ce, pour trois raisons.

La première est que personne à ma connaissance n'a fourni la moindre preuve démontrant que l'exocentricité des syntagmes est une propriété formelle hors de portée d'un cerveau humain et qu'elle doit être absente des modèles de représentation des propriétés formelles des langues naturelles.

La seconde, plus technique, est que l'insertion lexicale ayant été conçue à l'origine comme une opération de SUBSTITUTION de chaîne phonétique (Chomsky, 1965), rien n'empêche d'envisager que d'autres types d'opérations formelles puissent assurer le couplage du lexique et de la syntaxe de manière plus efficace et plus « explicative ». Il devient évident qu'une grammaire conçue en termes d'insertion d'entrées lexicales sous chaque nœud atomique d'une structure syntaxique exclut radicalement l'insertion de matériel sous les autres nœuds, que ceux-ci soient de niveau intermédiaire ou de niveau syntagmatique (maximal).

La troisième raison est que la contrainte de l'atomicité des entrées lexicales est incompatible avec le phénomène de la variation qui affecte les entrées lexicales polylexémiques et ce, malgré l'incidence du phénomène de figement qui les caractérise significativement[2]. Par exemple, si on déclare que l'expression *basse pression* est une entrée lexicale d'un corpus médical, comment la traiter dans l'expression *une alternance de basses et hautes pressions* ?

En définitive, l'opération d'insertion restreint considérablement l'accès direct au lexique de l'information syntaxique.

2. Le concept de la « syntaxe dérivationnelle »

Le concept de « syntaxe dérivationnelle » que nous allons maintenant esquisser justifie notre choix en faveur d'une grammaire de type G2 que nous estimons plus appropriée au traitement des Entrées Lexicales Polylexémiques du français (ELP). Comme l'indique cette étiquette, la syntaxe est une composante pourvoyeuse de mots,

2. Pour un état de la question en français dans le cadre des grammaires applicatives, voir la **Présentation** de Habert & Jacquemin (1993) au numéro de *Traitement automatique des langues* portant sur le thème des **Traitements automatiques de la composition nominale**, vol. 34 (2).

à l'instar de la morphologie dérivationnelle. Les règles qui en composent l'algorithme engendrent les structures adaptées à la formation des locutions verbales et des noms composés, c'est-à-dire des structures susceptibles d'être interprétées lexicalement, d'où l'importance que nous accordons à la manipulation de traits sémantiques. Nous articulons ainsi, dans un cadre de grammaire moderne, la substantielle proposition de Benveniste (1974) relative à la « synapsie », reprise notamment par Guilbert (1975), mais sans la limiter au seul domaine du SN, ce qui n'a aucune raison d'être aux plans théorique et empirique. La formation d'ELP répond à un besoin de néologie qu'on pourrait qualifier de « transcatégoriel », d'où l'enjeu d'une théorie X-barre correctement formulée.

Dans son architecture générale, le modèle que nous préconisons fait appel à quelques principes de base et à des opérations compatibles seulement avec un nombre restreint de modèles de grammaires d'unification. Commençons par les fondements.

2.1. Fondements théoriques

Nous posons les deux postulats suivants. Le premier se formule ainsi :

(11) **POSTULAT DE L'EXOCENTRICITÉ NOMINALE**
L'exocentricité du syntagme nominal est aussi une propriété des langues naturelles.

De ce premier principe, qui se veut une alternative à la contrainte de l'endocentricité des syntagmes des grammaires génératives, on tire la conclusion que la catégorie SN est aussi une catégorie primitive de la grammaire universelle, au même titre que la catégorie initiale Σ = phrase, puisqu'elle n'est pas toujours dérivée d'un pivot de catégorie N. Un SN n'est donc pas la nécessaire projection de N. Il s'agit là d'un débat dont l'issue ne saurait être validée que sur le plan empirique. À cet égard, les faits de composition lexicale dans les langues romanes soutiennent fortement notre position théorique. En anglais, ce sont les nominaux gérondifs qui supportent l'énoncé (11)[3].

Le second postulat s'énonce comme suit :

(12) **POSTULAT DE LA NEUTRALITÉ CATÉGORIELLE DES ENTRÉES LEXICALES**
Toute catégorie syntaxique est susceptible de définir la forme d'une entrée lexicale.

Ce second principe est une alternative à la servitude de l'atomicité des entrées lexicales inhérente aux grammaires génératives de type chomskyen. Ce principe requiert l'élaboration d'un dispositif d'interface bidirectionnel entre le lexique et la syntaxe et il doit être apte à être implanté dans un modèle applicatif. Un tel dispositif tient dans la proposition suivante :

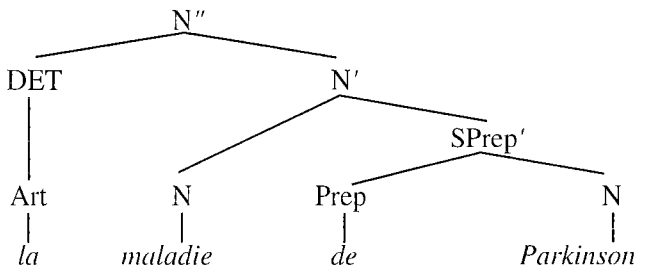
3. Comme Webelhuth (1995 : 89 n. 25) l'a fait récemment remarqué dans son manuel, les nominaux gérondifs font encore problème aujourd'hui puisque leur structure viole la contrainte d'endocentricité de la théorie X-barre.

(13) **PRINCIPE DE RÉTROPROJECTION**
Dans un modèle de syntaxe dérivationnelle, chaque SITE d'une structure syntaxique est susceptible de CORRESPONDRE à une entrée lexicale.

2.2. Sites et positions

Précisons brièvement de quoi il retourne. D'abord nous nous inspirons ici d'une distinction féconde de Milner (1987) entre « sites » et « positions » syntaxiques. Dans la configuration (14), on dénombre seulement quatre positions (nœuds terminaux encerclés) mais huit sites. En vertu du principe de fonctionnement (13), ce n'est plus un potentiel de seulement quatre entrées lexicales qui concourt à l'interprétation de cette structure mais un potentiel de huit, ce qui inclut non seulement les catégories préterminales (ou atomiques) mais aussi les catégories de niveau nodal X′ et X″. C'est certainement là un gain appréciable d'efficacité pour une même structure.

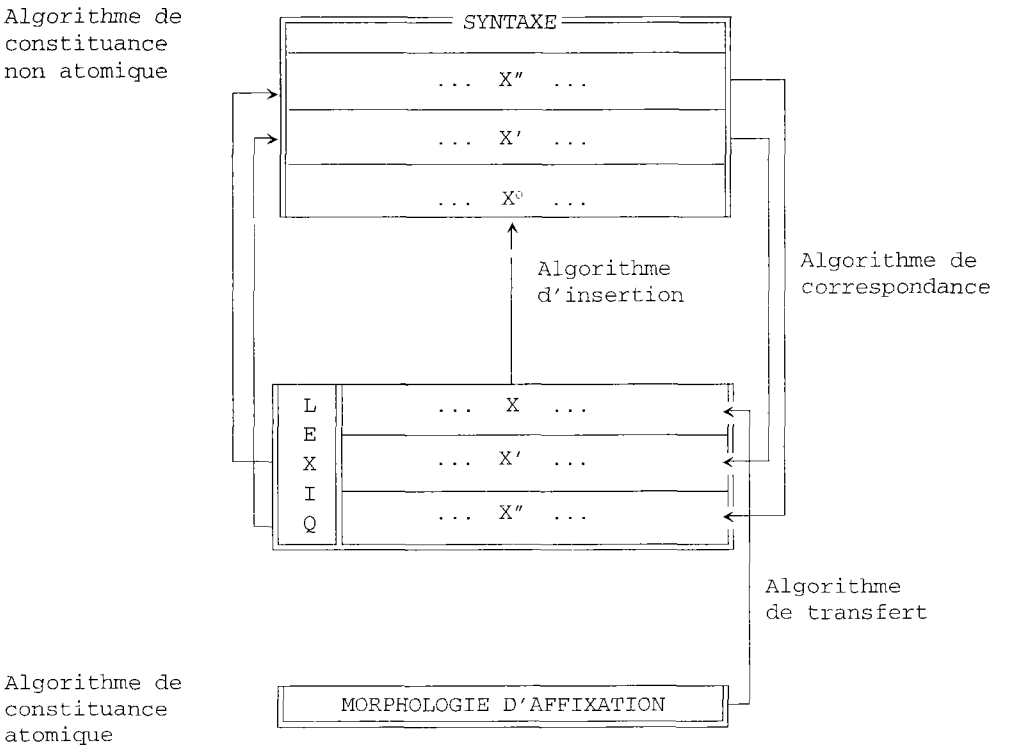
(14)



Ensuite, nous basons l'échange lexique \ syntaxe sur une opération de CORRESPONDANCE entre nœud syntaxique et signification lexicale et non de substitution entre atome et terme lexical. La distinction est cruciale parce que la correspondance n'implique pas l'insertion mécanique des entrées lexicales dans l'arbre mais elle ne l'exclut pas pour autant. La correspondance actualise alors la « rétroprojection » des catégories majeures dans le lexique qui se structure en conséquence. Par ailleurs, dans notre modèle comme dans les autres, toute position que contient un arbre quelconque définit le domaine d'application des règles d'insertion. En outre, une position ne peut être occupée que par un atome parce que le seul objet qu'une règle d'insertion puisse mettre en jeu, c'est l'atome défini comme catégorie lexicale. En revanche, toute position est un nœud qui se définit aussi comme un site mais l'inverse n'est pas vrai. Le principe (13) opérera donc à chaque niveau nodal d'une configuration syntaxique quelconque, ce qui revient à dire que les nœuds supérieurs peuvent aussi tenir lieu de catégories lexicales grâce à une opération de correspondance mais non pas d'insertion.

La bidirectionnalité du modèle est donc assurée à la fois par une opération d'insertion au niveau nodal X° dans le sens lexique → syntaxe et par une opération de rétroprojection aux autres niveaux nodaux dans le sens syntaxe → lexique, ce qui peut être illustré par le schéma suivant :

(15) **Fonctionnement d'une grammaire de type G2**



L'organisation du lexique n'est plus la même : il est maintenant polystratifié en fonction de trois dictionnaires, celui des atomes, des « synapses » et des syntagmes, ce qui garantit un accès direct aux entrées lexicales complexes de forme synaptique, majoritairement composées de mots composés et de locutions, et de forme syntagmatique, moins nombreuses mais tout aussi idiosyncrasiques, telles les dictons, proverbes et sentences du type *La montagne accouche d'une souris*. Celles-ci n'ont plus à être récupérées par l'intermédiaire d'un de leurs éléments constitutifs, comme dans les dictionnaires d'usage. Tout site fait-il ainsi l'objet d'une procédure de vérification constante (*checking device*) dans le lexique. C'est ce que nous voulons dire par « correspondance » sites \ lexique.

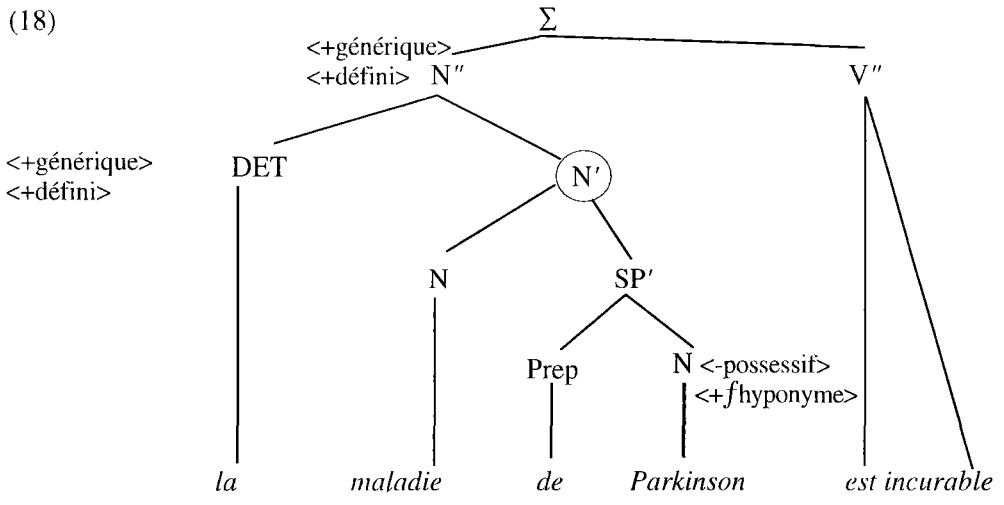
2.3. Traits lexicaux et isomorphie structurale

Notre approche est empiriquement justifiée par le phénomène de l'ambiguïté de nombreux mots composés selon qu'ils sont interprétés soit littéralement, c'est-à-dire par compositionnalité, ou soit non littéralement, c'est-à-dire par idiomaticité. Dans les deux cas, c'est la même structure qui tient lieu de domaine d'application des règles d'interprétation. C'est ce qu'illustre le contraste suivant :

(16) *La maladie de Parkinson est toujours incurable*
(17) *La maladie de Parkinson perturbe notre service des ventes*

Dans le premier exemple, la suite *maladie de Parkinson* s'interprète comme une entrée lexicale qui réfère à une maladie spécifique. De par sa position syntaxique dans la structure, le nom propre est un complément déterminatif qui s'interprète en tant que « foncteur hyponymique » du vocable générique *maladie* agissant comme pivot de la structure dans les langues romanes. Comme de fait, la *maladie de Parkinson* est une {SORTE DE} *MALADIE*, au sens de Levrat & Sabah (1990). Cette interprétation est propre à la sémantique lexicale et elle justifie le recours aux traits pertinents dans une description linguistique[4]. En outre, le mot *Parkinson* est dépourvu de son contenu référentiel en raison de l'absence significative de DET interne, suivant en cela les idées de Cadiot (1991). Cette propriété structurale se reflète directement, soutenons-nous, dans une version adéquate de la syntaxe X-barre des grammaires chomskyennes, (Barbaud, 1992 : 202 ; 1994 : 15), ce qui permet alors d'engendrer la structure grossièrement décrite en (18) :

(18)



Grâce au mécanisme de rétroprojection constante, il s'établira une correspondance entre le nœud N′ encerclé, qui est l'un des sites que nous qualifions de « synaptiques », et un objet du lexique préalablement encodé dans le lexique au moyen du trait <+N'> et dont la matrice contient une description structurale identique à celle du nœud concerné, ce qui inclut la suite des termes [*maladie de Parkinson*]. L'interprétation correcte est conférée par le trait <+hyperonyme> dont la valeur sémantique renvoie à un vocable approprié de la langue (lorsque disponible). Dans les composés endocentriques, il y a coïncidence par unification de traits entre ce vocable et le nom-pivot, ce qui rend compte de la variabilité en genre. À titre d'essai, le prototype d'une telle entrée est donné en (17)[5] :

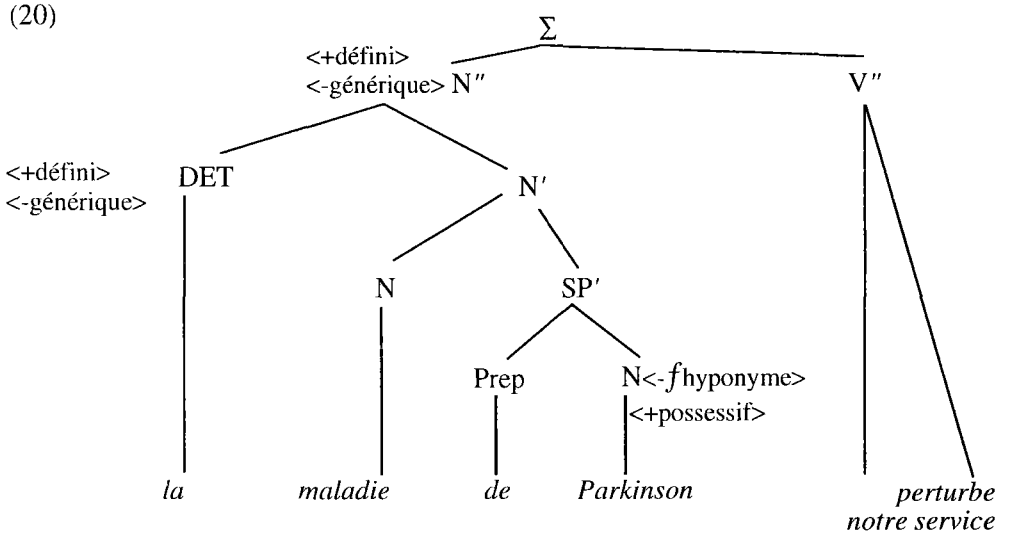
4 Pour être plus précis, nous mettons en parallèle les dénominations de catégorie SN et les propositions de catégorie Σ. Chaque domaine possède une interprétation fonctionnelle qui lui est particulière (voir le plaidoyer de Lerat (1990) pour une « hyperonymie fonctionnelle »). Alors que les phrases ont systématiquement recours au canevas des rôles (valeurs) thématiques pour être interprétées, les SN dénominatifs de leur côté ont systématiquement recours aux diverses valeurs d'hyper et d'hyponymie potentiellement associées aux mots. Le statut théorique de ces différentes valeurs interprétatives est le même et leur importance est déterminante dans l'architecture du modèle adopté.

5. Le traitement des locutions verbales tronquées du type *prendre parti, tirer parti, tenir tête*, etc., est similaire à la différence que ces entrées lexicales sont munies du trait <+V'>.

(19) **[maladie de Parkinson]** <+N'>
{ N1 = */maladie/*; N2 = */parkinson/*
N1 = <+hyperonyme> = {MALADIE}
N2 = <+*f*hyponyme> = {avec tremblements constants chez les personnes âgées}

À défaut de la présence d'une entrée lexicale existante, la compositionnalité des nœuds poursuit son cours. Cependant, chaque fois qu'un nœud synaptique s'apparie à une entrée lexicale, c'est la lecture idiomatique qui prévaut sur la lecture compositionnelle, laquelle est autrement assignée par défaut nonobstant la réelle faisabilité de la chose comme Kayser et Lerat (1990 : 120 *sqq.*) en discutent. C'est ce que nous obtenons dans le cas de l'exemple (17) qui est entièrement compositionnel. La structure est la même qu'en (18) mais l'interprétation diffère à cause des différences importantes de traits :

(20)

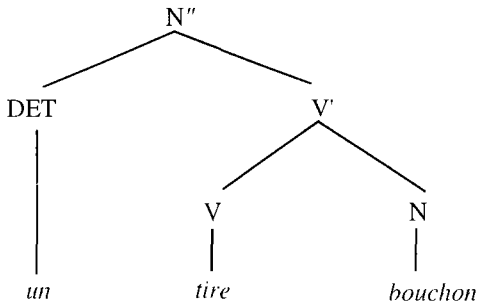


Ce traitement montre que l'interprétation dénominative, toute lexicale qu'elle est, se laisse appréhender en termes de traits proprement lexicaux et d'isomorphie structurale plutôt qu'en termes de catégories lexicales et de conversion. Une telle approche de l'ambiguïté lexicale en composition favorise les grammaires d'unification.

2.4. Grammaires d'unification et exocentricité

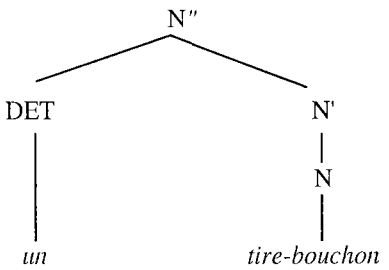
En prenant les NCPV en exemple, nous allons maintenant procéder à une analyse applicative de ces objets lexicaux. La rétroprojection et l'isomorphie structurale vont déterminer notre ligne de conduite. Linguistiquement parlant, nous disons que les NCPV sont des SV basiques nominalisés, en réalité des quasi-SV de catégorie V' (récursive) dominés par le nœud N'' en structure profonde. C'est donc la configuration typique, bien que fort incomplète, que nous donnons en (21), à mettre en parallèle avec la description (8c), qui est l'enjeu de notre analyse :

(21)



L'exocentricité de la configuration (21) contraste avec l'endocentricité de la configuration suivante, la seule actuellement permise en théorie syntaxique générative :

(22)



La validation de la structure (21) est assurée par le fait que la rétroprojection du V′ dans le lexique se voit ratifiée par l'existence d'une entrée lexicale de catégorie **<+V'>** dont la matrice contient le sens {INSTRUMENT} associé à la suite polylexémique */tire bouchon/*. En cas d'échec de la rétroprojection, la structure est rejetée parce que son exocentricité n'est pas interprétable par compositionnalité.

En ce qui concerne l'application, il faut convenir que l'implantation de ces idées dans un modèle de linguistique computationnelle ne peut pas être accomplie par n'importe quelle grammaire d'unification. À vrai dire, il n'y a que les **Grammaires d'Arbres Adjoints** (TAG = *Tree Adjoined Grammar*) qui autorisent un tel transfert en raison du formalisme particulier qui est le leur. En effet, parmi les quatre principaux modèles de grammaire d'unification exposés en détail dans Abeillé (1993), seules les TAG disposent de l'appareillage requis pour véritablement rendre compte de l'exocentricité nominale[6]. Dans ce modèle, la notion de « tête » (pivot) y est uniquement lexicale tandis que dans les modèles concurrents comme la GPSG de Gazdar, Klein, Pullum & Sag (1985) ou comme dans la HPSG de Pollard & Sag (1994), la tête identifie la structure (Abeillé, 1993 : 214). Toujours selon cet auteur, dans le

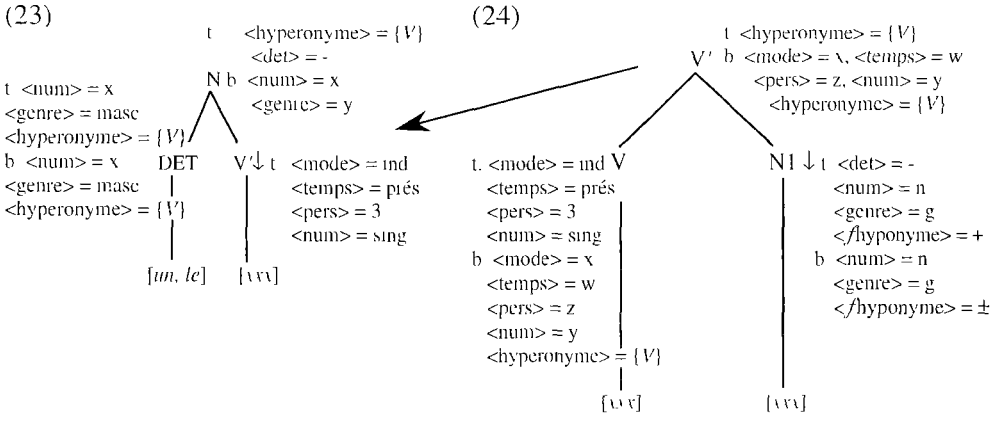
6 Ajoutons toutefois que de l'avis de Philippe Blache et Lorne Bouchard (communication personnelle), le modèle de la GPSG offre les ressources nécessaires pour traiter l'exocentricité formellement. Si c'est le cas, il convient de déterminer si la cohérence interne de la théorie linguistique qui le sous-tend peut s'accommoder du postulat énoncé en (11).

modèle basé sur la LFG de Bresnan (1982), les règles de réécriture hors contexte ne font jamais appel à l'exocentricité catégorielle des syntagmes sauf en ce qui concerne le symbole initial P (phrase). Enfin, précisons qu'aucune grammaire générative de type chomskyen comme celle par exemple du *Gouvernement et du Liage* (Chomsky, 1981) n'est dotée du pouvoir de générer des structures syntaxiques exocentriques (cf. Webelhuth, 1995 : 33).

En tout état de cause, les TAG offrent deux dispositifs qui permettent d'implanter cette propriété formelle des langues naturelles. D'abord, la manipulation d'unités de base sous forme d'arbres élémentaires (initiaux ou auxiliaires) définissant le domaine d'application d'opérations d'adjonction ou de substitution. Cela permet d'adjoindre directement un nœud de racine V' à un nœud de racine N pour former un nom. Ensuite, les TAG permettent de dissocier les traits d'un nœud quelconque parce qu'elles distinguent entre « traits aval » et « traits amont » lors du processus d'unification[7]. Grâce à cette dissociation, il devient possible d'exprimer un grand nombre de contraintes formelles sous forme de traits amont d'un nœud quelconque, ces derniers agissant notamment comme filtres vis-à-vis des lexèmes candidats.

