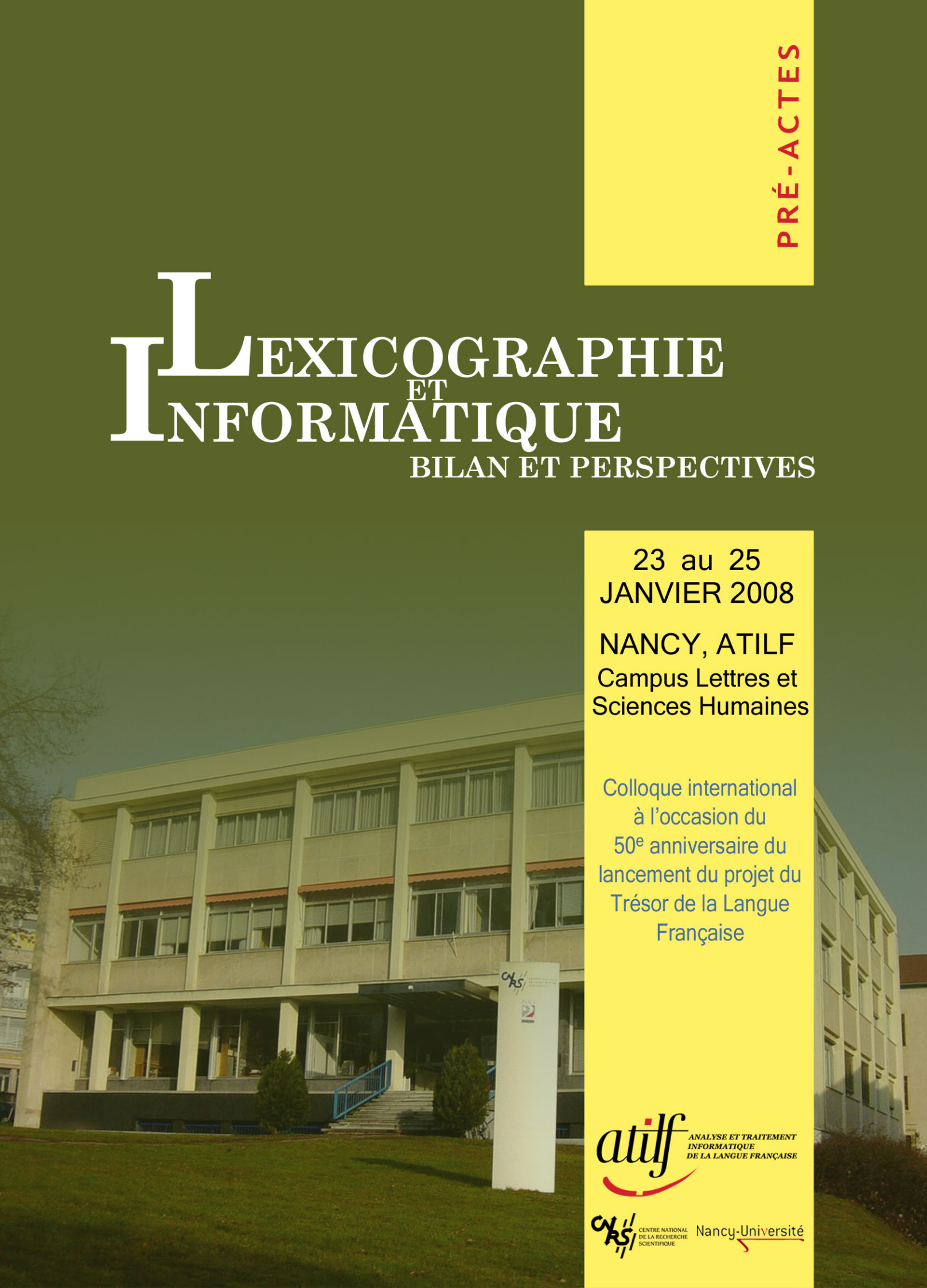
PRÉ-ACTES

LEXICOGRAPHIE ET INFORMATIQUE

BILAN ET PERSPECTIVES



23 au 25
JANVIER 2008

NANCY, ATILF
Campus Lettres et
Sciences Humaines

Colloque international
à l'occasion du
50e anniversaire du
lancement du projet du
Trésor de la Langue
Française

atilf
ANALYSE ET TRAITEMENT
INFORMATIQUE
DE LA LANGUE FRANÇAISE

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
Nancy-Université

Partenaires :

Sommaire

Informatisation et valorisation sur le Net : une deuxième vie pour le TLF

Jean-Marie Pierrel (1)
Jean-Marie.Pierrel@atilf.fr

(1) ATILF Nancy Université & CNRS

Mots-clés : dictionnaire électronique, informatisation, valorisation, ressources, mutualisation, accès structuré, interface, hyper-navigation.

Keywords: electronic dictionary, computerization, valorization, resources, mutualisation, structured access, interface, hyper navigation, XML encoding.

Résumé : Le *Trésor de la Langue Française* (TLF) est un grand dictionnaire de langue française en 16 volumes réalisé par l'Institut National de la Langue Française (INaLF, laboratoire du CNRS) entre le début des années 60 et le milieu des années 90. Ce dictionnaire était initialement conçu pour être édité uniquement sous forme papier. La décision de le transformer en dictionnaire électronique a été prise alors que le dictionnaire était pratiquement achevé. Après avoir rappelé l'importance d'une meilleure valorisation de nos productions de recherche, cet article présente les principales caractéristiques du *Trésor de la Langue Française informatisé* (TLFi), son insertion au sein du portail lexical du Centre National de Ressources Lexicales et Textuelles (CNRTL) et les impacts de ces versions informatisées du TLF sur sa diffusion internationale.

Abstract : The *Trésor de la Langue Française* (TLF) is a large 16 volumes of French language dictionary, released by the Institut National de la Langue Française (INaLF, former laboratory of the CNRS) between the beginning of the 60's and the middle of the 90's. This dictionary was intended to be published in a printed form. The decision to make an electronic version came as the dictionary was almost completed. Having to remind the importance of a better valorization of our research productions, this article presents the main characteristic of the *Trésor de la Langue Française informatisé* (TLFi), its insertion within the lexical common network of the National Center of Lexical and Textual Resources (CNRTL) and the impact of these TLF computerized versions on its international distribution.

Introduction

Dès que l'on s'intéresse à la langue, que cela soit pour un usage strictement humain ou pour une intégration dans une chaîne de traitement automatique, les informations lexicales liées aux mots de la langue occupent une importance primordiale (Laporte, 1997 ; Pierrel 2000). Pourtant, pour le français, il n'existe pas à ce jour de lexique ou de dictionnaire informatique optimal adapté tout à la fois à l'homme et à la machine. Pour les dictionnaires électroniques commerciaux et les lexiques spécifiques développés par telle ou telle équipe, une des questions essentielles porte sur la qualité et la couverture linguistique de tels outils : nombre d'entrées, richesse des informations disponibles, validité linguistique, facilité d'accès, etc. Dans ce contexte, le *Trésor de la Langue Française* (TLF) et sa version informatisée occupent une place de choix

Le *Trésor de la Langue Française* (TLF) est le dictionnaire de langue de référence réalisé entre le début des années 60 et le milieu des années 90 (CNRS, 1976-1994) par l'Institut National de la Langue Française (INaLF, laboratoire du CNRS) dont notre laboratoire ATILF est aujourd'hui le successeur nancéien. Dans son ouvrage sur les dictionnaires de la langue française, Jean Pruvost présente ainsi cet ouvrage : *« Ce projet, qui correspond à une entreprise publique ayant requis une centaine de chercheurs pendant 30 ans, avec un dépouillement de plus de 3 000 textes littéraires, scientifiques et techniques, a bénéficié des compétences nationales et internationales les plus éminentes […] Il en résulte, au-delà de la très grande qualité scientifique des articles, une description du fonctionnement de la langue qui ne manque pas d'être impressionnante : 23 000 pages, 100 000 mots, 450 000 entrées, 500 000 citations précisément identifiées. Le TLF relève pleinement d'une lexicographie philologique et historique, recourant aux citations-attestations qui permettent de fonder toutes les analyses morphologiques et sémantiques »* (Pruvost, 2002, page 78). Robert Martin, quant à lui, écrit « *De nombreux dictionnaires sont aujourd'hui disponibles sous un format électronique – des ouvrages encyclopédiques mais aussi des dictionnaires de langue. Un apport déterminant, en lexicographie française, est l'informatisation du Trésor de la Langue Française (TLF)* » (Martin, 2001 page 61).

L'objectif du présent article est de montrer comment, à travers ses versions CDROM (www.tlfi.fr) et web (www.atilf.fr/tlfi) d'une part, son intégration au sein du portail lexical du Centre National de Ressources Textuelles et Lexicales (www.cnrtl.fr) et ses produits dérivés tel le lexique morpho-syntaxique MORPHALOU (www.atilf.fr/morphalou) d'autre part, le *Trésor de la Langue Française informatisé* (TLFi) a permis de mieux valoriser le TLF et de lui offrir une seconde vie en terme d'usage effectif.

Après un premier paragraphe montrant l'importance d'une meilleure valorisation et mutualisation de nos résultats de recherche, nous rappelerons les principales caractéristiques du TLFi et de ses produits dérivés puis présenterons sa valorisation dans le cadre du portail lexical du CNRTL dont l'objectif est de regrouper le maximum d'informations sur le lexique français. Nous terminerons enfin par une rapide analyse des usages actuels du TLFi qui permit en quelques années de faire du TLF, dictionnaire de référence longtemps considéré comme un dictionnaire d'une élite pour une élite, un des dictionnaires français les plus utilisés sur le Net.

1. De la nécessité d'une meilleure valorisation des productions de recherche sur notre langue

Au-delà des besoins sociétaux de diffusion de connaissances sur notre langue sur lesquels nous reviendrons en fin de cet article, nous nous focalisons dans ce paragraphe sur les impératifs de recherche, primordiaux pour tous nos laboratoires, en particulier en sciences du langage.

1.1 Un enjeu pour la linguistique, la linguistique de corpus et le traitement automatique des langues

Une analyse de l'évolution de la linguistique au cours du dernier demi-siècle montre que sa confrontation avec l'informatique et les mathématiques lui a permis de se définir de nouvelles approches. C'est ainsi qu'au-delà d'une simple linguistique descriptive s'est développée une linguistique formelle, couvrant aussi bien les aspects lexicaux que syntaxiques ou sémantiques, qui tend à proposer des modèles s'appuyant sur une double validation, explicative d'un point de vue linguistique, opératoire d'un point de vue informatique. Par ailleurs, la disponibilité de ressources textuelles électroniques de grande taille (corpus, bases de données textuelles, dictionnaires et lexiques) et les progrès de l'informatique, tant en matière de stockage que de puissance de calcul, ont créé, au cours des années 1990, un véritable engouement pour les approches statistiques et probabilistes sur « corpus » (Habert, 1995). Ainsi se structura petit à petit un nouveau champ de recherche : la linguistique de corpus (Habert et col., 1997) permettant au linguiste d'aller au-delà de l'accumulation de faits de langue et de confronter ses théories à l'usage effectif de la langue.
Ces études et recherches en TAL et en linguistique de corpus nécessitent de plus en plus l'usage de vastes ressources linguistiques : textes et corpus, si possible annotés, dictionnaires, outils de gestion et d'analyse de ces ressources. Le coût de réalisation de telles ressources justifie pleinement des efforts de mutualisation pour permettre à la communauté de recherche de bénéficier, pour le français, de ressources comparables à celles existant pour d'autres grandes langues tel l'anglais.

1.2 Quelles ressources pour l'étude des langues aujourd'hui ?

1.2.1 Des corpus textuels

Le premier type de ressources, indispensable pour le développement de nombreuses études sur la langue, son analyse et son traitement, concerne les corpus textuels. Leur rôle est en effet central pour permettre la construction de modèles représentatifs de l'usage effectif de la langue. Il s'agit le plus souvent de faire émerger des invariants ou, au contraire, des comportements particuliers d'entités linguistiques. Si, pendant longtemps, ce type d'activité a pu se satisfaire des connaissances intrinsèques sur la langue qu'a le chercheur, les besoins de validation objective du monde scientifique nécessitent de plus en plus le maniement de vastes ensembles d'exemples attestés. La question fondamentale est alors de savoir comment recueillir des données fiables sur l'usage effectif de la langue. Le Web est aujourd'hui une source importante d'extraction de corpus, mais deux travers de taille caractérisent les textes qui y sont disponibles (Pierrel, 2005) :

- leur qualité est souvent très discutable.
- la pérennité de leur disponibilité n'est pas toujours assurée.

La question de la qualité et de la disponibilité de corpus de référence reste donc importante et, pour s'en convaincre, il suffit d'analyser certains projets nationaux ou internationaux. Ainsi, en France, le projet « technolangue »[1] lancé par le Ministère de la Recherche et des Nouvelles Technologies indiquait parmi ses quatre thèmes d'appel à proposition un volet sur les ressources linguistiques dont l'objectif était « *de stimuler la production, la validation et la diffusion de ressources linguistiques pour répondre aux besoins minimaux pour l'étude de la langue française, favoriser la réutilisabilité de ces ressources et diminuer le coût du « ticket d'entrée » dans le secteur* ». Les besoins sont en effet très diversifiés : que ce soit en terme de types de textes (littéraires, scientifiques ou techniques, mono et multilingues) ou en termes

[1] http://www.recherche.gouv.fr/appel/2002/technolangue.htm.

d'usages (professionnels ou grand public), la nécessité de vastes corpus normalisés, annotés et validés s'impose.

1.2.2 Des dictionnaires et des lexiques

Le second type de ressources concerne les dictionnaires et les lexiques. Bon nombre des arguments développés ci-dessus peuvent aussi s'appliquer à ce domaine. Or aucun traitement de la langue ne peut se passer du niveau lexical, et la disponibilité de ressources lexicales est absolument indispensable. Là encore les besoins sont très divers dans un contexte mono ou multilingue : dictionnaires spécialisés et dictionnaires généraux de langue, lexiques techniques ou bases terminologiques, par exemple.
Si, une fois de plus, la toile offre des réponses diversifiées à ce besoin, nombre de questions demeurent, concernant tout à la fois la qualité, la richesse, la couverture et la disponibilité de telles ressources. Nous sommes pour notre part convaincu qu'il importe de développer et partager des ressources de ce type et c'est cette conviction qui nous amena à proposer sur le Net une version informatisée du *Trésor de la Langue Française* et d'en dériver un lexique ouvert des formes fléchies du français (540 000 formes issues de 68 000 lemmes : http://www.cnrtl.fr/lexiques/morphalou/).

1.2.3 Des outils d'accès et de traitements

Un troisième type de ressources, complément des deux précédents, concerne les outils d'accès et de traitement de ces ressources. Deux types d'outils méritent une attention toute particulière :

- Les outils de gestion et d'exploitation des ressources textuelles, lexicales ou dictionnairiques. Que seraient en effet des ressources textuelles ou dictionnairiques du type de celles envisagées ci-dessus sans les logiciels d'exploration de ces ressources ?
- Les outils de base indispensables pour permettre à une équipe de recherche de proposer des avancées sur tel ou tel point : lemmatisation, conjugaison ou étiquetage morphosyntaxique.

Une fois de plus on ne peut que noter, tout en le regrettant, le manque de disponibilité d'outils fiables et généraux de ce type. Faute de cette disponibilité, la première tâche d'une équipe de recherche ou de développement travaillant sur des ressources linguistiques et plus généralement sur la langue consiste souvent, aujourd'hui, à redévelopper de tels outils !

1.3 Une nécessité : mutualiser les ressources et mieux prendre en compte leur production dans l'évaluation des chercheurs

En conclusion de ce paragraphe introductif, nous souhaitons faire partager notre conviction de la nécessité de mutualiser, au sein de la communauté francophone des sciences du langage, des ressources de références (corpus textuels, dictionnaires et lexiques, outils d'exploitation de ces ressources) pour la construction de modèles ou outils linguistiques, leur validation et leur comparaison.
Le coût de définition et de production de vastes ressources linguistiques de qualité (corpus, dictionnaires et lexiques) est important et c'est un gâchis énorme de vouloir, pour chaque projet, redéfinir l'ensemble des ressources dont on a besoin. A titre d'exemple, la construction d'un dictionnaire de langue tel le *Trésor de la Langue française* a nécessité près de cent personnes durant trente ans, et l'établissement d'une base de données textuelle tel FRANTEXT (www.atilf.fr/frantext) s'est chiffré aussi en dizaines d'hommes-an. Sans vouloir plaider ici pour une rentabilisation extrême de la recherche à travers une taylorisation de notre domaine, il convient néanmoins de prendre conscience que, sans une véritable mutualisation de telles ressources dans un domaine aussi vaste que les sciences du langage qui nécessite d'aborder

des aspects aussi divers que le lexique, la syntaxe, la sémantique, la pragmatique, chaque équipe de recherche ou chaque chercheur se verrait dans l'obligation de tout réinventer, alors même que nul ne peut être spécialiste de chacun de ces sous-domaines.
Un second point plaidant pour la mutualisation des ressources concerne l'évaluation, de plus en plus indispensable, de nos productions de recherche (analyseurs, systèmes de traitement) qui nécessite, pour des besoins de comparaison, la disponibilité de ressources de référence (corpus textuels, corpus d'exemples sur un phénomène de langue, ressources dictionnairiques) accessibles, partagées et clairement identifiables.
Enfin, il convient de noter qu'en termes de valorisation de la recherche et de partage de connaissances avec nos concitoyens, une disponibilité accrue, en particulier sur le Web, de nos productions de recherche est indispensable. Outre le fait que cela peut permettre un meilleur partage entre le monde de la recherche et la société civile, cela répond aussi à un besoin de plus en plus grand de connaissances chez nos concitoyens.
Mais ne nous leurrons pas, la constitution et la valorisation de telles ressources de qualité nécessitent des investissements en temps importants. Si l'on souhaite que des chercheurs puissent consacrer une partie de leur temps à de telles tâches au service de l'ensemble de la communauté scientifique, il convient de mieux prendre en compte cette activité de production de ressources numériques dans leur évaluation et de mettre en place une structure servant à la fois de validation et de diffusion de ces productions. C'est en partie du moins le rôle que le CNRS a confié aux Centres Nationaux de Ressources, dont le CNRTL.