2.5. Arbres initiaux des NCPV

Par conséquent, nous allons postuler deux arbres élémentaires initiaux, l'un de racine N donné en (23) et l'autre de racine V' donné en (24)[8]. C'est donc l'adjacence DET\V' qui instancie l'exocentricité des NCPV en TAG :



7 Rappelons que les traits amont d'un nœud donné permettent de spécifier les diverses relations de ce nœud avec ceux qui le dominent tandis que les traits aval permettent de spécifier les relations qu'entretient ce nœud avec ceux qu'il domine, plus particulièrement ceux qui sont associés aux termes lexicaux

8 Sur ce dernier point, nous divergeons d'opinion avec Abeillé (1993 : 207 n. 100). A l'instar de M. Gross, cet auteur écarte l'hypothèse que les arbres phrastiques élémentaires de catégorie P sont dépourvus de nœud SV intermédiaire pour le français. Nous invoquons au contraire la grande productivité des nœuds SN et SV « tronqués » en composition lexicale, c'est-à-dire dépourvus de position SPEC interne notamment dans quelques 200 locutions verbales et plusieurs milliers de noms composés, pour justifier le caractère récursif des nœuds intermédiaires V' et N', un argument qui interpelle aussi Corbin (1987).

Ce traitement appelle évidemment plusieurs commentaires.

1. L'arbre initial (24) de racine V′ se **substitue** (≠ adjoint) au nœud V′ de l'arbre initial (23) de racine N. Ce choix est justement dicté par la nécessité de rendre compte de l'exocentricité catégorielle dans le domaine de la racine nominale et non dans celui d'une autre racine. Grâce à la substitution (notée ↓) requise par le nœud V′ de l'arbre **initial** (23), la configuration du DET devient réceptrice de l'arbre (24) selon des conditions très strictes. Cela n'aurait pas été le cas si on avait choisi un arbre **auxiliaire** étant donné que ce type d'arbre commande une opération d'adjonction qu'on aurait notée V′*, ce qui aurait eu pour effet d'insérer l'arbre (23) dans un autre arbre de racine éventuellement différente, par exemple de racine P, d'où l'impossibilité de réinsérer le tout sous un nœud N ;
2. Le trait amont de **genre <masc>** du nœud DET de l'arbre (23) prédit que les NCPV sont toujours masculins[9] ;
3. L'invariabilité de la **flexion verbale** des CNPV est prédite par les traits amont du nœud V′ de l'arbre (23) ;
4. Le **caractère tronqué** du COD pro forma des NCPV est prédit par la valeur négative du trait amont <det> de l'arbre (24) ; la même valeur en amont du nœud racine de l'arbre (23) évite que l'arbre dérivé puisse être adjoint à un arbre auxiliaire DET ;
5. L'interprétation du COD en tant que **foncteur hyponymique** est requise par la présence du trait amont <*f*hyponyme> spécifié positivement dans l'arbre (24) ;
6. La **contrainte d'hyperonymie** nécessaire à l'interprétation des synapses nominaux exocentriques se traduit par la présence du trait <hyperonyme> en amont des nœuds DET et N de l'arbre (23). Il s'agit d'un trait d'accord destiné à s'unifier avec un trait identique dans le nœud V′ adjacent ;
7. Si le nœud V′ s'interprète effectivement en fonction d'un hyperonyme quelconque, sa variable notée {*V*} (pour « vocable hyperonymique ») sera remplacée par la valeur lexicale inscrite dans l'entrée lexicale synaptique, *i.e.* {INSTRUMENT} pour [*tire-bouchon*]. L'unification se fait par accord DET\V′.

Par conséquent, le partage des traits qui résulte de la substitution de l'arbre (24) au nœud V′ de l'arbre (23) donne en principe non seulement une structure bien formée dans ce type de grammaire mais aussi rend compte simultanément des propriétés prédictibles des NCPV tant sur le plan morphologique que syntaxique.

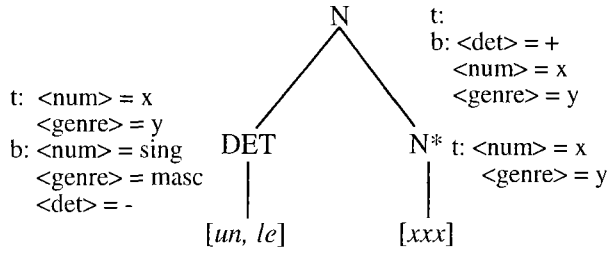
2.6. Canonicité

Toutefois, bien que ce traitement soit à première vue conforme aux quatre principes de bonne formation des arbres élémentaires en TAG, certaines difficultés doivent être résolues. Nous en décelons deux. La première a trait à l'arbre (23) qui n'est pas canonique parce que la substitution est censée être réservée aux noms et aux verbes à

9. Il est possible cependant de se débarrasser de ce trait dans la mesure où le trait <genre> est un « trait d'accord », cf. Abeillé (1993 : 219), dans le domaine du SN et que la grammaire comprend une règle d'accord entre catégories adjacentes selon le genre non marqué en français. Puisque le nœud substitué V′ ne contient aucune mention de trait de genre, le trait <genre> du DET prendra automatiquement la valeur non marquée <+masc>. Un tel traitement n'est pas sans rappeler l'accord SPEC\tête (flexionnelle) ou encore l'accord sujet-verbe dans le domaine de Σ

compléments nominaux tandis que l'adjonction l'est aux adjectifs, adverbes et autres lexèmes modifieurs, y compris les déterminants, cf. Abeillé (1993 : 206). Ici, l'arbre (23) est un arbre à substitution assigné à DET, ce qui n'est pas permis. La seconde est que si l'on a affaire à un arbre à substitution, ce dernier est censé afficher un nœud racine de même catégorie que le nœud feuille où a lieu la substitution. Ce n'est pas le cas de l'arbre (23) qui affiche trois nœuds différents par comparaison avec la représentation canonique des DET donnée en (25), cf. Abeillé (1993 : 226) :

(25)



La résolution de ces deux difficultés ne pose guère de problème majeur, nous semble-t-il. D'une part, la restriction relative à l'identité du nœud-feuille et du nœud-racine dans les arbres à substitution n'est pas absolue : le nœud P en TAG ne domine pas de nœud-feuille de catégorie P. D'autre part, un modèle de TAG doit pouvoir incorporer des structures à substitution sans égard à la nature des éléments impliqués. On ne voit pas pourquoi la substitution serait interdite aux déterminants. En outre, il n'y a aucune raison de principe à diviser le travail dérivationnel en attribuant l'exclusivité de l'une ou l'autre des deux opérations à un ensemble de catégories plutôt qu'à un autre. Ce n'est pas la catégorie d'un nœud qui commande une opération d'adjonction ou de substitution mais la dynamique de la récursivité dans une langue donnée. C'est ce qu'Abeillé (1993 : 209) a tenté d'illustrer en distinguant structuralement les composés du type *verre de vin* par opposition à ceux du type *verre à vin*. Toutefois, bien que nous souscrivions à son analyse, nous n'entérinons pas le traitement qu'elle propose et qui consiste en fin de compte à traiter l'hyperonymie par le biais de différences structurales et dérivationnelles, c'est-à-dire par une structure à adjonction dans le premier cas (arbre auxiliaire) et à substitution dans le second (arbre initial). Il y a un danger réel à attribuer aux opérations qui manipulent des formes catégorielles certaines propriétés de nature interprétative.

La différence de distribution entre ces deux types de composés sur laquelle s'appuie Abeillé pour justifier une différence de traitement syntaxique est attribuable bien plus à une différence lexicale qu'à une différence dérivationnelle. En réalité, il y a une différence d'hyperonyme entre les deux expressions nominales bien que leur structure soit la même. Celle des SP de forme [N Prép N] ou [N Prép DET N] du français exhibe toujours la même configuration dans le domaine du SN parce que la récursivité du syntagme prépositionnel ne varie pas en fonction des différences lexicales du pivot : un SP se branche toujours de la même manière par rapport à son pivot. Dans *verre à vin*, nous n'avons qu'un seul hyperonyme référentiel, à savoir une sorte de {*VERRE*}, autrement dit un contenant sans contenu référentiel, et cet hyperonyme coïncide lexicalement avec le pivot *verre*, comme c'est le cas dans tous les composés endocentriques. Comme de fait, on peut remplir son verre à vin d'eau Perrier par exemple. Dans *verre de vin* par contre, nous avons deux hyperonymes référentiels, le

{*VERRE*} et le {*VIN*}, c'est-à-dire le contenant et le contenu. Comme de fait, lorsqu'on ajoute de l'eau Perrier à son verre de vin, on obtient un contenu dilué, ce qui n'implique pas que l'on boive dans un verre à vin. La distinction relève non pas d'une différence de structure résultant d'une différence dérivationnelle mais bien d'une différence d'interprétation lexicale du SP. Incidemment, une telle différence de structure s'avère redondante par rapport à celle que marque déjà le jeu des deux prépositions. Dans l'expression *verre à vin*, le SP tronqué s'interprète en tant que **foncteur** hyponymique exerçant sa rection sur le pivot qui le précède tandis que dans l'expression *verre de vin*, le SP tronqué n'est pas un foncteur hyponymique car il ne contribue pas à faire de *vin* une {SORTE DE} *VERRE*. Cela se reflète dans l'alternance des prépositions : dans le domaine du SN, la préposition *à* est presque toujours un **marqueur** d'hyponymie mais pas la préposition *de*. Il est donc préférable de marquer cette différence interprétative au moyen de traits fonctionnels plutôt que de structures dérivées différemment en syntaxe, d'où le « commutateur » <*f*hyponyme>= +.

C'est aussi la raison pour laquelle nous avons opté pour la spécification d'un trait d'hyperonymie muni d'une variable dans les arbres (23) et (24), trait qui sollicite l'unification des deux structures par accord. Encore une fois, il s'agit d'un trait de sémantique lexicale amplement justifié par les faits de composition[10]. Un tel trait fait partie de l'encodage des entrées lexicales synaptiques. Comme on a pu le remarquer avec la description (17), le vocable qui sert d'hyperonyme est encodé dans les entrées lexicales synaptiques. L'encodage du NCPV *tire-bouchon*, par exemple, se présentera grossièrement comme suit :

(26) [**tire-bouchon**]
$$\begin{cases} <+V'> \\ V = /tire/ \\ V = <\text{hyperonyme}> = \{\text{INSTRUMENT}\} \\ N = /bouchon/ \end{cases}$$

Nous appliquons alors aux composés exocentriques du type NCPV la même règle qu'aux composés endocentriques, à savoir que c'est le pivot de la structure qui véhicule le sens de l'hyperonyme. C'est donc le nœud V de la description (26) qui se verra attribuer le trait « d'accord » <hyperonyme>, lequel partagera sa valeur lexicale explicite avec le nœud DET de l'arbre dérivé après substitution. Ce partage de traits fera que la variable hyperonymique du DET prendra la valeur {INSTRUMENT} du nœud V', ce qui assure l'interprétation correcte des NCPV.

3. Conclusion

Dans notre analyse de certains faits de composition lexicale, nous avons rigoureusement appliqué la logique grammaticale que dicte une saisie directe de certaines données telles qu'un parseur ou un analyseur pourrait les traiter. En prenant appui sur certains cas patents d'exocentricité catégorielle, cette logique conduit à prendre acte de la nature syntaxique des suites polylexémiques que l'on interprète comme entrées lexi-

10. Précisons que la loi de l'hyperonymie en composition lexicale est implicitement postulée dans Boresdon & Tamba (1991) et Cadiot (1991). Elle l'est explicitement dans Corbin (1992).

cales. Les implications théoriques de cette approche n'ont rien de trivial parce qu'elles sollicitent un réaménagement profond de l'architecture d'une grammaire générative. Le concept de syntaxe dérivationnelle que nous venons de mettre de l'avant est une avenue prometteuse, croyons-nous, pour le traitement informatisé des mots composés du français. L'idée de base est que l'isomorphie des structures syntaxiques fait reporter sur la manipulation de certains traits sémantiques la tâche de générer la partie prédictible (sémantique) de l'interprétation lexicale. En définitive, la néologie par composition ne procède pas au coup par coup au gré des mots de la langue. La néologie par composition est au contraire fortement dépendante de la structure parce qu'elle procède en utilisant l'information syntaxique pertinente aux règles d'interprétation sémantiques qui sous-tendent la formation des entrées lexicales polylexémiques.

Outil d'intégration de bases de connaissances lexicales aux analyseurs syntaxiques

Philippe Blache et Mireille Delpui

2LC - CNRS, Sophia-Antipolis, France

1. Introduction

Les lexiques informatisés constituent le cœur de tout système de traitement automatique du langage naturel. La qualité de leur conception, mais aussi leur exhaustivité est un prérequis indispensable en particulier pour les outils nécessitant large couverture et robustesse : étiqueteurs de corpus, analyseurs couvrants, etc.

Par ailleurs, nous assistons ces dernières années à un vaste courant de lexicalisation des théories linguistiques qui consiste à représenter au niveau lexical un grand nombre d'informations syntaxiques (par exemple, la structure argumentale ou les relations de spécification) ou sémantiques (s'appuyant évidemment sur la sémantique lexicale). Cette propriété est caractéristique des théories basées sur les contraintes. Nous nous intéressons plus particulièrement ici à la théorie HPSG qui entre toutes est celle utilisant les lexiques parmi les plus complexes.

Nous avons donc besoin non seulement d'importantes ressources lexicales à proprement parler, mais aussi d'outils facilitant l'utilisation de ces ressources en termes de conception, d'accès, d'intégration à des systèmes ainsi que de maintenance et d'évolution du lexique lui-même.

On remarque cependant qu'il n'existe qu'une très faible réutilisation des outils lexicaux. En d'autres termes, le développement de systèmes de traitement du langage naturel (notamment d'analyseurs syntaxiques) conduit généralement à la mise au point de nouveaux outils, en particulier concernant l'accès lexical. Le premier objectif du gestionnaire décrit ici est très concret et porte sur la mise à disposition d'un système autonome et facilement intégrable à un analyseur. Mais l'intérêt essentiel de cette approche repose sur le fait que le lexique source utilisé est considéré comme étant en quelque sorte un lexique générique. Il ne s'agit donc pas d'adapter un lexique au formalisme requis par un analyseur donné, mais au contraire de proposer un environnement capable de filtrer ce lexique en fonction de spécifications particulières.

Pratiquement, cet environnement permet la gestion du lexique lui-même, et propose de plus un utilitaire d'accès efficace. Par ailleurs, le module de génération et de maintenance du format lexical HPSG permet de créer, à partir d'une entrée lexicale générique, le schéma de l'entrée lexicale HPSG correspondante. Ce type d'outil présente un double intérêt : il permet, d'une part, de faciliter considérablement la tâche de la constitution d'un lexique HPSG et constitue, d'autre part, un véritable module de filtrage à partir d'un lexique générique. Nous nous situons ainsi dans la perspective de nombreux projets nationaux et internationaux qui ont pour objectif la mise au point de ressources génériques en fournissant un outil de mise en œuvre pour un cadre formel donné[1].

Cet article comporte trois parties : (*i*) description des problèmes posés par les lexiques HPSG, (*ii*) présentation du gestionnaire et (*iii*) intégration à un analyseur.

2. Le lexique en HPSG

HPSG représente toutes les informations à l'aide de traits. Ces données sont hiérarchisées à l'intérieur de structures de traits permettant la description de relations complexes entre les éléments. Le mécanisme explicatif de base en HPSG repose en effet sur le partage de structure (exprimé sous forme de coindexation de sous-structures). La complexité des entrées lexicales HPSG (quelle que soit la catégorie syntaxique) provient donc à la fois de la richesse des informations, mais aussi de la représentation du partage de valeur.

La figure 1 correspond à l'entrée lexicale du déterminant *un*. Les informations principales contenues dans cette structure concernent les aspects syntaxiques et sémantiques et sont de plusieurs types. Les informations *endocentriques* portent sur les caractéristiques internes à l'item. Il s'agit, par exemple, de la partie du discours ou de la structure sémantique (dans ce cas, la quantification). On remarquera au passage que la valeur du trait QMEM gérant la portée de la quantification est partagée avec celle de CONTENU. Mais on trouve également des informations *exocentriques* portant sur les relations de ce déterminant avec un autre constituant, en l'occurrence le nom. Ceci se fait par l'intermédiaire du trait SPEC indiquant que le déterminant spécifie un nom. La valeur du trait CONTENU de ce nom (représentée par l'indice $\boxed{1}$) est partagée avec la valeur du trait sémantique RESTIND. On remarque de plus qu'une restriction est appliquée au nom sélectionné : celui-ci doit être au masculin singulier. En d'autres termes, le déterminant ne porte pas directement de traits d'accord, mais stipule simplement des propriétés de l'objet sélectionné.

1. Il existe en particulier sur ces questions un groupe de travail *Lexique* dans le cadre du GDR-PRC Communication Homme-Machine. Contact : Damien.Genthial@imag.fr

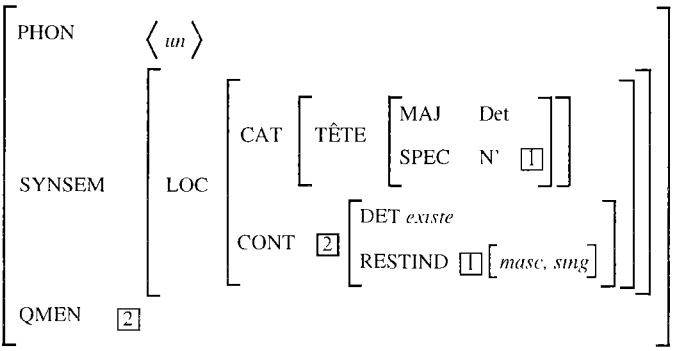


FIGURE 1 : Entrée lexicale de *un*

L'entrée lexicale d'un nom comme celle décrite dans la figure 2 comporte plus d'informations directes. On y remarque en effet des traits sémantiques plus détaillés ainsi que des traits d'accord (faisant partie en HPSG des traits sémantiques). La structure argumentale de cette tête est quant à elle décrite dans le trait VALENCE qui spécifie la sélection d'un déterminant (trait SPR).

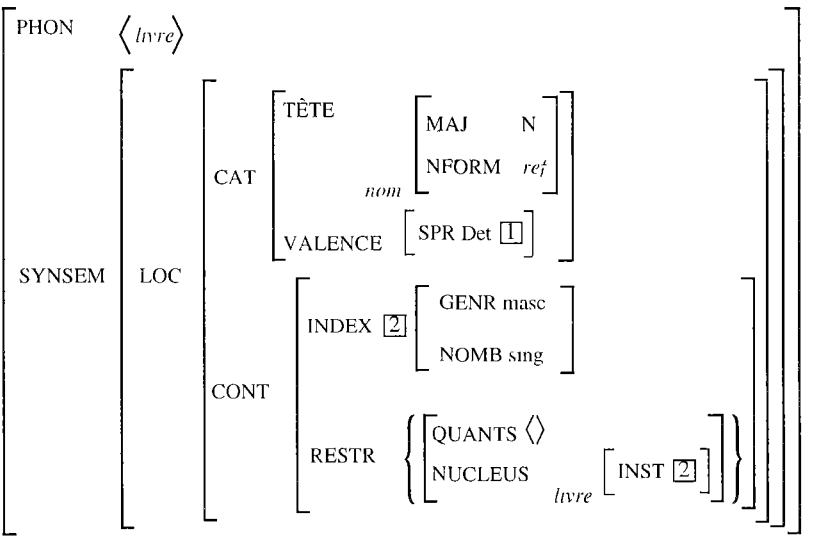


FIGURE 2 : Entrée lexicale de *livre*.

D'une facon générale, ces exemples révèlent deux types de difficultés rencontrées lors de la création d'un lexique HPSG :

- la structuration des données ;
- la représentation du partage des structures.

Il n'est donc pas envisageable de créer « manuellement » un véritable lexique HPSG. Il est en revanche intéressant de tirer partie de certaines propriétés et notamment d'utiliser l'héritage comme solution (désormais classique, cf. Briscoe *et al.*,

1993) de représentation générique. Il est en effet possible (nous le verrons dans la dernière section) de définir un certain nombre de schémas lexicaux constituant la base du lexique. Le système que nous décrivons propose ainsi le formatage d'entrées HPSG à partir d'un lexique général.

3. Présentation du système

L'objectif principal de ce système est de fournir aux analyseurs syntaxiques la possibilité de réutiliser d'importantes bases lexicales. Ceci nécessite une gestion optimale du lexique permettant sa maintenance, son évolution, un accès rapide à ses données ainsi qu'une adaptation efficace des formats du lexique de l'analyseur. Cette adaptation se présente, nous y reviendrons dans la section suivante, comme un module de génération et de maintenance des entrées utilisées par l'analyseur à partir d'entrées lexicales génériques.

D'une facon générale, le système propose :

- la recherche dans le lexique des mots à analyser. Cette recherche s'effectue via le gestionnaire lui-même ;
- une heuristique proposant en fonction de critères de fréquence et de contexte la meilleure probabilité de catégorisation ;
- l'adaptation des formats du lexique générique au formalisme utilisé.

L'outil que nous présentons utilise comme source le lexique LGE (Lexique Général Étendu), développé dans le cadre du projet BDLEX (bases de données et de connaissances lexicales du francais écrit et parlé, cf. Perennou, 1988). Chaque entrée lexicale contient des informations phonologiques et syntaxiques enregistrées dans plusieurs champs. Elles se présentent sous la forme suivante :

GRAPH_ACC	HG	PHON_SYLL	FPH	HP	CL_PHON	NS	FREQ	CS	GN	CFLB	DEFLIEN	CONST
aimer	11	&/m&	r"	11	E/NE	2	C2B0TR	V	01	060	**	Ta
amibe	11	A/mib	e	11	A/NID	2	-	N	Fn	81	-	-
elle	11	&/l	-	32	EL	1	C0B0	P	FS	s3	elles	-

Sans entrer dans les détails, ces champs représentent respectivement la représentation graphique de l'entrée, son numéro d'homographe, la représentation phonologique, le fonctionnement de la finale, le numéro d'homophone, la représentation phonologique en classes majeures, le nombre de syllabes, la fréquence, la catégorie, la variation genre/nombre, la catégorie flexionnelle, la défection temps/personne, la structure argumentale.

Le gestionnaire décrit ici a été intégré notamment à un analyseur syntaxique HPSG. L'adaptation consiste donc à générer des structures de traits à partir des informations lexicales contenues dans les entrées. L'outil a été développé en Prolog III (cf. Delpui, 1993), il est basé sur des règles de réécriture qui effectuent l'adaptation. Il pourrait être étendu afin de permettre à l'utilisateur de définir les champs utiles dans les entrées lexicales, et de spécifier les renseignements nécessaires au codage des structures de traits. Ce dernier point est décrit dans la section suivante.

3.1. Gestionnaire du lexique

La nécessité pour les systèmes de traitement du langage naturel de disposer d'une bonne couverture impose l'utilisation de lexiques très complets. Un des problèmes posés provient du fait qu'il s'agit de ressources volumineuses. Leur gestion doit être optimale, tant pour l'accès aux données que pour leur maintenance et l'évolution du lexique lui-même.

La structure interne du lexique est importante, car de ce choix va dépendre la rapidité d'accès, de modification, et d'ajout d'entrées lexicales. Nous avons choisi d'utiliser une représentation arborescente particulière appelée *Arbres 2-3*. Ceux-ci sont des cas particuliers des B-arbres :

- chaque nœud intérieur possède 2 ou 3 fils ;
- chaque chemin de la racine aux feuilles est de même longueur ;
- seules les feuilles contiennent les informations des mots ;
- si un élément A est à gauche d'un élément B dans l'arbre, alors la relation clé (A) < clé(B) est vraie. La clé est un champ particulier qui permet de définir l'ordre de classement dans l'arbre.

Ces propriétés font de cet arbre, un arbre équilibré possédant à chaque nœud au plus 3 fils. Ceci autorise un accès rapide aux données pour les consultations et les modifications. L'ajout d'un nouvel élément peut nécessiter un équilibrage de l'arbre. Ce travail s'effectue rapidement en « remontant » les données sur les nœuds internes.

L'outil que nous présentons est basé sur le lexique LGE contenant 22 930 mots. Son chargement dans l'arbre s'effectue en 9 secondes[2]. Les mots sont ordonnés suivant leur représentation graphique. Le gestionnaire a été développé en langage C. Une interface conviviale[3] permet de manipuler aisément ce lexique. Tous les accès fournis (consultation, ajout, modification) utilisent directement l'arbre.

4. Intégration à un analyseur

L'intégration du gestionnaire lexical à un analyseur repose à la fois sur la généricité des outils développés (leur portabilité, le format des données, etc.) et sur la capacité d'adaptation au format de l'analyseur lui-même. Nous avons vu dans la section précédente les aspects concernant l'accès et la gestion du lexique. Cette partie est consacrée à la présentation du filtrage nécessaire à la constitution d'un lexique dans un format donné.

4.1. Le filtrage

Le format du lexique source est considéré comme générique. Seules certaines informations seront utiles à la construction d'une entrée lexicale dans un format donné.

2. Ce résultat est obtenu sur une station DEC 5000
3. L'interface est écrite sous X11-XToolkit et Athena Widget

Pour ce qui concerne l'analyseur HPSG évoqué ici, les informations nécessaires, comme nous l'avons vu dans la première section, sont de deux types et concernent soit des informations que nous qualifierons de *discriminantes* concernant les propriétés inhérentes à la forme rencontrée, soit des informations *génériques* portant notamment sur les relations entre les structures de traits représentées par la coindexation.

L'exemple de la figure 3 indique les champs de l'entrée du lexique source pour le verbe *aimer* utilisés pour la construction de l'entrée HPSG. Nous remarquons ainsi que seuls quelques champs du source sont effectivement récupérés. Il s'agit de PHON_SYLL, CS et CONST utilisés respectivement en tant que valeurs des traits PHON, MAJ et COMPS de la structure HPSG. Le champ GRAPH_ACC du source permet quant à lui d'informer le type de la relation sémantique du trait NUCLEUS.

Ce type d'opération de filtrage est similaire pour les autres catégories syntaxiques. Il est donc possible de définir de véritables *schémas de filtrage* qui indiqueront les champs du lexique source à utiliser, leur destination dans le lexique cible et le partage des structures de l'entrée construite. Ainsi, le schéma de filtrage pour le verbe se présentera sous la forme de la figure 4.

La structure de traits décrite dans la figure 4 indique à l'aide d'un astérisque les champs du lexique utilisés en tant que valeur de trait. On remarquera de plus l'utilisation de variables pour la représentation de la coindexation.

Le sous-ensemble de champs utilisés par le filtrage est bien entendu extensible. En particulier, toutes les informations phonologiques doivent être également utilisées pour la construction d'entrées plus complètes. Ces champs sont sujets à modification en fonction du lexique source utilisé. Mais le principe du filtrage reste le même.

GRAPH_ACC	HG	PHON_SYLL	FPH	HP	CL_PHON	NS	FREQ	CS	GN	CFLB	DEFLIEN	CONST
aimer	11	*&/m&*	1"	11	E/NE	2	C2B0TR	V	01	060	1+	Ta

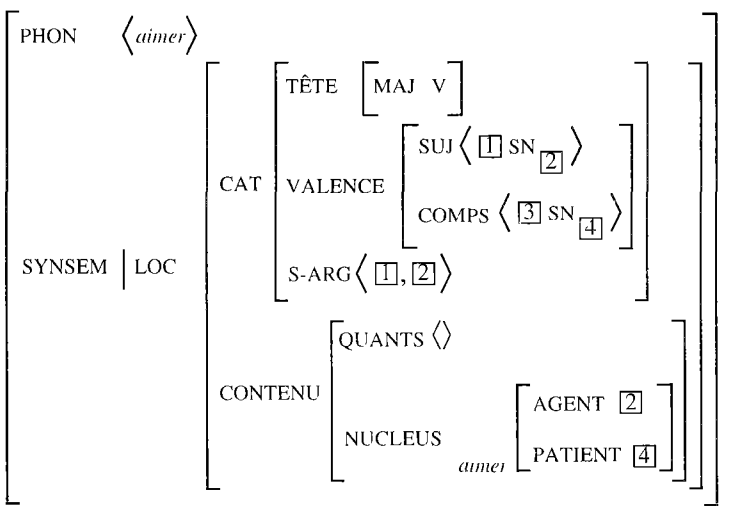


FIGURE 3 : Filtrage de l'entrée lexicale *aimer*

4.2. Implantation

L'implantation de ces schémas de filtrage se fait à l'aide d'un utilitaire permettant de gérer de façon autonome ces informations. Ce module est en effet distinct de l'analyseur lui-même dans la mesure où les schémas sont stockés sous forme de fichier textes composés de la description des structures en question. L'analyseur accède à ces informations via un ensemble de fonctions permettant leur interrogation, mais également leur manipulation. Elles sont représentées par un arbre binaire qui contient à chaque nœud :

- le trait concerné ;
- un arbre fils qui représente les sous-traits du trait ;
- un arbre frère qui représente les traits qui ont même trait père que le trait du nœud.

Ainsi, pour la structure suivante :

$$\begin{bmatrix} \text{PHON} & \\ \text{SYNSEM} & \begin{bmatrix} \text{LOC} \\ \text{NONLOC} \end{bmatrix} \end{bmatrix}$$

le nœud correspondant à SYNSEM aura :

- pour fils l'arbre composé des nœuds LOC et NONLOC ;
- pour frère l'arbre composé du nœud PHON.

Le gestionnaire permet de générer ces structures à partir des structures syntaxiques du modèle. Il fournit également une interface pour modifier, créer, visualiser les règles.

Cette méthode présente plusieurs avantages. Tout d'abord, nous avons ainsi une séparation claire des données et de l'analyseur. De plus, cet outil permet la maintenance du format des structures. On sait en effet que celles-ci, en fonction des évolutions de la théorie, peuvent être sujettes à des remaniements plus ou moins importants.

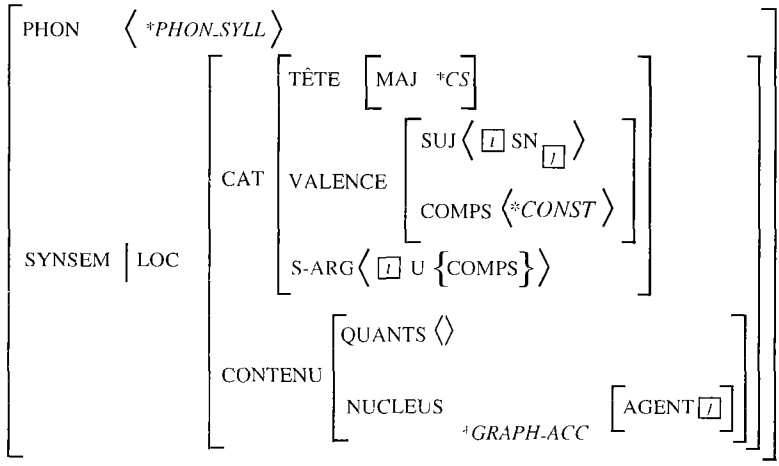


FIGURE 4 Filtrage de l'entrée lexicale *aimer*.

5. Conclusion

L'intégration du gestionnaire a été expérimentée sur plusieurs systèmes. Il a, dans sa première version, été utilisé par un correcteur grammatical reposant sur le cadre formel des GPSG (cf. Meyer, 1993). D'un point de vue pratique, le gestionnaire est écrit en C et le système en Object Pascal. Sa seconde utilisation a été expérimentée sur un analyseur HPSG écrit en Prolog III. Son intégration à un nouvel analyseur écrit en **LIFE** est en cours de développement. L'avantage dans ce dernier cas vient du fait que **LIFE** permet de manipuler directement les structures de traits et propose un mécanisme d'héritage sous forme de contraintes. Les schémas d'instanciation décrits plus haut sont donc implantés sous forme de propriétés de catégories syntaxiques héritées par les entrées lexicales elles-mêmes. L'accès au lexique se contentera de retourner les valeurs utiles.

Cette approche permet d'envisager l'utilisation d'un lexique véritablement générique. Il s'agit dans ce cas de proposer un lexique comportant un très grand nombre d'informations ainsi qu'un ensemble de modules de filtrages permettant l'adaptation des entrées à des formats donnés. Il serait ainsi possible soit de construire directement des lexiques adaptés à un formalisme donné (*i.e.* de compiler le lexique générique source) soit de maintenir un filtrage dynamique permettant une utilisation dans des plate-formes plus complexes mettant en jeu par exemple plusieurs formalismes.

Un réseau lexico-sémantique des verbes construit à partir du dictionnaire pour le traitement automatique du français

Karim CHIBOUT et Nicolas MASSON

Groupe Langage et Cognition, LIMSI-CNRS, Orsay, France

Survoler, c'est « voler au-dessus de » ;
donc survoler est en dessous de voler !

• *Abstract* •

Our work turns on computational modelling of verbal polysemy based on a hierarchical representation of verbs (in French) The final goal is to integrate into a parser a specific module which detect verbs used in a figurative sense and interpret them in order to remove semantic incoherences.

We propose a systematic and comprehensive representation of verbs in a hierarchical structure. Our rationale for a classification seen as the keystone of an interpretation system is the following : the different senses of a verb are variations semantically related ; they share a common basis, that is, a common thematic relation structure, and a common basic action. Different senses derive from inflexions in the thematic relation structure, and/or from the sense of the hyperonym, as it is recursively defined in terms of its hyperonym and its own thematic relation structure. The hierarchy of verbs depends on the thematic relations associated with them.

From this representation we have determined three major types of heuristics to help interpreting the different meanings conveyed by a verb and its hyponyms.

The verb network and the specific module which detect verbs used in a figurative sense are used in a scientific texts structuring system. In this application, structuring consists in determining rhetorical relations between sentences or sets of sentences. Some other applications are also discussed.

Introduction

Notre travail porte sur la résolution des incohérences sémantiques liées à la polysémie des verbes. Le principal but est de réaliser un module informatique « expert » dans la

détection et le traitement de sens figurés au niveau de la phrase. On se limitera à l'étude des verbes d'action.

La modélisation de la polysémie est menée selon deux axes complémentaires :

- la représentation des connaissances lexicales ;
- les processus d'interprétation des différents sens susceptibles d'être associés à un prédicat et les modalités de discrimination entre ces significations dans un énoncé donné.

Après un aperçu de quelques études sur les métaphores qui justifient notre approche, nous présenterons les liens entre les verbes au sein d'un réseau lexical et la structure sémantique associée à chacun d'eux. Nous définirons ensuite le modèle de la polysémie proposé à partir de ce formalisme de représentation. Nous terminerons par les autres applications possibles du formalisme élaboré, et en particulier l'aide à la structuration de textes scientifiques.

1. D'une figure de sens aux sens figurés

Nous partons de l'étude d'énoncés mettant en jeu des verbes employés métaphoriquement. La métaphore verbale a été étudiée en particulier par Wilks (1978), qui a réalisé un programme destiné à résoudre les incohérences sémantiques propres aux métaphores verbales via une analogie pertinente entre les termes. La phrase suivante montre le principe général de son système :

My car drinks gasoline
(Ma voiture « boit » de l'essence)

nous sommes en présence d'une anomalie de sens dans la mesure où le verbe *drink* attend un *animé* comme cas agent et un cas objet de type *liquide*. Une métaphore est mise en évidence par la comparaison du verbe *drink* et d'un sens de *use* qui prend comme objet *gasoline*.

Une primitive sémantique commune aux deux verbes (Expend) permet le remplacement de *drink* par *use* dans la phrase. [Expend] désigne une action de base signifiant **dépenser, consommer (énergie, argent...)**, qui s'applique tout aussi bien à *boire* pour un animé qu'à *consommer* pour une voiture. Le verbe *boire* peut prendre le sens d'*ingurgiter* ou de *consommer*. C'est ce deuxième sens qui convient pour cet exemple, l'acte d'*ingurgiter* étant propre aux animés. Par conséquent, le verbe *use* est retenu pour l'interprétation :

My car uses (consomme) gasoline.

D'autres programmes sont fondés sur un principe similaire ; en particulier le système Met* (Metstar) de D. Fass (1991) qui résout, dans sa forme actuelle, les métaphores verbales fondées sur un lien d'analogie entre deux termes et un nombre réduit de métonymies.

Ce type d'analyse permet d'accepter certains énoncés métaphoriques dans lesquels les préférences de sens sont violées.

Nous constatons que dans une organisation hiérarchique de verbes, plus on descend dans la hiérarchie plus les relations casuelles vont être spécifiques et imposer une signification précise (p. ex. *drink*), et *a contrario*, plus on s'élève dans la hiérarchie, moins ces contraintes casuelles vont jouer (p. ex. *use*). Un des modes de résolution des métaphores va donc consister à relâcher certaines des contraintes imposées sur les relations casuelles d'un verbe en remontant dans l'arbre.

Notre propre étude, qui a donné lieu à une implantation sous forme d'un module d'interprétation d'énoncés métaphoriques (Chibout, 1994), montre que ce processus ne concerne qu'un nombre restreint de métaphores verbales.

Les verbes métaphorisés contenant dans leur description **la manière** particulière dont une action est effectuée (exprimée généralement par un adverbe) imposent une interprétation différenciée. En plus de la relaxation de contraintes par la récupération d'un hyperonyme, il est nécessaire, pour une interprétation correcte de la figure, de rattacher l'adverbe de manière à l'hyperonyme. L'exemple suivant répond à ce mécanisme :

La mer, devenue calme, ***berçait*** *l'embarcation de fortune.*

Le verbe *bercer* a pour description grossière :

balancer (hyperonyme immédiat) *lentement* (manière) *pour endormir* (but).

Bercer a pour cas agent un *humain* et comme cas objet un *humain*. Le système détecte une incohérence sémantique sur les deux liens casuels.

La résolution consiste à rattacher à l'hyperonyme immédiat (que nous appellerons **action de base**) le trait manière afin d'obtenir un énoncé plus sensé :

La mer, devenue calme, ***balançait lentement*** *l'embarcation de fortune*

Cette idée de « glissements » de sens à partir d'une description sémantique nous a amené à ne plus modéliser une figure particulière isolée du reste des faits de langue mais à aborder la **polysémie** dans son ensemble.

Wilks (1978) et Fass (1991) se sont attachés davantage à l'interprétation des métaphores, qu'à la représentation des connaissances nécessaires à leur résolution. Or il nous semble, au regard des exemples cités, qu'une représentation lexicale suffisamment élaborée est le déterminant majeur pour l'interprétation, c'est-à-dire que les mécanismes à mettre en œuvre en découlent.

Par conséquent, dans notre démarche, nous avons recherché, pour la représentation des connaissances lexicales, un formalisme suffisamment souple pour rendre compte de la polysémie, et suffisamment puissant en termes d'opérations sur les représentations, pour pouvoir déduire aisément la dynamique d'interprétation de la polysémie verbale.

En effet, dans quelle mesure la description sémantique d'un verbe ne pourrait-elle pas être utilisée pour traiter les sens figurés portés par celui-ci ?

Dans les exemples donnés, la description d'un verbe qui semble pertinente pour le traitement des métaphores implique son action de base et des composants de sens qui vont spécifier ce verbe.

Nous reprenons donc ces éléments pour l'étude de la polysémie en créant, d'une part, un réseau lexical hiérarchique pour rendre compte des relations d'hyponymie/hyperonymie entre verbes et, d'autre part, une représentation détaillée associée à chaque verbe sous la forme d'une structure de cas sémantiques.

2. Représentations des verbes

2.1. Quelques travaux

Les classifications du lexique fondées sur des bases sémantiques sont relativement nombreuses en intelligence artificielle. Les réseaux sémantiques sont un des principaux modes de représentation des connaissances lexicales ; mais ils concernent essentiellement les substantifs.

Il existe quelques tentatives d'organisation hiérarchique fine des verbes pour l'anglais (Talmy, 1985), (Miller et Fellbaum, 1991 ; Miller *et al.*, 1989). Les auteurs soulignent par ailleurs la complexité de la tâche, en raison notamment des différents champs sémantiques impliqués dans les relations entre les verbes étant apparentés au niveau du sens.

Talmy (1985) fournit certains critères sémantiques utiles à la constitution d'une taxonomie. Son analyse des verbes de mouvements est la plus significative de ce point de vue. Il met en évidence une primitive de mouvement commune à tous les verbes et d'autres composants sémantiques comme la *manière* et la *cause*, à l'instar de *slide* (glisser) et *pull* (tirer) respectivement. À ces composants peuvent s'ajouter la *rapidité* (*run, stroll*) ou le *moyen de transport* utilisé (*bus, truck, bike,* etc.).

D'une manière similaire, les verbes qui dénotent l'action de frapper (*hit*) peuvent exprimer le *degré de force* utilisé par l'agent (*chop, slam, whack, swat,* etc.).

D'autres verbes font référence au *degré d'intensité* d'une action ou d'un état (*drowse, doze, sleep,...*).

Les travaux en psycholinguistique de Miller et ses collaborateurs (*op. cit.*) portent sur les relations sémantiques liant les mots de trois catégories lexicales : les noms, les adjectifs et les verbes. Les mots de chaque catégorie sont reliés entre eux par un ensemble de relations telles que synonymie, antonymie, hyperonymie/hyponymie, etc.

Concernant les verbes, les auteurs reprennent les travaux de Talmy et définissent une relation de particularisation permettant de réunir les différents composants sémantiques qui distinguent un verbe de son superordonné (hyperonyme). Cette relation

entre deux verbes (V1 hyponyme de V2) est appelée **troponymie** (du grec *tropos* : manière, façon) et est exprimé par la formule « accomplir l'action V1, c'est accomplir l'action V2 d'une manière particulière ».

Le verbe *fight* a pour troponymes *battle, war, tourney, duel,...* Les troponymes des verbes de communication impliquent l'intention du locuteur ou sa motivation à communiquer, comme dans *examine, confess, preach,* ou le moyen de communication utilisé : *fax, email, phone, telex,...*

Leur étude a donné lieu à une mise en œuvre informatique d'un réseau lexical (Wordnet) qui organise les mots des différentes catégories lexicales en termes de signifiés.

Pour la catégorie qui nous intéresse, ce travail constitue, à notre connaissance, une des rares tentatives de hiérarchisation systématique de verbes dans un réseau sémantique.

2.2. Notre classification

Partant des travaux exposés ci-dessus, un dictionnaire de type terminologique nous semble le seul outil qui puisse expliciter les composants de sens nécessaires à la constitution d'un réseau lexico-sémantique des verbes d'action.

Les définitions sont en effet des descriptions assez précises qui incluent en général l'hyperonyme et font des renvois aux verbes qui ont une parenté de sens avec celui défini. On y retrouve également les différents composants sémantiques (moyen, manière, but, etc.) qui vont spécifier le verbe par rapport à son superordonné.

Les définitions ne sont cependant pas toujours homogènes dans leur structure et leur contenu de sorte qu'on ne peut pas s'appuyer totalement sur le dictionnaire pour réaliser notre réseau. Des dissimilarités se retrouvent dans les définitions selon le dictionnaire utilisé.

N. Wurbel (1995 : 50), qui a étudié un grand nombre de dictionnaires pour définir un automate de génération de définitions pour les substantifs, montre en particulier que les « dictionnaires courants [...] ont des niveaux de description d'informations extrêmement variés ».

Ce phénomène se retrouve pour les verbes. Ainsi le sens habituel de *couper* est formulé de différentes façons selon le dictionnaire consulté :

- **Collins Cobuild English Dictionary** (Collins, 1995) :
  Cut (something) : ...use a knife or a similar tool to divide it into pieces,...
- **Hachette, dictionnaire de notre temps** (Hachette, 1993) :
  diviser avec un instrument tranchant.
- **Le Petit Littré** (Beaujean, 1990) donne une définition proche de la précédente :
  diviser un corps avec un instrument tranchant.
- **Le Petit Robert** (Robert, 1994) :
  Diviser (un corps solide) avec un instrument tranchant.

Au travers de ces définitions, on constate principalement des différences dans la précision des descriptions du prédicat. Chacune donne l'action générique (*diviser*) et le moyen propre à l'acte de couper (*instrument tranchant*). Mais en ce qui concerne l'objet de l'action, seule la définition donnée par *Le Petit Robert* comporte une caractéristique importante de l'objet (*corps solide*).

L'exemple suivant montre de manière encore plus significative les disparités qui peuvent exister entre les dictionnaires.

Le sens propre de *élaguer* est défini comme suit :

- **Le Petit Robert :**
  dépouiller (un arbre) des branches superflues sur une certaine hauteur.
- **Le Petit Littré** :
  couper les branches, principalement les branches inférieures d'un arbre.
- **Collins Cobuild English Dictionary :**
  prune a tree or bush : cut off some of the branches so that it will grow better the next year.
- **Hachette, dictionnaire de notre temps :**
  débarrasser (un arbre) des branches nuisibles à son développement, à sa fructification, etc.

Les dissemblances entre les dictionnaires sont ici plus marquées : ils ne font pas référence à la même action (*couper, débarrasser, dépouiller*), ni au même qualificatif pour l'objet de l'action (branches *superflues, inférieures, nuisibles,...*)

L'hétérogénéité dans la caractérisation des termes nous a conduit à ne nous appuyer que partiellement sur cet outil terminologique et à proposer une réorganisation des définitions.

Nous suggérons une systématisation des définitions en s'imposant des règles explicites : les verbes sont décrits par l'intermédiaire d'une structure canonique de traits sémantiques. Plus précisément, tout verbe est exprimé par son **action de base** (hyperonyme immédiat) et un certain nombre de cas sémantiques qui la spécifient. Outre les liens casuels habituels (agent, objet, moyen,...), 4 cas apparaissent essentiels pour la description fine d'un verbe d'action :

– la **manière** dont l'action est réalisée (exprimée par un adverbe) ;
– la **méthode** par laquelle elle s'effectue (exprimée par une action) ;
– son **résultat** (représenté par un état) ;
– le **but** intrinsèque à cette action (représenté par un verbe d'action).

La procédure consiste à récupérer les différentes définitions (sens propre) d'un verbe dans différents dictionnaires avec l'hypothèse que le sens maximal est disséminé dans ceux-ci.

Quand un composant vient à manquer dans les définitions mais qu'il est présent sous la forme de sens figuré et qu'il a une pertinence par rapport à la définition, nous l'intégrons dans notre représentation.

C'est le cas du verbe *couper* (ci-après), pour lequel la **méthode** et le **but** n'apparaissent dans aucun des dictionnaires consultés :

couper : *diviser* (action de base) *objet solide* (objet) *en plusieurs parties* (résultat) à *l'aide d'un instrument tranchant* (moyen) *en traversant l'objet* (méthode) *pour en enlever une partie ou en réduire la taille* (but).

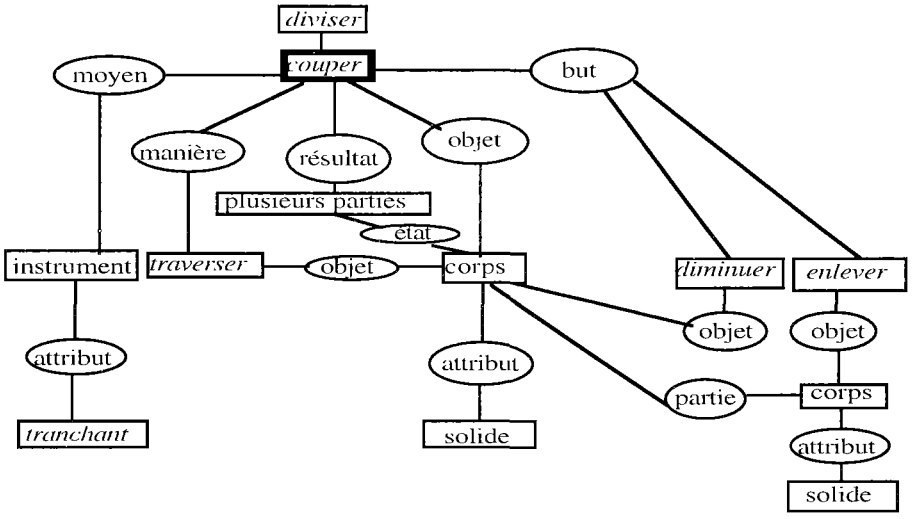


FIGURE 1 Représentation (action de base + structure casuelle) du verbe *couper*.

De plus, ces cas sémantiques permettent, nous allons le voir, de définir plus précisément des critères de hiérarchisation des verbes.

La hiérarchisation des verbes est faite en fonction des relations casuelles qui leur sont associées. Un verbe est hyperonyme (respectivement hyponyme) d'un autre s'ils ont une action de base commune et s'il y a dans la structure casuelle de ce verbe (cf. figure 2) :

1. **absence** (respectivement **présence**) d'une valeur définie pour un cas particulier ;
2. présence d'un cas à **valeur multiple** (respectivement à **valeur unique**) ;
3. présence d'un cas à **valeur générique** (respectivement à **valeur spécifique**).

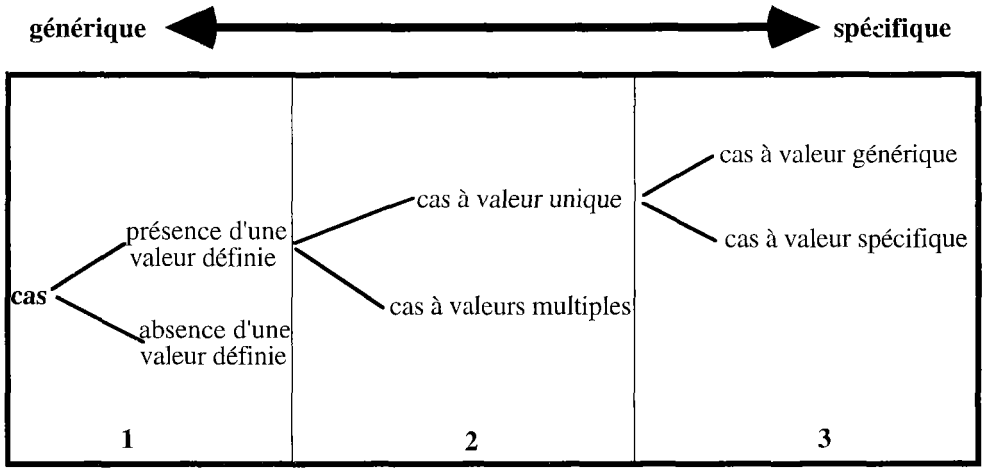


FIGURE 2 : Les valeurs portées par les relations casuelles comme critères de classification des verbes.