2. Le TLF, une ressource inestimable valorisée à travers le TLFi

2.1 Caractéristiques du TLFi

Reflet fidèle de la version papier, jusque dans sa présentation typographique à l'écran, le TLFi (www.atilf.fr/tlfi) se caractérise, comme le TLF, par la richesse de son matériau et la complexité de sa structure :

- Importance de sa nomenclature : 100 000 mots avec leur étymologie et leur histoire, et 270 000 définitions.
- Richesse des objets méta-textuels inclus dans chaque article (vedettes, codes grammaticaux, indicateurs sémantiques ou stylistiques, indicateurs de domaines, définitions, exemples référencés…).
- Richesse des 430 000 exemples, tirés de plus de deux siècles de production littéraire française.
- Diversité des rubriques : une rubrique synchronie couvrant la période 1789 à nos jours, une rubrique étymologie et histoire, et une rubrique bibliographie pour les principaux articles.

La version informatique du TLF (Dendien et Pierrel, 2003) intègre, de plus, des accès à très haut niveau de tolérance permettant une insensibilité aux accents, une tolérance aux fautes d'orthographe courantes, un traitement phonétique et un traitement morphologique. Ainsi, on peut offrir une correction automatique des fautes, permettre des accès à partir de formes et non plus uniquement de lemmes ou de vedettes et proposer des procédures d'accès diversifiées pour une consultation humaine.

Nous ne reviendrons pas ici sur les étapes d'informatisation du TLF traitées par ailleurs (Dendien et Pierrel 2003), mais nous nous contenterons ici de rappeler, essentiellement par l'exemple, les accès offerts par la version informatisée du TLF.

2.2 Quels accès au TLFi ?

Le TLFi correspond à une rétro-conversion de la version papier du TLF pour laquelle, par des procédures de repérage semi-automatique des objets textuels composant les articles du dictionnaire original, nous avons introduit un balisage fin, tant typographique (de manière à conserver une image 100 % fidèle du TLF) que sémantique (repérage des principaux objets textuels au sein de chaque article). Quelques chiffres peuvent donner un aperçu de la finesse de ce balisage : après validation sur l'ensemble des seize tomes, 36 613 712 balises XML ont été positionnées : 17 364 854 balises typographiques, 1 070 224 balises décrivant la hiérarchie, 18 178 634 balises repérant les objets textuels, dont 92 997 entrées et 64 346 locutions faisant l'objet de 271 166 définitions et illustrées par 427 493 exemples.

C'est ce balisage fin du TLF et l'exploitation du document XML correspondant qui nous permet de proposer des accès à l'ensemble du dictionnaire, cumulant les avantages d'un dictionnaire avec ceux d'une ressource textuelle et d'une véritable base de données lexicales :

- Recherche d'un mot, d'une expression ou d'une forme lexicale plus ou moins bien orthographiés, avec possibilité, via un « panneau de réglage », de mettre en évidence divers champs dans le résultat de la recherche (définition, code grammatical, domaine spécifique, exemple, auteur d'exemple, construction, indicateur, etc.).
- Possibilité d'hyper-navigation à l'intérieur du dictionnaire permettant en un clic-souris de passer d'un mot à sa définition.
- Interrogations assistées ou requêtes complexes exploitant l'ensemble de la structure du dictionnaire à travers le croisement de multiples critères.

2.3 Exemples de recherches dans le TLFi

On peut trouver à l'adresse www.tlfi.fr une présentation et des démonstrations sur les modes de recherche offerts dans le TLFi, mais la meilleure façon de se rendre compte de l'intérêt d'une telle transformation du TLF en document numérique consiste soit à accéder au Cédérom du TLFi (ATILF, 2004), soit à se connecter directement à l'adresse : www.atilf.fr/tlfi. Trois principaux types d'accès sont alors proposés : la recherche d'un mot, la recherche assistée et la recherche complexe.

2.3.1 Recherche d'un mot ou d'une expression

Cette recherche permet un accès à un mot à travers un système de correction automatique (forcée ou non) : ainsi, en introduisant la recherche de la forme *etique* (sans accent), on accède aux deux articles correspondant aux mots *étique* ou *éthique ;* de même un accès à partir de la forme *sussiez* permet d'obtenir automatiquement l'article *savoir*. Elle donne aussi la possibilité d'obtention directe des définitions et conditions d'usage d'expression tel « le trompette » en focalisant la réponse sur l'élément pertinent demandé et en offrant la possibilité, à l'aide d'une sorte de « stabilo boss » électronique, de surligner tel ou tel objet textuel. Ici, par exemple, la définition :

Objets de la recherche : 1Paragraphe1
H TROMPETTE, subst.

II. — *Subst. masc.* Personne qui joue de la trompette.
1A. — Soldat chargé d'exécuter les sonneries. ***Le trompette*** *de l'escadron, d'un régiment de cavalerie. Tu seras capitaine, avec une nuée de trompettes courant et sonnant devant toi* (HUGO, *Légende*, t. 3, 1877, p. 390).1

—*Loc. fam., vieilli. Il est bon cheval de trompette*. Il ne se laisse ni effrayer, ni intimider. *Son air, un air de bon cheval de trompette qui ne craignait pas le bruit* (A. DAUDET, *Tartarin de T.*, 1872, p. 13).
B. —Musicien jouant dans une fanfare, un orchestre. Synon. *trompettiste* (*infra* dér.). *Le trompette noir du dancing* (BEAUVOIR, *Mandarins*, 1954, p. 306).
noir du dancing (BEAUVOIR, *Mandarins*, 1954, p. 306).

2.3.2 Recherche assistée

Ce second type d'accès permet par exemple de rechercher des expressions composées d'une forme : ainsi, en demandant les mots contenant la forme *queue* on obtient 35 réponses dont :

COURTE-QUEUE, adj. et subst.
DEMI-QUEUE, subst. fém.
HOCHEQUEUE, HOCHE-QUEUE, subst. masc.
PAILLE-EN-CUL, PAILLE-EN-QUEUE, subst. masc.
PORTE-QUEUE, subst. masc.
QUEUE(-)D'ARONDE, voir ARONDE.
Etc.

ou de rechercher *« les verbes qui, en marine, concernent le maniement des voiles »*. Il suffit de préciser que l'on recherche dans la classe des verbes ceux qui, dans le domaine de la marine, correspondent à une définition incluant une forme du mot *voile*, soit dans une structure plus compacte : [code grammatical : *verbe* ; domaine : *marine* ; type d'objet : *définition*, contenu : &m*voile*[2]]. Voici un extrait des 61 réponses que l'on obtient :

ABRIER, ABREYER, verbe trans.
3Empêcher le vent, en l'interceptant, de passer jusqu'à (une autre voile) :3
AGRÉER[2], verbe trans.
3,,Préparer ou travailler à la garniture, aux agrès d'un bâtiment, fourrer les dormans, estroper les poulies, garnir voiles, vergues, etc. : `` (WILL. 1831) :3
AMURER, verbe.
3Fixer l'amure d'une voile pour l'orienter selon le vent :3
ETC……

Autre exemple : pour l'ensemble des mots dont la définition utilise le mot *liberté* [type d'objet : définition, contenu : &mliberté], on obtient 306 réponses dont :

Objets de la recherche : 1Définition1

ABUSER, verbe trans.
1Exagérer dans l'usage d'une possibilité, d'une liberté :1
AFFRANCHI, IE, part. passé, adj. et subst.
1(Celui) à qui on a donné la liberté.1

[2] &m*subs* permet de tester toutes les formes d'un *substantif*, de même que &c*verbe* toutes les formes d'un *verbe*.

AISE[1], subst. fém.
1Grande liberté.1
ALIÉNANT, ANTE, part. prés. et adj.
1Qui prive l'homme de son humanité, de sa liberté :1
Etc…..

2.3.3 Recherche complexe

Les interrogations possibles au sein de ce dictionnaire peuvent prendre des formes encore plus complexes. Ainsi, il est possible de répondre à la requête suivante : « *Quels sont les substantifs empruntés à une langue étrangère (non précisée) et qui, lorsqu'ils sont employés dans le domaine de l'art culinaire, sont illustrés par une définition empruntée au dictionnaire de l'Académie ?* ». Il convient pour cela d'utiliser l'onglet « recherche complexe » et de préciser :

Objet 1 : type "*Entrée*" ; Objet 2 : type "*Code grammatical*", contenu "*substantif*", lien "*inclus dans l'objet 1*" ; Objet 3 : type "*Domaine technique*", contenu "*art culinaire*", lien "*dépendant de l'objet 1*" ; Objet 4 : type "*Définition*", lien "*dépendant de l'objet 3*" ; Objet 5 : type "*Source*", contenu "*Académie*", lien "*inclus dans l'objet 4*" ; Objet 6 : type "*Langue empruntée*", lien "*dépendant de l'objet 1*".
Le lien "inclus dans l'objet 1" de l'objet 2 exprime que l'entrée est un substantif, le lien "dépendant de l'objet 1" de l'objet 3 exprime que l'indication de domaine technique est dans la portée de l'objet 1, le lien "dépendant de l'objet 3" de l'objet 4 exprime que la définition est valable dans le domaine de l'art culinaire, le lien "inclus dans l'objet 4" de l'objet 5 exprime que la source de la définition est le dictionnaire de l'Académie, et le lien "dépendant de l'objet 1" de l'objet 6 exprime que l'objet est dans l'article dont l'entrée est l'objet 1.
Une telle interrogation nous fournit quatre résultats dont :

Objets de la recherche : 1Entrée1 3Domaine technique3 4Définition4 5Source5 6Langue empruntée6

MORTIFICATION, subst. fém.
1MORTIFICATION, subst. fém.1 3ART CULIN.3 1 4„Action de garder certaines viandes pour qu'elles deviennent tendres et gagnent du fumet`` (Ac. 1878, 1935).4 5Ac. 1878, 1935 5 6Empr. au lat.6
NAPOLITAIN, -AINE, adj. et subst.
1NAPOLITAIN, -AINE, adj. et subst.1 3ART CULIN.3 2 4Gros gâteau cylindrique ou hexagonal fait d'une pâte à base d'amandes et fourré de confiture d'abricots et de gelée de groseilles (d'apr. Ac. Gastr. 1962).4 5d'apr. Ac. Gastr. 1962 5 6Empr. à l'ital.6
ETC.

3. Le portail lexical du CNRTL : un outil de mutualisation de résultats en lexicographie

3.1 Objectifs du CNRTL

Créé par le CNRS en 2005, le Centre National de Ressources Textuelles et Lexicales (CNRTL : www.cnrtl.fr) est adossé au laboratoire Analyse et Traitement Informatique de la Langue Française (ATILF / CNRS - Nancy Université). Son objectif est de réunir au sein d'un

portail unique le maximum de ressources informatisées et d'outils de consultation pour l'étude, la connaissance et la diffusion de la langue française.
Grâce à une mutualisation de connaissances issues des travaux de différents laboratoires, le CNRTL se propose d'optimiser la production, la validation, l'harmonisation, la diffusion et le partage de ressources, qu'il s'agisse de données textuelles et lexicales informatisées ou d'outils permettant un accès intelligent à leur contenu.
La décision de création du CNRTL s'inscrit dans la politique du CNRS visant à la création de nouvelles infrastructures indispensables aux travaux de recherche menés par l'ensemble de la communauté scientifique et résulte d'une action commune à la Direction de l'Information Scientifique et au Département Homme et Société du CNRS. Il est aujourd'hui l'une des composantes du très grand équipement d'accès unique aux documents numériques en sciences humaines et sociales ADONIS.
L'expertise scientifique reconnue ainsi que les nombreux projets coopératifs nationaux et internationaux des laboratoires auxquels il est adossé ont permis en outre au CNRTL de se positionner au niveau européen à travers :

- Des collaborations directes avec des centres partenaires, en Grande-Bretagne (Université d'Oxford), en Allemagne (Centres de compétence de Trèves et de Würzburg, DFKI à Sarrebruck, MPI) et aux Pays-Bas (Université de Nimègue).
- La participation au réseau européen CLARIN (http://www.mpi.nl/clarin/) des centres de gestion de ressources et de technologies linguistiques.

3.2 Les ressources gérées au sein du CNRTL

Le CNRTL se structure autour de cinq pôles de compétence : un portail lexical sur le français ; des corpus et données textuelles, annotés ou non ; des dictionnaires encyclopédiques et linguistiques (anciens et modernes) ; des lexiques phonétiques, morphologiques, syntaxiques, sémantiques ; des outils linguistiques (étiqueteurs, analyseurs, aligneurs, concordanciers, outils d'annotation). Parmi les ressources déjà intégrées au CNRTL, outre le portail lexical sur lequel nous allons revenir dans le paragraphe suivant, il convient de noter, entre autres :

- Les corpus de textes libres de droit d'auteur et d'éditeur (dans un premier temps 500 textes issus de Frantext) : à travers une sélection par auteurs, titres, dates ou genres, nous offrons la possibilité de télécharger les textes sélectionnés au format XML dans une DTD respectant les recommandations de la TEI (www.tei-c.org)[3], l'utilisateur récupèrant une archive contenant la DTD et le codage XML/TEI des textes. A notre connaissance, le CNRTL est le premier site offrant un ensemble de corpus français normalisés XML/TEI d'environ 150 millions de caractères.

- Le lexique Morphalou, dérivé de la nomenclature du TLF en accès libre tant en consultation qu'en téléchargement : lexique ouvert des formes fléchies du français qui fournit 524 725 formes fléchies, appartenant à 95 810 lemmes, linguistiquement valides (responsabilité d'un comité éditorial) et respectant les propositions de normalisation pour les ressources lexicales de l'ISO (TC37/SC4).

- Des versions informatisées de dictionnaires tant modernes (TLFi ; Dictionnaire de l'Académie française : 8ème et 9ème éditions) qu'anciens (Dictionnaires de R. Estienne (1552), de Jean Nicot (1606), de Bayle (1740), de Ferraud (1787-1788), de

[3] Notons que Nancy, à travers une association entre l'ATILF, l'INIST et le LORIA, est aujourd'hui centre support européen de la TEI.

l'Académie (1ère édition, 1694 ; 4ème édition, 1762 ; 5ème édition, 1798 ; 6ème édition, 1835)), ainsi que de l'Encyclopédie de Diderot et d'Alembert[4].

Le CNRTL se propose également de mettre à disposition de la communauté des outils linguistiques utilisables directement sur le site Web à partir d'un simple navigateur Internet. Parmi les différents projets en cours ou à venir, nous comptons offrir aux utilisateurs un accès simple et convivial à des outils comme :

- FLEMM : outil d'analyse flexionnelle de textes en français au préalable étiquetés, au moyen de l'un des deux catégorisateurs : Brill ou TreeTagger.
- POMPAMO : outil de détection de candidats à la néologie formelle et catégorielle basé sur l'utilisation de lexiques d'exclusion. Ce projet exploite des ressources lexicales comme Morphalou et permet d'en constituer de nouvelles.

3.3 Un exemple d'intégration de données lexicographiques : le portail lexical

Le portail lexical a pour vocation de valoriser et de partager, en priorité avec la communauté scientifique, un ensemble de données issues des travaux de recherche sur le lexique français. Projet évolutif, cette base de connaissances lexicales propose, à partir d'une forme lexicale, d'intégrer un maximum de connaissances disponibles.

Informations lexicographiques

Au premier rang de ces connaissances se placent les informations lexicographiques. A ce jour nous avons intégré dans ce portail les informations issues du TLF (www.atilf.fr/tlfi) qui apparaissent par défaut lorsqu'on demande des informations lexicographiques. Elles sont complétées par des informations facilement accessibles via un menu et issues :

- du dictionnaire de l'Académie Française (4ème, 8ème et 9ème éditions) (www.atilf.fr/academie) dont l'informatisation a été réalisée au sein du laboratoire dans le cadre d'un partenariat avec l'Académie.

- de la Base de Données Lexicographiques Panfrancophone (BDLP : http://www.tlfq.ulaval.ca/bdlp/), projet d'envergure internationale qui s'inscrit dans l'entreprise du Trésor des vocabulaires français lancée par le professeur Bernard Quemada dans les années 1980. L'objectif de la BDLP est de constituer et de regrouper des bases représentatives du français de chacun des pays et de chacune des régions de la francophonie. Les bases de données sont conçues de façon à pouvoir être interrogées de façon séparée ou comme un seul corpus et à servir de complément au *Trésor de la Langue Française informatisé*. Dans sa dimension internationale, le projet de la BDLP est patronné par l'Agence Universitaire de la Francophonie qui l'appuie à travers son réseau d'étude du français en francophonie (http://www.eff.auf.org/).

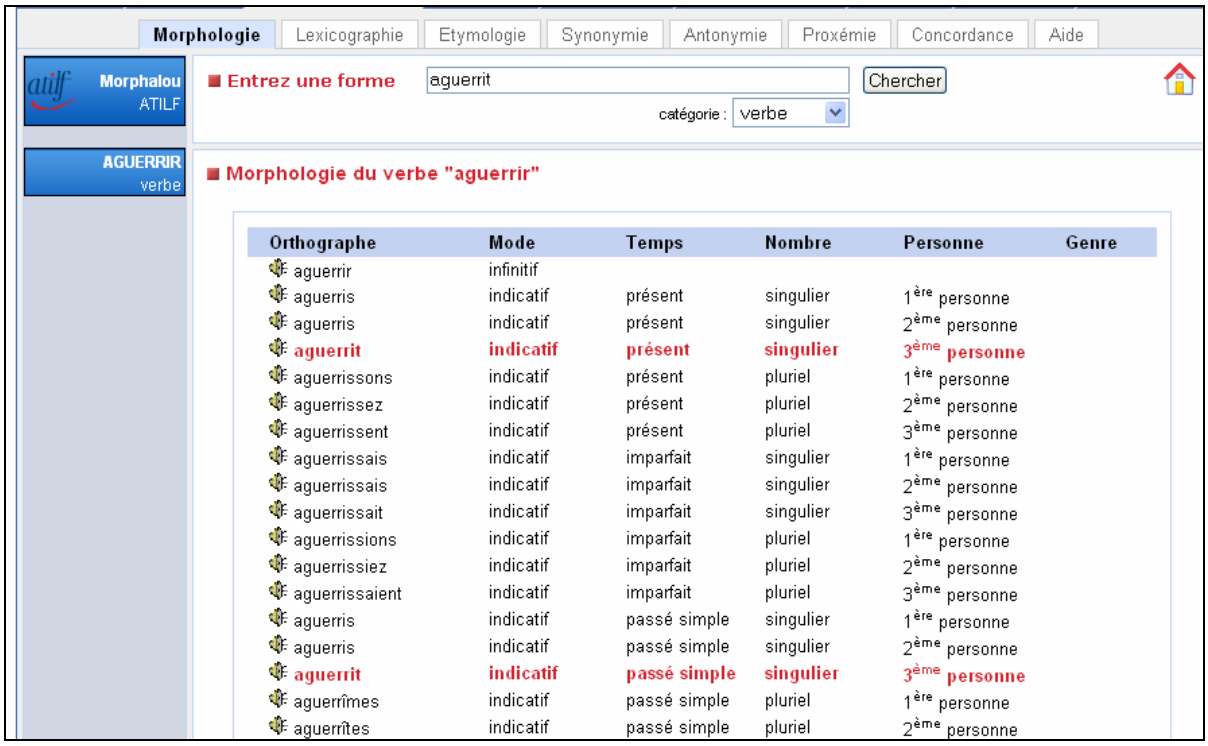
- De la Base Historique du Vocabulaire Français (Datations et Documents Lexicographiques : www.atilf.fr/ddl) constituée de datations du vocabulaire français, s'appuyant sur des données des 48 volumes de la collection *Datations et Documents Lexicographiques*.