1. Exemple : pour la paire **diviser/couper** ; couper, c'est diviser avec un instrument tranchant (cas moyen défini), donc *diviser* est hyperonyme de *couper* (noté *diviser* > *couper*) ;
2. Exemple : pour **couper/élaguer** ; *couper>élaguer* parce que le cas but du premier peut être *enlever une partie* (couper les bords de pages), *réduire la taille* (couper du bois) ; alors que le but de *élaguer* est unique : *enlever* (les branches inutiles d'un arbre) ;
3. Exemple : **couper/décapiter**, couper a un cas objet à valeur générique : *objet solide* / alors que l'objet de décapiter est relativement spécifique (*tête*), donc *couper*> *décapiter*.

Le réseau apparaît comme une structure en arbre (cf. figure 3 ci-dessous). Les verbes décrivant des situations du monde, nous plaçons en haut de la hiérarchie **l'événement** qui définit les deux grands types de verbes (d'action et d'état).

Pour la catégorie verbes d'**action**, **faire** est hyperonyme de toutes les primitives sémantiques (p. ex. **faire cesser**) à partir desquelles est déterminée la hiérarchie des concepts verbaux, des plus généraux (p. ex. **séparer**) aux plus spécifiques (p. ex. **guillotiner**).

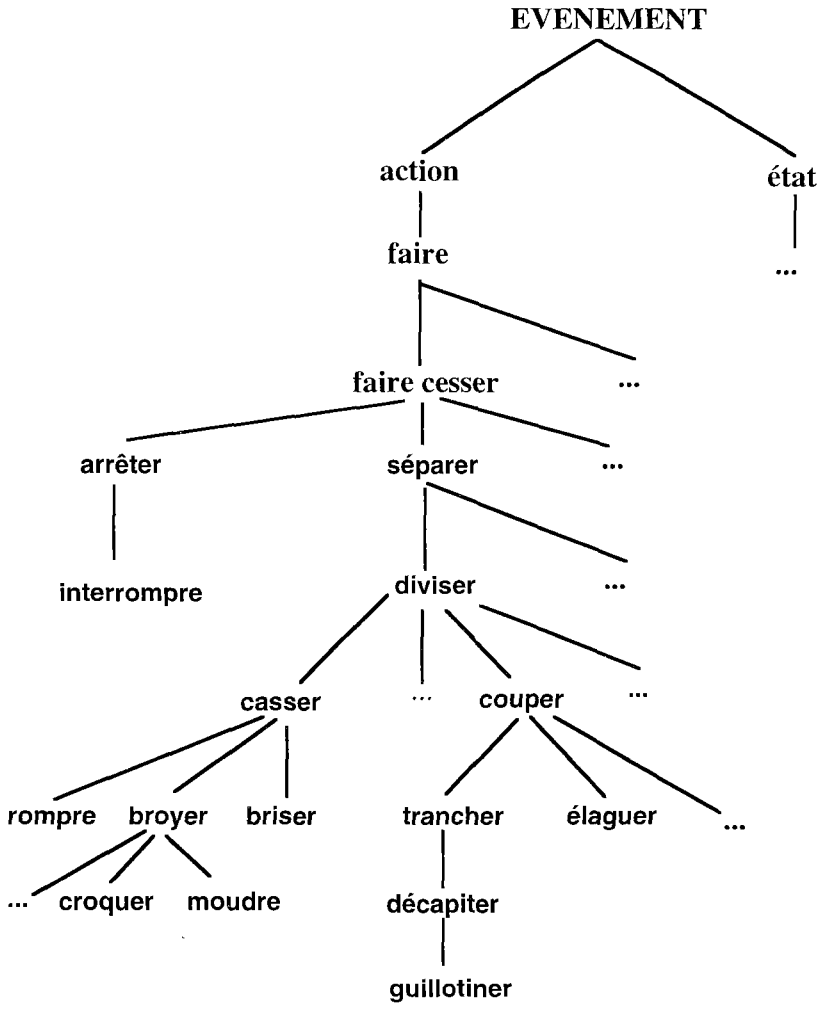


FIGURE 3 . Vue locale du réseau.

Le réseau et les schémas descriptifs des verbes sont principalement destinés au traitement de la polysémie verbale. La partie qui suit explicite le modèle de la polysémie envisagé et notamment les heuristiques mises en évidence pour l'automatisation de l'interprétation des sens figurés.

3. Application à la polysémie des verbes

3.1. Point de vue sur la polysémie

Nous formulons un certain nombre d'hypothèses à la base de notre étude

- Il existe des règles explicites (ou explicitables) sous-jacentes à la construction de

la polysémie des verbes.
- Ces règles sont en nombre limité.
- Les règles déterminant la polysémie des verbes prennent appui sur la description de l'action et les différentes caractéristiques (modalités de réalisation) propres à cette action. Ces modalités correspondent aux relations casuelles que nous avons définies auparavant.

Les sens figurés sont principalement contenus dans la définition du sens propre, telle peut être en résumé notre conception de la polysémie. De ce point de vue, une description sémantique fine d'un verbe permettra de retrouver l'ensemble, ou tout au moins une grande partie de ses sens.

Les différentes significations d'un verbe sont vues comme des variations sémantiquement apparentées ; elles partagent un support commun : la structure casuelle et hiérarchique du verbe telle que nous l'avons définie. Ces significations dérivent de l'ossature casuelle du verbe et/ou de la signification de son (ou d'un de ses) hyperonyme(s). De par l'organisation hiérarchique du réseau, un hyperonyme est lui-même défini récursivement par son hyperonyme et sa propre structure casuelle.

Or, si les différents sens d'un mot dérivent d'une même représentation, quels sont les **modes d'interprétation** à l'origine de cette dérivation de sens ?

3.2. Heuristiques de résolution

À partir de cette représentation, nous avons déterminé trois catégories d'heuristiques pour l'interprétation (sans toutefois prétendre à l'exhaustivité) des sens multiples véhiculés par un verbe et ses sous-ordonnés :

- **récupération d'un hyperonyme**
  - l'action de base
    Les coulées de lave marchaient vers le village (SE DÉPLACER)
  - un hyperonyme plus général
    Jean coupa l'eau (ARRÊTER)

Arrêter est récupéré à partir de la primitive sémantique **faire cesser** commune à *couper* et *arrêter* (cf. le réseau local au verbe couper figure 3). Notons que dans l'exemple la substitution de *couper* par *arrêter* n'est pas suffisante du fait de la présence d'une figure elliptique (arrêter l'écoulement de l'eau).

- **attachement d'un cas à l'hyperonyme récupéré**
  - l'action de base
    Les vagues couraient jusqu'aux rochers (se déplacer RAPIDEMENT)
  - un hyperonyme plus général
    Il brisa leur conversation (interrompre BRUSQUEMENT)

De la même manière que pour couper l'eau, interrompre est accessible via la primitive **faire cesser** qu'ils partagent.

- **transfert d'un cas comme action effective en remplacement du verbe polysème traité.**

  - Le cas méthode

Ce mode de résolution est illustré par l'exemple suivant :

Le paysan coupa par le champs (TRAVERSER)

dans lequel *traverser* se substitue à *couper*, parce que ce dernier contient dans sa description détaillée le cas méthode « en traversant l'objet » (cf. figure 1).

  - Le cas manière

*Le vieil arbre **tranche** avec le paysage*

Trancher a, en particulier, pour action de base *couper* et pour cas manière *distinctement.* En redonnant une forme verbale au cas manière, on obtient après substitution :

*Le vieil arbre se **distingue** du paysage*

  - Le cas but

Il a élagué son exposé (ENLEVER LES PARTIES INUTILES)

élaguer : *couper* (action de base) enlever les branches inutiles (but).

Ces heuristiques représentent autant de « glissements » de sens à partir de la structure unique qui définit le verbe.

Polysémie locale récurrente

Il est apparu que certains sens portés par un verbe sont également véhiculés par des verbes localement proches (*i.e.* ayant un hyperonyme en partage) dans le réseau.

Par exemple, les verbes *couper, entrecouper, briser, hacher* qui relèvent tous de la primitive sémantique **faire cesser** ont un sens figuré commun :

couper la parole (INTERROMPRE)
briser une conversation (INTERROMPRE BRUSQUEMENT)
entrecouper ses phrases de sanglots (INTERROMPRE FRÉQUEMMENT)
hacher un discours (INTERROMPRE DE FAÇON RÉPÉTÉE)

Ces heuristiques, qui répondent à la catégorie **récupération d'un hyperonyme**

avec en sus pour les trois derniers exemples, l'attachement du cas manière spécifique à chacun des verbes, valident partiellement le principe d'une hiérarchisation du lexique.

Cette récurrence de sens locale nous conforte donc dans l'adoption d'une structure arborescente pour représenter les concepts verbaux.

3.3. Les limites du modèle

Les heuristiques proposées ne permettent pas de résoudre la totalité des modes polysémiques des verbes. Mais si nous ne prétendons pas à l'exhaustivité, compte tenu de la complexité du phénomène, nous souhaitons parvenir à une meilleure couverture de la polysémie.

En outre, il reste à déterminer précisément les conditions de choix entre les différentes heuristiques. En effet dans quelle(s) condition(s) appliquer telle ou telle des heuristiques définies ?

Nous savons d'ores et déjà que les catégories sémantiques des valeurs portées par les cas agent et objet (quand ils existent) vont servir de filtres pour mettre en évidence un éventuel sens figuré du verbe (incohérence sémantique).

L'incohérence si elle porte sur l'un ou l'autre de ces cas ne met-elle pas en jeu des heuristiques différenciées ?

Les catégories sémantiques de ces éléments n'impliquent-elles pas des modes de résolution particuliers ? On pense notamment aux deux dimensions sémantiques : **abstrait/concret** et **animé/inanimé.**

La détection de sens figurés pose elle-même problème dans la mesure où la violation des préférences casuelles n'est pas systématique. L'exemple qui suit est significatif de ce problème :

Il a coupé les cartes.

Malgré le fait que les contraintes casuelles sont respectées, il subsiste une ambiguïté sur le sens du verbe *couper.* S'agit-il de couper effectivement les cartes ou de diviser le paquet, comme il se fait d'ordinaire au début d'une partie de cartes ?

Le modèle est limité au traitement des sens figurés induits par une incohérence sémantique entre les valeurs attendues par les cas du verbe et les valeurs en présence dans l'énoncé. Les ambiguïtés de sens vont être considérées comme relevant d'un niveau d'interprétation plus poussé fondé sur des règles sémantico-pragmatiques qui restent à préciser.

Il existe des expressions dont l'interprétation est subordonnée à l'application d'une suite de traitements. Nous les appellerons **sens figurés en chaîne**, généralisant ainsi la notion de chaîne proposée par Reddy (1979) pour la figure métonymique.

Ce type d'expressions nécessite parfois des structures de représentation pour les **substantifs** afin de mettre en œuvre des règles propres à traiter des sens figurés liés à ceux-ci.

décapiter une entreprise

Pour retrouver le sens de *décapiter*, il faut non seulement récupérer le cas but (détruire) de ce verbe, mais également conserver la valeur du cas objet (*tête*) utilisé dans un de ses sens métaphoriques et le traiter comme tel.

4. Un réseau utile pour d'autres applications du TAL

Aide à la structuration de textes scientifiques

Le réseau de verbes sera utilisé dans le cadre d'un système de structuration de textes scientifiques en vue de réaliser la génération automatique de résumés. La structuration consiste ici à déterminer et étiqueter les relations rhétoriques (de type argumentatives et logiques) entre des phrases ou des ensembles de phrases du texte. Ces relations sont en nombre fini et expriment la *causalité*, le *contraste*, le *renforcement d'idée*, l'*illustration*, etc. La structuration réalisée s'appuie sur deux modules :

- un module de repérage des bornes thématiques du texte. Celui-ci utilise des techniques statistiques qui portent sur l'analyse de la distribution des occurrences nominales dans le texte ;

- un module qui détermine le rôle des phrases les unes par rapport aux autres, c'est-à-dire établir les liens rhétoriques ou argumentatifs entre les phrases. Ce module utilise les résultats du précédent et met en œuvre une analyse linguistique de surface fondée sur un modèle linguistique d'analyse de textes scientifiques que nous avons développé (Masson, 1995). À ce niveau nous analysons différents phénomènes linguistiques tels que les *portées temporelles*, les *connecteurs*, certaines *expressions remarquables* (p. ex. *En conclusion...)* ainsi que les *verbes référents à l'argumentation*. C'est pour l'analyse de ces derniers que l'on utilise le réseau de verbes et le module de résolution de la polysémie précédemment explicités.

À chaque relation sont associés des verbes qui expriment la signification de la relation (p. ex. **renforcement d'idée** : *confirmer*). La détection, dans une phrase du texte, d'un verbe attaché à une relation (p. ex. *confirmer, étayer, renforcer, appuyer...*) donne le rôle argumentatif de la phrase. L'exemple suivant montre une relation de renforcement d'idée entre la première et la deuxième phrase.

> *(...) les peuples de ces régions l'utilisent [l'asphalte] dans le bâtiment, pour la réalisation de jouets, d'objets d'art, d'objets d'usage domestique. L'analyse détaillée des échantillons de bitume des diverses sources géologiques par des méthodes isotopiques et moléculaires* ***a confirmé*** *cette hypothèse.* [« Les bitumes de Suse », *Pour la Science*, n° 204, octobre 1994].

Dans cet exemple, le verbe (*confirmer*) est un verbe d'argumentation strictement, c'est-à-dire qu'il est utilisé au sens littéral. Imaginons maintenant qu'à la place de confirmer ont ait eu le verbe *étayer*, soit une phrase du type :

> *L'analyse détaillée des échantillons de bitume des diverses sources géologiques par des méthodes isotopiques et moléculaires* ***étaye*** *cette hypothèse.*

Dans ce cas, le verbe *étayer* est utilisé au sens figuré, soit « *renforcer une hypothèse* », et non au sens littéral, soit « *supporter par des étais* ». Pour allouer une fonction argumentative ou rhétorique à certains verbes, il faut donc pouvoir résoudre les problèmes de polysémie.

Le réseau de verbes, accompagné du module de résolution de la polysémie verbale, permet de repérer les verbes utilisés au sens figuré susceptibles alors de référer à une fonction argumentative ou rhétorique.

Lors de la tâche de résumé, il sera également possible d'utiliser le réseau de verbes comme « dictionnaire des synonymes » pour des reformulations ou paraphrases (étape de « lissage ») dans le résumé engendré.

4.2. Autres applications

D'autres applications sont envisageables pour le réseau et les schémas descriptifs des verbes.

Les réseaux sémantiques des noms sont en général établis indépendamment des verbes. Puis, on tente de rattacher aux prédicats verbaux les catégories sémantiques ainsi définies comme valeurs de relations casuelles. Cette manière de procéder semble erronnée au regard des restrictions imposées par les verbes.

Considérons par exemple le verbe *verser* auquel on associe généralement un *liquide* comme objet de l'action. Ce verbe admet, avec une signification similaire, des concepts de catégories sémantiques autres que liquide, *poudres* ou *granulés* par exemple. Il est donc essentiel d'envisager la construction des catégories des concepts nominaux relativement aux restrictions de sélection imposées par les verbes de manière à obtenir une cohérence entre les noms (leurs catégories et leurs propriétés) et les valeurs des relations casuelles attendues par les verbes. La difficulté de cette approche vient de ce qu'il n'existe pas toujours de termes dans la langue pour désigner des concepts aussi hétérogènes que *liquide* et *granulés*. Il faudra par conséquent créer des types conceptuels à l'aide de termes d'un métalangage ou d'un pseudo-langage qui regrouperaient les catégories en question.

Le réseau peut être utilisé comme outil de génération automatique de définitions. Il suffit pour cela de lier un descripteur (entre crochets dans l'exemple qui suit) à chaque trait sémantique du verbe à définir, afin de reconstruire une définition correcte du point de vue syntaxique.

Pour le verbe *couper* : diviser **[un]** objet solide **[en]** plusieurs parties **[au moyen d'un]** instrument tranchant...

La structure en arbre autorise aussi des renvois à tous les verbes ayant un apparentement de sens et offre donc la fonction de dictionnaire des synonymes. Associé à la description de chaque verbe, le réseau permet d'accéder à la paraphrase (*i.e.* exprimer de plusieurs manières un même contenu sémantique) : le verbe *décapiter* par exemple, peut être reformulé en *couper la tête* ou encore *trancher la tête*.

Conclusion

En nous fondant sur l'étude des modes d'interprétation de la métaphore verbale, nous avons proposé un formalisme de représentation des connaissances lexicales pertinent pour le traitement de la polysémie des verbes (descriptions maximalisées des verbes et construction d'un réseau hiérarchique). Nous avons montré qu'une hiérarchisation des verbes est d'emblée possible en fonction de la présence de valeurs pour les cas et de leur degré de spécificité. Enfin nous avons fait apparaître quelques heuristiques d'interprétation des sens figurés des verbes et de leurs sous-ordonnés.

Il reste à définir l'ensemble des règles pour le traitement de la polysémie et les conditions de choix entre les différents sens d'un verbe donné.

Nous souhaitons élargir l'étude aux autres catégories grammaticales (noms, adjectifs, adverbes) parce que ces représentations sont à notre sens indispensables à l'analyse sémantique de phrases et de textes : fréquence des phénomènes de paraphrases et de sens figurés dans les différents écrits (scientifiques, littéraires, journalistiques...). La robustesse des systèmes informatiques de traitement des langues (en analyse comme en génération) dépend grandement de leur capacité à intégrer la dimension polysémique du lexique.

Comme l'évoque G. Sabah (1988), la polysémie reste encore un des aspects du langage naturel qui pose le plus de difficultés en informatique linguistique.

Références

ABEILLÉ, A. (1993) : *Les nouvelles syntaxes. Grammaires d'unification et analyse du français*, Paris, Armand Colin.

ABEILLÉ, A. et Y. SCHABES (1989) : « Parsing Idioms in lexicalized TAGs », *Proceedings of EACL-89*, Manchester, pp. 1-9.

Acad (1694) : *Dictionnaire de l'Académie française*, 1re éd., Paris, Coignard, 1694 ; 2e éd., Paris, Coignard, 1718 ; 3e éd., Paris, Coignard, 1740 ; 4e éd., Paris, Brunet, 1762 ; 5e éd., Paris, Smits, 1798 ; 6e éd., Paris, Firmin-Didot, 1835 ; 7e éd., Paris, Firmin-Didot, 1878 ; 8e éd., Paris, Hachette, 1932-1935.

AGARWAL, R. (1994) : « (Almost) Automatic Semantic Feature Extraction from Technical Text », *Proceedings of the Human Language Technology Worskshop*, Plainsboro, (N. J.), 8-11 mars, pp 378-382

AGIRRE, E., ALEGRIA, I., ARREGI, X., ARTOLA, X., DIAZ DE ILARRAZA., SARASOLA, K. et M. URKIA (1989) : « Aplicación de la morfología de dos niveles al euskara », *SEPLN*, vol. 8, Barcelona, pp. 87-102.

AHA, D. W., KIBLER, D. et M. K. ALBERT (1991) : « Instance-Based Learning Algorithms », *Machine Learning*, 6, pp. 37-66.

AHLSWEDE, T. & M. EVENS (1988) : « Generating a Relational Lexicon from a Machine-Readable Dictionary », *International Journal of Lexicography*, 1 (3), pp 214-237.

AHO, A. V. & M. J. CORASICK (1975) : « Efficient String Matching : An Aid to Bibliographic Search », *Comm. of the ACM*, 18 (6), pp 330-340.

AHO, A. V., KERNIGHAN, B. W. et P. J. WEINBERGER (1988) : *The AWK Programming Language*, AT&T Bell Laboratories, Murray Hill, New Jersey

ALI, Nabil (1988) : *Al-Lugha Al -'Arabiyya wa Al Hâsûb*, Dâr Taarîb.

ALI, Nabil (1994) : *Al- 'Arab wa 'Asr Al Ma'lûmât*, Al Koweït, 'Alam Al Ma'rifa, 48.

ALLEN, M. (1978) : *Morphological Investigations*, unpublished Doctoral dissertation, University of Connecticut, Storrs

ALLERTON, David J. (1982) : *Valency and the English Verb*, London, Academic Press.

ALONSO RAMOS, M. (1993) : *Las funciones lexicales y el modelo lexicografico de I. Mel'čuk*, Universidad Nacional de Educacion a Distancias, Madrid, thèse de doctorat.

ALONSO RAMOS, M. et S. MANTHA (1996) : « Description lexicographique des collocatifs dans un *Dictionnaire Explicatif et Combinatoire (DEC)* : articles de dictionnaire autonomes ? », *Actes de Lexicomatique et dictionnairiques*, Lyon, 28-30 septembre 1995.

ALONSO RAMOS, M. et A. TUTIN (1996) : « A Classification and Description of the Lexical Functions for the Treatment of LF Combinations », L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam, Benjamins.

ALONSO RAMOS, M., TUTIN, A. et G. LAPALME (1995) : « Lexical Functions of the *Explanatory Combinatorial Dictionary* for the Lexicalization in Text Generation », P. Saint-Dizier et E. Viegas (Eds), Cambridge University Press.

ALSHAWI, H. (Ed.) (1992) : *The Core Language Engine*, Cambridge (MA), The MIT Press, 322 p.

ALSHAWI, H., ARNOLD, D.-J., BACKOFEN, R., CARTER, D.-M., LINDOP, J., NETTER, K., PULMAN, S.-G., TSUJII, J. et H. USZKOREIT (1991) : *Eurotra ET6/1 : Rule Formalism and Virtual Machine Design Study (Final Report)*, Luxembourg, CEC.

AMSLER, R. A. (1980) : *The Structure of the Merriam-Webster Pocket Dictionary*, Ph.D. Thesis, University of Texas at Austin, Austin.

ANDERSON, Stephen (1977) : « On the Formal Description of Inflection », *Papers from the XIIIth Regional Meeting of the Chicago Linguistic Society*, Chicago, pp. 15-44.

ANDERSON, Stephen (1985) : « Inflectional Morphology », T. Shopen (Ed.), *Language Typology and Syntactic Description*, Part III, Grammatical Categories and the Lexicon, Cambridge University Press, pp. 150-201

ANDERSON, Stephen (1985) : « Typological Distinction in Word Formation », T. Shopen (Ed), *Language Typology and Syntactic Description*, Part III, Grammatical Categories and the Lexicon, Cambridge University Press, pp. 3-56

ANGLUIN, D. (1980) : « Inductive Inference of Formal Languages from Positive Data », *Information and Control*, 45, pp. 117-135.

ANTONY-LAY, M.-H., FRANCOPOULO, G. & L. ZAYSSER (1994) : « A Generic Model for Reusable Lexicons : The GENELEX Project », *Linguistics and Literary Computing*, vol. 8.

ANTWORTH, E. L. (1990) : *PC-KIMMO : A Two-Level Processor for Morphological Analysis*, Dallas (TX), Summer Institute of Linguistics, 273 p.

APPELT, D. E., HOBBS, J. R., BEAR, J., ISRAEL, D. et M. TYSON (1993) : « FASTUS : A Finite-state Processor for Information Extraction from Real-world Text », *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambéry (France), 28 août-3 septembre, pp 1172-1178.

ARMSTRONG-WARWICK, S. (1993) : « Acquisition and Exploitation of Textual Resources for NLP », *Proceedings of Knowledge Base and Knowledge Systems*, Tokyo.

ARNTZ, R. (1993) : « Terminological Equivalence and Translation », Sonneveld, H. et K. Loening (Eds), *Terminology Applications in Interdisciplinary Communication*, Amsterdam et Philadelphia, John Benjamins Publishing Company, pp 5-19.

ARONOFF, Mark (1974) : *Word-Structure*, unpublished Doctorate dissertation, MIT.

ARONOFF, M. (1976) : *Word Formation in Generative Grammar*, Linguistic Inquiry, Monograph one, Cambridge (MA), The MIT Press

ARONOFF, Mark (1978) : « Lexical Representations », *Papers from the Parasession on the Lexicon · XIVth Regional Meeting of the Chicago Linguistic Society*, Chicago, pp. 12-26.

ASSAL, Allal et al. (1992) : « Sémantique et terminologie . sens et contextes », *Terminologie et Traduction*, n° 2/3, éd. Commission des Communautés européennes, pp. 411-421.

ATKINS, B. T. & A. DUVAL (1978) : *Robert & Collins Dictionnaire Français-Anglais, Anglais-Français*, Paris, Le Robert/Glasgow, Collins.

ATKINS, B. et A. ZAMPOLI (Eds) (1994) : *Automating the Lexicon*, Oxford University Press.

ATKINS, B. T. & A. ZAMPOLLI (Eds) (1994) : *Computational Approaches to the Lexicon*, Oxford University Press.

ATTALI, A., BOURQUIN, G. *et al.* (1992) : « Aide au transfert lexical dans une perspective de TAO : expérimentation sur un lexique non terminologique », *Meta*, 37 (4), pp. 770-790.

AUROUX, S. (juin 1994) : « L'hypothèse de l'histoire et la sous-détermination grammaticale », *Langages*, 114, Paris, Larousse, pp. 25-40.

AZOUGARH, M. (1992) : *Lexique berbère structures et signification*, thèse de DES, Faculté des Lettres et des Sciences Humaines, Oujda.

BACHIMONT, B. (1995) : « Ontologie régionale et terminologie : quelques remarques méthodologiques et critiques », *La Banque des Mots*, Numéro spécial du Centre de Terminologie et de Néologie du CNRS, Actes de la Première journée "Terminologie et Intelligence Artificielle", Paris Villetaneuse.

BAEZA-YATES, R. et G. GONNET (1992) : « A New Approach to Text Searching », *Communication of the ACM*, October 1992, 35 (10), pp. 74-82.

BANKS, David (1994) : « Clause Organization in the Scientific Journal Article », *Unesco ALSED-LSP Newsletter*, 17 (2), pp. 4-16.

BARBAUD, Philippe (1992) : « Recycling Words », Christiane Lauefer et Terrell A. Morgan (Eds), *Theoretical Analyses in Romance Linguistics*, coll. « CILT 74 », Philadelphie/ Amsterdam, John Benjamins, pp. 197-217.

BARBAUD, Philippe (1994) : « Conversion syntaxique », *Linguisticæ Investigationes*, XVIII, pp. 1-26

BARBAUD, Philippe (à paraître) : « La nominalisation d'un participe passé : la suppléance **mettre\mise** en composition lexicale », *Revue canadienne de linguistique\Canadian Journal of Linguistics*.

BARKER, K., COPECK, T., DELISLE, S. et S. SZPAKOWICZ (1993) : *A Case System for Interactive Knowledge Acquisition from Text,* Technical Report TR-93-08, Computer Science Department, University of Ottawa, février, 32 p.

BARKER, K. et S. SZPAKOWICZ (1995) : « Interactive Semantic Analysis of Clause-Level Relationships », à paraître dans *Proceedings of the 1995 PACLING (Pacific Association for Computational Linguistics) Conference*, University of Queensland, Brisbane (Australie), 19-22 avril

BASILI, R., PAZIENZA, M. T. et P. VELARDI (1992) : « Computational Lexicons : the Neat Examples and the Odd Exemplars », *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento (Italy), 31 mars-3 avril, pp. 96-103

BASSET, A. (1929) : *La langue berbère. Morphologie. Le verbe – Étude des thèmes*, Paris, Leroux, LII + 269 p.

BATEMAN, J. A. (1991) : « The Theoretical Status of Ontologie in Natural Language Processing », *KIT Report 97*, Technical University Berlin.

BATEMAN, J. A. (1993) : « Ontology Construction and Natural Language », *International Workshop on Formal Ontology*, Padova (Italy), LADSEB-CNR National Research Council, pp. 83-93.

BAUER, D., SEGOND, F. & A. ZAENEN (1995) : *Enriching an SGML-Tagged Bilingual Dictionary for Machine-Aided Comprehension*, Rank Xerox Research Centre Technical Report, Meylan

BAUER, Daniel, SEGOND, Frédérique et Annie ZAENEN (1994) : « Enriching an SGML-Tagged Dictionary for Machine-Aided Comprehension », *Technical Report MLTT-011,* Rank Xerox Research Centre, Grenoble

BAUER, Daniel, SEGOND, Frédérique et Annie ZAENEN (1995) : « LOCOLEX: the Translation Rolls off your Tongue », *Proceedings of the ACH-ALLC Conference*, Santa Barbara, pp 6-8

BEAUJEAN, A. (1990) : *Le Petit Littré, dictionnaire de la langue française*, coll. « classiques modernes », Paris, La Pochothèque/Le Livre de Poche.

BÉJOINT, H. (1993) : « La définition en terminologie », P. J. L. Arnaud & Ph. Thoiron (dir.), *Aspects du vocabulaire*, Lyon, Presses universitaires de Lyon, pp 18-25.

BÉJOINT, H. et Ph. THOIRON (1987) : « Compte rendu de Benson *et al* (1986a) », *Les langues modernes*, 81 (3-4), pp. 152-159.

BENMRAD, M. (1994) : *Agregats et composition de requêtes dans les hypertextes virtuels*, thèse de doctorat n 1284, Département d'informatique, École Polytechnique Fédérale de Lausanne, Octobre 1994.

BENMRAD, M., CORAY, G. et C. VANOIRBEEK (1995) : « Designing Virtual Hypertexts with Aggregates », *Proceedings of IWHD'95*, Montpellier, France, June 1995

BENSON, M. (1989) : « The Structure of the Collocational Dictionary », *International Journal of Lexicography*, 2 (1), pp 1-14.

BENSON, M. (1990) : « Collocations and General-Purpose Dictionaries », *International Journal of Lexicography*, 3 (1), pp. 23-34.

BENSON, M., BENSON E. et R. ILSON (1986a) : *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*, Amsterdam/Philadelphia, John Benjamins, 286 p.

BENSON, M., BENSON E. et R. ILSON (1986b) : *The Lexicographic Description of English*, Amsterdam, John Benjamins, 288 p.

BENVENISTE, E. (1939) : « Nature du signe linguistique », *Acta Linguistica*, 1, Copenhague, 1939, et *Problèmes de linguistique générale*, 1, Gallimard, 1974.

BENVENISTE, Émile (1966) : « Formes nouvelles de la composition nominale », *Problèmes de linguistique générale*, Paris, Gallimard, pp. 163-173.

BENVENISTE, Émile (1974) : *Problèmes de linguistique générale*, Paris, Gallimard.

BENVENISTE, Émile (1974) : *Problèmes de linguistique générale 2*, coll. « TEL », Paris, Gallimard.

BERNARD, P. et F. AL (1991) : « Dictionnaire bilingue et ordinateur », *Wörterbucher, Ein internationales Handbuch zur Lexikographie*, Berlin/New York, Walter de Gruyter, pp 2804-2813.

BERWICK, R. C. (1986) : « Learning from Positive-Only Examples: The Subset Principle and Three Case Studies », R. S. Michalski, J. G Carbonell & T. M. Mitchell (Eds), *Machine Learning : An Artificial Intelligence Approach, Vol. II*, Los Altos (CA), Morgan Kaufmann, pp. 625-645.

BIERWISCH, M. et K. E. HEIDOLPL (Eds) (1979) : *Progess in Linguistics*, The Hague, Mouton.

BLACHE, P. (1995a) : « Contraintes et Héritage pour l'analyse syntaxique : pour une programmation multi-paradigmes », Actes de *TALN'95*.

BLACHE, P. (1995b) : *Introduction à HPSG*, Rapport Technique #PB-9501, 2LC-CNRS.

BLACK, A., RITCHIE, G., PULMAN, S. et G. RUSSELL (1987) : « Formalisms for Morphographemic Description », *EACL-3*.

BLAMPAIN, D. (1992) : « Traduction et écosystèmes terminologiques », *Terminologie & traduction*, 2 (3), pp. 457-466.

BLAMPAIN, D. (1993) : « Notions et phraséologie. Une nouvelle alliance ? », Phraséologie. *Actes du Séminaire international, Terminologies nouvelles*, 10, pp. 43-49.

BLAMPAIN, D., PETRUSSA, P. & M. VAN CAMPENHOUDT (1992) : « À la recherche d'écosystèmes terminologiques », *L'environnement traductionnel. La station de travail du traducteur de l'an 2001. Journées scientifiques du Réseau thématique de recherche « Lexicologie, Terminologie et Traduction »*, Actes du colloque (Mons, 25-27 avril 1991), Presses de l'Université du Québec et AUPELF-UREF, pp. 273-282.

BLANCHON, H. (1994a) : « Perspectives of DBMT for Monolingual Authors on the Basis of LIDIA-1, an Implemented Mock-up », *15th International Conference on Computational Linguistics*, COLING-94, Kyoto, August 5-9

BLANCHON, H. (1994b) : *LIDIA-1 : Une première maquette vers la TA interactive « pour tous »*, thèse de Doctorat, Université Joseph Fourier, Grenoble, (N. P.)

BOGURAEV, B. (1988) : « A Natural Language Toolkit : Reconciling Theory with Practice », U. Reyle & C. Rohrer (Eds), *Natural Language Parsing and Linguistics*, Dordrecht, D. Reidel, pp 95-130.

BOGURAEV, B. (1991) : « Building a Lexicon : The Contribution of Computers », *International Journal of Lexicography*, 4 (3), pp. 227-260.

BOGURAEV, B. (1994) : « Machine-Readable Dictionaries and Computational Linguistics Research », Zampolli, Calzolari & Palmer (Eds), *Current Issues in Computational Linguistics : in Honour of Don Walker*, Series « Linguistica Computazionale », IX-X, pp. 119-154.

BOGURAEV, B. & T. BRISCOE (1989) : *Computational Lexicography for Natural Language Processing*, London and New York, Longman.

BOGURAEV, B., BRISCOE, T., CARROLL, J. & A. COPESTAKE (1992) : « Database Models for Computational Lexicography », *EURALEX'90 Proceedings*, Barcelona, Biblograf, pp. 59-78

BOGURAEV, B. et J. PUSTEJOVSKY (1990) : « Knowledge Representation and Acquisition from Dictionary », *Coling Tutorial*, August 16-18, 1990, Helsinki, Finland.

BOISVERT, R. (1989) : *Programme de concordance*, rapport de projet de synthèse en informatique #89 1 06, Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières.

BOITET, C. & H. BLANCHON (1993) : « Dialogue-Based MT for Monolingual Authors and the LIDIA Project », H. Nomura (Ed.), *Proceedings NLPRS'93 (Natural Language Processing Rim Symposium)*, Fukuoka, 6-7/12/93, Kyushu Institute of Technology, pp. 208-222.

BORESDON, Bernard et Irène TAMBA (1991) : « Verre à pied, moule à gaufres . préposition et noms composés de sous-classe », *Langue française*, 91, pp. 40-55.

BOUCHARD, L. H. & L. EMIRKANIAN (1990) : « Développement d'un analyseur morpho-syntaxique pour le français », *Actes du colloque « Les industries de la langue . Perspectives des années 1990 »*, Montréal, Office de la langue française et Société des traducteurs du Québec, pp. 115-130.

BOUCHARD, L. & L. EMIRKANIAN (1994) : « The Organization of the Lexicon in GSF : Structure and Implementation », Kiefer, Kiss & Pajzs (Eds), *Papers in Computational Lexicography – COMPLEX'94*, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, pp. 13-22.

BOUCHARD, L., EMIRKANIAN, L. & F. GROS D'AILLON (1991) : « Extracting French Morphological and Syntactic Information from a Machine-Readable Dictionary », Kiefer (Ed.), *Computational Lexicography*, [Balatonfüred, Hungary 8-11 September 1990], Research Institute for Linguistics, Hungarian Academy of Sciences, pp. 9-24.

BOUCHARD, L., EMIRKANIAN, L. & J.-Y. MORIN (1992) : « Computational Grammar as Knowledge Representation », *Proceedings Sixth International Conference on Systems Research Informatics and Cybernetics* (Volume II), Baden-Baden, International Institute for Advanced Studies in System Research and Cybernetics, pp. 121-132.

BOUCHARD, L. H., EMIRKANIAN, L. & S. RATTÉ (1989) : « L'analyse MLR une application au français », *ICO : Intelligence Artificielle et Sciences Cognitives au Québec*, 1(4), pp. 50-60

BOUILLON, P. (1995) : « Le lexique génératif : une alternative au traitement de la polysémie. Le cas de *commencer* », communication présentée au cours du Colloque

BOURIGAULT, D. (1992) : « Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases », *Proceedings of the Fourteenth International Conference on Computational Linguistics*, COLING-92, Nantes, pp 977-981.

BOURIGAULT, D. (1993a) : « An Endogenous Corpus-based Method for Structural Noun Phrase Disambiguation », *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, Utrecht.

BOURIGAULT, D. (1993) : « Analyse syntaxique locale pour le repérage de termes complexes dans un texte », *T.A.L.*, 34 (2), pp 105-118

BOURIGAULT, D. (1994) : *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes*. Thèse de doctorat de l'École des hautes études en sciences sociales, Paris, juin 1994

BOURIGAULT, D. (1995) : « Lexter, a Terminology Extraction Tool for Knowledge Acquisition from Texts », *Proceedings of the 9th Knowledge Acquisition for KnowledgeBased Systems Workshop (KAW'95)*, Banff.

BOURIGAULT, D. et P. LÉPINE (1995) : « Utilisation d'un logiciel d'extraction de terminologie (Lexter) pour l'acquisition des connaissances à partir de textes », *Acquisition et ingénierie des connaissances : tendances actuelles*, Eds. N. Aussenac-Gilles, P. Laublet, C. Reynaud, Toulouse, Cépaduès.

BOUTIN-QUESNEL, R. *et al.* (1985) : « Vocabulaire systématique de la terminologie », *Cahiers de l'Office de la langue française*, Québec.

BRESNAN, Joan (1982) : *The Mental Representation of Grammatical Relations*, Cambridge (MA), MIT Press.

BRILL, E. (1992) : « A Simple Rule-Based Part of Speech Tagger », *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento (Italy), 31 mars-3 avril, pp. 152-155.

BRISCOE, T., DE PAIVA, V. et A. COPESTAKE (Eds) (1993) : *Inheritance, Defaults and the Lexicon*, Cambridge University Press.

BRISCOE, T. et A. COPESTAKE (1993) : *Default Inheritance in the Lexicon*, Cambridge, CUP.

BROWN, P., CHURCH, K., GODBY, J., LEWIS, D., REIGHART, R. et F. ZHOU (Eds.) (1993) : *Very Large Corpora : Academic and Industrial Perspectives*, Proceedings of the Workshop Sponsored by the Association for Computational Linguistics, American Chemical Society Chemical Abstracts, Mead Data Central, Inc., OCLC Online Computer Library Center, Inc. June 22, 1993, Ohio State University, Columbus, USA.

BROWN, P. *et al.* (1991) : « Word-sense Disambiguation Using Statistical Methods », *Proceedings of ACL'91.*

BROWN, P. *et al.* (1993) : « The Mathematics of Statistical Machine Translation : Parameter Estimation », *Computational Linguistics*, 19 (2), pp. 261-311.

BRUANDET, M. F. (1980) : « A Conceptual Network for Automatic and Dynamic Thesaurus Updating in Information Retrieval System », *Proceedings of COLING-80*, 30 sept.-3 oct , Tokyo.

BRUNDAGE, Jennifer, KRESSE, Maren, SCHWALL, Ulrike et Angelika STORRER (1992) : « Multiword Lexemes: A Monolingual and Contrastive Typology for NLP and MT IBM Deutschland GmbH », Institut fur Wissensbasierte Systeme, Heidelberg, *IWBS Report 232*, September 1992, IBM TR-80.92-029.

BUVET, P.-A. (1995) : « Lexicalisation et domaines d'emplois », communication présentée au cours du Colloque.

BYRD, R. (1989) : « Discovering Relationships among Word Senses », *Dictionaries in the Electronic Age – Proceedings of the Fifth Annual Conference of the UW Centre for the New Oxford English Dictionary*, Oxford, pp. 67-79.

BYRD, R., CALZOLARI, N., CHODOROW, M., KLAVANS, J., NEFF, M. & O. RIZK (1987) : « Tools and Methods for Computational Lexicology », *Computational Linguistics*, 13 (3/4), pp. 219-240.

CADIOT, Pierre (1991) : « **À la hache** ou **avec la hache** ? Représentation mentale, expérience située et donation du référent », *Langue française*, 91, pp. 7-23.

CALZOLARI, Nicoletta et Remo BINDI (1990) : « Acquisition of lexical information from a large textual Italian corpus », *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki, Finlande.

CAP-College of American Pathologists (1993) : *SNOMED International*, Introduction

CARDIE, C. (1993) : « A Case-Based Approach to Knowledge Acquisition for Domain-Specific Sentence Analysis », *Proceedings of the 11th National Conference on Artificial Intelligence*, Washington (DC), 11-15 juillet, pp 798-803.

CARRÉ, R., DEGREMONT, J.-F., GROSS, M. et G. SABAH (1991) : *Langage humain et machine*, Paris, Presses du CNRS.

CARROL, J. (1993) : *Lexical Database System, User Manual*, ESPRIT BRA 3030 Computer Laboratory, University of Cambridge, United Kingdom.

CEUSTERS W., DEVILLE G., MOUSEL P., STREITER O. et G. THIENPONT (1994) : *Functional Specification of the ANTHEM Prototype*, ANTHEM Deliverable n D1-1.

CHAKER, S. (1984) : *Textes en linguistique berbère*, Paris, éditions du CNRS, 291 p.

CHAKRAVARTHY, A. S. (1995) : « Sense Disambiguation Using Semantic Relations and Adjacency Information », *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge (MA), 26-30 juin, pp 293-295 (student session).

CHANGEUX, Jean-Pierre (1983) : *L'homme neuronal*, Paris, Fayard.

CHANOD, J.-P. & P. TAPANAINEN (1995) : « Statistical and Constraint Based Taggers for French », *Proceedings EACL'95*, Dublin, 17 p.

CHAUCHÉ, Jacques (1984) : « Le Système Sygmart », *Actes de COLING'84*, Stanford.

CHEN, S. F. (1993) : « Aligning Sentences in Bilingual Corpora Using Lexical Information », *Proceedings of the 13th Annual Meeting of ACL '93*, pp. 9-16.

CHIBOUT, K. (1994) : *Traitement automatique des tropes*, document interne LIMSI, n° 94-04, Orsay, 66 p.

CHOMSKY, Noam (1965) : *Aspects of the Theory of Syntax*, Cambridge (MA), The MIT Press.

CHOMSKY, Noam (1970) : « Remarks on Nominalization », *Readings in English Transformational Grammar*, Rodeock Jacobs & Peter Rosenbaum (Eds), Waltman (MA), Ginn, pp. 184-221

CHOMSKY, Noam (1981) : *Lectures on Government and Binding*, Dordrecht, Foris Publications.

CHOUEKA, Y. (1988) : « Looking for Needles in a Haystack », *Proceedings of the RIAO Conference on User-Oriented Context Based Text and Image Handling*, Cambridge, Ma.

CHOUEKA, Y. (1988) : « Looking for Needles in a Haystack or Locating Interesting Collocationnal Expressions in a Large Textual Database », *Actes de colloque du RIAO 88*, Cambridge, Cambridge University Press, pp. 609-623.

CHOUEKA, Y., KLEIN, T. et E. NEUWITZ (1983) : « Automatic Retrieval of Frequent Idiomatic and Collocationnal Expressions in a Large Corpus », *ALLC Journal*, Grande-Bretagne, 4 (1), pp 34-39.

CHUKWU, U. (1993) : *Le repérage des termes dans un corpus bilingue anglais/français*, Thèse de doctorat, Université Lumière Lyon II.

CHUKWU, U. et Ph. THOIRON (1989) : « Reformulation et repérage des termes », *La Banque des mots*, numéro spécial, pp. 23-50.

CHURCH, K. & P. HANKS (1990) : « Word Association Norms, Mutual Information and Lexicography », *Computational Linguistics*, 16 (1), pp 22-29.

CHURCH, K. & R. PATIL (1982) : « Coping with Syntactic Ambiguity or How to Put the Block in the Box on the Table », *Computational Linguistics*, 8 (3-4), pp. 139-149.

CHURCH, K., GALE, W., HANKS, P. et D. HINDLE (1991) : « Using Statistics in Lexical Analysis », *Lexical acquisition : Using On-Line Resources to Build a Lexicon*, U. Zernik (Ed.), Lawrence Erlbaum.

CHURCH, K., GALE, W., HANKS, P., HINDLE, D. & R. MOON (1994) : « Lexical Substitutability », Atkins & Zampolli (Eds), *Computational Approaches to the Lexicon*, Oxford University Press, pp. 153-177.

CLAVIER, V. (1995) : *Modélisation de la suffixation pour le traitement automatique du français : Application à la recherche d'informations*, Thèse de doctorat, Grenoble.

CLAVIER, V. et G. LALLICH-BOIDIN (1994) : « Modélisation linguistique de la suffixation en vue de l'analyse automatique », *T.A.L.*, 35 (2), pp. 129-143.

CNET (1990) : Communication personnelle.

COHEN, B. (1986) : *Lexique de cooccurrents. Bourse – Conjoncture économique*, Montréal, Linguatech.

COHEN, D. (1993) : « Racines », *À la croisée des études libyco-berbères Mélanges offerts à Paulette Galang-Pernet et Lionel Galand. Comptes rendus de G.L.E.C.S.*, supplément 15, Paris, Librairie orientaliste Paul Geuthner, pp. 161-175.

Collins Cobuild English Dictionary (1995): London, Harper Collins Publishers Ltd.

Collins Cobuild English Language Dictionary (1987) : London and Glasgow, Collins Publishers, xxiv + 1703 p.

CONDAMINES, A. (1995) : « Terminology . New Needs, New Perspectives », *Terminology*, Vol. 2 (2).

COPECK, T., DELISLE, S. et S. SZPAKOWICZ (1992) : « Parsing and Case Analysis in TANKA », *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes (France), 23-28 juillet, pp. 1008-1012

COPESTAKE, A., SANFILIPPO, A. BRISCOE, T. et V. DE PAIVA (1993) : « The Acquilex LKB : An Introduction », T. Briscoe, V. De Paiva and A. Copestake (Eds), *Inheritance, Defaults and the Lexicon*, Cambridge University Press.

CORBIN, D. (1987) : *Morphologie dérivationnelle et structuration du lexique*, 2 volumes, Tübingen, Niemeyer.

CORBIN, Danielle (1987) : « Contre une transposition de la théorie X' à la morphologie dérivationnelle », *Acta Linguistica Academiæ Scientiarum Hungaricæ*, T. 37 (1-4), pp. 73-92.

CORBIN, Danielle (1992) : « Hypothèse sur les frontières de la composition nominale », *Cahiers de grammaire 17*, Université de Toulouse-le-Mirail, pp. 25-55.

CORREARD, M.-H. & V. GRUNDY (Eds) (1994) : *The Oxford-Hachette French Dictionary (French-English, English-French)*, Hachette et Oxford University Press.

COWIE, A. P. (1986) : « Collocational Dictionaries – A Comparative View », Murphy (Ed.), *Fourth Joint Anglo-Soviet Seminar*, London, British Council, pp 61-69

COWIE, A. P. (Ed.) (1989) : *Oxford Advanced Learner's Dictionary of Current English*, 4th edition, Oxford University Press.

CROFT, W. (1984) : *The Representation of Adverbs, Adjectives and Events in Logical Form*, Technical Note 344, SRI International.

CROFT, W. (1993) : « The Semantics of Mental Verbs », Pustejovsky (Ed.), *Semantics and the Lexicon*, Dordrecht, Kluwer.

CRUSE, D. A. (1986) : *Lexical Semantics*, Cambridge Textbooks in Linguistics, Cambridge, London, New York, etc., Cambridge University Press.

DACHELET, R. (1990) : « État de l'art de l'informatique documentaire », *Le Document électronique*, Cours INRIA dirigé par C Bornès, 11-15 juin 1990, Châtelailon, pp. 107-132.

DAGAN, I. *et al.* (1991) : « Two Languages Are More Informative Than One », *Proceedings of ACL'91.*

DAGAN, I. et I. ALON (1994) : « Word Sense Disambiguation Using a Second Language Monolingual Corpus », *Computational Linguistics*, vol. 20, pp. 563-596.

DAHLBERG, I. (1981a) : « Conceptual Definitions for INTERCONCEPT », *International Classification*, 5 (3), pp 142-151.

DAHLBERG, I. (1981b) : « A Referent-Oriented, Analytical Concept Theory for INTERCONCEPT », *International Classification*, Francfort, INDKS Verlag, 8 (1), pp. 16-22.

DAHLBERG, I. (1982) : « Terminological Definitions Characteristics and Demands », *Problèmes de la définition et de la synonymie en terminologie*, Actes du colloque international de terminologie, Université Laval, Québec, mai 1982, pp 15-28.

DAHLGREN, K. (1988) : *Naive Semantics for Natural Language Understanding*, Boston, Kluwer, 257 p.

DAHLGREN, K. (1993) : A Linguistic Ontology, *International Workshop on Formal Ontology*, Padova (Italy), LADSEB-CNR National Research Council, pp 165-174.

DAHLGREN, K. & J. McDOWELL (1986) : « Using Commonsense Knowledge to Disambiguate Preposition Phrase Modifiers », *AAAI'86*, Philadelphia, pp. 589-593.

DAILLE, B. (1993) : « Extraction automatique de terminologie monolingue », *Actes du colloque Informatique et langue naturelle*, Nantes, 21 p.