L'ensemble de ces informations lexicographiques sont également accessibles directement, pour une forme donnée, par *http://www.cnrtl.fr/lexicographie/suivi de la forme que l'on souhaite interroge.* Ainsi :
http://www.cnrtl.fr/lexicographie/aguerrit permet d'accéder aux informations lexicographiques du verbe *aguerrir*.

[4] La plupart des versions informatisées de dictionnaires anciens, tout comme celle de l'Encyclopédie de Diderot et d'Alembert, sont le fruit d'un partenariat avec l'ARTLF (http://humanities.uchicago.edu/orgs/ARTFL/).

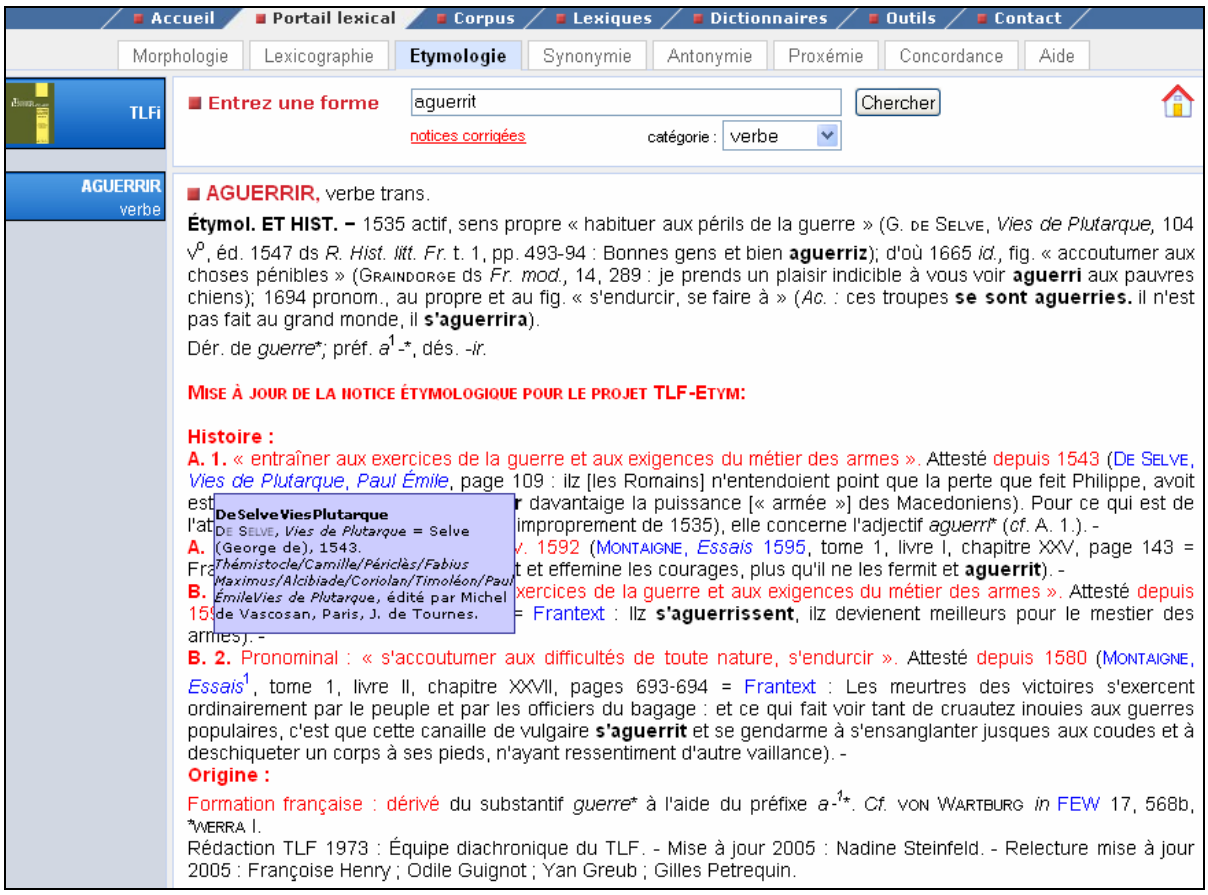
Informations morphosyntaxiques

Ces informations morphologiques sont issues de la base Morphalou (www.atilf.fr/morphalou), construite au départ à partir de la nomenclature du TLFi. Elles sont aussi accessibles directement pour la forme, telle *aguerrit,* par : http://www.cnrtl.fr/morphologie/*aguerrit*



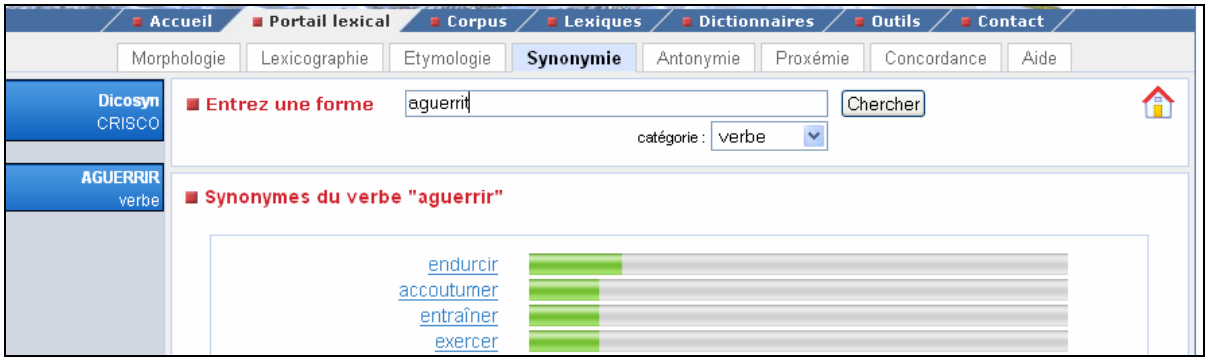
Informations étymologiques

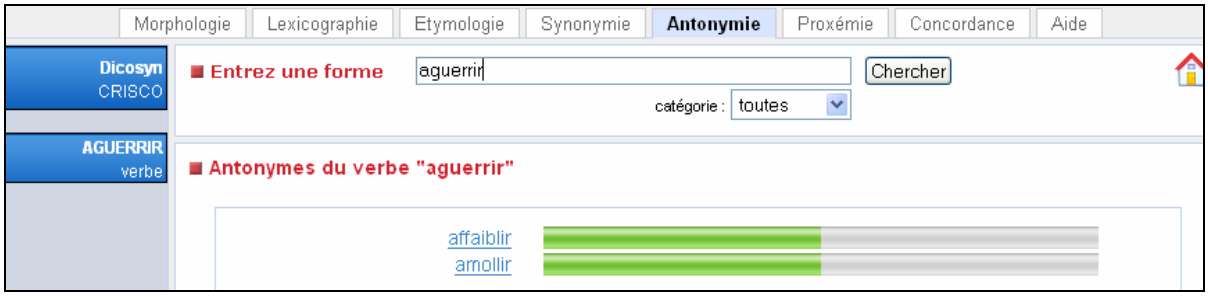
Ces informations étymologiques sont issues du TLF (www.atilf.fr/tlfi) et du projet TLF-Etym de mise à jour des rubriques étymologiques du TLF (www.atilf.fr/tlf-etym). Elles sont accessibles directement, pour une forme, par : http://www.cnrtl.fr/etymologie/*aguerrit*



Synonymies et antonymies

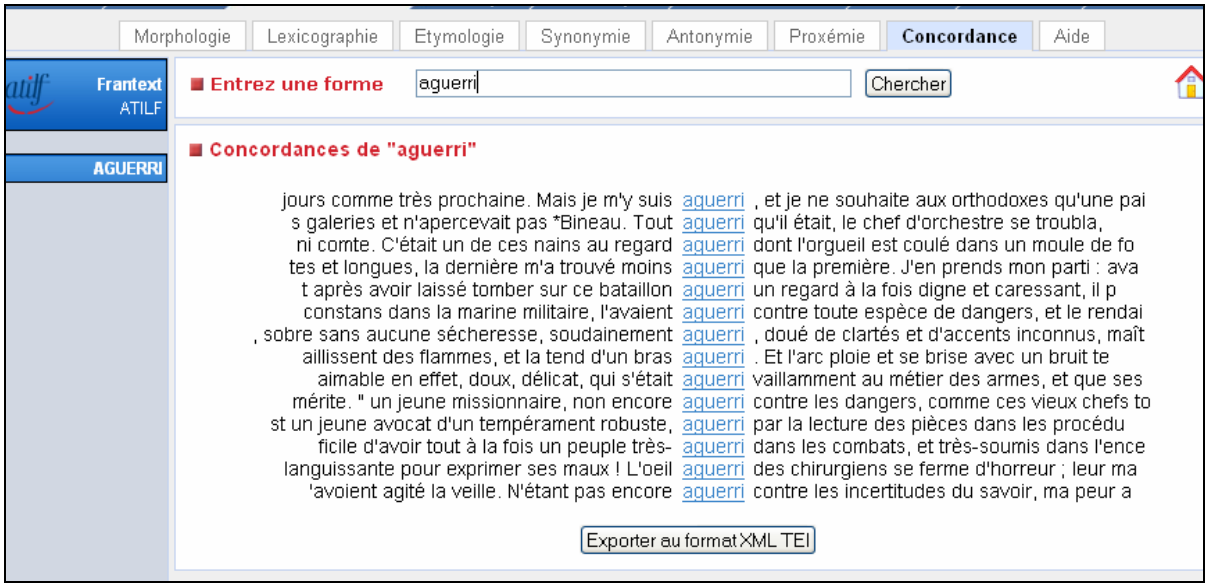
Ces informations de synonymie et d'antonymie proviennent du dictionnaire de synonymes de Caen (http://www.crisco.unicaen.fr/), construit à partir de données issues de l'INaLF. Ces informations sont aussi directement accessibles par : http://www.cnrtl.fr/synonymie/*aguerrit* ou http://www.cnrtl.fr/antonymie/*aguerrit*





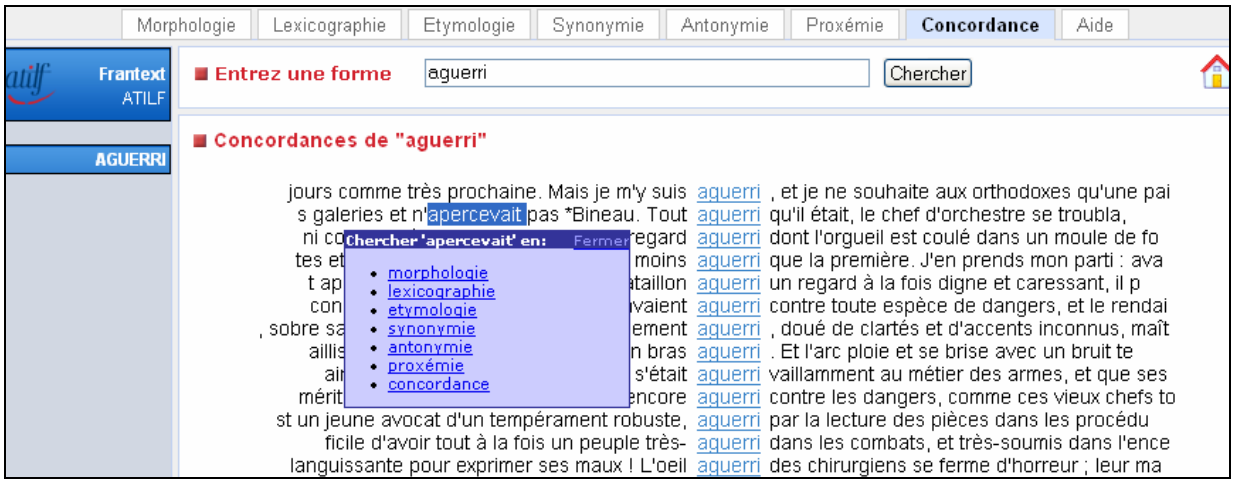
Concordance

Cette concordance utilise le corpus des textes de la base Frantext (www.atilf.fr/frantext) qui offre aussi la possibilité d'exporter les résultats du concordancier au format XML/TEI. C'est à notre connaissance le seul site permettant à un utilisateur d'exporter dans un format normalisé un concordancier français d'une telle importance. Ces concordances sont aussi directement accessibles par : http://www.cnrtl.fr/concordance/*aguerri*



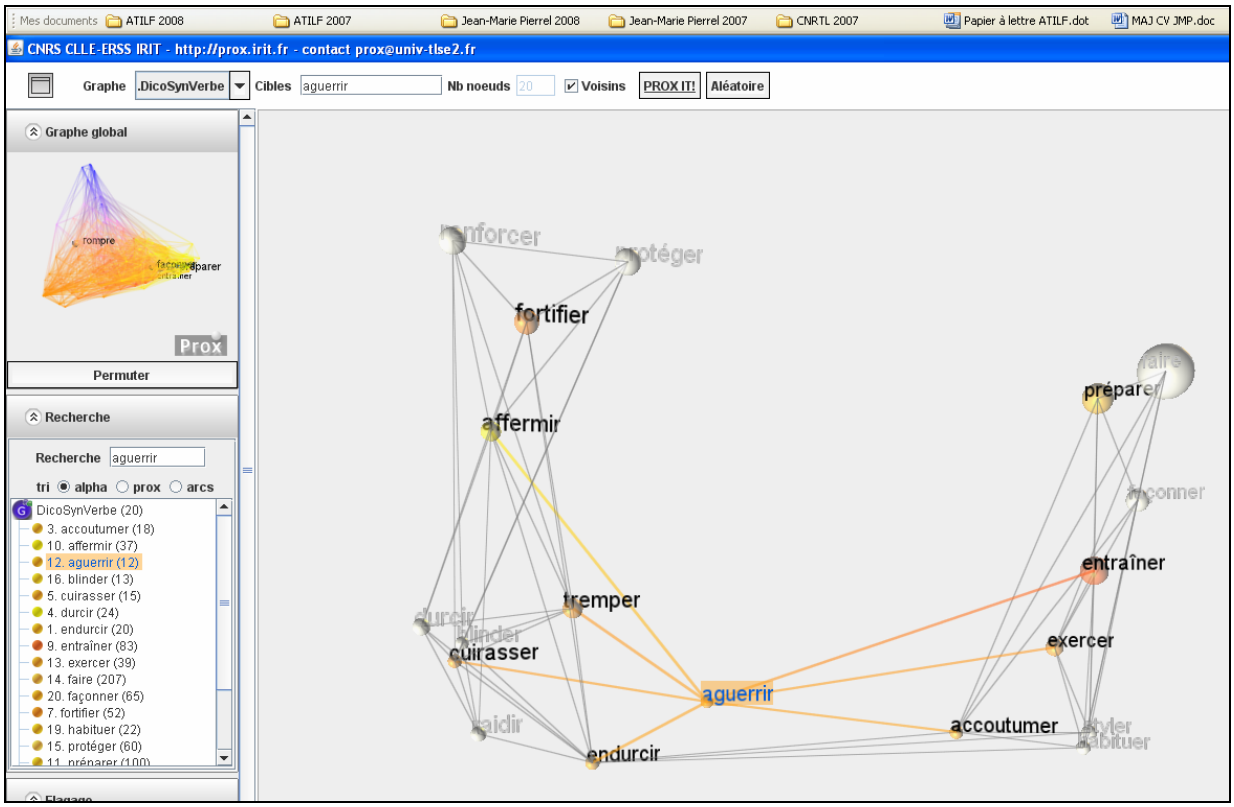
De plus, un simple clic droit sur un des exemples permet d'obtenir la référence complète de l'exemple sélectionné. Ainsi pour le premier exemple :

Le portail lexical permet également, à partir d'un simple double clic sur un mot, une hyper-navigation vers toutes les informations lexicales disponibles pour ce mot. Par exemple, si l'on veut obtenir des informations sur la forme « apercevait » du deuxième exemple de concordance, un double clic sur le mot affiche un menu qui permet d'hyper-naviguer vers l'ensemble des informations disponibles sur cette forme :



Proxémie

Une représentation en trois dimensions de la proxémie des mots de langues réalisée en coopération entre l'IRIT et l'ERSS (http://Prox.irit.fr) (Gaume 2006) est accessible elle aussi directement par :
http://www.cnrtl.fr/proxemie/aguerrir.



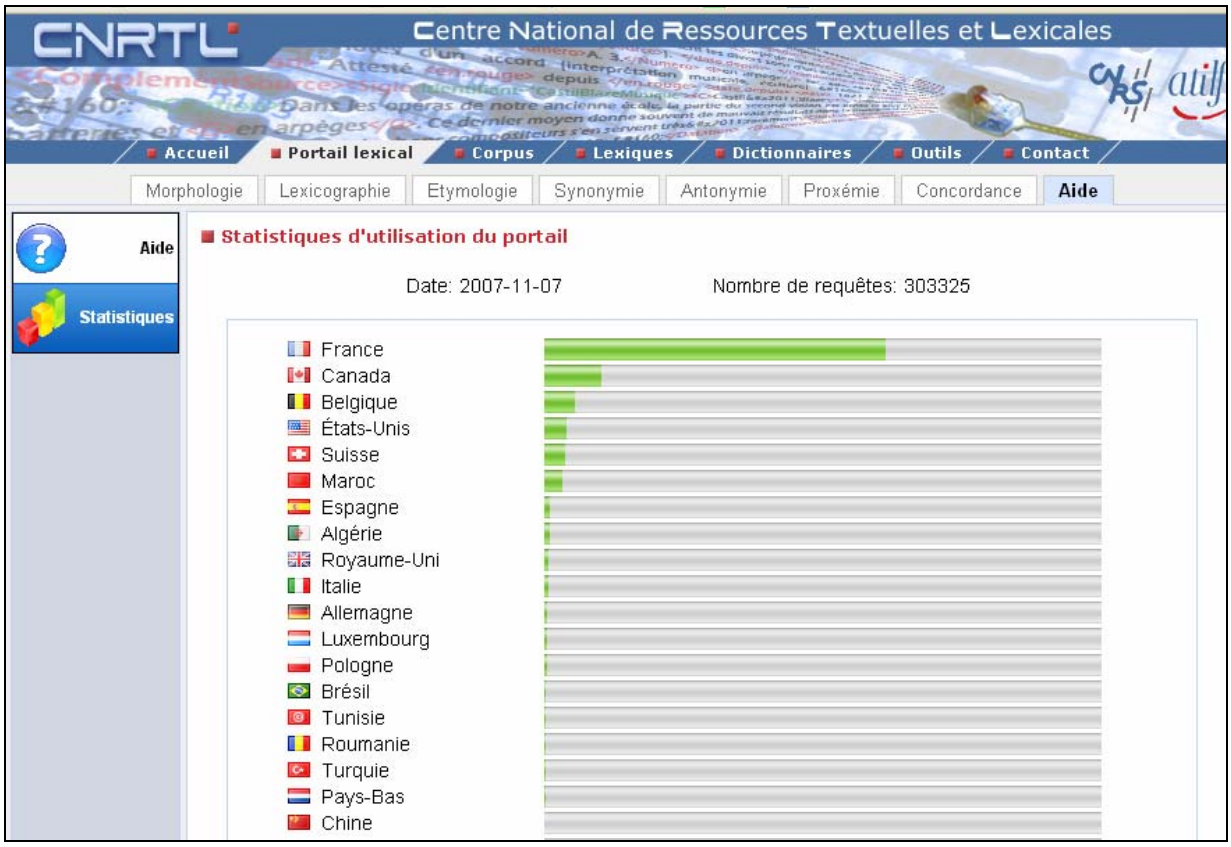
4. Impact de la valorisation de ressources lexicales sur le web : une seconde vie pour le TLF

Sans aucun doute le plus grand dictionnaire informatisé consacré à la langue française, le TLFi, grâce à la richesse de son contenu entièrement encodé en XML, a ouvert des perspectives intéressantes. Le TLF a eu pendant longtemps la réputation tenace d'être un dictionnaire réservé à une élite. Cette perception du TLF pouvait s'expliquer par au moins trois caractéristiques de sa version papier :

- Sa taille : 16 volumes de plus de 1 000 pages chacun.
- Sa richesse de description qui parfois nuisait à sa lecture, au moins pour les articles les plus lourds : l'article « aimer » se développe ainsi sur 12 pages soit 24 colonnes, et il n'est pas toujours aisé pour un non-spécialiste d'appréhender cette information très riche.
- Son coût, environ 1 500 euros, qui ne le rendait pas facilement accessible à tous.

S'il a su se positionner comme une référence en lexicographie française, la diffusion de sa version papier s'est néanmoins limitée à quelques millers d'exemplaires au sein d'une intelligentsia somme toute limitée.
Sa version informatique sous forme de Cédérom (environ 15 000 exemplaires vendus en moins de 4 ans) ou de ressources librement accessibles sur le Web a rencontré un succès important tant auprès du grand public que des utilisateurs universitaires ou des professionnels de la langue. Sa version web fait l'objet d'environ 300 000 connexions quotidiennes en provenance de tous les continents, et il est référencé par d'innombrables sources. La notoriété qu'il a acquise en fait un outil de promotion appréciable de la langue française.
Son intégration plus récente encore au sein du portail lexical du CNRTL et ses interconnexions avec d'autres types de ressources sur le vocabulaire français le positionnent au cœur d'un ensemble de ressources sur la langue française au sein desquelles il joue un rôle

actif et prépondérant, démontrant ainsi que sa réputation élitiste est devenue largement injustifiée et sa diffusion au sein du portail lexical du CNRTL, qui fait l'objet d'environ 300 000 requêtes par jour provenant d'horizons très divers, en fait aujourd'hui l'un des dictionnaires les plus exploités sur le Web



Conclusion

Notons pour conclure que le partage d'une telle version informatisée d'une production scientifique de référence offre aujourd'hui des modes nouveaux de valorisation de ressources ou de résultats de recherche. Au-delà du seul monde universitaire, ces techniques permettent de mettre à disposition de l'ensemble de la société nos résultats de recherche. On peut, pour s'en convaincre, analyser les commentaires apparaissant sur le web dans des sites institutionnels (http://www.terminometro.info/article.php?ln=fr&lng=fr&id=4546 par exemple) ou professionnels (http://www.entreprisesaletranger.org/archive-01-05-2007.html). La généralisation de telles exploitations et valorisations de versions électroniques est ainsi en train de modifier notablement les modes de travail et d'échanges scientifiques au sein des communautés de recherche SHS.

Bibliographie

ATILF *Trésor de la Langue Française informatisé*, CNRS Editions, Livre d'accompagnement 591 p. et CD du texte intégral, Version PC, ISBN 2-271-06273-X, 2004, Version Mac OX X, ISBN 2-271- 06365-5, 2005.

CNRS, *TLF, Dictionnaire de la langue du 19e et 20e siècle*, CNRS, Gallimard, Paris, 1976-1994.

Dendien J., Pierrel J.M. Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence, *TAL* Vol 44 – n° 2/2003, Hermès Sciences Edition, p. 11-37.

Gaume B., Cartographier la forme du sens dans les petits mondes lexicaux, in *JADT* 2006, p 541-465.

Habert B., Traitements probabilistes et Corpus, *TAL* Vol 36- N°1-2, Paris, Hermès, 1995.

Habert B., Nazarenko A. et Salem A. *Les linguistiques de corpus*, Paris, Armand Colin, 1997.

Laporte E., Mots et niveau lexical, TAL, vol. 38, n°2, 1997.

Martin R., *Sémantique et automate*, PUF, Paris 2001.

Pierrel J.M., *Ingénierie des langues*, Hermès Editions, 2000, 354 p.

Pierrel J.M., Un ensemble de ressources de référence pour l'étude du français : TLFi, Frantext et le logiciel Stella , *Revue Québécoise de Linguistique*, volume 32/1, TAL Web et Corpus, p. 155-176, 2005.

Pruvost J., *Les dictionnaires de la langue française*, collection Que Sais-je ?, PUF, Paris 2002.

TLF et TLFi. Naissance et évolution d'un dictionnaire

Pascale Bernard (1)
pascale.bernard@atilf.fr
Christiane Jadelot (1)
christiane.jadelot@atilf.fr

(1) ATILF Nancy Université & CNRS (Axe Lexiques)

Mots-clés : lexicographie, lexicologie, lexicographie informatisée, méthodologie, historique

Keywords: lexicography, lexicology, computerized lexicography, methodology, historical background

Résumé : Les *Mélanges de Linguistique Française et de Philologie et Littérature Médiévales* remis à M. Imbs, le 11 mai 1973, nous donnent l'occasion d'un retour non pas nostalgique mais historique sur quelques principes fondateurs (contenu sémantique du mot, choix des exemples) ayant présidé à la rédaction du dictionnaire au cours des années dites de mise en route et sur son évolution constante.

Nous montrerons combien le contenu et la présentation des articles du *TLF* ont évolué entre le premier et le dernier tome, même si les normes de rédaction concernant la définition et le choix des exemples ont été finement définies au départ. La décision de l'informatisation du *TLF* fut prise en 1990, nous en décrirons les difficultés ainsi que les avantages apportés au lecteur par rapport à la version papier. Notre bilan s'achèvera sur une note optimiste:bien qu'imparfait le *TLF* demeure un vaste patrimoine sur lequel le laboratoire peut s'appuyer pour développer des recherches dans le domaine lexical ou la didactique des langues.

Abstract : The *Miscellany of French Medieval Linguistics and Philology and Literature* delivered to Mr Imbs, on 11 May 1973, give us the occasion of return not nostalgic but historic on some founding principles (semantics of the word, the choice of the examples) which have governed the writing of the dictionary during the years of starting up and on its constant evolution.

We shall show how much the contents and the presentation of the articles of the *TLF* evolved between the first one and the last volume, even if the standards of writing the definition and the choice of the examples were finely defined at first. The decision of the computerization of the *TLF* was taken in 1990, we shall describe the difficulties as well as the advantages brought to the reader with regard to the paper version. Our assessment will

end on an optimistic note: although imperfect the *TLF* remains a vast heritage on which the laboratory can lean to develop researches in the lexical domain or the didactics of languages.

Introduction

Les *Mélanges de Linguistique Française et de Philologie et Littérature Médiévales* remis à Paul Imbs le 11 mai 1973, nous donnent l'occasion d'un retour non pas nostalgique mais historique sur quelques principes fondateurs (i.e. le contenu sémantique du mot, le choix des exemples) qui ont présidé à la rédaction du dictionnaire au cours des années dites de mise en route et sur son évolution constante.

1. Historique de l'évolution des normes du *TLF*

Dans une note intitulée «Note sur la structure lexicale immanente du français » publiée dans un *Bulletin TLF*, Paul Imbs déclarait « Cette note est un document distribué aux rédacteurs du dictionnaire des XIXe et XXe siècles, préparé au *Centre pour un Trésor de la langue française*. D'autres notes internes, techniques ou scientifiques, seront publiées à cette place ; leur publication voudrait susciter un débat ouvert sur un ensemble de problèmes dont l'actualité n'échappera pas aux spécialistes. »

Cette note concluait sur la délicate tâche de la définition : « Au moment de conclure cette introduction, il convient de préciser encore une fois ceci : avec le découpage des sens, le travail le plus délicat de l'analyse lexicologique est la définition ». On y trouve les conseils de Paul Imbs sur la rédaction de la définition et sur l'optimisation de la définition : « En effet une définition, pour ne convenir qu'au seul défini, doit donner de celui-ci une description minutieuse, avec un grand nombre de composantes sémiques permettant de cerner le sens sans rien laisser de pertinent dans l'ombre ».

Dans les conclusions du colloque « Lexicologie et Lexicographie françaises et romanes » organisé à Strasbourg en novembre 1957, il est conseillé d'achever la définition par un choix d'exemples pertinents : ils ne seront jamais forgés par le lexicographe et seront aptes à évoquer l'atmosphère « stylistique » du mot. Dans l'*Encyclopédie internationale de lexicographie* éditée en 1989 Robert Martin décrit les fonctions de l'exemple. Elles sont linguistiques, philologiques, encyclopédiques, idéologiques, esthétiques. Il cite à plusieurs reprises la préface de Paul Imbs décrivant le souci philologique qui entoure le choix des exemples « Est philologie tout ce qui concerne les références et dans une large mesure le choix même des exemples ; est philologique le principe, posé dans ce dictionnaire, que les énoncés servant d'exemples ne sont pas l'œuvre des rédacteurs, mais d'auteurs usant de la langue sans préoccupation linguistique directe et donc non suspects de gauchir les matériaux de la preuve dans le sens de la thèse à prouver ». Tout ce qui est décrit est attesté dans le corpus et le corpus est décrit de manière exhaustive.

Si les normes de rédaction concernant la définition et le choix des exemples étaient clairement définies pour l'ensemble des rédacteurs, la lecture des seize tomes du *TLF* montre que le contenu et la présentation des articles ont évolué entre le premier et le dernier tome en raison de leurs différentes normes de rédaction et de présentation. Au fur et à mesure de l'avancement de la rédaction, le nombre de volumes augmentait. Aucun volume n'était défini au départ, puis quatorze volumes furent un premier objectif, mais l'ensemble des données a

nécessité seize volumes. Les derniers volumes présentent un texte plus condensé sans exemples détachés (exemples longs avec alinéas), moins aéré avec de très nombreuses abréviations et des paragraphes plus compacts, une police de caractères plus petite sur du papier plus fin.

Les grandes étapes de rédaction des normes comportent trois dates essentielles bien qu'il y eût plusieurs adaptations des normes tout au long des années de rédaction. Tout d'abord en 1970 lors de la rédaction du premier volume puisque celui-ci est paru en 1971. La deuxième étape datant d'octobre 1972 touche six grands domaines que nous expliciterons plus en détails. Et enfin, la dernière étape de rédaction d'un cahier des normes date de février 1979, sous la direction de Bernard Quemada.

De nombreux séminaires, animés par M. Robert Martin, abordaient les difficultés rencontrées lors de la rédaction des articles afin d'aboutir à une nouvelle version des normes qui remplaçaient certaines pages du cahier des normes. Nous dresserons l'inventaire de ces modifications successives.

2. Bilan de l'évolution des articles

Force nous est de constater que la présentation des articles a beaucoup fluctué : les articles des premiers tomes ne contiennent pas d'exemples enchaînés (exemples sans alinéas) mais en revanche ils contiennent de nombreux exemples détachés, 318 exemples pour le mot AMOUR, ce qui est tout à fait à l'opposé du mot VOIR figurant dans le seizième et dernier volume et qui, avec une analyse sémantique plus riche et complexe que le mot AMOUR, ne contient que 5 exemples détachés mais 590 exemples enchaînés.

Au travers de quelques exemples nous montrerons que cette évolution progressive n'est, en fait, pas néfaste à l'homogénéité des données du dictionnaire.
Tout d'abord en ce qui concerne la présentation de la rubrique « ETYMOLOGIE ET HISTOIRE », celle-ci a beaucoup évolué au cours des articles des premiers tomes. Nous montrerons en détail son évolution du début de la lettre A aux mots commençant par CAL-, à partir desquels elle a atteint sa présentation définitive.

Les synonymes et les antonymes n'ont pas toujours bénéficié du même traitement dans l'ensemble des volumes. Ils sont présents du début à la fin mais leur fréquence varie. Quant aux antonymes, fréquents dans les premiers tomes, ils commencent à être moins fréquemment cités à partir de la lettre L pour devenir assez rares dans les derniers volumes. Leur place dans l'article a aussi un peu évolué. Les normes de 1979 précisent que la place du synonyme doit être « immédiatement derrière la définition » ce qui s'avère exact dans les derniers volumes tandis que dans les premiers, le synonyme est placé entre les syntagmes et l'exemple détaché, ce qui a rendu difficile leur reconnaissance automatique par les automates de rétroconversion lors de l'informatisation du dictionnaire et ceci peut également rendre difficile un traitement automatique de récupération et d'analyse des données dans le but de travaux futurs sur le *TLF*.

Nous montrerons grâce à certains articles comment leur présence et leur place ont évolué dans le dictionnaire. On pourrait penser a priori que pareille évolution est une gêne pour la lecture des tomes qui sont si différents, mais en réalité le lecteur qui consulte un article ne semble pas troublé par l'apparence des autres articles qui ont évolué selon d'autres normes. Son intérêt se

concentre essentiellement sur les données lexicographiques. Au-delà de la présentation non homogène des articles, il réorganise les informations en fonction de sa pratique du dictionnaire ou de ses besoins de recherche.

Nous citerons un autre exemple de présentation dont la pratique s'est accrue au fil des tomes et qui peut provoquer une réelle gêne dans la consultation des articles : dans un but de gain de place, on a eu recours, dans les sources bibliographiques des exemples, des syntagmes et des définitions, à l'emploi de « Id. » pour remplacer un auteur qui vient d'être cité, à « ibid. » pour remplacer un ouvrage ou une revue, seul ou accompagné d'une date, et/ou une page, éventuellement un numéro de tome qui vient d'être immédiatement cité et à « op. cit. » pour un ouvrage ou une revue cités à un endroit quelconque dans l'article.
Nous montrerons comment cette pratique est difficile à lire dans la version papier.

3. Apport de l'informatisation du TLF

La décision d'informatiser le *TLF* fut prise en 1990, elle était notamment définie ainsi : « En simplifiant, on peut dire qu'informatiser le dictionnaire c'est le rendre lisible après coup sur ordinateur, non pas seulement en substituant à sa forme éditoriale classique celle d'écrans-pages, mais en transformant son contenu en base de données relationnelles qui permettra d'afficher des données que la consultation manuelle ne pourrait parfaitement rassembler… » in Dictionnairique et Lexicographie p. 195, 1990.

3.1 Les difficultés rencontrées lors de l'informatisation

L'informatisation du *TLF* a surmonté cette évolution méthodologique grâce à des programmes qui ont aplani de nombreuses difficultés.

Si certaines fautes de frappe ne sont pas un obstacle lors de la lecture d'un article sur la version papier, certaines ont été un véritable barrage aux programmes automatiques d'informatisation. L'ordre alphanumérique des paragraphes hiérarchiques n'a pas toujours été respecté, on passe du A au C, on trouve deux grand I et on passe au III, sans omettre de citer le désordre dans l'alphabet grec. Les parenthèses ouvrantes sont parfois fermées par des crochets ou inversement et on peut noter bien d'autres fantaisies sur la version papier qui sont toutes corrigées sur la version électronique du fait de la rigueur des programmes.

Les données des articles sont organisées en blocs hiérarchiques, ces blocs sont matérialisés sur la version papier par des espaces typographiques qu'on appelle « doubles blancs ». Ces espaces, souvent invisibles par le lecteur, ont été primordiaux pour l'informatisation des articles, et plus particulièrement des articles des derniers tomes qui ont été fortement condensés en paragraphes denses. L'informatisation a dû les prendre en compte puisqu'ils différencient les informations lexicographiques du bloc informationnel et que des balises vont permettre d'interroger le texte sur la portée des blocs les uns par rapport aux autres.
Nous donnerons un aperçu des difficultés rencontrées lors de l'informatisation au travers de différents exemples.

3.2 Rôle des programmes de recherche

Les programmes d'informatisation ont su se jouer de la place aléatoire des objets, et ils ont été adaptés pour retrouver les synonymes quelle que soit leur place dans l'article, à la suite du

définition, entre les syntagmes et les exemples enchaînés. Le texte du *TLFi* n'a pas été modifié dans ce cas.