DAILLE, B. (1994) : « Extraction de noms composés terminologiques du domaine des télécommunications », *5^es^ Journées ERLA-GLAT (Études et Recherches Lexicales Appliquées)*, Brest, 13 p.

DAILLE, Béatrice (1994) : « Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques », thèse de doctorat, Université Paris 7.

DAILLE, B. (1995) : « ACABIT : une maquette d'aide à la construction automatique des banques de données terminologiques », *Actes des IV^es^ Journées scientifiques du réseau LTT de l'AUPELF-UREF*, Lyon, sept. 1995.

DAILLE, Béatrice, GAUSSIER, Éric et Jean-Marc LANGÉ (1994) : « Towards Automatic Extraction of Monolingual and Bilingual Terminology », *COLING-94*, Kyoto, Japon.

DAILLE, Béatrice, GAUSSIER, Éric et Jean-Marc LANGÉ (1995) : « An Evaluation of Statistical Scores for Word Association », *The Tbilisi Symposium on Language, Logic and Computation*, octobre, Tbilisi, Georgia.

DALLET, J.-M. (1982) : *Dictionnaire kabyle-français*, Paris, SELAF, XL + 1052 p.

DANLOS, L. et P. SAMVELIAN (1992) : « Translation of the Predicative Element of a Sentence : Category Switching, Aspect and Diathesis », *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, pp. 21-34.

DAOUST, F. (1992) : *SATO . Système d'analyse de texte par ordinateur, Manuel de référence*, Centre ATO, Université du Québec à Montréal.

DAVID, S. (1990) : « Le progiciel TERMINO : de la nécessité d'une analyse morphosyntaxique pour le dépouillement des textes », *Acte du colloque Les industries de la langue : perspective des années 1990*, 21 au 24 novembre 1990, Montréal, Office de la langue française et Société des traducteurs du Québec, pp. 71-89.

DAVID, S. et P. PLANTE (1990) : « De la nécessité d'une approche morpho-syntaxique dans l'analyse de texte », *Revue ICO*, 2 (3).

DE BESSÉ, B. (1991) : « Des fichiers terminologiques aux bases de connaissances », Clas, A. et Safar, H. (dir.), *L'environnement traductionel. La station de travail du traducteur de l'an 2001 – Journées scientifiques du Réseau thématique de recherche « Lexicologie, terminologie, traduction »*, Mons (Belgique), 25-27 avril 1991, Presses de l'Université du Québec, pp 283-300.

DELHEURE, J. (1984) : *Dictionnaire mozabite-français*, Paris, SELAF.

DELISLE, S. (1994) : *Text Processing without A-Priori Domain Knowledge : Semi-Automatic Linguistic Analysis for Incremental Knowledge Acquisition*, Ph.D. thesis, Department of Computer Science Ottawa-Carleton Institute for Computer Science, TR-94-02, University of Ottawa, janvier

DELISLE, S. (1995) : *The Reattachment Module Design Document*, rapport technique, Université du Québec à Trois-Rivières, Département de mathématiques et informatique, 2 avril, 32 p.

DELISLE, S., BARKER, K., COPECK, T. et S. SZPAKOWICZ (à paraître) : « Interactive Semantic Analysis of Technical Texts : Case Pattern Acquisition », à paraître dans le numéro de mai 1996 (vol.12, #2) de *Computational Intelligence*.

DELISLE, S., COPECK, T., SZPAKOWICZ, S. et K. BARKER (1993) : « Pattern Matching for Case Analysis : A Computational Definition of Closeness », O. Abou-Rabia, C. K. Chang and W. W. Koczkodaj (Eds), *Proceedings of ICCI-93*, Sudbury (Canada), 27-29 mai, pp. 310-315.

DELISLE, S. et S. SZPAKOWICZ (1991) : « A Broad-Coverage Parser for Knowledge Acquisition from Technical Texts », *Proceedings of the 5th International Conference on Symbolic and Logical Computing*, Madison (S.D), avril, pp. 169-183.

DELISLE, S. et S. SZPAKOWICZ (1995) : « Realistic Parsing : Practical Solutions of Difficult Problems », à paraître dans *Proceedings of the 1995 PACLING (Pacific Association for Computational Linguistics) Conference*, University of Queensland, Brisbane (Australie), 19-22 avril, 11 p.

DELL, F. (1970) : *Les règles phonologiques tardives de la morphologie dérivationnelle du français*, Thèse de doctorat, Cambridge (MA), The MIT Press.

DELPUI, M. (1993) : *Intégration d'une base de connaissances lexicales pour un analyseur syntaxique*, Rapport de stage, ESSI, Sophia-Antipolis.

DESCLÉS, J.-P. (1987) : « Réseaux sémantiques : La nature logique et linguistique des relateurs », *Langages*, n° 87, pp. 55-78.

DESCLÉS, J.-P. (1990) : *Langages applicatifs, langues naturelles et cognition*, Paris, Hermès.

DESCLÉS, J. P. (1993) : *Représentation des connaissances : archétypes cognitifs, chaînes conceptuelles et schémas grammaticaux*, Centre d'analyse et de mathématique sociales, CNRS EHESS-Sorbonne.

DESCLÉS, J.-P., H. ABAAB, J. DICHY, D. E. KOULOUGHLI, M. S. ZIADAH (1983) : *Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement assisté par Ordinateur*, Rapport rédigé sous la direction de J -P. Desclés.

DESCLÉS, J.-P. et C. JOUIS (1993) : « L'exploration contextuelle : une méthode linguistique et informatique pour l'analyse automatique de textes », *Actes du colloque Informatique et Langue Naturelle, ILN'93*, Nantes, 2 & 3 déc. 1993, pp. 339-350

DEVILLE, G. (1989) : *Modelization of Task-Oriented Utterances in a Man-Machine Dialogue System*, Ph.D. Thesis, Universitaire Instelling Antwerpen.

DEVILLE, G. et E. HERBIGNIAUX (1994) : *Methodological Principles for the Elaboration of Multilingual Corpora of Medical Diagnostic Expressions*, ANTHEM Deliverable n. D2-2 – Part I.

DICHY, J. et M. O. HASSOUN (dir) (1989) : *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe - Travaux SAMIA I*, Paris, Conseil international de la langue française.

Dictionnaire du français contemporain (DFC) (1971) : Paris, Librairie Larousse.

DIK, S. (1989) : *A Theory of Functional Grammar*, Dordrecht, Foris.

DIXON, R. M. W. (1991) : *A New Approach to English Grammar, On Semantic Principles*, Oxford University Press.

DOSTIE, G., MEL'ČUK I. A. et A. POLGUÈRE (1992) : « Méthodologie d'élaboration des entrées lexicales du Dictionnaire Explicatif et Combinatoire (REPROCHER, REPROCHE et IRRÉPROCHABLE) », *International Journal of Lexicography*, 5 (3), pp. 165-198.

DUBOIS, Jean et al. (1994) : *Dictionnaire de linguistique et des sciences du langage*, Paris, Larousse, 514 p.

DUNNING, Ted (1993) : « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, 19 (1).

EDR (1993) : *EDR Electronic Dictionary Technical Guide*, Japan Electronic Dictionary Research Institute Ltd, Project report n° TR-042, August 16, 144 p.

EJERHED, E. I. (1988) : « Finding Clauses in Unrestricted Text by Finitary and Stochastic Methods », *Second Conference on Applied Natural Language Processing*, pp. 219-227.

EL-BÈZE, Marc (1993) : « Les modèles de langage probabilistes · quelques domaines d'applications », *Habilitation à diriger des recherches*, LIPN, Université Paris-Nord.

EMIRKANIAN, L. & L. H. BOUCHARD (1989) : « La correction des erreurs d'orthographe d'usage dans un analyseur morphosyntaxique du français », *Langue française*, 83, pp. 106-122.

EMIRKANIAN, L. & L. H. BOUCHARD (1992) : *Approche computationnelle aux phénomènes morphologique et syntaxique du français*, Rapport de recherche, Université du Québec à Montréal.

ENGELKE, Sabine (1994) : *Eigenschaften von Phraseolexemen : Eine Untersuchung zur syntaktischen Variabilitat und internen Modifizierbarkeit von somatischen verbalen Phraseolexemen*, Magisterarbeit, Universität Tübingen, April 1994

ENGUEHARD, C. (1992) : *ANA, Apprentissage Naturel Automatique d'un Réseau Sémantique,* thèse de Doctorat en Contrôle des Systèmes, Université de Technologie de Compiègne.

ENGUEHARD, C. (1993) : « Acquisition automatique de terminologie à partir de gros corpus », *Actes du colloque Informatique & Langue Naturelle, ILN'93*, Nantes, 2 & 3 déc 1993, pp. 373-384.

ERNST, T. B. (1984) : *Towards an Integrated Theory of Adverb Position in English*, Indiana Linguistics Club, Indiana.

FARWELL, D., GUTHRIE, L. & Y. WILKS (1993) : « Automatically Creating Lexical Entries for ULTRA, a Multilingual MT System », *Machine Translation*, 8, pp. 127-145.

FASS, D. (1991) : « MET* : A Method for Discriminating Metonymy from Metaphor by Computer », *Computational Linguistics*, 17 (1), pp. 49-88.

FELBER, H. (1987) : *Manuel de terminologie*, Paris, UNESCO.

FELBER, H. (1994) : « Terminology Research : Its Relation to the Theory of Science », *ALFA*, vol. 7/8, pp 163-172.

FELLBAUM, C. (1990) : « English Verbs as a Semantic Net », *International Journal of Lexicography*, 3 (4), pp. 278-301.

FILLMORE, C. (1968) : « The Case for Case », E. Bach and R.T. Harms (Eds), *Universals in Linguistic Theory*, New York, Holt, Rinehart and Winston, pp. 1-88.

FLEISCHER, Wolfgang (1982) : *Phraseologie der deutschen Gegenwartssprache*, Leipzig, VEB Bibliographisches Institut.

FONTENELLE, T. (1992) : « Collocation Acquisition from a Corpus or from a Dictionary : a Comparison », Tommola, Varantola, Salmi-Tolonen & Schopp (Eds), *EURALEX'92 Proceedings I-II, Fifth EURALEX International Congress*, Studia Translatologica, Ser. A, Vol. 1, University of Tampere, pp. 220-228.

FONTENELLE, T. (1994) : « Towards the Construction of a Collocational Database for Translation Students », *Meta*. Presses de l'Université de Montréal, 39 (1), pp. 47-56.

FONTENELLE, T. (1995) : *Turning a Bilingual Dictionary into a Lexical-semantic Database*, thèse de Doctorat, Université de Liège, ms

FOSTER, Georges (1991) : « Statistical Lexical Desambiguation », Master's thesis, School of Computer Science, McGill University.

FOUCAULD, Ch. de. (1951) : *Dictionnaire touareg-français (dialecte de l'Ahaggar)*, Paris, Imprimerie Nationale.

FOUCAULT, Michel (1969) : *L'archéologie du savoir*, Gallimard, 275 p.

FOUQUERÉ, C. (1994) : Communication personnelle.

FRADIN, B. (1993a) : *Organisation de l'information lexicale et interface morphologie/syntaxe dans le domaine verbal*, Thèse de doctorat d'État, Université Paris 8.

FRADIN, B. (1993b) : « La théorie morphologique face à ces choix », *Cahiers de Lexicologie*, 63, pp. 5-42.

FRADIN, B. (1994) : « L'approche à deux niveaux en morphologie computationnelle et les développements récents de la morphologie », B. Fradin (Éd.), *La morphologie computationnelle, TAL (Traitement automatique des langues)*, 34 (2), Paris, Association pour le Traitement Automatique des Langues, pp. 9-48.

FRAMIS, F. R. (1994) : « An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus », *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto (Japon), 5-9 août, pp. 769-774.

FRAWLEY, W. (1988) : « New Forms of Specialized Dictionaries », *International Journal of Lexicography*, 1 (3), pp 189-213.

FUCHS, C. (1988) : « Représentation linguistique de la polysémie grammaticale », *TA. Informations*, Revue internationale du traitement automatique du langage, bulletin semestriel de l'ATALA, 29 (1-2), pp. 7-20.

FUNG, P. (1995) : « A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora », *Proceedings of ACL'95*.

GALAND, L. (1974) : « Signe arbitraire et signe motivé en berbère », *Actes du premier congrès international de linguistique sémitique et chamito-sémitiques*, Paris, The Hague, Mouton, pp. 90-101

GALAND, L. (1979a) : « Berbère et "traits sémitiques communs" », *Comptes rendus du G.L.E.C.S. tomes XVII-XXIII années 1973-1979*, Paris, Librairie orientaliste Paul Geuthner, pp. 463-478.

GALAND, L. (1979b) : « La langue berbère existe-t-elle ? », *Mélanges linguistiques offerts à Maxime Rodinson. Comptes rendus du G.L.E.C S., supplément 12*, Paris, Librairie orientaliste Paul Geuthner, pp. 175-184.

GALAND, L. (1984) : « Le comportement des schèmes et des racines dans l'évolution de la langue : exemples touaregs », *Current Progress in Afro-Asiatic Linguistics. Third International Hamito-Semitic Congress*, Londres, Benjamains Publishing Company, pp. 304-315.

GALAND, L. (1988) : « Le berbère », *Les langues dans le monde ancien et moderne, troisième partie : les langues chamito-sémitiques*, Paris, éditions du CNRS, pp. 207-242.

GALE, William A. et Kenneth W. CHURCH (1991) : « Concordances for parallel texts », *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research, Using Corpora*, Oxford (U.K.), pp. 40-62

GAMBIER, Yves et François GAUDIN (dir.) (1993) : « numéro spécial Socioterminologie », *Le langage et l'homme*, XXVIII (4), DeBoeck Université.

GARCIA, D. et A. JACKIEWICZ (1995) : « Aide à l'acquisition des connaissances causales par exploration de textes », *Actes des 6es Journées Acquisition, Validation, (JAVA 95)*, Grenoble, pp. 147-158.

GARDIN, Bernard *et al.* (Eds) (1994) : *Aspects terminologiques des pratiques langagières au travail*, cahier n° 7, réseau Langage et travail, 74 p.

GAUDIN, François (1992) : « Terminologie et démocratisation du savoir : à propos de dictionnaires scientifiques », *Le langage et l'homme*, XXVII (2-3), Bruxelles, Institut Libre Marie Haps, pp. 123-129.

GAUDIN, François (1993) : « Socioterminologie : propos et propositions épistémologiques », Y. Gambier et F. Gaudin (dir.), « numéro spécial Socioterminologie », *Le langage et l'homme*, XXVIII (4), DeBoeck Université, pp. 247-257.

GAUDIN, François (dir.) (1995) : « numéro spécial : Usages sociaux des termes . théories et terrains », *Meta*, 40 (2), juin 1995, Montréal, Presses de l'Université de Montréal, pp. 193-329.

GAUDIN, François et Allal ASSAL (dir) (1991) : « Terminologie et sociolinguistique », *Cahiers de linguistique sociale*, n° 18, éd. URA CNRS 1164/Université de Rouen, 213 p

GAZDAR, G., KLEIN, E., PULLUM, G. & I. SAG (1985) : *Generalized Phrase Structure Grammar*, Cambridge (MA), Harvard University Press, 276 p.

GIREIL (1994) : *La sous-catégorisation et la cliticisation*, Rapport de recherche, Université du Québec à Montréal

GOETSCHALCKX, J. (1992) : « Terminologie et phraséologie », *Terminologie & traduction*, 2 (3), pp 477-484.

GOMEZ, F., HULL, R. et C. SEGAMI (1994) : « Acquiring Knowledge from Encyclopedic Texts », *Proceedings of the 4th Conference on Applied Natural Language Processing*, Stuttgart (Germany), 13-15 octobre, pp. 84-90

GOUADEC, D. (1994) : « Traduction et informatique : les implications pour la formation », *Langages*, n° 116, Paris, Larousse, pp. 59-74.

Grand Robert (1985) : *Le grand Robert de la langue française. Dictionnaire alphabétique et analogique de la langue française*, de P. Robert, 2e édition entièrement revue et enrichie par A. Rey, 9 vol., Paris, Le Robert

GREFENSTETTE, G. (1994a) : « Corpus-Derived First, Second and Third-Order Word Affinities », *EURALEX'94 Proceedings*, Vrije Universiteit Amsterdam, pp. 279-290.

GREFENSTETTE, G. (1994b) : *Explorations in Automatic Thesaurus Discovery*, Boston, Kluwer Academic Press.

GRISHMAN, R. (1994) : « Whither Written Language Evaluation? », *Proceedings of the Human Language Technology Worskshop*, Plainsboro (N. J.), 8-11 mars, pp. 120-125.

GRISHMAN, R. et J. STERLING (1994) : « Generalizing Automatically Generated Selectional Patterns », *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto (Japon), 5-9 août, pp. 742-747.

GRISHMAN, R., MACLEOD, C. et A. MEYERS (1994a) : « Complex Syntax : Building a Computational Lexicon », *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto (Japon), 5-9 août, pp. 268-272.

GROSS, G. (1990) : « Définition des noms composés dans un lexique-grammaire », *Langue française*, n° 87, Paris, Larousse.

GROSS, G. (1990) : « Les mots composés », *Modèles Linguistiques*, 12, pp. 47-63.

GROSS, G. (1992) : « Forme d'un dictionnaire électronique », *L'environnement traductionnel. La station de travail du traducteur de l'an 2001*, Actes du colloque de Mons, Sillery (Québec), Presses de l'Université du Québec et AUPELF-UREF, pp. 255-271.

GROSS, G. (1994) : « Classes d'objets et description des verbes », *Langages*, n° 115, septembre 1994, Paris, Larousse, pp. 15-30

GROSS, G. (1995) : « Une sémantique nouvelle pour la traduction automatique Les classes d'objets », *La tribune des industries de la langue et de l'information électronique*, n° 17-18-19, Paris.

GROSS, Gaston, CHAURAND, Jacques, VIVÈS, Robert, MATHIEU-COLAS, Michel et Pierre BILLY (1986) : « Typologie des noms composés », *Rapport technique A.T.P.-Nouvelles Recherches sur le Langage*, Université Paris 13.

GROSS, Maurice (1975) : *Méthode en syntaxe*, Hermann

GROSS, M. (1986) : « Lexicon grammar. The representation of compound words », *COLING-86*, pp 1-6.

GROSS, M. (1988) : « Sur les phrases figées complexes du français », *Langue française*, n° 77, Paris, Larousse.

GROSS, M. (1990) : « Le programme d'extension des lexiques électroniques », *Langue française*, n° 87, Paris, Larousse.

GROSS, M. (1990) : « Sur la notion harissienne de transformation et son application au français », *Langages*, 99, Paris, Larousse, pp 39-56.

GROSS, Maurice (1990) : *Grammaire transformationnelle du français Syntaxe de l'adverbe*, Paris, ASSTRIL.

GUERARD, F. (Ed.) (1989) : *Le dictionnaire de notre temps*, Paris, Hachette, 1714 + 48 p.

GUILBERT, L. (1965) : *La formation du vocabulaire de l'aviation*, Paris, Larousse.

GUILBERT, Louis (1975) : *La créativité lexicale*, Paris, Larousse, 285 p

GUILLET, A. (1990) : Communication personnelle.

GUIRAUD, G. (1900) : *Vocabulaires des dialectes Sango, Bakongo et A'zande*, Paris, Challamel, 58 p.

HABAILI, Hussein (1976) : *Contraintes de structure morphématique en arabe*, mémoire de Maîtrise ès art en linguistique, Montréal, Université de Montréal.

HABAILI, Hussein (1990) : *Phonologie et morphologie flexionnelle et dérivationnelle de l'arabe* thèse de Doctorat d'État, Paris, Université de la Sorbonne-Nouvelle, Paris III.

HABERT, Benoît et Christian JACQUEMIN (1993) : « Présentation », *Traitement automatique des langues*, 34 (2).

HABERT, B. et C. JACQUEMIN (1993) : « Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques », *TAL*, 34 (2), ATALA, pp. 5-42.

Hachette (1993) : *Dictionnaire de notre temps*, Paris, Hachette.

HALLE, Moris (1973) : « Prolegomena to a Theory of Word Formation », *Linguistic Inquiry*, 4 (1), pp. 3-16.

HALLIDAY, Michael A. K. (1988) : « On the Language of Physical Science », *Registers of Written English : Situational Features and Linguistic Features*, M. Ghadassy (Ed.), London, Pinter.

HANSE, J. et D. BLAMPAIN (1994) : *Nouveau dictionnaire des difficultés du français moderne*, Louvain-la-Neuve, De Boeck/Duculot, 983 p

HarperCollins German-English Dictionary (1991) : P. Terell, V. Schnorr, W. V A. Morris, R. Breitsprecher (Eds), 2nd Edition, Glasgow, HarperCollins Publishers.

HARRIS, Mary Dee (1985) : *Introduction to Natural Language Processing*, Reston, Reston Publishing Company, 312 p.

HAUSMANN, F. J. (1979) : « Un dictionnaire des collocations est-il possible ? », *Travaux de linguistique et de littérature*, 17 (1), pp. 187-195.

HAUSMANN, F. J. (1985) : « Kollokationen im Deutschen Wörterbuch. Ein Beitrag zur Theorie des Lexikographischen Beispiels », Bergenholtz & Mugdan (Eds), *Lexikographie und Grammatik*, Tübingen, Niemeyer, pp. 118-129.

HAUSMANN, F. J. (1989) : « Le dictionnaire des collocations », Hausmann, Reichmann, Wiegand & Zgusta (Eds), *Wörterbücher, Dictionaries, Dictionnaires : An International Encyclopedia of Lexicography*, Berlin and New York, Walter de Gruyter, pp. 1010-1019.

HEID, U. (1991) : *EUROTRA-7. Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerized Applications. Intermediate report* (non publié).

HEID, U. (1992) : « Décrire les collocations Deux approches lexicographiques et leur application dans un outil informatisé », *Terminologie et traduction*, 2/3, pp 523-548

HEID, U. (1992a) : « Décrire les collocations : deux approches lexicographiques et leur application dans un outil informatisé », *Terminologie et Traduction*, Commission des Communautés européennes, pp. 523-548.

HEID, U. (1992b) : « Notes on the Use of Lexical Functions for the Description of Collocations in an NLP Lexicon », K. Haenelt & L. Wanner (Eds), *International Workshop on the Meaning-Text Theory*, Darmstadt, Schloss Birlinghoven, G.M.D., pp. 217-229.

HEID, U. (1994) : « On Ways Words Work Together - Topics in Lexical Combinatorics », *EURALEX'94 Proceedings*, Vrije Universiteit Amsterdam, pp. 226-257

HEID, U. et G. FREIBOTT (1991) : « Collocations dans une base de données terminologique et lexicale », *Meta*, 36 (1), pp. 77-91

HEYLEN, A., MAXWELL, A. K. et S. WARWICK-ARMSTRONG (1989) : « Collocations, Dictionaries and Machine Translation », *Proceedings of the AAAI Symposium on Machine Translation and the Lexicon*, Stanford (CA).

HEYLEN, D. et K. MAXWELL (1994) : « Lexical Functions and the Translation of Collocations », *Proceedings of EURALEX '94*, Amsterdam, pp 298-305.

HOCKEY, S. (1988) : « Creating and Using Large Text Databases for Scholarly Research in the Humanities Some Practical Issues », Gignoni et C. Peters (Eds), *Computational Lexicology and Lexicography*, Pisa, L. Giardini Editori e Stampatori n° 1, coll. « Linguistica computazionale », vol. VI

HUDSON, R. (1988) : « The Linguistic Foundation for Lexical Research and Dictionary Design », *International Journal of Lexicography*, 1 (4), pp. 287-312

ICHIKAWA, S. (1990) : *New Japanese-English Dictionary*, Kenkyuusha

IDE, N., LE MAITRE, J. & J. VERONIS (1994) : « Outline of a Model for Lexical Databases », Zampolli, Calzolari & Palmer (Eds), *Current Issues in Computational Linguistics : in Honour of Don Walker*, Series « Linguistica Computazionale », IX-X, pp. 283-320.

IDE, N., VERONIS, J., WARWICK-ARMSTRONG, S. et N. CALZOLARI (1992) : « Principles for Encoding Machine-readable Dictionaries », *EURALEX'92 Proceedings*, Tampere (Finlande), pp. 239-246

IORDANSKAJA, L., KIM, M. et A. POLGUÈRE (1996) : « Some Procedural Problems in the Implementation of Lexical Functions for Text Generation », L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam, John Benjamins.

ISO 1087 (1990) : *Terminology - Vocabulary = Terminologie - Vocabulaire*, Genève, Organisation internationale de normalisation.

ISO 1951 (1973) : *Symboles lexicographiques particulièrement pour l'emploi dans les vocabulaires systématiques à définitions*, Genève, Organisation internationale de normalisation.

ISO 704 (1987) : *Principes et méthodes de la terminologie*, Genève, Organisation internationale de normalisation.

ISO R 1087 (1969) : *Vocabulaire de la terminologie*, Genève, Organisation internationale de normalisation

JACKENDOFF, Ray (1975) : « Morphological and Semantic Regularities in the Lexicon », *Language*, vol. 31, pp. 639-671.