En ce qui concerne le traitement des sources bibliographiques, on remarque que la restitution des références a été effectuée dans le *TLFi*.

L'informatisation présente aussi l'avantage de combiner des éléments proches, qui ne posent pas de problèmes de compréhension au lecteur de la version papier puisqu'il ne fait qu'une lecture linéaire des données mais qui permettent à l'utilisateur de la version électronique de faire des requêtes transversales plus cohérentes sur l'ensemble des seize volumes.

Ainsi, dans le cadre 3 du formulaire de recherche assistée, permettant de faire une recherche par domaine technique, l'utilisateur averti peut constater qu'en interrogeant le domaine « Cuisine » il obtient également les résultats des domaines « Art(s) Culinaire(s) », « Cuisine » et « Gastronomie ». En effet, ces trois domaines étaient si proches (CROUSTADE appartient au domaine « Cuisine », CHIPOLATA et ANDOUILLETTE sont marquées du domaine « Gastronomie » alors que SAUCISSE et BOUDIN appartiennent aux « arts culinaires ») qu'il était préférable de les regrouper en une seule interrogation sans changer le texte du *TLF*. L'utilisateur peut toujours dans le cadre 5 de la recherche assistée interroger séparément chacun des trois domaines.

Le logiciel regroupe ainsi dans quelques recherches des éléments voisins afin de donner de meilleurs résultats et de gommer l'effet disparate de l'évolution des normes sur autant d'années de rédaction.

Conclusion

Le TLF est parfois critiquable, mais nous sommes fières d'avoir contribué à l'élaboration de cette oeuvre, car associé à la base Frantext, le *TLF* reste un vaste patrimoine sur lequel le laboratoire peut s'appuyer pour développer des recherches dans le domaine lexical ou la didactique des langues. La mission de l'après *TLF* se perpétue : le *Supplément* demandé par tous est en cours d'achèvement. Le projet DILAN s'appuie sur le *TLF* pour sa recherche sur la proximité sémantique. Morphalou s'est appuyé sur la nomenclature du *TLF*. En traitement automatique du langage, le *TLFi* est une ressource qui est loin d'être entièrement exploitée.

Bibliographie

Dictionnairique et Lexicographie, 1990.

Paul Imbs : Préface au TLF. TLF T1. Paris 1971, IX-XLVII.

Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles. Colloque Strasbourg 1957. Paris 1960.

Mélanges de Linguistique Française et de Philologie et Littérature Médiévales à M. Imbs, 11 mai 1973.

Le traitement automatique : un moteur pour l'évolution des dictionnaires de synonymes

Jean-Luc Manguin (1)
jean-luc.manguin@unicaen.fr
Lonneke Van der Plas (2)
vdplas@let.rug.nl
Jörg Tiedemann (2)
tiedeman@let.rug.nl

(1) CRISCO, Université de CAEN (FR)
(2) Alfa-Informatica, Université de GRONINGEN (NL)

Mots-clés : lexicographie, extraction de synonymes, corpus multilingues.

Résumé : Nous présentons ici un aperçu des traitements automatiques liés aux dictionnaires de synonymes et devant servir à l'enrichissement lexicographique. Nous décrivons tout d'abord les méthodes endogènes qui se basent sur une modélisation en graphe de la relation de synonymie, et qui ont pour but secondaire de vérifier la qualité des relations. En seconde partie, nous abordons les méthodes exogènes qui extraient les relations synonymiques d'une analyse de corpus textuels monolingues ou multilingues. Nous insisterons plus particulièrement sur ce second type de corpus qui apporte une précision et un rappel supérieurs à la méthode monolingue, et dégage de bonnes perspectives lexicographiques.

Introduction : la lexicographie instrumentée

La préface du récent « Dictionnaire des combinaisons de mots » paru aux éditions Le Robert [Le Fur, 2006] nous révèle l'importance actuelle d'une lexicographie « à l'instrument »[1], y compris dans l'élaboration de produits destinés à l'édition papier. Dans le domaine des dictionnaires électroniques, la partie instrumentale devient parfois un but en soi, concomitant à l'élaboration de la ressource, comme dans les travaux qui visent à construire des ressources terminologiques.
Dans le cas des dictionnaires de synonymes, l'instrumentation telle que nous venons de la décrire aura aussi pour but la construction de la ressource, ou d'une partie complémentaire de celle-ci. Cependant, il importe de distinguer deux aspects dans cette construction, qui servent d'ailleurs à disposer les dictionnaires selon deux points de vue classiquement admis[2] : d'une part la recherche des substituts d'un mot (qui oriente la ressource constituée vers les dictionnaires cumulatifs), d'autre part la recherche des conditions de substitution (qui la place

[1] En extrapolant l'expression de B. Habert [Habert, 2005].
[2] Voir par exemple [Quemada, 1968], ou [Pruvost, 2007].

cette fois du côté des dictionnaires distinctifs). Il va de soi que le second aspect exige de recourir à un matériau externe à la ressource que l'on bâtit, qu'il soit un corpus de textes ou un dictionnaire complémentaire. Par contre, et c'est ce que nous allons développer ici, l'ajout de nouvelles relations synonymiques à un dictionnaire déjà existant peut non seulement faire appel à des apports extérieurs, mais aussi fonctionner de manière autarcique. Dans les deux cas, la théorie sous-jacente diffère sérieusement en raison de la nature des données à traiter, et du modèle qui leur est appliqué.

1. Les forces et les faiblesses d'un dictionnaire électronique

Il existe peu de dictionnaires électroniques des synonymes du français, tandis que l'anglais bénéficie des développements du projet WordNet, qui n'est pas à proprement parler un dictionnaire de synonymes, puisqu'il organise les mots en réseau au moyen de différents types de relations (entre autres synonymie, hyponymie, hyperonymie). Celui que nous avons utilisé dans ce travail est mentionné par Habert [op. cit.], et plus connu sous le nom de « Dictionnaire Electronique des Synonymes du CRISCO ». Sa particularité essentielle, voulue dès le départ par ses concepteurs, est d'être purement cumulatif, c'est-à-dire de ne donner pour chaque entrée qu'une liste de synonymes, sans faire de distinction d'emploi ni de regroupement de sens.

Cette démarche radicale s'explique par le fait que ce dictionnaire ne devait constituer qu'un support réel à une démarche de modélisation du sens formulée par B. Victorri [Victorri et Fuchs, 1996] ; la mise en ligne de cette ressource et le succès qui a suivi ont révélé par ailleurs la forte demande des internautes à l'égard d'une telle ressource, et ont déclenché des travaux complémentaires visant à doter la liste des synonymes d'informations supplémentaires, comme la barre de classement des synonymes.

Cela dit, le contenu initial du dictionnaire, issu de la fusion de fichiers contenant les liens synonymiques présents dans sept autres ressources[3], et complété par le travail d'harmonisation des lexicographes du CRISCO, reste néanmoins essentiellement tributaire des relations posées par les auteurs des dictionnaires compilés. Par exemple, la relation de synonymie entre *curieux* et *insolite*, mentionnée par le TLFi, n'est présente dans aucun des dictionnaires sources. Cet exemple simple montre, comme l'avait déjà signalé Kahlmann [Kahlmann, 1975], que d'une part le travail des lexicographes peut être entaché d'oublis, et que d'autre part, même la compilation de plusieurs ouvrages ne met pas à l'abri de telles lacunes.

D'un autre côté, la simplicité structurelle de ce dictionnaire de synonymes, dans lequel les mots sont simplement reliés par une relation symétrique de type booléen, permet de le faire entrer parmi les graphes. Placé ainsi dans une catégorie d'objets mathématiques dont le substrat théorique est déjà abondamment développé et toujours en cours d'enrichissement, il peut être soumis à des analyses et à des transformations bien connues, et dont les résultats pourront être examinés en tenant compte cette fois de la valeur sémantique de la relation présente dans le graphe. En d'autres termes, l'objectivité des méthodes mathématiques de la théorie des graphes fait contrepoids à la critique précédente concernant la subjectivité des liens présents dans la ressource.

[3] Dictionnaires de Guizot, Lafaye, Bailly, Bénac, du Chazaud, et renvois synonymiques du Grand Larousse et du Grand Robert ; pour plus de détails voir [Ploux, 1997].

2. Enrichissement endogène

Ce premier type d'enrichissement ne consiste bien entendu qu'à ajouter des liaisons dans le graphe de synonymie, puisqu'il se base sur l'exploitation de ce graphe par des méthodes diverses ; celles-ci ne peuvent pas faire apparaître de nouveaux nœuds dans la structure.
Le principe général de ces méthodes consiste à explorer le graphe de synonymie autour d'un nœud, en allant plus loin que les voisins directs de ce nœud de départ. Les premiers travaux de ce genre sont probablement ceux de Brodda et Karlgren [Brodda et Karlgren, 1969], qui ont proposé un modèle « thermodynamique » fondé sur une analogie entre le réseau synonymique et un réseau conducteur de chaleur avec des pertes de transmission. Les auteurs ont réalisé une version informatique d'un dictionnaire de synonymes suédois dans lequel, pour trouver les synonymes proches d'un mot, il suffisait de « chauffer » le mot en question, puis de laisser le système arriver à un état d'équilibre tout en maintenant constante la température du mot « chauffé » ; bien entendu, tous les changements de température au sein du réseau étaient calculés par l'ordinateur, et l'utilisateur n'avait plus qu'à consulter une liste des mots les plus « chauds », qui selon le modèle implémenté, s'avérait correspondre aux synonymes les plus proches du mot activé. L'intérêt du modèle était d'offrir la possibilité de chauffer plusieurs mots à la fois, par exemple dans le cas où l'un d'eux est polysémique, et de pouvoir faire varier la transmission de chaleur entre les nœuds du graphe ; mais la finalité n'était pas de modifier le réseau de départ.
Plus récemment, JL Manguin [Manguin, 2004] a proposé d'ajouter de nouvelles liaisons par « transitivité conditionnelle » ; cette méthode se base d'une part sur la transitivité (si B est synonyme de A, et C synonyme de B, alors C est synonyme de A), et d'autre part sur la similitude entre deux mots du graphe, mesurée par l'indice de Jaccard. En l'occurrence, cet indice de communauté est égal au nombre de synonymes communs aux deux mots, divisé par le nombre total de synonymes qu'ils possèdent à eux deux. La transitivité est dite « conditionnelle » si pour B synonyme de A et C synonyme de B, on peut ajouter une liaison entre A et C qu'à certaines conditions, notamment si la similitude entre A et C est strictement supérieure à 0,5 . Dans l'étude faite à partir du dictionnaire de Bailly, les liaisons ajoutées étaient très pertinentes, et si des liaisons proposées étaient aberrantes, cela révélait même des erreurs dans le dictionnaire d'origine. L'exemple mentionné précédemment de *curieux* et *insolite* peut être résolu de cette manière dans le DES du CRISCO.
Dans un but presque similaire, Bruno Gaume [Gaume, 2006] a exploité le graphe de synonymie issu du même dictionnaire de synonymes, pour trois catégories (verbe, nom et adjectif), par une méthode mathématique tenant compte de la totalité du graphe, afin de trouver quels sont les mots les plus proches de celui choisi comme point de départ[4]. Les nouvelles liaisons ainsi créées ont été baptisées « proxémies », et si cette proxémie peut être souvent confondue avec la synonymie (car elle relie deux mots généralement considérés comme substituables), elle aboutit également à des relations au sein d'un même champ sémantique, comme entre les verbes *déshabiller* et *éplucher*. L'intérêt du travail de Bruno Gaume, sur le lexique verbal par exemple, est de révéler par ces relations des substitutions qui apparaissent en production chez des sujets dont le lexique mental n'est pas encore fixé[5].

[4] A ce point de vue, on peut considérer son travail comme analogue à celui de Brodda & Karlgren.
[5] La raison de cet inachèvement peut être normale (chez des petits enfants) ou anormale ; c'est l'une des applications de ce travail, qu'il serait trop long de détailler ici.

3. Enrichissement exogène

Ce type d'enrichissement consiste à développer un programme d'extraction de synonymes applicable à divers corpus ; à vrai dire, la finalité de ce travail n'est pas forcément l'enrichissement d'un dictionnaire existant, mais plus souvent la production d'un dictionnaire ex nihilo. En outre, les principes mis en œuvre vont notablement différer selon qu'ils s'appliqueront à un corpus monolingue ou à un corpus multilingue.

Dans le cas du traitement d'un corpus monolingue, on commence généralement par analyser syntaxiquement le corpus en question, avant d'effectuer une analyse distributionnelle sur les unités repérées. L'idée qui sous-tend cette méthode est que les unités qui partagent des contextes distributionnels semblables sont sémantiquement proches. Ainsi, la mesure des similarités distributionnelles permet par la suite de faire apparaître les unités liées sémantiquement, comme l'a fait Didier Bourigault [Bourigault, 2002] avec un corpus de 10 années du journal « Le Monde » ou bien avec les textes des romans provenant de la base Frantext. Cependant, comme il l'a lui même montré [Bourigault et Galy, 2005], cette méthode rapproche assez peu d'unités synonymes, et quand bien même on appliquerait un filtrage catégoriel sur les résultats, les « voisins » obtenus compteraient parmi eux de nombreux antonymes, hyponymes ou hypernonymes.

La méthode que nous détaillerons ici améliore grandement la précision des résultats, grâce au choix d'un corpus multilingue aligné. Cette fois, l'idée sous-jacente est que si deux mots sont souvent traduits de la même manière dans de nombreuses langues, il y a une forte probabilité pour qu'ils soient synonymes. Ainsi, en utilisant les traductions, on trouve moins de mots apparentés parmi les unités similaires, car typiquement la traduction ne s'étend pas aux hyperonymes, (co)hyponymes ou antonymes. Par exemple, les mots « vin », « boisson » et « bière » ne se traduisent pas avec le même mot dans une autre langue. En outre, comme nous l'avons montré dans une étude précédente qui concernait le néerlandais, l'emploi de plusieurs langues donne de meilleurs résultats qu'un corpus bilingue, même dans le cas des langues les mieux apparentées [Van der Plas & Tiedemann, 2006].

Dans notre démarche, nous utilisons le corpus Europarl dont les textes proviennent des actes du Parlement Européen en 11 langues différentes[6], et dont nous avons tiré les traductions par des techniques dérivées de la traduction statistique automatique (outil open-source GIZA++). Plus précisément, les textes sont tout d'abord alignés phrase à phrase selon les techniques développées par Gale et Church [Gale & Church, 1993], puis mot à mot par GIZA++. Chaque mot se trouve ainsi pourvu de ses traductions dans les 10 autres langues avec leurs fréquences, ce qui constitue son vecteur caractéristique. Les similarités individuelles entre les mots sont ensuite calculées par comparaison de ces vecteurs, en tenant compte des fréquences des traductions, puisque l'indice prend en compte l'Information Mutuelle[7]. Finalement nous pouvons, pour les mots de la langue cible (ici le français), obtenir une liste de mots sémantiquement proches dont chacun est pourvu d'une similarité comprise entre 0 et 1. Un filtrage catégoriel est appliqué aux listes de synonymes proposés avant d'évaluer les résultats, comme dans le cas du corpus monolingue[8].

Les résultats pour le néerlandais ayant été satisfaisants [Van der Plas & Tiedemann, op. cit.], nous avons renouvelé l'opération avec la langue française, en changeant la référence d'évaluation des synonymes proposés en raison de certaines difficultés rencontrées avec la version néerlandaise de EuroWordnet. Pour le français, nous avons donc choisi le DES du

[6] Précisons qu'il s'agit des débats qui ont lieu au Parlement Européen, et non des textes issus de la Commission Européenne dont le technolecte n'est pas suffisamment riche.

[7] Pour plus de détails, on pourra se reporter à [Van der Plas & Tiedemann, 2006].

[8] Le filtrage élimine comme dans l'autre méthode certains résultats erronés, mais dont l'origine réside dans les problèmes d'alignement.

CRISCO. L'évaluation des résultats montre que les synonymes proposés par le système peuvent atteindre une précision double et un rappel triple de ceux obtenus avec un corpus monolingue, si l'on considère le dictionnaire des synonymes comme référence. Mais l'autre intérêt de la méthode, sur lequel nous voulons insister ici, c'est que ce traitement automatique fait apparaître des paires de synonymes qui devraient en principe se trouver dans le dictionnaire. Ainsi, les paires « affection – pathologie » ou « documentaire – reportage » sont clairement mises en évidence alors qu'elles sont absentes du dictionnaire. Par exemple, pour des similarités entre termes supérieures à 0,3, environ les deux tiers des relations proposées sont déjà présentes dans le dictionnaire ; mais parmi les relations qui en sont absentes et que le processus de traitement propose, un tiers forment des paires synonymiques parfaitement acceptables. La précision observée passe ainsi de 67 % à 77 %, et le dictionnaire peut par cette voie s'enrichir de nouvelles relations.

En outre l'apport des résultats ne se limite pas là, puisque le système produit aussi des paires dont les membres ne figurent pas forcément tous les deux parmi les entrées du dictionnaire. Par exemple, la détection de couples comme « adaptation – *reformulation* », « affaiblissement – *fragilisation* » ou « diversité – *pluralisme* » introduisent dans le dictionnaire les mots placés ici en italique. Cette faculté de détection prouve que la lexicographie synonymique trouve là un intérêt non négligeable.

Cela dit, deux problèmes liés, l'un d'ordre technique, l'autre d'ordre linguistique demeurent en suspens. Tout d'abord, les méthodes d'alignement statistique ne sont pas toujours à même de réaliser des appariements obéissant à l'organisation des énoncés ; c'est la raison pour laquelle nous avons appliqué à nos résultats un filtrage catégoriel pour mettre à l'écart des paires comme « majorité – majoritairement », celle-ci résultant bien sûr d'un alignement incorrect de la locution « en majorité ». Nous avons commencé à travailler sur ce problème en dotant le système d'alignement d'un dictionnaire de locutions, ce qui améliore grandement les résultats. La présence de ces paires hétérogènes (au point de vue catégoriel) révèle aussi un second problème qui pourrait se définir d'une manière globale par « la question des paraphrases ». En effet, le premier synonyme proposé par exemple pour « ville » est l'adjectif « urbain », et provient des difficultés d'alignement d'un mot sur des paraphrases comme « milieu urbain », « zone urbaine » ou « domaine urbain » ; mais il est relativement difficile de répertorier les paraphrases qui forment un ensemble ouvert.

Enfin, le fait que le rappel n'atteigne pas 50 % est un peu décevant, mais s'explique d'une part par le fait qu'une substitution d'un mot par un de ses synonymes va parfois s'accompagner d'un changement de niveau de langue, et que ce changement va aussi se retrouver dans les traductions. Par exemple, « pompe » est synonyme familier de « chaussure » selon notre dictionnaire, mais dans notre corpus ce mot n'est jamais utilisé dans ce sens, mais dans celui de l'appareil de pompage. Cela nous amène à l'autre explication de la faiblesse du rappel : la richesse du corpus ; il est évident que les débats entre parlementaires européens se cantonnent dans un niveau de langue assez soutenu, malgré la diversité des sujets abordés et la variété des avis formulés, et que cette forme de discours n'atteint pas la richesse d'un corpus journalistique de 10 années. Il est donc nécessaire pour nous de poursuivre ce travail par une mise à l'épreuve avec un corpus plus vaste et plus varié ; cette expérience de variété a débuté avec les sous-titrage des films, mais nous ne disposons pas à l'heure actuelle des sous-titres en français.

Conclusion : vers un dictionnaire des substituts ?

Comme nous l'avons vu, les traitements automatiques, qu'ils soient endogènes ou exogènes, constituent un jeu d'instruments intéressants pour le contrôle et surtout l'enrichissement des dictionnaires de synonymes. En outre, les perspectives qui apparaissent lors du traitement des corpus multilingues alignés laissent entrevoir des possibilités d'évolution pour cette sorte de dictionnaire. En effet, le traitement des unités complexes, et l'évolution des méthodes multilingues vers un alignement « fonctionnel » et non plus mot à mot, permettraient d'accéder à la construction de dictionnaires des substituts (ou de paraphrases) qui seraient d'un grand intérêt pour les apprenants d'une langue étrangère.

Bibliographie

BOURIGAULT Didier (2002) : « UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus », *Actes TALN 2002*, Nancy.

BOURIGAULT Didier & GALY, Edith (2005) « Analyse distributionnelle de corpus de langue générale et synonymie », Lorient, *Actes JLC 2005*.

BRODDA Benny & KARLGREN Hans (1969) : « Synonyms and synonyms of synonyms », *SMIL*, 5, pp. 3-17. Stockholm.

GALE, W. & CHURCH, K. (1993) : « A program for aligning sentences in bilingual corpora » In *Computational Linguistics*, 19(1).

GAUME Bruno (2006) : « Cartographier la forme du sens dans les petits mondes Lexicaux », *Actes JADT 2006*, Besançon.

HABERT Benoît (2005), *Instruments et ressources électroniques pour le français*, Paris, Ophrys.

KAHLMANN André (1975), *Traitement automatique d'un dictionnaire de synonymes*, Stockholm, Université de Stockholm.

LE FUR Dominique et al. (2006), *Dictionnaire des combinaisons de mots*, Paris, Editions le Robert.

MANGUIN Jean-Luc (2004) : « Transitivité partielle de la synonymie : application aux dictionnaires de synonymes », *CORELA – Cognition, Représentation, Langage*, Vol 2, n° 2.

PLOUX Sabine (1997). « Modélisation et traitement informatique de la synonymie ». *Linguisticae Investigationes*, XXI (1), Amsterdam, John Benjamins.

PRUVOST Jean (2006), *Les dictionnaires français outils d'une langue et d'une culture*, Paris, Ophrys.

QUEMADA Bernard (1968), *Les Dictionnaires du français moderne (1539-1863). Étude sur leur histoire, leurs types et leurs méthodes*, Paris, Didier.

VAN DER PLAS Lonneke & TIEDEMANN Jörg (2006), « Finding synonyms using automatic word alignment and measures of distributional similarity » *Actes de ACL/Coling 2006*, Sydney.

VICTORRI Bernard & FUCHS Catherine (1996), *La polysémie : une construction dynamique du sens*, Paris, Hermès.

L'apport méthodologique du TLF et les orientations d'aujourd'hui

Robert Martin (1)
eveline.martin@wanadoo.fr

(1) ATILF Nancy Université & CNRS

Le TLF apparaît incontestablement, par son extraordinaire richesse et par la réussite de son informatisation, comme le dictionnaire le plus important du français. Mais la *méthodologie* du TLF est-elle innovante et, si oui, en quoi ? Dans une entreprise d'aujourd'hui qui prend le TLF pour modèle (même beaucoup plus restreinte comme celle du DMF), quelles *orientations lexicographiques* peuvent renouveler la méthode ? Ce sont les questions auxquelles, successivement, on s'efforcera de répondre.

1. L'apport méthodologique du TLF

a) L'apport méthodologique du TLF porte moins sans doute sur les principes que sur l'environnement de l'élaboration lexicographique. Comme on sait, les *principes* sont exposés, au reste sans excessive contrainte, dans la magnifique *Préface* du premier volume (1971, p. IX – XLVII). Soucieux, à cette étape décisive, de fixer les orientations, Paul Imbs évoque le "profil général", puis la "procédure lexicographique" et conduit ensuite de la "théorie" à la "pratique". On observera qu'à aucun moment, dans cet exposé lumineux, l'initiateur du TLF ne présente tel ou tel principe comme novateur en soi. L'idée est plutôt (prolongeant en cela le Colloque de 1957) qu'une entreprise lexicographique est confrontée à un ensemble d'options, et que sa réussite tient avant toute chose à la cohérence des choix qui sont faits. Ce sont des choix que la *Préface* détermine, justifiés un à un par les finalités qui sont assignées à l'ouvrage.

Avec le recul cependant, certaines des options prises apparaissent à bon droit comme novatrices et fécondes. En voici quelques-unes.

- Tout d'abord le principe qui sous-tend la macrostructure : l'idée de fonder la nomenclature sur un corpus, avec les exigences d'homogénéité que cela suppose et les correctifs inévitables qu'il y faut apporter (p. XXVI – XVIII), apparaît, à la date du premier volume, comme incontestablement neuve et assurément pertinente ; elle émerge déjà à diverses reprises au Colloque de 1957 ; Paul Imbs a le mérite de la cerner avec la rigueur et la prudence qui s'imposent. Dans le domaine des termes techniques p. ex., le parti (formulé par la suite dans le Cahier dit "des normes" en 1978) de retenir dans un dictionnaire général comme le TLF les vocables, et eux seuls, qui appartiennent aux manuels scolaires et aux ouvrages du premier cycle universitaire constitue un choix explicite et du fait même objectivement appréciable ; il y a là une importante avancée.

- On citera aussi un principe de microstructure, auquel à vrai dire Paul Imbs n'accorde pas toute l'importance qui lui revient, mais qui porte en lui une innovation de très grande portée ; ce principe consiste à séparer nettement les contenus définitoires et les "conditions d'emploi" (p. XXXIV) ; les "conditions d'emploi" sont les conditions contextuelles qui doivent être satisfaites pour que se réalise discursivement tel ou tel sens. C'est là, quoique à peine suggérée, une idée de première importance ; il ne suffit pas en effet, dans un dictionnaire, d'énumérer les sens et d'en spécifier le contenu ; il faut dire aussi ce qui est nécessaire pour qu'un sens émerge et non pas tel autre ; si la portée d'une telle suggestion n'a pas été évaluée à sa juste valeur, et si le TLF, il faut en convenir, est très loin de la réaliser de manière satisfaisante, l'innovation, au moins suggérée, n'en est pas moins considérable.

- Le principe par ailleurs, toujours en microstructure, que les sens (du moins les "acceptions") doivent être reliés par un système cohérent d' "indicateurs sémantiques" (identifiés dans le *Cahier des normes* : p. ext., en partic., p. méton., p. anal., au fig.), aucun dictionnaire, avant le TLF, ne l'avait aussi clairement systématisé.

- Il s'y ajoute aussi le principe, peut-être le plus important, qu'un dictionnaire synchronique n'est pas indépendant, et ne saurait l'être, de l'histoire de la langue : un mot, tout mot, porte en lui la trace de son histoire, et certains faits ne trouvent leur explication que dans la diachronie ; cette idée éloigne fortement de la tradition et plus encore du structuralisme régnant dans les années soixante ; certes, de très grands dictionnaires antérieurs sont, dans leur fondement, des dictionnaires historiques (le NED ; le Littré) ; Paul Robert, au Colloque (*Actes*, p.108-109), insiste avec force sur le rôle de l'histoire dans les classements sémantiques et Georges Matoré (*Actes*, p.91) estime, à rebours, que la préoccupation historique encombre la méthode lexicologique. Mais là où le TLF incontestablement innove, c'est dans la portée explicative, en synchronie, qu'il accorde à l'histoire ; à une époque où le structuralisme triomphant impose ses vues et fait table rase du passé, Paul Imbs, souvent dans la tourmente, a su maintenir le cap avec une belle constance, non seulement en plaidant pour des "historiques", mais en défendant le point de vue fécond que l'explication synchronique ne saurait, en lexicographie même synchronique, rendre compte de tout et suffire à l'intelligence des vocables.

b) Bref, il y a là des idées assurément novatrices. Mais l'apport méthodologique le plus visible du TLF est ailleurs : le TLF a innové avant tout par l'*environnement lexicographique* qu'il a créé.

Le TLF est tout d'abord le premier dictionnaire scientifique de nature institutionnelle. En lexicographie "institutionnelle", on ne connaissait jusque-là que les dictionnaires de l'Académie : des dictionnaires d' "honnêtes gens", soucieux du "bon usage", influant sur la norme, infléchissant et régulant l'usage, mais sans véritable visée scientifique. Le TLF s'écarte d'emblée de la préoccupation normative. Dès le départ, il s'apparente au NED, mais en s'inscrivant avec plus de force encore dans la recherche publique. Au Colloque de 1957, la présence de Michel Lejeune, éminent linguiste et à l'époque Directeur adjoint du CNRS, la part qu'il a prise aux conclusions et les engagements que dès cette date il a le mérite de formuler, l'influence des *Actes* du Colloque que, sous l'impulsion de Paul Imbs, le CNRS a patronné et surtout l'action déterminante de Paul Imbs pour la création du "Centre de recherche pour un Trésor de la langue française", tout cela a constitué, dans l'histoire de la lexicographie, un véritable tournant. On estimera peut-être que l'on s'éloigne ici du fond. En fait, l'orientation institutionnelle a rendu possible, par les moyens exceptionnels qu'elle a

procurés, l'essentiel de l'apport méthodologique du TLF : le TLF a contribué de manière déterminante à l'émergence de la lexicographie informatisée.

Avant le TLF, il n'existait, en matière d'automatisation, que de maigres réalisations mécanographiques : le Colloque de 1957 les donne en modèle, faute de disposer encore des perspectives technologiques d'avenir qui, à travers le monde, s'esquissaient à peine. C'est au TLF qu'est revenu le privilège extraordinaire de l'ouverture informatique. Le CNRS a su, dans notre discipline, prendre très tôt la direction qu'il fallait, et le Gamma 60 a représenté pour la méthode lexicographique, en 1964, une innovation capitale. Certes, les tâtonnements ont été nombreux (les ayant vécus, j'aurais beaucoup à dire là-dessus !). Mais dès les débuts, certaines initiatives ont infléchi sensiblement la méthode lexicographique. Ainsi on s'est orienté très vite du côté de ce que l'on a appelé depuis les "analyseurs morphologiques" (par la construction, dès 1965, d'un "dictionnaire des formes flexionnelles" et d'un "dictionnaire des homographes") ; on a mis à profit, sous la conduite de Charles Muller, les techniques statistiques (en particulier dans la recherche dite des "groupes binaires") ; on a élaboré aussi des "organigrammes fonctionnels" en vue du repérage (dans les mots fréquents, en particulier les mots grammaticaux) des exemples les plus significatifs (cette recherche, il est vrai, est restée expérimentale et n'a pas été mise à profit systématiquement dans la rédaction, mais elle a ouvert des perspectives novatrices).

Plus généralement, l'expérience du TLF a accrédité l'idée que dans une entreprise lexicographique, au-delà des résultats publiés, il fallait rendre disponible, pour la recherche, la documentation (en l'occurrence gigantesque) que le projet a suscitée. Dès le Colloque de 1957, cette idée est mainte fois esquissée. Mais c'est évidemment la base Frantext, grâce au logiciel Stella de Jacques Dendien, qui l'a pleinement réalisée. Une date très importante a été par la suite, en 2002, quand le TLF informatisé a été relié à Frantext : le TLF est le premier dictionnaire qui, au-delà de son contenu, est mis directement en relation avec les bases textuelles qui le fondent. Désormais, le consultant du dictionnaire peut compléter la matière à son gré, la critiquer en toute connaissance de cause et la réorienter au besoin. C'est déterminant pour la recherche. En matière d'informatisation, le TLF avait certes un prédécesseur prestigieux, le NED ; mais il est le premier à faire le lien avec les bases documentaires, et il y a là, tout porte à le penser, une donnée essentielle en lexicographie.

2. Quelques orientations méthodologiques d'aujourd'hui

a)Les acquis méthodologiques du TLF profitent pleinement aux projets d'aujourd'hui, et tout particulièrement, parmi d'autres, à celui du DMF (c'est celui que je connais le mieux).

Qui douterait désormais que la *documentation lexicographique* doit être largement informatisée ? On notera au passage que les dépouillements du TLF ont montré que la courbe, hyperbolique, d'accroissement du vocabulaire tend à devenir plate une fois rassemblé un corpus homogène de quelque quatre à cinq millions d'occurrences. Au-delà, les textes nouveaux apportent de moins en moins ; une fois le seuil franchi, tout en poursuivant modérément les saisies intégrales pour maintenir la base vivante, on gagne donc, par la lecture de textes supplémentaires, à sélectionner les faits qui paraissent lexicologiquement les plus pertinents et, moyennant la saisie des passages qui les illustrent, à constituer une base de "partiels" ; celle-ci, au reste, peut être interrogeable, comme les bases intégrales, sur tout l'ensemble des faits qu'elle comporte.

Par ailleurs, une double visée devrait guider désormais tout projet lexicographique : l'*informatisation du dictionnaire* par le balisage de son contenu et le *lien hypertextuel avec les bases qui ont servi à le construire*. Ici et là, le TLF sert de modèle, étant entendu cependant que le balisage n'est plus opéré désormais a posteriori, par rétroconversion, mais au moment-même de l'élaboration du dictionnaire.

b) Il s'ajoute à tous ces acquis de nouvelles orientations méthodologiques ; on peut les résumer sous trois chefs.

- Tout d'abord, il y a tout à gagner (l'expérience du DMF le prouve amplement grâce aux outils élaborés par Gilles Souvay) à pratiquer une *rédaction assistée par ordinateur* ; l'ordinateur se place alors au coeur de l'élaboration lexicographique et fournit incessamment une aide technique à la rédaction ; le dictionnaire se construit à l'écran ; naturellement, les problèmes scientifiques restent strictement inchangés ; mais on trouve dans l'automate un appui technique considérable. Une telle rédaction – sous masque de saisie – suppose que soit mise au point une *grammaire du métatexte* ; les balises s'ouvrent dès lors selon un ordre intangible ; le lexicographe les remplit au fur et à mesure ; la rédaction est en somme guidée par une grammaire à états finis, qui a l'immense avantage de garantir un résultat homogène, structuré, où des "blocs" strictement définis résolvent, à l'interrogation, les problèmes de portée, et où aucune des informations utiles, correctement hiérarchisées, ne peut plus être omise. Un "*correcteur lexicographique*" peut contribuer par ailleurs à repérer certaines erreurs (par exemple dans l'ordre chronologique des exemples, dans la structuration de l'article ou encore dans la gestion des renvois). Le lien avec les bases textuelles (et avec la bibliographie et ses abréviations) permet des enrichissements constants et immédiats.

- Grâce à l'informatique, la lexicographie d'aujour'hui présente aussi l'avantage de la *modularité*. On peut rédiger le dictionnaire, non plus dans l'ordre alphabétique, mais par une suite de modules dont chacun possède sa propre cohérence tout en restant ouvert sur les étapes ultérieures. Ainsi la rédaction devient évolutive, dynamique si l'on préfère. Qu'il y ait des inconvénients à cette pratique, il ne faut pas en disconvenir (l'objet étant instable). Mais les avantages l'emportent de beaucoup, le plus important étant que les manques peuvent être comblés, les erreurs corrigées, les structures repensées. Le dictionnaire devient en somme indéfiniment perfectible.

- L'instrument peut même devenir *modulable* au gré du consultant. L'objet infiniment complexe et mutidimensionnel du "signifié de langue" ne peut se satisfaire d'une représentation bidimensionnelle unique, comme la proposent forcément les dictionnaires sur papier. Il existe toujours, en représentation, une pluralité de possibles. L'informatique permet de ménager cette pluralité. On renvoie sur ce point au *Bull. Soc. Ling. de Paris* 102, 2007, 17-33.

Il va sans dire qu'un autre aspect encore, déterminant, est l'ouverture du dictionnaire informatisé sur le traitement automatique des langues. Mais n'ayant plus travaillé moi-même la question depuis *Sémantique et automate* (Paris, PUF, 2002), je n'y toucherai pas, sinon pour rappeler après d'autres l'extrême importance qu'elle revêt.

.

Définition et exemple : quelle complémentarité ? L'illustration du concept dans le « Dictionnaire alphabétique et analogique du français des activités physiques et sportives » (à paraître, 2009)

Pierluigi Ligas (1)
pierluigi.ligas@univr.it

(1) Université de Vérone (Italie), Département d'Études Romanes

---

Mots-clés : lexicographie, dictionnaire spécialisé, sport, concept, définition, exemple

Key-words : lexicography, specialized dictionary, sport, concept, definition, example

Résumé :

Cet article s'inscrit dans le cadre des méthodologies en lexicographie et en constitution de ressources dictionnairiques. Il y est question d'un dictionnaire de langue de spécialité, le « Dictionnaire alphabétique et analogique du français des activités physiques et sportives », en cours d'élaboration à l'Université de Vérone, et, plus particulièrement, de l'articulation entre définition et exemple dont dépend la bonne appréhension (et compréhension) du concept dénoté par l'entrée lexicale. L'optique choisie est compatible avec les méthodes actuelles de confection des dictionnaires, qui s'appuient principalement sur les corpus et qui adoptent une démarche sémasiologique.