JACQUEMIN, Christian (1991) : *Transformations des noms composés*, thèse de doctorat en informatique fondamentale, Université Paris 7.

JACQUEMIN, Christian (1994) : « FASTR: A unification-based front-end to automatic indexing », *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'94)*, New York.

JANSEN, J. & T. FONTENELLE (1994) : *Short Description of an Implementation of the Robert & Collins English-French Dictionary under Database Format*, Deliverable D-2a of the DECIDE Project, Liège.

JOUIS, C. (1993) : *Contributions à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype : le système SEEK*, Thèse de doctorat de l'École des hautes études en sciences sociales, Paris, mars 1993.

JOUIS, C. (1994) : « Contextual Approach: SEEK, a Linguistic and Computational Tool for Use in Knowledge Acquisition », *Proceeding of the First European Conference "Cognitive Science in Industry"*, 28th-30th September 1994, Luxembourg, pp. 259-274.

JOUIS, C. (1995) : « SEEK, un logiciel d'acquisition des connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe », *Actes des 6es Journées Acquisition, Validation, (JAVA 95)*, Grenoble, Avril 1995, pp. 159-172.

JOUIS, C., COMPAGNON, F., GAUDINAT, B., ROUSSEAU, J.-M. et C. TORA (1991) : « Metodac : Une méthodologie pour l'acquisition et la modélisation des connaissances », *Actes du 8e congrès RFIA*, AFCET, Villeurbanne, nov. 1991, vol. 1, pp. 35-44.

JOUIS, C. et W. MUSTAFA-ELHADI (1995) : « Conceptual Modeling of Database Schema Using Linguistic Knowledge. Application to Terminological Databases », *First Workshop on Application of Natural Language to Databases (NLDB'95)*, Versailles, Juin 1995, pp. 103-118.

JUDGE, Anne et Solange LAMOTHE (1995) : *Stylistic Developments in Literary and Non-literary French Prose*, coll. « Studies in French Literature », Vol 19., Lampeter, The Edwin Mellen Press Ltd.

KARP, D., SCHABES, Y., ZAIDEL, M. et D. EGEDI (1992) : « A Freely Available Wide Coverage Morphological Analyzer for English », *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes (France), 23-28 juillet, pp. 950-955.

KARTTUNEN, L. (1993) : *Finite-State Lexicon Compiler*, Research Report, XEROX Palo Alto Research Center.

KARTTUNEN, Lauri et Kenneth BEESLEY (1992) : « Two Level Rules Compiler », *Technical Report ISTL-92-2*, Xerox Palo Alto Research Center.

KARTTUNEN, Lauri et Todd YAMPOL (1993) : « Interactive Finite-State Calculus », *Technical Report ISTL-NLTT-1993-04-01*, Xerox Palo Alto Research Center

KAVI, M., (en préparation) : *Ontology Development : Ideology and Methodology*, Computing Research Laboratory, New Mexico State University.

KAY, M. et M. ROSCHEISENJ (1993) : « Text Translation Alignment », *Computational Linguistics*, 19 (1), pp. 121-142.

KAYE, Jonathan Derek (1974) : « Morpheme Structure Conditions Live », *Recherches Linguistiques Montréal*, no 3, pp. 55-62.

KAYSER, D. (1987) : « Une sémantique qui n'a pas de sens », *Langages*, no 87, Paris, Larousse, pp. 33-45

KAYSER, Daniel et Pierre LERAT (1990) : « La notion de définition dans les systèmes de traitement du langage naturel », *La définition*, coll. « Langue et Langage », Paris, Larousse, pp. 113-124.

KEMBLE, I. R. (1991) : « Lexicography », *Computers as a Tool in Language Teaching*, Brierly, W and I. R. Kemble (Eds), Chichester, Ellis Horwood.

KEY SUN CHOI YOUNG, S. HAN (1992) : « Syntactic Analysis Based Automatic Indexing for Korean Texts », *International Conference on Terminology, Standardization and Technology Transfer*, Pekin, Science Press.

KIEFER, F. (1973) : « Morphology in Generative Grammar », M. Gross *et al.* (Eds), *The Formal Analysis of Natural Languages*, The Hague.

KIPARSKY, P. et C. KIPARSKY (1979) : « Fact », Bierwisch et Heidolph (Eds), *Progess in Linguistics*, The Hague, Mouton.

KITTREDGE, R. (1983) : *Sublanguage – Specific Computer Aids to Translation – a survey of the most promising areas*, Contract n° 2-5273, Université de Montréal et Bureau des traductions, mars 1983.

KLEIBER, G. et I. TAMBA (1990) : « L'hyponymie revisitée : inclusion et hiérarchie », *Langages*, n° 98, pp 7-32

KNOWLES, Frank E. (1993) : « Review of *English Adverbial Collocations* », (voir Kozlowska), *International Journal of Lexicography*, 6 (4), pp 300-302.

KOCOUREK, R. (1991) : *La langue française de la technique et de la science : vers une linguistique de la langue savante*, 2e édition, Wiesbaden, Oscar Brandsetter Verlag, 259 p.

KOINE, Y. (1990) : *New English-Japanese Dictionary*, Kenkyuusha.

KOSKENNIEMI, K. (1983) : *Two-Level Morphology : A General Computational Model for Word-Form Recognition and Production*, Research Report, University of Helsinki, University of Helsinki, Department of General Linguistics. Publications 11.

KOZLOWSKA, Christian D. (1991) : *English Adverbial Collocations*, PWN, Varsovie.

KRIEGER, H.-U. (1994) : « Derivation without lexical rules », C. Rupp *et al.* (Eds), *Constraints, Language and Computation*, Academic Press, pp. 277-313

KRIEGER, H.-U. et J. NERBONNE (1993) : « Feature-Based Inheritance Networks for Computational Lexicons », *Research Report 31*, Saarbrücken, Deutsches Forschungszentrum für Künstliche Intelligenz.

KUPIEC, J. (1992) : « Robust Part-of-Speech Tagging Using a Hidden Markov Model », *Computer Speech and Language*, 6, pp 225-242.

LADL (1993) : *Outils de traitement linguistique, applications à l'analyse documentaire*, Rapport Technique LADL #43.

LAFON, Pierre (1984) : *Dépouillements et statistiques en lexicométrie*, Genève, Slatkine et Champion.

LAKOFF, G. (1987) : *Women, Fire and Dangerous Things*, Chicago.

LALLICH-BOIDIN, G., HENNERON, G. et R. PALERMITI (1990) : *Analyse du français. Achèvement et implantation de l'analyseur morpho-syntaxique*, Grenoble, Cahiers du CRISS, n° 16.

LAUER, M. (1994) : « Conceptual Association for Compound Noun Analysis », *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Student Session*, juin, Las Cruces, 6 p.

LAURISTON, A. (1994) : « Automatic Recognition of Complex Terms : Problems and the TERMINO Solution », *Terminology : International Journal of Theoretical and Applied Issues in Specialized Communication*, 1 (1), John Benjamins Publishing Company, pp 147-170.

LE NY, J.-F. (1989) : *Science cognitive et compréhension du langage*, Paris, PUF.

LEBART, L. et A. SALEM (1988) : *Analyse statistique des données textuelles . questions ouvertes et lexicométrie*, Paris, Dunod, 210 p.

LECLÈRE, C. (1990) : « Organisation du lexique-grammaire des verbes du français », *Langage*, 87, pp. 112-122.

LEHRER, A. (1990) : « Polysemy, Conventionality and the Structure of the Lexicon », *Cognitive Linguistics*, 1-2, pp. 207-246.

LELUBRE, X. (1992) : *La terminologie arabe contemporaine de l'optique : faits - théories - évaluation*, thèse de nouveau doctorat, Université Lyon-2, 546 p.

LERAT, P. (1988) : « Terminologie et sémantique descriptive », *La banque des mots*, numéro spécial, pp. 11-30.

LERAT, P. (1990) : « L'hyperonymie dans la structuration des terminologies », *Langages*, n° 98, pp. 79-86.

LERAT, P. (1990) : « Sélection et analyse de termes nouveaux dans une base de données prédictionnairiques », *Cahiers de lexicologie*, n° 56-57, Paris, Didier Érudition, pp 255-260.

LERAT, P. (1995) : *Les langues spécialisées*, coll. « Linguistique nouvelle », Paris, PUF, 206 p.

LEROUX, D., MINEL, J.-L. et J. BERRI (1994) : « Seraphin Project (Expert System for Automatic Marking of Important Sentences in a Text) the Industrial Approach », *Proceeding of the First European Conference "Cognitive Science in Industry"*, 28-30 Sept. 1994, Luxembourg.

LEROY-TURCAN, I. (1994a) : « L'informatisation du *Dictionnaire étymologique ou Origines de la langue françoise* de Gilles Ménage (1694) », I. Lancashire et T. R. Wooldridge (Éd.), *Early Dictionary Databases*, University of Toronto, Centre for Computing in the Humanities, pp. 131-142. = T. R. Wooldridge (Éd.), *Informatique et dictionnaires anciens*, Paris, Didier Érudition, 1995, pp. 131-142.

LEROY-TURCAN, I. (1994b) : « Intérêt d'une base informatisée pour le *Dictionnaire étymologique ou Origines de la langue françoise* de Ménage (*DEOLF* 1694) ; les modalités de mise en œuvre », *Actes* du Séminaire sur *Dictionnaires historiques et dictionnaires anciens : problèmes de méthode et d'édition*, CNRS-HESO, Ivry-sur-Seine, déc. 1994. À paraître

LEROY-TURCAN, I. et T. R. WOOLDRIDGE (1995) : « L'informatisation des premiers dictionnaires de langue française : les difficultés propres à la première édition du *Dictionnaire de l'Académie française* », J Pruvost (dir.), *Actes de la Journée des dictionnaires*, Université de Cergy-Pontoise (1995).

LEVIN, B. (1993) : *English Verb Classes and Alternations (A Preliminary Investigation)*, The University of Chicago Press.

LEVRAT, B. (1993) : *Le problème du sens dans les systèmes de traitement automatique du langage naturel : une approche alternative au travers de la paraphrase*, thèse de doctorat d'État, Université de Paris-Nord Villetaneuse, 214 p

LEVRAT, B. et G. SABAH (1990) : « 'Sorte de', une façon de rendre compte de la relation d'hyponymie/hyperonymie dans les réseaux sémantiques », *Langages*, n° 98, pp. 87-102.

LEXIS (1987) : *Dictionnaire de la langue française Lexis*, Paris, Larousse.

LIEBERT, W. A. (1994) : « Lascaux – a Hypermedia lexicon of Metaphor Models for Scientific Imagination », W. Martin *et al.* (Eds), *Euralex 1994*, Amsterdam, pp. 494-500

LUN, S. (1983) : « A Two-Level Morphological Analysis of French », *Texas Linguistic Forum*, 22, pp. 271-278.

LYONS, J. (1970) : *Linguistique générale. Introduction à la linguistique théorique*, Paris, Larousse, coll. « Langue et langage ».

LYONS, J. (1977) : *Semantics*, vol. I, Cambridge, London, New York, Melbourne, Cambridge University Press.

MACLEOD, C. et R. GRISHMAN (1994) : *COMLEX Syntax Reference Manual.*

MANNING, C. D. (1993) : « Automatic Acquisition of a Large Subcategorization Dictionary from Corpora », *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus (Ohio), 22-26 juin, pp. 235-242.

MARCUS, M., KIM, G., MARCINKIEWICZ, M. A., MACINTYRE, R., BIES, A., FERGUSON, M., KATZ, K. et B. SCHASBERGER (1994) : « The Penn Treebank : Annotating Predicate Argument Structure », *Proceedings of the Human Language Technology Worskshop*, Plainsboro (N. J.), 8-11 mars, pp 114-119.

MARTIN, John R. (1993) : *Life as a Noun . Arresting the Universe in Science and Humanities*, London, Falmer Press Ltd.

MARTIN, R. (1994) : « Dictionnaire informatisé et traitement automatique de la polysémie », E. Martin (Éd.), *Textes et informatique*, coll. « Études de sémantique lexicale », Paris, CNRS-INaLF, Didier Érudition, pp. 77-113.

MASSON, N. (1995) : « An Automatic Method for Text Structuring », *Proccedings of the 18th Annual International ACM SIGIR Conference on Research an Development in Information Retrieval*, Edward A. Fox, Peter Ingwersen and Raya Fidel (Eds), Seattle (WA), July 9-13, pp. 372-373.

MATHIEU, J.-P., KASTLER, A. et P. FLEURY (1985) : *Dictionnaire de physique*, 2e édition, Paris, Masson/Eyrolles, 568 p. (MKF)

MATHIEU-COLAS, Michel (1988) : « Typologie des noms composés », *Rapport technique n° 7, Programme de Recherches Coordonnées « Informatique et Linguistique »*, Université Paris 13.

MATHIEU-COLAS, M. (1994) : *Les mots à trait d'union. Problèmes de lexicographie informatique*, coll « Études de sémantique lexicale », Paris, CNRS-INaLF, Didier Érudition, 351 p.

MATTHEWS, PH. (1974) : *Morphology · An Introduction to the Theory of Word-structure*, Cambridge, Cambridge University Press.

MAUREL, Denis (1993) : « Passage d'un automate avec tables d'acceptablilité à un automate lexical », *Actes du colloque Informatique et langue naturelle*, Nantes, pp. 269-279.

Mc CARTHY, J. J. (1979) : *Formal Problems in Semantic Phonology and Morphology*, Ph.D., MIT, Inédit

McCAWLEY, N. A. (1976) : « On Experiencer Causatives », Shibatani (Ed.), *The Grammar of Causative Constructions*, coll. « Syntax and Semantics », vol. 6, New York.

McNAUGHT, J., NKWENTI-AZEH, B., MARTIN, W. et E. TEN PAS (1991) : *EUROTRA-7 Study DOC-11 Feasibility of Standards for Terminological Description of Lexical Items*, Final Version (non publié).

MEL'ČUK, I. (1981) : « Meaning-Text Models · a Recent Trend in Soviet Linguistics », *Annual Review of Anthropology*, 10, pp. 27-62.

MEL'ČUK, I. A. (1982) : « Lexical Functions in Lexicographic Description », *Proceedings of the Eighth Annual Meeting of the Berkeley Linguistics Society*, Berkeley, UCB, pp. 427-444.

MEL'ČUK, I. A. (1982a) : *Towards a Language of Linguistics, A System of Formal Notions for Theoretical Morphology*, München, Wilhem Fink Verlag.

MEL'ČUK, I. A. (1982b) : « Élaboration d'un langage formel pour la morphologie », C. Bertaux, J -P. Desclés, D. Dubarle *et al.*, *Linguistique et mathématiques Peut-on construire un discours cohérent en linguistique ?*, Berne, Peter Lang Verlag, pp. 99-119

MEL'ČUK, I. A. (1988) : « Principes et critères de description sémantique dans le DEC », *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II*, Montréal, Les Presses de l'Université de Montréal, pp. 27-39.

MEL'ČUK, I. A. (1994) : « Collocations and Lexical Functions », *Proceedings of the Leeds Colloquium on Collocations.*

MEL'ČUK, I. A. (1994) : « Typologie des phrasèmes et leur présentation dans un dictionnaire de langue », Conférence plénière au Colloque International *La Locution*, Paris, ENS St-Cloud.

MEL'ČUK, I. A. (1995a) : « Phrasemes in Language and Phraseology in Linguistics », M. Everaert et R. Schrender (Eds), *Idioms : Structural and Psychological Perspectives*, Hillsdale/Hove, Lawrence Elbaum Associates, pp. 167-232.

MEL'ČUK, I. A. (1995b) : « The Future of the Lexicon in Linguistic Description . The Explanatory Combinatorial Dictionary », *Linguistics in the Morning Calm 3. Selected Papers of Seoul International Conference on Linguistics 1992*, Seoul, Hanshin Publishing Company, pp. 181-270.

MEL'ČUK, I. A. (1996) : « Lexical Functions : A Powerful Tool for the Description of Lexical Relations in a Lexicon », L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam, John Benjamins, pp. 37-102.

MEL'ČUK, I. A., CLAS A. et A. POLGUÈRE (1995) : *Introduction à la lexicologie explicative et combinatoire*, coll. « Universités francophones » et « Champs linguistiques », Louvain-la-Neuve, AUPELF-UREF et Duculot, 256 p.

MEL'ČUK, I. et A. POLGUÈRE (1987) : « A Formal Lexicon in the Meaning-Text Theory (or how to do lexica with words) », *Computational Linguistics*, 13, pp 261-275.

MEL'ČUK, I. *et al.* (1984) : *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques I*, Montréal (Québec), Canada, Presses de l'Université de Montréal

MEL'ČUK, I. *et al.* (1988) : *Dictionnaire explicatif et combinatoire du français contemporain Recherches lexico-sémantiques II*, Montréal, Presses de l'Université de Montréal

MEL'ČUK, I. *et al.* (1992) : *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques III*, Montréal (Québec), Canada, Presses de l'Université de Montréal.

MEUNIER, J.-G. (1970) : Communication personnelle.

MEYER, C. (1993) : *Correction orthographique et grammaticale : vers une machine à dicter,* Rapport de diplôme, Université de Neuchâtel.

MEYER, I. et B. MCHAFFIE (1993) : « De la focalisation à l'amplification . nouvelles perspectives de représentation des données terminologiques », *TA-TAO : recherches de pointe et applications immédiates Actes du colloque de Montréal.* Beyrouth, AUPELF-UREF et FMA, pp. 425-440

MEYER, I., ONYSHKEVYCH, B. et L. CARLSON (1990) : *Lexicographic Principles and Design for Knowledge-based Machine Translation,* Technical Report CMU-CMT-90-118, Carnegie Mellon University.

MICHIELS, A. (1995) : « Feeding LDOCE Entries into HORATIO », Alberto & Bennett (Eds), *Lexical Issues in Machine Translation*, Studies in Machine Translation and Natural Language Processing, Vol 8. Luxembourg, European Commission, pp 93-115.

MICHIELS, A. & J. NOËL (1984) : « The Pro's and Con's of a Controlled Defining Vocabulary in a Learner's Dictionary », *LEXeter'83 Proceedings*, Tübingen, Max Niemeyer, pp. 386-394.

MICLET, L. (1980) : « Regular Inference with a Tail-Clustering Method », *I.E.E.E. Trans. on Systems, Man and Cybernetics*, 10, pp. 737-743.

MILLER, A. G., FELLBAUM, C. et D. GROSS (1989) : « WORDNET a Lexical Database Organised on Psycholinguistic Principles », Zernik (Ed.), *Proceedings of the First International Lexical Acquisition Workshop, IJCAI*, Détroit.

MILLER, G. A. (1990) : « Nouns in Wordnet . a Lexical Inheritance System », *International Journal of Lexicography*, 3 (4), pp. 245-264.

MILLER, G. A. (dir.) (1990) : « WordNet : An On-Line Lexical Database », *International Journal of Lexicography*, special issue, 3 (4), pp. 235-312.

MILLER, G. A. et C. FELLBAUM (1991) : « Semantic Networks of English », *Cognition*, n° 41, pp. 197-229.

MILNER, Jean-Claude (1987) : *Introduction à une science du langage*, coll. « Des Travaux », Paris, Le Seuil.

MONTEIL, Vincent (1960) : *L'arabe moderne*, Paris, Klincksiek (thèse de doctorat)

MOODY, M. (1978) : « Some Preliminaries to a Theory of Morphology », *Glossa*, 12, pp. 16-38

MOUSEL, P. et G. THIENPONT (1994) : *Technical Specification of the ANTHEM Prototype*, ANTHEM Deliverable n. D1-2.

MULTEXT (1994) : *Common Specifications and Notation for Lexicon Encoding*, MULTEXT Report WP1.6.

MUSTAFA-ELHADI, W. (1989) : *Terminologie arabe des télécommunications : théorie et faits de variation*, Thèse de doctorat en linguistique, Université Lyon II.

MUSTAFA-ELHADI, W. (1990) : « The Contribution of Terminology to the Theoretical Conception of Classificatory Languages and Thesaurus Indexing », *Tools for Knowledge Organization and the Human Interface, Advances in Knowledge Organization*, vol. 1, Frankfurt, Index VERLAG, pp. 98-106.

MUSTAFA-ELHADI, W. (1993) : *État de l'art de l'extraction de concepts à partir du langage naturel*, Rapport SGDN/STS/VST/5 & CREDO (CNRS-URA 1423), Université Charles De Gaulle-Lille III.

MYAENG, S. H., KHOO, C. et M. LI (1994) : « Linguistic Processing of Text for a Large-Scale Conceptual Information Retrieval System », *Proceedings of the 2nd International Conference on Conceptual Structures*, Maryland (USA), août, dans *Lecture Notes in Artificial Intelligence*, n° 835, Springer-Verlag, Tepfenhart, W. M., Dick, J. P. et Sowa, J. F. (dir.), pp. 69-83.

NIDA, E. A. (1975) : *Componential Analysis of Meaning and Introduction to Semantic Structure*, The Hague, Mouton, 272 p.

NIRENBURG, S., RASKIN, V. et B. ONYSHKEVYCH (1994) : *Apologiae ontologiae*, Memoranda in Computer and Cognitive Science MCCS-95-281, New Mexico State University, Computing Research Laboratory

NIWA, Yoshiki et Yoshihiko NITTA (1995) : « Co-occurence vectors from corpora vs. distance vectors from dictionaries », *The Computation and Language E-Print Archive*, sur Internet.

NOALLY, Michèle (1990) : *Le substantif épithète*, Paris, PUF.

Nouveau Petit Robert (1993) : *Le nouveau petit Robert Dictionnaire alphabétique et analogique de la langue française. Nouvelle édition remaniée et amplifiée*, Paris, Le Robert

NUNBERG, Geoff, WASOW, Thomas et Ivan SAG (1994) : « Idioms », *Language*.

Oxford-Hachette French Dictionary (1994) : Oxford University Press.

Observatoire wallon des industries de la langue (1993) : *Des industries de la langue pour quoi faire ? Technologies, usages et marchés*, A. Moulin (dir.), Liège, Université de Liège.

OGONOWSKI, A., HERVIOU, M. L. et E. DAUPHIN (1994) : « Tools for Extracting and Structuring Knowledge from Texts », *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto (Japon), 5-9 août, pp. 1049-1053.

OH, H.-G., JOUIS, C., MAIRE-REPPERT, D. et J. BERRI (1992) : « Analyse de textes par exploration contextuelle : application aux problèmes des temps et à l'extraction des connaissances », *ECCO-92*, Orsay, 29 juin-1er juillet 1992, pp. 193-210

ONYSHKEVYCH, B. et S. NIRENBURG (1994) : *The Lexicon in the Scheme of KBMT Things*, Technical Report MCCS-94-277, Computing Research Laboratory, New Mexico State University.

ORACLE Corporation (1993) : *ORACLE, SQL*Forms, SQL*Plus, PL/SQL, Text*Retrieval and so on*, Redwood City.

OTMAN, G. (1989) : « Terminologie et intelligence artificielle », *La Banque des mots*, pp. 63-95.

OTMAN, G. (1995) : *Les représentations sémantiques en terminologie*, thèse de doctorat, Paris IV Sorbonne, 357 p

PAASCH, H. (1901) : *From Keel to Truck = De la quille à la pomme de mât = Vom Kiel zum Flaggenknopf. Dictionnaire de marine en anglais, français et allemand illustré de nombreux dessins explicatifs [...]*, 3e édition, Anvers, H. Paasch, Hamburg, Eckardt & Messtorff.

PAJZS, J. (1990) : « Számítógép és lexikográfia » (Ordinateur et lexicographie), *Linguistica*, Series A., 4, Budapest, MTA Nyelvtudományi Intézet (Institut de Linguistique de l'Académie des Sciences de Hongrie).

PERENNOU, G. (1988) : « Le projet BDLEX de base de connaissances lexicales et phonologiques », *Actes des Premières Journées du GDR-PRC Communication Homme-Machine*

PETITPIERRE, D., ROBERT, D. & S. WARWICK-ARMSTRONG (1994) : « DICO : A Network-Based Dictionary Consultation Tool », poster presented at the EURALEX'94 International Congress, Vrije Universiteit Amsterdam.

PETITPIERRE, P. et G. ROBERT (1995) : *DICO, Technologiestandort Schweiz*, CeBIT 1995, Hanover.

PETITPIERRE, P., ROBERT, G. et S. ARMSTRONG (1994) : *Design of an On-line Dictionary Consultation Tool.*

PICABIA, L. (1978) : *Les constructions adjectivales en français. Systématique transformationnelle*, Genève et Paris, Droz.

PICCHI, E., PETERS, C. & E. MARINAI (1992) : « The Pisa Lexicographic Workstation : The Bilingual Components », Tommola, Varantola, Salmi-Tolonen & Schopp (Eds), *EURALEX'92 Proceedings I-II, Fifth EURALEX International Congress*, Studia Translatologica, Ser A, Vol. 1, University of Tampere, pp. 277-285.