Abstract:

This paper deals with a specialized dictionary, the « Dictionnaire alphabétique et analogique du français des activités physiques et sportives », under elaboration at the University of Verona, and, more specifically, with the relationship between the definition and the examples, on which the correct interpretation of the concept denoted by the lexical entry largely depends. Our approach is compatible with the most recent dictionary compilation methods which start from linguistic corpora, adopting a semasiological view.

Introduction

La théorie générale de la terminologie place le concept au cœur de l'activité terminologique[1], et c'est autour de concepts que normalement les dictionnaires spécialisés organisent leurs

[1] La théorie générale de la terminologie de Wüster a été remise en question notamment par Cabre, qui en dénonce la standardisation excessive, qui conduit à négliger l'aspect communicatif des termes (Cf. Cabre, M. T.

entrées, puisqu'ils considèrent les termes comme étant l'expression linguistique de l'organisation de connaissances dans un domaine donné. Ce genre de travail, mû la plupart du temps par des impératifs au départ extérieurs à la langue, n'est pas sans avoir des conséquences sur le plan linguistique et même, le cas échéant, pédagogique. Tant et si bien que notre « Dictionnaire alphabétique et analogique du français des activités physiques et sportives » – qui comprend également l'anatomie, la médecine du sport, la condition physique, la presse sportive, les installations, les matériels, etc. (au total 10000 entrées) – est le résultat d'une démarche qui se situe à mi-chemin entre celle du terminologue et celle du lexicographe et qui consiste à appréhender les termes dans leur fonctionnement linguistique sans pour autant perdre de vue l'optique conceptuelle.

1. Entre langue générale et langue de spécialité

1.1 La définition

Une forme linguistique devient terme lorsqu'on parvient à cerner et à expliquer la place du concept qu'elle dénote dans le système conceptuel d'un domaine [L'Homme, 2005], et un terme n'est un terme que dans la mesure où il peut recevoir une définition, car la définition est à la base même de la terminologie [Béjoint, 1997]. Condition nécessaire, mais non suffisante. Si la particularité du terme, par rapport aux autres unités lexicales, est d'avoir un sens spécialisé, c'est-à-dire un sens qui peut être mis en rapport avec un domaine de spécialité, il n'est pas toujours exclusif au domaine étudié : il arrive en effet que certains termes spécifiques à un domaine soient largement utilisés ailleurs, y compris dans la langue générale. Et encore faut-il s'entendre sur le sens de définition. Pour les logiciens de Port Royal, la définition est un remède à la confusion qui naît dans nos pensées et dans nos discours de la confusion des mots ; étymologiquement, le mot « définition » provient du latin classique *definitio*. En français, c'est un substantif verbal de 'définir', composé du radical 'finir' qui renvoie aux sens de la finitude et du bornage, autrement dit, le fait de mettre un terme à quelque chose [Rey, 1977]. Finitude et bornage concernent bien évidemment la place du terme et du concept qu'il dénote dans le champ sémantique et le système conceptuel du domaine d'une part, et ses relations avec les autres termes et concepts appartenant au même domaine d'autre part.

Nous avons commencé notre travail, voici une vingtaine d'années, dans le cadre des cours de langue de spécialité dispensés à l'UFR STAPS de l'Université de Bologne en constituant un corpus que nous avons constamment enrichi et mis à jour – 'corpus ouvert', selon la typologie proposée par Pearson [1998] – pour y puiser, au fur et à mesure de l'avancement du travail de confection du dictionnaire, des formes linguistiques susceptibles de dénoter des concepts là où elles étaient réalisées, à savoir : ouvrages spécialisés, quotidiens et magazines sportifs, dictionnaires généraux, encyclopédies, lexiques, glossaires, radio, télévision, Internet (plus récemment), enquêtes sur le terrain (interviews, conversations etc.…). Après avoir isolé les unités lexicales à charge spécialisée et/ou pouvant être rattachées à un domaine spécifique, nous avons analysé les contextes pour en extraire des énoncés d'intérêt définitoire, des exemples, des synonymes, des collocations etc., et ce pour chacune des unités lexicales relevées. En dépouillant le corpus, nous nous sommes aperçu que la tâche la plus compliquée consistait justement à formuler des énoncés définitoires[2] obéissant un tant soit peu aux principes uni-

(2000) : *Elements for a theory of terminology : Towards an alternative paradigm, Terminology*, Amsterdam, Benjamins, vol. 6, n. 1, p. 35-57).

[2] Nous utilisons ici 'énoncés définitoires' comme synonyme de définitions (ou de parties de définitions), non pas dans le sens de structures, au niveau du corpus, où s'actualisent les relations sémantiques que les mots entretiennent entre eux, et qui sont suffisamment stables

versellement admis de rédaction des définitions, compte tenu des contraintes de grammaticalité et de conformité à la norme syntactique.
Comme chacun sait, on peut reconnaître deux niveaux dans une définition : le niveau conceptuel et le niveau linguistique, qui ne coïncident pas mais qui sont en relation. S'agissant de constituer une ressource dictionnairique en langue de spécialité, l'approche lexicographique que nous avons adoptée a produit nécessairement, dans la plupart des cas, des définitions qui prennent en compte les relations chose-chose ou signe-chose [Rey-Debove, 1966]. Rédigées à partir des traits pertinents des notions relevés dans les ouvrages consultés et les ressources terminologiques constituant le corpus, nos définitions sont presque toutes construites en une seule phrase. D'une manière générale, dans le cas de termes non uninotionnels, nous nous sommes donné comme règle de faire en sorte que la zone sémantique, par ses 'blocs définitionnels' [Pruvost, 2006], permette de faire le lien entre les différents sens du même terme.

Dans notre dictionnaire sont représentés les procédés définitoires suivants : 1) définition par genre prochain et différence spécifique ; 2) définition partitive ; 3) définition synonymique ; 4) définition par description ; 5) définition mixte par synonymie et par description ; 6) définition opératoire ou définition par fonction. Le seul procédé définitoire non retenu est la définition par démonstration, qui aurait consisté à fournir une représentation visuelle du concept (dessin, illustration…)3.

Très tôt nous avons été amené à constater que la définition ne parvenait pas toujours à appréhender la réalité d'une manière pleinement satisfaisante, surtout dans le cas d'objets concrets qu'il est nécessaire d'envisager dans différents contextes, et nous nous sommes posé la question de savoir où commence et où s'arrête la pertinence dans l'énumération des traits sémantiques, conscient du fait que la longueur de la définition n'est pas fonction de la complexité du concept à illustrer, même si l'on s'accorde à dire qu'une définition minimale devrait comprendre au moins deux mots et que, pour une définition maximale, théoriquement il n'y a pas de limite au nombre de mots. Or, au lieu d'alourdir la zone sémantique, nous avons décidé d'intervenir au niveau des exemples – qui occupent les zones syntagmatique et phraséologique (combinatoires) de l'article – les seules structures lexicales capables de pallier convenablement les insuffisances de la définition. Tout comme l'entrée, l'exemple forme une séquence autonyme [Rey, 1979 - Rey-Debove, 2005]; cette séquence autonyme rétablit l'usage de l'entrée en discours. Qu'il soit bref ou long, signé ou forgé, « l'exemple est cité à l'intérieur de l'article en tant que macrosigne, et c'est ce signifié autonymique qui fait la preuve de la définition du signe en la signifiant » [Rey, 1979].

1.2 L'exemple

Ce qui suit va nous permettre de revenir sur le constat que certains termes spécifiques à un domaine sont largement utilisés dans d'autres, y compris dans la langue générale, et surtout de montrer que l'exemple est en mesure de fournir un complément d'information, au cas où celle

---

pour que leur repérage automatique puisse être envisagé dans différents discours (Cf. Rebeyrolle, J. et Tanguy, L. (2000) : *Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires*, Cahiers de Grammaire, 25, Sémantique et Corpus, p. 153-174), structures qui sont pour nous des énoncés 'd'intérêt' définitoire [Auger, 1997] pouvant donner lieu à des définitions. L'énoncé définitoire correspond à ce que R. Martin désigne comme étant une *définition naturelle*, qui s'oppose à la définition lexicographique qu'il désigne par *définition conventionnelle* [Martin, 1990].

[3] La typologie de procédés définitoires que nous avons adoptée est celle proposée par l'équipe du Pavel, didacticiel de terminologie (http://www.termiumplus.gc.ca/didacticiel_tutorial/francais/lecon3/page3_5_4_f.html, consulté le 27.08.2007).

apportée par la définition se révèlerait insuffisante ou non suffisamment claire, et de rattacher le terme à un domaine de spécialité dans le cas d'une définition trop abstraite ou généralisante. De plus, l'exemple est le seul énoncé pouvant témoigner du comportement en langue de la vedette ainsi que des liens sémantiques entre la vedette et les éventuelles expressions figurées qui la contiennent. Ce qui compte, c'est que l'exemple, choisi pour le contenu du mot et son contexte situationnel [Rey-Debove, 2005], ne fasse pas double emploi avec la définition, qu'il n'y ait pas de redondance. Le fait de privilégier le contexte (donc l'exemple) au lieu de la définition dans le cas de termes à faible teneur en spécificité ou dans le cas d'ambiguïté sémantique a permis une meilleure et plus sûre accessibilité au concept et au système conceptuel du domaine évoqué, et, par le réseau des renvois, de voir d'autres termes et concepts à l'œuvre en discours[4]. Le passage de la langue au discours est fondamental pour la connaissance du mot [Rey-Debove, 2005].

Prenons *jument* et *pouliche*, mots de la langue générale (parce que les concepts qu'ils dénotent sont au départ extérieurs au domaine spécialisé qui nous intéresse, en l'occurrence les sports équestres). *Le Robert des sports*, de G. Petiot [1984], pour citer un dictionnaire de sport généraliste proche du nôtre, ne leur consacre aucun article. *Le Petit Robert* [2000] donne comme définition de *jument* : 'femelle du cheval' ; suivent des synonymes et des phrasèmes. De ces derniers, le seul pouvant en quelque sorte ramener le concept de *jument* à l'activité sportive ou au sport loisir de la randonnée à cheval est *monter une jument.* Quant à *pouliche* : 'jument qui n'est pas encore adulte (mais qui n'est plus un poulain)' ; suivent le syntagme *pouliche de courses*, renvoyant au sport, et une citation : « Une ancienne pouliche, encore fort belle, un peu couronnée seulement » (Flaubert). Or, dans *Le Petit Robert*, dictionnaire de la langue générale, le sens de *jument* et de *pouliche* ne se distingue pas, au niveau de la définition, du sens que ces mots véhiculent dans d'autres contextes, comme le contexte sportif, car il s'agit toujours respectivement de la *femelle du cheval* et d'une *jument qui n'est pas encore adulte*. Nous verrons que, dans le cas du « Dictionnaire alphabétique et analogique du français des activités physiques et sportives », la spécificité, au niveau du concept, émerge dans la zone syntagmatique, qui réunit des exemples cités et/ou forgés, et dans la zone phraséologique (collocations, locutions…) si elle est présente dans l'article. De plus, l'exemple (contexte) convoque des co-textes spécialisés.

Voici les articles *jument* et *pouliche* de notre dictionnaire:

jument

[n.f.] SEQ

Femelle du cheval.

a) «Plusieurs jeunes, dont Olivier Guillon, trente ans, avec sa *jument* Baladine du Mesnil, poussent la porte de l'équipe de France» *(EQU)*; *b)* Descendante d'excellents chevaux d'Auteuil, Princesse d'Armos a le profil type de la très bonne *jument* d'obstacle* *(INT)*.

Syn.: *cavale*, *pouliche*.

pouliche

[n.f.] SEQ

Jeune jument qui n'a pas procréé et qui est généralement âgée de moins de trois ans.*

a) «À l'occasion d'une course de trot attelé*, la *pouliche* drivée* par le Français a changé de ligne» (ad.) *(EQU)*; *b)*

[4] Les renvois, matérialisés par des astérisques, établissent des liens avec d'autres entrées lexicales et d'autres concepts favorisant ainsi une appréhension globale du champ notionnel du domaine dans lequel se situe une forme. Ce cheminement notionnel de renvoi en renvoi, cette circularité obligée, indiquera que l'on a fait le tour des développements possibles de la notion de départ.

«Petite *pouliche* au grand coeur, qui ne cesse de progresser depuis le début de l'année, Poltava remporte sa première victoire de groupe, c'est beau» *(INT)*.

Syn.: *jument*.

[légende - n.f. : nom féminin ; (ad.) : adapté ; SEQ : sports équestres ; *(EQU)* : chaîne thématique *Equidia*) ; *(INT)* : Internet ; Syn. : synonymes (zone paradigmatique) ; * renvoi].

On remarquera au passage que des deux définitions, celle de *jument*, bien que minimale, appartient aux définitions par genre prochain (cheval) et différence spécifique (femelle), tandis que celle de *pouliche* appartient aux définitions mixtes, par synonymie (*jeune jument*) et description (*qui n'a pas procréé et qui est généralement âgée de moins de trois ans*).

Dans une définition, c'est à l'ensemble des concepts que renvoie l'ensemble des signifiants. Il est évident que dans le cas de ces deux mots, ce qui les rattache au domaine spécialisé et qui en fait des termes, ce ne sont pas les définitions, ce sont les exemples : on peut dire pertinemment que l'exemple vole la vedette à la définition... Cela se traduit par une meilleure appréhension et compréhension du concept au sein du système conceptuel du domaine des sports équestres : dans le cas de *jument*, par les syntagmes *équipe de France*, ex. *a)*, et *jument d'obstacle*, ex. *b)* ; dans le cas de *pouliche*, par les mots *course*, *trot attelé*, *drivée*, ex. *a)*, et par le syntagme *victoire de groupe*, ex. *b)*. De plus, l'exemple *b)* de l'article *jument* nous apprend que les *juments* peuvent disputer des concours d'obstacles, l'exemple *a)* de l'article *pouliche* que les *pouliches* peuvent disputer des courses de trot attelé, et dans l'exemple *b)* le syntagme 'victoire de groupe' renvoie incontestablement à la compétition.

Dans le cas de termes à forte charge spécialisée, la description sémantique ne devrait pas viser aux définitions absolues, mais aux définitions qui délimitent le(s) sens du terme par rapport à celui (ceux) de ses quasi-synonymes[5]. L'exemple peut alors soit se limiter à fournir une simple attestation du mot en contexte, soit apporter des renseignements complémentaires.

Prenons l'article *bon-plein*, terme de marine utilisé en sports nautiques et en navigation de plaisance.

bon(-)plein

[n.m.] SPN

Allure de près, c'est-à-dire assez proche du lit du vent*, mais plus abattue* que le près serré*, comprise entre ce dernier et le travers* (INT).*

«L'allure* la plus efficace pour la navigation et la plus confortable pour le marin est désignée sous le terme de *bon-plein* ; les voiles sont écartées de l'axe du bateau* pour maintenir l'écoulement laminaire* optimal» (ad.) *(WIK)*.

Syn.: *près*, *près serré*, *travers*.

♦ Gouverner* *bon-plein*, porter* *bon-plein*, se positionner *bon-plein*: barrer* de manière à présenter les voiles du navire directement à l'action du vent (ad.) *(ALEX)* ♦ Naviguer au *bon plein* - EX.: «12h30: nous coupons le moteur et *naviguons au bon plein*, le vent vient maintenant du SE» *(INT)*.

[5] Ce qui permet entre autres de fixer les nuances d'emploi. La compréhension des ces termes et des conditions de leur emploi en contexte est favorisée par la présence, dans le bloc-entrée (qui comprend également des informations métalinguistiques : catégorie grammaticale, genre, et, éventuellement, langue et/ou niveau de langue) d'indicateurs de domaine, classés par ordre alphabétique. Dans le cas de termes appartenant à plusieurs domaines, l'ordre des définitions suit l'ordre des indicateurs (v. les articles *cuiller* et *looping* ci-dessous).

[légende - n.m. : nom masculin ; spn : sports nautiques ; (WIK) : Wikipédia ; (ad.) adapté ; (ALEX) : Alexandria ; ♦ : début de la zone phraséologique et séparation entre les expressions].

La définition proprement dite, qui occupe la zone sémantique, se présente ici sous la forme d'une définition par compréhension, privilégiée en terminologie [Dubuc, 1992], qui permet essentiellement de reconnaître : 1) l'appartenance de l'objet bon-plein à la classe conceptuelle des allures (incluant ou mot d'ancrage) ; 2) les caractéristiques (traits spécifiques) permettant de distinguer cet objet des autres objets équivalents du même système conceptuel. Quant à l'unique exemple, la première partie (jusqu'au point-virgule), véritable énoncé définitoire, fournit, plus par sa forme (il prend en compte la relation chose-signe par la présence du verbe métalinguistique désigner) que par son contenu, une définition 'possible' de bon-plein : 'allure la plus efficace pour la navigation et la plus confortable pour le marin'. Cependant, cet exemple ne saurait suffire du point de vue conceptuel, car il ne répond au critère de spécificité que d'une manière fort incomplète, bien qu'il respecte la forme canonique 'genre prochain et différence spécifique' : nous pouvons en effet reconnaître l'incluant allure et les traits spécifiques 'la plus efficace pour la navigation' et 'la plus confortable pour le marin'. Il apporte donc un complément d'information qu'il eût été parfaitement inutile d'intégrer à la définition. En revanche, la deuxième partie de l'exemple, qui n'est pas un énoncé définitoire : « les voiles sont écartées de l'axe du bateau pour maintenir l'écoulement laminaire* optimal », non seulement a un sens pour le spécialiste, mais introduit le terme technique 'écoulement laminaire' (qui, certes, représente une complication pour les non-spécialistes, mais il fait l'objet d'un renvoi) lequel, avec 'axe du bateau', fournit des renseignements mieux adaptés au système conceptuel du domaine évoqué. La zone phraséologique, avec ses collocations suivies du sens recouvert par l'ensemble base + collocatif, apporte des éléments très utiles pour une meilleure illustration du concept dénoté. L'exemple phraséologique, cité, montre le fonctionnement en discours d'une des collocations.

Il est aisé d'observer que dans les articles reproduits ci-dessus les exemples sont tous cités (la source est indiquée en forme abrégée à la fin, entre parenthèses) : nous en avons fait presque une règle, car comme Potter [1998], nous sommes d'avis que les exemples forgés ne reflètent pas toujours l'usage langagier courant, c'est-à-dire le langage actuellement parlé ou écrit.