PICHT, H. (1987) : « Terms and their LSP Environment – LSP Phraseology », *Meta*, 23 (2), pp. 149-155.

POLLARD, C. & I. A. SAG (1987) : *Information-Based Syntax and Semantics*, Stanford, CSLI Lecture Notes No. 13, 233 p.

POLLARD, C. et I. SAG (1994) : *Head-Driven Phrase Structure Grammar*, CSLI/University of Chicago Press.

PROCTER, P. (Ed.) (1978) : *Longman Dictionary of Contemporary English*, 2nd edition edited by D. Summers, Harlow, Longman Group Ltd.

PROCTER, P. (Ed.) (1995) : *Cambridge International Dictionary of English*, Cambridge University Press.

PUGEAULT, F., SAINT-DIZIER, P. et M.-G. MONTEIL (1994) : « Knowledge Extraction from Texts : A Method for Extracting Predicate-argument Structures from Texts », *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto (Japon), 5-9 août, pp 1039-1043.

PUSTEJOVSKY, J. (1991) : « The Generative Lexicon », *Computational Linguistics*, 17(1)

PUSTEJOVSKY, J. (Ed.) (1993) : *Semantics and the Lexicon*, Dordrecht, Kluwer.

PUSTEJOVKY, J. (1995 à paraître) : *The Generative Lexicon*, Cambridge, MIT.

PUSTEJOVSKY, J. et P. BOUILLON (1995) : « Aspectual Coercion and Logical Polysemy », *Journal of Semantics*, vol. 2

PUSTEJOVSKY, J. et S. BERGLER (1991) : « Lexical Semantics and Knowledge Representation », *Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the Association for Computational Linguistics*, 17 June 1991, University of California, Berkeley, California, USA

QUILLIAN, M. R. (1967) : « Word Concepts : a Theory and Simulation of some Basic Semantic Capabilities », *Behavioral Science*, 12 (5), pp. 410-443.

QUILLIAN, R. (1968) : « Semantic Memory », *Semantic Information Processing*, M. Minsky (Ed.), Cambridge (Mass.), MIT Press, pp. 227-270.

QUIRK, R., GREENBAUM, S., LEECH, G. et J. SVARTVIK (1994) : *A Comprehensive Grammar of the English Language*, London et New York, Longman, 12e édition

RAHMSTORF, G. (1993) : « Role and Representation of Terminological Definitions », *Actes* du colloque *Terminology & Knowledge Engineering*, Cologne, 25-27 août 1993, pp. 39-48.

RAO, R., PEDERSEN, J. O., HEARST, M. A., MACKINLAY, J. D., CARD, S. K., MASINTER, L., HALVORSEN, P.-K. et G. G. ROBERTSON (1995) : « Rich Interaction in the Digital Library Communication », *ACM*, April 1995, 38 (4), pp. 29-39.

RAPP, R. (1995) : « Identifying Word Translations in Non-Parallel Texts », *Proceedings of ACL '95.*

RASTIER, François (1987) : *Langages : sémantique et intelligence artificielle*, Bernard Willerval Jouve, 14192 édition.

RASTIER, F. (1987) : *Sémantique interprétative*, Paris, Presses Universitaires de France.

RASTIER, F. (1991) : *Sémantique et recherches cognitives*, Paris, Presses Universitaires de France, 262 p.

RASTIER, F. (1995) : « Le terme entre ontologie et linguistique », *La Banque des Mots*, Numéro spécial du Centre de Terminologie et de Néologie du CNRS, actes de la Première Journée « Terminologie et Intelligence Artificielle », Paris, Villetaneuse.

RASTIER, F., CAVAZZA, M. et A. ABEILLÉ (1994) : *Sémantique pour l'analyse – De la linguistique à l'informatique*, Paris, Masson

REDDY, M. J. (1979) : « The Conduit Metaphor – a Case of Frame Conflict in our Language about Language », Andrew Ortony (Ed.), *Metaphor and Thought*, Cambridge University Press, pp. 284-324.

REN, X. et F. PERRAULT (1992) : « The Typology of Unknown Words : An Experimental Study of Two Corpora », *Proceedings of COLING-92*, Nantes.

RESNIK, Philip (1995) : « Using Information Content to Evaluate Semantic Similarity in a Taxonomy », *Actes de IJCAI'95*.

REY-DEBOVE, J. (1971) : *Étude linguistique et sémiotique des dictionnaires français contemporains*, La Haye/Paris, Mouton.

RILOFF, E. (1993) : « Automatically Constructing a Dictionary for Information Extraction Tasks », *Proceedings of the 11th National Conference on Artificial Intelligence*, Washington (DC), 11-15 juillet, pp. 811-816.

RITCHIE, G. D., RUSSELL, G. J., BLACK, A. W. & S. G. PULMAN (1992) : *Computational Morphology : Practical Mechanisms for the English Lexicon*, Cambridge (MA), The MIT Press, 291 p.

Robert (1994) : *Le Nouveau Petit Robert, dictionnaire alphabétique et analytique de la langue française*, Paris, Dictionnaires Le Robert.

ROCHETTE, A. (1988) : *Semantic and Syntactic Aspects of Romance Sentential Complementation*, PhD thesis, Massachusetts Institute of Technology

ROMAN, A. (1993) : « La voie des hypertextes ? », P. J. L. Arnaud & Ph. Thoiron (dir), *Aspects du vocabulaire*, Lyon, Presses universitaires de Lyon, pp. 103-132.

RONDEAU, G. (1981) : *Introduction à la terminologie*, Centre éducatif et culturel, Québec

ROULEAU, M. (1994) : *La traduction médicale (une approche méthodique)*, Brossard (Canada), Linguatech.

RUESSINK, H.-A. (1990) : « Two-level Formalisms », Coopmans, P., Schouten, B , Zonneveld, W (Eds), *OTS Yearbook*, University of Utrecht.

RUSSELL, G., BALLIM, A., CARROLL, J. et S. WARWICK-ARMSTRONG (1992) : « A Practical Approach to Multiple Default Inheritance for Unification-based Lexicons », *Computational Linguistics*, 18 (3).

SABAH, G. (1988/89) : *L'intelligence artificielle et le langage*, vol 1 : *Représentation des connaissances* (1988), vol 2 : *Processus de compréhension* (1989), Paris, Hermès.

SAGER, J. C. (1982) : « Definitions in Terminology », *Problèmes de la définition et de la synonymie en terminologie*, Actes du colloque international de terminologie, Université Laval, Québec, mai 1982, pp 113-140

SAGER, J. C. (1990) : *A Pratical Course in Terminology Processing*, Amsterdam, John Benjamins Publishing Company, 254 p.

SALTON, Gerard (1988) : *Term-weighting Approaches in Automatic Text Retrieval.*

SALTON, Gerard et J. M. MCGILL (1983) : *Introduction to Modern Information Retrieval*, McGraw-Hill Computer Science Series, New York, McGraw-Hill.

SANFILIPPO, A. (Ed.) (1992) : *The (Other) Cambridge Acquilex papers*, Technical Report n° 253, University of Cambridge Computer Laboratory, New Museums Site

SANFILIPPO, A. (1994) : « Word Knowledge Acquisition, Lexicon Construction and Dictionary Compilation », *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto (Japon), 5-9 août, pp. 273-277.

SANFILIPPO, A. et V. POZNANSKI (1992) : « The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources », *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento (Italy), 31 mars-3 avril, pp. 80-87.

SCHMITT, L., OLIVAN, E., LANDI, B., ROYAUTÉ, J. et J. DUCLOY (1992) : « STDI : Une Station de travail pour une indexation assistée », *Natural Language Processing and its Applications*, Avignon

SEFFAH, A. et J.-G. MEUNIER (1995) : « Un atelier génie logiciel orienté objets pour l'analyse cognitive de texte », *Actes du congrès JADT*, Rome.

SEGOND, F. & A. ZAENEN (1994) : « Multi-word Expressions in Bilingual Dictionaries and in Compass », paper read at the workshop on « *The Future of the Dictionary* » co-sponsored by Rank Xerox Research Centre and Acquilex-II, Grenoble.

SEGOND, Frédérique et Pasi TAPANAINEN (1995) : « Using a Finite-state Based Formalism to Identify and Generate Multiword Expressions », *Technical Report MLTT-019*, Rank Xerox Research Centre, Grenoble, July 1995

SEKINE, S., CARROLL, J. J., ANANIADOU, S. et J. TSUJII (1992) : « Automatic Learning for Semantic Collocation », *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento (Italy), 31 mars-3 avril, pp. 104-110

SÉRASSET, G. (1994) : *SUBLIM : un système universel de bases lexicales multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions*, Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1, 194 p.

SÉRASSET, G. (1994a) : « Internal Lexical Organization for Multilingual Lexical Databases », *15th International Conference on Computational Linguistics, COLING-94*, Kyoto, August 5-9.

SHIBATANI, M. (Ed.) (1976) : *The Grammar of Causative Constructions*, coll. « Syntax and Semantics », vol. 6, New York.

SHIEBER, S. M. (1986) : *An Introduction to Unification-based Approaches to Grammar*, Stanford (CA), Stanford University CSLI, 105 p

SIEGEL, D. (1974) : *Topics in English Morphology*, unpublished Doctoral dissertation, MIT.

SILBERZTEIN, M. (1993) : *Dictionnaires électroniques et analyse automatique de textes . le système INTEX*, Paris, Masson, XIV + 233 p.

SINCLAIR, John (Ed.) (1987) : *Collins Cobuild English Language Dictionary*, London et Glasgow, Collins.

SINCLAIR, J. (1987) : *Looking UP. An account of the COBUILD project in lexical computing*, London, Collins Cobuild.

SINCLAIR, J. (1991) : *Corpus, concordance, collocations*, Oxford, Oxford University Press.

SLODZIAN, M. (1994) : « La doctrine terminologique, nouvelle théorie du signe au carrefour de l'universalisme et de du logicisme », *Actes de langue française et linguistique*, volume 7/8

SLODZIAN, M. (1995) : « Comment revisiter la doctrine terminologique aujourd'hui ? », *La Banque des Mots*, Numéro spécial du Centre de Terminologie et de Néologie du CNRS, actes de la Première Journée "Terminologie et Intelligence Artificielle", Paris, Villetaneuse.

SMADJA, F. (1991) : « Macrocoding the Lexicon with Co-occurrence Knowledge », Zernik (Ed.), *Lexical Acquisition : Using On-Line Resources to Build a Lexicon*, Hillsdale, Lawrence Erlbaum Associates, pp 165-189.

SMADJA, F. (1993) : « Retrieving Collocations from Text . Xtract », *Computational Linguistics*, 19 (1), pp. 143-177.

SMADJA, Frank A. et Kathleen R. MCKEOWN (1990) : « Automatically extracting and representing collocations for language generation », *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 252-259

SMADJA, F. et K. MCKEOWN (1991) : « Using Collocations for Language Generation », *Computational Intelligence*, 7 (4), pp. 229-239.

SNELL-HORNBY, M. (1988) : *Translation Studies . An Integrated Approach*, Amsterdam/ Philadelphia, John Benjamins.

SODERLAND, S., FISHER, D., ASELTINE, J. et W. LEHNERT (1995) : « CRYSTAL : Inducing a Conceptual Dictionary », *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montréal (Canada), 20-25 août, pp. 1314-1319.

SOMERS, H. L. (1987) : *Valency and Case in Computational Linguistics*, Edinburgh University Press.

SOWA, J. (1984) : *Conceptual Structures . Processing in Mind and Machine*, Reading (MA), Addison-Wesley.

SPERBERG-MCQUEEN, C. M. et L. BURNARD (Eds) (1994) : *Guidelines for Electronic Text Encoding and Interchange.*

STRAUSS, Steven L. (1979) : *Some Principles of Word Structure in English and German*, unpublished Doctoral dissertation, City University of New York, Graduate Center.

STRAUSS, Steven L. (1982) : *Lexicalist Phonology of English and German*, Dordrecht-Holland/ Cinnaminson-USA, Foris publications.

STREITER, O., HALLER, J., SHARP, R., SCHMITT-WIGGER, A. et C. PEASE (1994) : « Aspects of a Unification Based Multilingual System for Computer Aided Translation », *Proceedings of the 14th International Conference « Avignon '94 »*, May 30th-June 3rd 1994.

SURRIDGE, Marie E. (1985) : « Le genre grammatical des composés en francais », *Revue canadienne de linguistique/Canadian Journal of Linguistics*, 30 (3), pp. 246-271.

SWALES, John M. (1990) : *Genre Analysis*, Cambridge University Press.

TAIFI, M. (1988) : « Problèmes méthodologiques relatifs à la confection d'un dictionnaire du tamazight », *Awal, Cahiers d'Études Berbères*, n° 4, Paris, Awal, pp. 15-26

TAIFI, M. (1989) : *Le lexique berbère (parlers du Maroc central) : formes, sens et évolution*, thèse de doctorat d'État, Université de Paris III, Sorbonne nouvelle, XLIX + 940 p

TAIFI, M. (1990a) : « Pour un théorie des schèmes en berbère », *Études et Documents Berbères*, n° 7, Paris, La Boîte à Documents, pp 92-110.

TAIFI, M. (1990b) : « L'altération des racines berbères : la diachronie dans la synchronie », *Awal. Cahiers d'Études Berbères, numéro spécial en hommage à Mouloud Mammeri*, Paris, Awal, pp. 219-232.

TAIFI, M. (1992) : *Dictionnaire tamazight-français (parlers du Maroc central)*, Paris, L'Harmattan/ Awal, XXII + 879 p.

TAIFI, M. (1995) : « Unité et diversité du berbère : détermination des lieux linguistiques d'intercompréhension », *Études et Documents Berbères*, n° 12, Paris, La Boîte à Documents/Edisud, pp 119-138.

TAKEBE, Y. *et al.* (1976) : *The Japanese Thesaurus*, Sanseidou.

TALMY, L. (1985) : « Lexicalization Patterns : Semantic Structure in Lexical Forms », Shopen, T. (Ed.), *Language Typology and Syntactic Description*, vol 3, Cambridge University Press.

TANAKA, K. et K. UMEMURA (1994) : « Constuction of a Bilingual Dictionary Intermediated by a Third Language », *Proceedings of the International Conference for Computational Linguistics '94*, pp. 293-393.

TAPANAINEN, Pasi (1994) : « RXRC Finite-State Rule Compiler », *Technical Report MLTT-020*, Rank Xerox Research Centre, Grenoble.

TARSKI, A. (1936) : « Der Wahrheitsbegriff in den formalisierten Sprachen », *Studia philosophica*, I.

TESNIÈRE, Louis (1959) : *Éléments de syntaxe structurale*, Paris, Librairie C. Klincksieck.

THIELE, Johannes (1987) : *La formation des mots en français moderne*, Montréal, Presses de l'Université de Montréal.

THOIRON, Ph. (1994) : « La terminologie multilingue . une aide à la maîtrise des concepts », *Meta*, 39 (4), déc. 1994, pp. 765-773

THOIRON, Ph. et H. BÉJOINT (1991) : « La place des reformulations dans les textes scientifiques », *Meta*, 36 (1), pp 101-110.

THOMAS, Patricia (1995) : *Orientation in Multiple Lexical Terms and Verb Phrases: A Model for Special Language Combinants*, Ph.D. Thesis, University of Surrey, Guildford, U.K.

TOKUNAGA, T. et H. TANAKA (1990) : « The Automatic Extraction of Conceptual Items from Bilingual Dictionaries », *PRICAI*.

TOMITA, M. (1984) : « Disambiguating Grammatically Ambiguous Sentences by Asking », *Proceedings 22nd Annual Meeting of the ACL*, Stanford, pp 476-480.

TOURETZKY, D. (1994) : « Continuity, Polysemy and Representation : Understanding the Verb *Cut* », Fuchs, C. and Victorri, B. (Eds), *Continuity in Linguistic Semantics*, Amsterdam, John Benjamins, pp. 231-240.

TROST, H. (1990) : « The application of two-level morphology to non-concatenative German morphology », *Proceedings of COLING-90*, vol. 2, Helsinki, pp. 371-376.

UTSURO, T. *et al.* (1994) : « Bilingual Text Matching Using Bilingual Dictionary and Statistics », *Proceedings of the International Conference for Computational Linguistics '94*, pp. 1076-1082.

VACHON-L'HEUREUX, Pierrette (1995) : « Table ronde sur les marques lexicographiques. Compte rendu », *Terminogramme*, n° 75, Office de la langue française, pp 1-6.

VAN CAMPENHOUDT, M. (1991) : « *TI*, le logiciel d'expérimentation notionnelle de Termisti », *Terminologies nouvelles*, n° 5, pp. 11-14

VAN CAMPENHOUDT, M. (1994) : *Un apport du monde maritime à la terminologie notionnelle multilingue : étude du dictionnaire du capitaine Heinrich Paasch « De la quille à la pomme du mât »* (1885-1901), thèse de doctorat en sciences du langage, Université de Paris XIII.

VERLINDE, S., BINON, J. & J. VAN DYCK (1992) : *Dictionnaire contextuel du français économique – Tome A : L'entreprise*, Louvain, Garant.

VERONIS, J. & N. IDE (1994) : « From Dictionaries to Knowledge Bases... and Back », paper read at the workshop on « *The Future of the Dictionary* » co-sponsored by Rank Xerox Research Centre and Acquilex-II, Grenoble (abstract published under the title « Machine-Readable Dictionaries : Have we wasted our time? », *Cambridge Language Reference News*, Cambridge University Press, Number 4, p. 1).

VIEGAS, E. (1995) : *Lexicon Development : Pro a Semi-automated Approach*, Technical Report (in preparation), Computing Research Laboratory, New Mexico State University

VIEGAS, E. et M. GONZALES (1995) : *Derivational Morphology Rules to Enhance Lexicons Acquisition*, Technical Report (in preparation), Computing Research Laboratory, New Mexico State University.

VIEGAS, E. et S. NIRENBURG (1995) : « The Semantic Recovery of Event Ellipsis: its Computational Treatment », *Proceedings of the Workshop on Context and Natural Language*, IJCAI 95, Montréal.

VIEGAS, E. et S. NIRENBURG (1996) : « The Ecology of Lexical Acquisition : Computational Lexicon Making Process », Soumis à Euralex'96.

WANG, William S. Y. (1977) : *The Lexicon in Phonological Change*, Paris, The Hague, Mouton.

WANNER, L. (Ed.) (1996) : *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam, Benjamins.

WARWICK, S. (1994) : « Automated Lexical Resources in Europe : A Survey », D Walker & A Zampolli (Eds.), *Automating the Lexicon*, Clarendon Press, forthcoming.

WARWICK, S., J. HAJIC et G. RUSSELL (1990) : « Searching on Tagged Corpora, Linguistically Motivated Concordance Analysis », *Electronic Text Research. Proceedings of the Conference*, Waterloo, UW Centre for the New OED and Text Research.

WASOW, T., SAG, I. et G. NUNBERG (1994) : « Idioms », *Language*, volume 70, pp. 491-538.

WASTON, Thomas (1977) : « Transformations and the Lexicon », *Formal Syntax*, P. W. Culicover, Thomas and Adrian Akinajian (Eds), New York, Academic Press.

WEBELHUTH, Gert (Ed.) (1995) : *Government and Binding Theory and the Minimalist Program*, Oxford (UK) & Cambridge (USA), Blackwell.

WEHRLI, E. (1985) : « Design and Implementation of a Lexical Data Base », *Proceedings of the 2nd European ACL Conference*, pp. 146-153

WILKS, Y. (1978) : « Making Preference more Active », *Artificial Intelligence*, 11, pp. 197-223.

WILKS, Y., HUANG, X. & D. FASS (1985) : « Syntax, Preference and Right Attachment », *Proceedings IJCAI*, Los Angeles (CA), pp. 779-784.

WITTGENSTEIN, L. von (1969) : *Philosophische Untersuchungen*, Frankfurt.

WOOLDRIDGE, T. R. (1977) : *Les Débuts de la lexicographie française · Estienne, Nicot et le* Thresor de la langue françoyse *(1606)*, Toronto/Buffalo, University of Toronto Press.

WOOLDRIDGE, T. R. (1988) : « Les vocabulaire et fréquence métalinguistiques du discours lexicographique des principaux dictionnaires généraux monolingues français des XVI^{e}-XX^{e} siècles », *Travaux de linguistique et de philologie*, 26, Paris, Klincksieck, pp. 305-313.

WOOLDRIDGE, T. R. (1993) : « Le flou en informatique textuelle », *Texte*, 13/14, Toronto, Éds Paratexte, pp. 275-289.

WOOLDRIDGE, T. R. (1994) : « Projet d'informatisation du Dictionnaire de l'Académie (1694-1935) », B. Quemada (dir.), *Actes du Colloque sur le* Dictionnaire de l'Académie française *et la lexicographie institutionnelle européenne*, Institut de France (1995). À paraître.

WOOLDRIDGE, T. R. (à paraître) : « Le mot métalinguistique du discours dictionnairique », *Cahiers de lexicologie*, Paris, Didier.

WU, S. et U. MANBER (1992) : « Fast Text Searching Allowing Errors », *Communications of the ACM*, 35 (10), pp. 83-91.

WURBEL, N. (1995) : *Dictionnaires et bases de connaissances : traitement automatique des données dictionnairiques de langue française,* thèse de Doctorat en informatique, Université d'Aix-Marseille III, 262 p.

WÜSTER, E. (1968) : *Dictionnaire multilingue de la machine-outil. Notions fondamentales, définies et illustrées, présentées dans l'ordre systématique et l'ordre alphabétique. Volume de base anglais-français = The Machine Tool An Interlingual Dictionary of Basic Concepts comprising an Alphabetical Dictionary and a Classified Vocabulary with Definitions and Illustrations. English-French Master Volume*, London, Technical Press.

WÜSTER, E. (1971) : « Les classifications de notions et de thèmes. Différences essentielles et applications » = « Begriffs- und Themaklassifikationen. Unterschiede in ihrem Wesen und ihrer Anwendung », *Nachrichten fur Dokumentation*, vol. 22 (3), pp. 98-104 et n° 4, pp. 143-150, traduit par INFOTERM, Bibliothèque d'INFOTERM.

WÜSTER, E. (1981) : « L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et la science des choses », Rondeau, G. et H. Felber (dir.), *Textes choisis de terminologie. Vol. I . Fondements théoriques de la terminologie*, Québec, Université Laval - GIRSTERM, pp. 55-113.

XEROX (1993) : *Lexical Tools : French Lexicon*, Palo Alto (CA), XEROX Corporation.

XEROX (1995) : *Part of Speech Disambiguator and Text Tokenizer Reference : French Version 1.0*, Palo Alto (CA), XEROX Corporation.

YAOLIANG, J. & D. ZHENDONG (1991) : « As a Participant in CICC MMT (ODA) Project », *Proceedings International Symposium on Multilingual Machine Translation,* (MMT-91), Beijing, 19-21 August 1991, vol. 1, pp. 13-15

YAROWSKY, David (1992) : « Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora », *Actes de COLLING-92*, Nantes, 23-28 août 1992.

YAROWSKY, D. (1995) : « Unsupervised Word Sense Disambiguation Rivaling Supervised Methods », *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge (MA), 26-30 juin, pp. 189-196

ZAMPOLLI, A., CALZOLARI, N. & M. PALMER (Eds) (1994) : *Current Issues in Computational Linguistics : in Honour of Don Walker*, Series « Linguistica Computazionale », IX-X, Pisa and Dordrecht, Giardini Editori and Kluwer Academic Publishers.

ZERNIK, U. (1991) : *Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon*, Lawrence Erlbaum Associates.

ZWICKY, A. M. (1992) : « Some Choices in the Theory of Morphology », R. Levine (Ed.), *Formal Grammar Theory and Implementation*, Oxford, Oxford University Press

ZYZOMYS (1989) : Paris, ACT Informatique.

Achevé d'imprimer sur
les presses de la SIEL (Beyrouth)
en décembre 1996

La collection **Universités francophones** créée en 1988 à l'initiative de l'UREF, propose des ouvrages de référence, des manuels spécialisés et des actes de colloques scientifiques aux étudiants des 2e et 3e cycles universitaires ainsi qu'aux chercheurs francophones et se compose de titres originaux paraissant régulièrement.

Leurs auteurs appartiennent conjointement aux pays du Sud et du Nord et rendent compte des résultats des recherches et des études récentes entreprises en français à travers le monde. Ils permettent à cette collection pluridisciplinaire de couvrir progressivement l'ensemble des enseignements universitaires en français.

Enfin, la vente des ouvrages à un prix préférentiel destinés aux pays du Sud tient compte des exigences économiques nationales et assure une diffusion adaptée aux pays francophones.

Ainsi la collection **Universités francophones** constitue une bibliothèque de référence comprenant des ouvrages universitaires répondant aux besoins des étudiants de langue française.

Prix : 140 FF • Prix préférentiel UREF (Afrique, Asie, Amérique du Sud, Moyen-Orient) : 60 FF

9 782920 021709 ISSN 0993 - 3948