Notre approche du corpus et l'élaboration des articles du dictionnaire au niveau des zones sémantique, syntagmatique et phraséologique, dont on vient de voir quelques spécimens, nous a permis de dénombrer, toujours dans l'optique de la complémentarité définition/exemple – qui seule peut conduire à une illustration satisfaisante du concept et au respect du principe de cohérence au niveau de l'article tout entier (le même principe que celui préconisé par les auteurs du DECFC [Mel'čuk *et al.*, 1984, 1988, 1992, 1999] et qui consiste à mettre en correspondance les composantes sémantiques de l'entrée, les actants syntaxiques, les co-occurrents lexicaux)[6] – quatre types d'exemples : *a)* exemple comme simple attestation de l'entrée et de son fonctionnement en discours ; *b)* exemple à vocation définitoire ; *c)* exemple d'intérêt définitoire ; *d)* exemple à vocation (ou d'intérêt) définitoire par alliance conceptuelle (fournit une attestation du terme vedette tout en en définissant un autre appartenant au même système conceptuel). Par brièveté, il ne sera question ici que des exemples de type *b)* et *c)*.

[6] Des dix zones présentes dans le DECFC nous en avons retenu cinq : zone vedette (avec variantes orthographiques éventuelles); zone morphologique (suivie des indicateurs de domaine); zone sémantique (définitions); zone syntagmatique (exemples); zone paradigmatique (synonymes, antonymes, mots de sens voisin) ; zone phraséologique (collocations, combinaisons de mots, locutions...).

L'exemple à vocation définitoire est en fait une définition, qui au lieu d'investir la zone sémantique de l'article se situe dans la zone syntagmatique et qui contient le terme à définir. Ce type d'énoncé, à condition qu'il soit extrait d'un corpus, offre une description du sens lexical plus naturelle, en ce sens qu'elle est plus proche de la pratique spontanée de la langue, et véhicule en même temps des informations sur le signe et sur le référent[7], comme la première partie de l'unique exemple de l'article *bon-plein* ci-dessus – «L'allure* la plus efficace pour la navigation et la plus confortable pour le marin est désignée sous le terme de *bon-plein* » – et les exemples *b)* des articles *cuiller* et *looping* ci-dessous, dont les vedettes dénotent des concepts appartenant à deux domaines différents.

cuiller (cuillère) [2]
[n.f.] RUG, VLB
1) Action de faire perdre l'équilibre à un joueur en lui attrapant ou en lui bousculant un pied. 2) Frappe de type service effectuée en serrant les doigts et en creusant la main.
a) « La *cuillère*, le soulèvement du joueur et les plaquages* au-dessus du bassin sont interdits » RUG *(INT)*; *b)* Le service des débutants est communément appelé *cuillère* VLB *(INT)*.
♦ *Cuillère* de bois: expression de rugby qui désigne une récompense virtuelle pour l'équipe qui perd tous ses matchs lors du Tournoi des Six Nations - EX.: « Le pays de Galles, grâce à une énorme entame de match et à son ouvreur* James Hook, a évité l'humiliation de la *cuillère de bois* en battant les Anglais 27 à 18 » RUG *(INT)* ♦ Cueillir le ballon en *cuillère*: technique consistant à pointer les mains vers le sol et à ramener le ballon sur la poitrine FTB ♦ Passe en *cuiller* (syn. *coup frappé*) HGA: action de pousser la balle avec la crosse* et de l'envoyer en l'air ♦ Projection en *cuiller* (syn.: *sukui-nage*) JUJ.

looping
[n.m.] angl. ESL, SPA
1) Étrier à quatre marches utilisé en escalade* artificielle. 2) Manoeuvre acrobatique aérienne qui consiste à faire une boucle* dans un plan vertical.*
a) « Sangles* larges rigidifiées avec renforts d'usure, le *looping* se mousquetonne* directement sur l'ancrage* ou sur le crochet* FIFI® » ESL *(INT)*; *b)* « Appelé aussi inversion ou tonneau* barriqué, le *looping* découle de l'inversion de virage durant une série de 360° asymétrique* » PAR (parapente) *(INT)*.
Syn.: 1) *étrier* 2) *boucle*, *inversion*, *tonneau*.

[légende - RUG : rugby ; VLB : volley-ball ; FTB : football ; HGA : hockey sur gazon ; JUJ : jiu-jitsu ; angl. : anglais ou anglicisme ; ESL : escalade ; SPA : sports aériens ; PAR : parachute (parapente est un sous-domaine)]

L'exemple *b)* de l'article *looping* apporte, en plus des équivalents lexicaux précédés du marqueur métalinguistique *appelé*, des détails techniques s'adressant aux spécialistes du

[7] C'est ce que Rey-Debove appelle 'connotation autonymique', système sémiotique consistant à faire en même temps usage et mention du défini [Rey-Debove, 2005].

domaine, alors que la définition correspondante, moins technique, est accessible aux non-spécialistes.
Quant à l'exemple d'intérêt définitoire, il contient bien évidemment, mais ne désigne pas, le terme à définir et, de même que l'exemple à vocation définitoire, il est censé apporter des informations complémentaires pour une meilleure compréhension du concept, comme les exemples *a)* des articles *cuiller* et *looping*. Le premier nous apprend qu'en rugby la *cuiller* est interdite, le second renseigne sur la composition et le mode d'emploi de l'objet *looping* dans le sens illustré à la définition *1)* qui se rapporte à l'escalade.

2. Conclusion

Les exemples disposent d'un atout considérable : ils échappent aux principes de rédaction des définitions. Le principe de simplicité, par exemple : ils peuvent en effet préciser des caractéristiques intrinsèques (nature, matière…) et extrinsèques (forme, fonction, origine, destination...) de l'objet représenté par le concept ; le principe de non-circularité : l'exemple doit forcément contenir le terme dénotant le concept à définir8. Ils peuvent en outre se présenter sous la forme d'énoncés négatifs, inclure la définition d'un autre terme, contenir des éléments subjectifs et des informations extérieures à la notion... Pour les définitions comme pour les exemples, ce qui est certain, c'est que le lexicographe jouit d'une grande liberté dans le choix, à partir d'un corpus, des traits accessoires qu'il juge utiles à la compréhension du concept et/ou au fonctionnement en discours du terme qui le dénote.

Même si des travaux comme ceux de Mel'čuk [1984] ou de Cruse [1986] peuvent aider à rendre les définitions plus fiables, surtout celles du lexique spécialisé, la précision des définitions sera toujours relative et l'exemplification demeurera fondamentale pour une meilleure connaissance de la chose nommée. « L'exemple est censé élucider l'information fournie par la définition et illustrer sous forme d'énoncés réels les propriétés les plus typiques du mot-vedette, ainsi que les contextes qui doivent permettre d'entrevoir les relations syntagmatiques ou collocationnelles et grammaticales que le mot entretient au niveau de la phrase. En outre, les exemples peuvent nous renseigner sur le registre ou niveau stylistique approprié d'un mot donné » [Dugardin, 2000]. Lorsque définition et exemple sont complémentaires et en parfait alignement, le concept s'en trouve doublement illustré. C'est pourquoi le choix (ou la rédaction) des exemples représente une tâche particulièrement délicate qui mérite toute l'attention du lexicographe.

Une typologie exhaustive de l'exemple a été proposée par Rey-Debove aux 1ères Journées allemandes des dictionnaires (2005). En ouverture de colloque, elle situe la problématique de l'exemple en faisant ressortir son statut autonymique et en examinant le problème du bon et du mauvais exemple. Elle conclut sur le statut particulier de l'exemplification, qui recouvre le vaste domaine des rapports entre langue et discours.

Une étude ultérieure resterait à faire qui mettrait en relief d'autres aspects de la complémentarité définition/exemple, afin de définir des 'niveaux', ou 'degrés', de complémentarité et d'illustrer une problématique explicite en posant des hypothèses théoriques et méthodologiques et, éventuellement, en proposant des parcours pédagogiques.

[8] « La seule obligation, impérative, est que l'exemple présente une occurrence du mot-entrée » [Martin, 1989].

Bibliographie

Auger, A. (1997) : « Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles », Thèse, Université de Neuchâtel.

Béjoint, H. (1997) : *Regards sur la définition en terminologie*, Cahiers de lexicologie, 70, 1, p. 19-26.

Cruse, D.A. (1986) : *Lexical Semantics*, Cambridge, London, New York, etc., Cambridge University Press (Cambridge Textbooks in Linguistics).

Dubuc, R. (1992) : *Manuel pratique de terminologie*, Montréal, Linguatech.

Dugardin, K. (2000) : *La problématique des phrases-exemples dans les dictionnaires d'apprentissage*, Romaneske (http://www.kuleuven.be/vlr/001dico.htm, consulté le 30.08.2007).

L'Homme, M.-Cl. (2005) : *Sur la notion de 'terme'*, Méta, I, 4, p. 1112-1132

Martin, R. (dir.) 1982 : *Regards sur la lexicographie*, Le français moderne, 50e année, n. 4, CILF.

Martin, R. (1989) : *L'exemple lexicographique dans le dictionnaire monolingue*, dans Hausmann, F. J. *et al.* (éd.), *Dictionnaires : Encyclopédie internationale de lexicographie*, Berlin [et] New York, De Gruyter, p. 599-607.

Martin, R. (1990) : *La définition 'naturelle'*, dans Chaurand, J. et Mazière, F. (dir.) (1990): *La Définition.* Actes du Colloque *La définition* organisé par le CELEX (Centre d'Etudes du Lexique) de l'Université Paris-Nord (Paris 13, Villetaneuse) à Paris, les 18 et 19 novembre 1988, Paris, Larousse, p. 86-95.

Mel'čuk, I. A., *et al.* (1992) : *Dictionnaire explicatif et combinatoire du Français contemporain*. Recherche lexico-sémantique III, Montréal, Les Presses de l'Université de Montréal.

Mel'čuk, I. A. *et al*. (1995) : *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve.

Pearson, J. (1998) : *Terms in Context*, Amsterdam, John Benjamins Publishing.

Potter, L. (1998) : *Setting a good example. What kind of examples best serve the users of learners' dictionaries*, dans *EURALEX '98 Proceedings. Papers submitted to the eighth EURALEX International Congress on Lexicography in Liege, Belgium*, p. 357-362. University of Liège, English and Dutch Departments.

Pruvost, J. (2006) : *Les Dictionnaires français outils d'une langue et d'une culture*, Paris, Ophrys.

Rey, A. (1977) : *L'impossible définition*. Le lexique images et modèles: du dictionnaire à la lexicologie, p. 98-113.

Rey A. (1979) : *La terminologie, noms et notions*, Collection « Que Sais-je ? », n°1780, PUF, Paris.

Rey-Debove, J. (1966) : *La Définition lexicographique: recherches sur l'équation sémique*, Cahiers de lexicologie, VIII, I, Didier-Larousse, p. 71-94.

Rey-Debove, J. (2005) : *Statut et fonction de l'exemple dans l'économie du dictionnaire*, dans Heinz, M. (dir.) : *Entre définition et citation : l'exemple. L'exemple lexicographique dans les dictionnaires français contemporains.* Actes des 1ères Journées allemandes des dictionnaires.

Colloque international de lexicographie, Klingenberg am Main, 25-27 juin 2004, Tübingen, Niemeyer, p. 15-20.

Wüster, E. (1976) : *La théorie générale de la terminologie - un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et les sciences des objets*, dans Dupuis, H. (éd.), *Essai de définition de la terminologie. Actes du colloque international de terminologie*, Québec, Manoir du lac Delage, 5-8 octobre 1975, Québec, Régie de la langue française, p. 49-57.

Métalangage et définition de substantif dans le T.L.F.: *le cas des «entre-crochets»*

Paolo Frassi (1)
frassipaolo@tiscali.it

(1) Université de Vérone

Mots-clés : définition, métalangage, marqueurs métalinguistiques, entre-crochets, approche modulaire.

Keywords: definition, metalanguage, metalinguistic markers, entre-crochets, modular approach.

Résumé : Les articles du *Trésor de la Langue Française* contiennent, à côté des définitions proprement dites, un ensemble d'informations placées entre crochets droits. Dans notre communication nous nous proposons de présenter les *entre-crochets* du point de vue du métalangage. Pour ce faire, après avoir défini la fonction des *entre-crochets* telle qu'elle a été envisagée dans les différents documents paratextuels édités et inédits qui accompagnent le *T.L.F.*, et après avoir tracé une ligne nette de démarcation entre approche typologique et approche métalinguistique, nous nous attarderons sur les *adjuvants démarcatifs* et *stylistiques*, en fournissant, pour chaque type d'adjuvant, et dans les limites des seules définitions de substantif, le type d'information qu'il véhicule ainsi que le niveau de métalangage auquel il appartient.

Abstract : The articles of the *Trésor de la Langue Française* contain, beyond the definitions, a set of information placed between square brackets. In my communication I suggest presenting the *entre-crochets* from a metalinguistic approach. To do this, having defined the function of the *entre-crochets* as it was considered in the various paratextual published and unpublished documents which accompany the *T.L.F.*, and having drawn a sharp line of demarcation between typological approach and metalinguistic approach, I shall linger on *adjuvants démarcatifs* and *adjuvants stylistiques*, by supplying, for every type of *adjuvant*, referring to the only definitions of nouns, the information which it conveys as well as the level of metalanguage to which it is up.

Introduction

Les articles qui composent le *Trésor de la Langue Française* ne consistent pas uniquement en un ensemble d'énoncés définitoires mais ils offrent un exemple d'approche modulaire (s'inscrivant dans le sillage des études menées par [Putnam, 1975a et 1975b]), où les informations stéréotypées contenues dans les définitions sont accompagnées d'informations de

types différents, généralement placées entre crochets droits, qu'il convient d'aborder du point de vue du métalangage.
Dans notre communication nous nous attacherons à démontrer que:

1. L'exploitation d' *entre-crochets* a bien été prévue dans les documents paratextuels édités (*Préface* au Tome Premier) et inédits («Normes de rédaction», «Pour un nouveau cahier de normes. Documents à discuter les 22, 23 [et 24] février 1979 lors de la "réunion des experts"»; «Cahier de normes») qui accompagnent le *T.L.F.*, bien que dans tous ces cas l'approche à ce sujet soit purement descriptive ou normative.
2. Le problème du métalangage a été traditionnellement abordé dans le cadre de l'étude typologique de la définition et pourtant, de par sa nature, il nécessite d'en être détaché.
3. Le *T.L.F.* constitue une référence et une source importantes, en raison de la grande exploitation des *entre-crochets*, dont il offre une variété à partir de laquelle nous allons proposer, dans les limites de la seule catégorie grammaticale du substantif, un classement dans le cadre de la question du métalangage et à partir de critères indépendants des études typologiques traditionnelles.

1. Documents paratestuels

Dans sa *Préface* au tome premier[1] du *T.L.F.*, Paul Imbs présente, à côté de la définition proprement dite, un ensemble d'informations qu'il appelle *adjuvants*: généralement placées entre crochets droits, ces informations sont classées en *adjuvants rhétoriques*, *adjuvants stylistiques* et *adjuvants démarcatifs*. Si les *adjuvants rhétoriques* relèvent généralement des figures de style, les *adjuvants stylistiques* portent éminemment sur la fonction pragmatique du langage. Quant aux *adjuvants démarcatifs*, ils placent l'entrée lexicale du point de vue de sa distribution, qui concerne tantôt les relations syntagmatiques qu'elle entretient avec les mots ou les parties du discours qui précèdent ou qui suivent, tantôt les variations morphologiques.
Les «Normes de rédaction» et les documents qui composent «Pour un nouveau cahier de normes. Documents à discuter les 22, 23 [et 24] février 1979 lors de la "réunion des experts"» contiennent un ensemble d'observations qui visent plutôt à la normalisation de la rédaction du dictionnaire qu'à une véritable théorie de la définition et du métalangage appliquée au *T.L.F.* et, en général, à tout ouvrage lexicographique. Ainsi ces mêmes remarques se trouvent-elles dans les «Normes de rédaction» à côté des prescriptions qui concernent la définition proprement dite ou encore dans «Pour un nouveau cahier de normes. Documents à discuter les 22, 23 [et 24] février 1979 lors de la "réunion des experts"» (section III, «Structure de l'article et métalangage»), où le problème du métalangage est abordé à l'instar de la définition, dans le cadre d'une esquisse de formalisation typologique.
À aucun moment il n'est question de la conscience d'une ligne nette de démarcation entre étude typologique et étude métalinguistique de la définition.

2. Métalangage et définition

Les études typologiques sur la définition classent, pour la plupart des cas, la définition métalinguistique (celle où l'hyperonyme est un terme métalinguistique) à côté d'autres types de définitions: c'est le cas de « Pour un nouveau cahier de normes. Documents à discuter les 22, 23 [et 24] février 1979 lors de la "réunion des experts" » et de [Martin, 1983], où le critère

[1] Cf. Imbs, P. (1971): *Préface* in C.N.R.S. (1971-1994): *Trésor de la langue française. Dictionnaire de la langue du XIX^e^ et du XX^e^ siècle (1789-1960)*, Tome Premier, Gallimard, Paris, p. XXXII.

pertinent choisi pour la classification des définitions oppose les définitions métalinguistiques aux définitions paraphrastiques; dans [Rey-Debove, 1998] cette même opposition ne constitue plus un critère pertinent de classification: les définitions métalinguistiques représentent une sous-catégorie des définitions avec faux hypéronyme.
Or, nous ne voulons pas nier la catégorie des définitions métalinguistiques mais déceler les enjeux de ce type de définitions et les raisons qui nous permettent de justifier la séparation entre métalangage et typologie.
L'enjeux d'une analyse typologique est représenté par la nature des relations existant entre l'entrée et l'ensemble des informations qui composent la définition: celle-ci contient un (ou plusieurs) hypéronyme(s) et un certain nombre de sèmes spécifiques ; seule la relation entre l'entrée et l'hypéronyme est de contenu à contenant, l'hypéronyme sélectionnant le genre ou l'espèce qui est ensuite spécifié par les sèmes qui composent la différence spécifique. Le critère pertinent de classification des études typologiques citées plus haut repose sur l'association entre la nature d'hypéronyme et la nature du référent: ainsi, une définition comme *Aiguillette: nom populaire de l'orphie* (*T.L.F.*), est métalinguistique alors que *Chaise: siège à dossier sans bras* est paraphrastique. De fait, la nature d'hypéronyme est indépendante de la nature du référent car le renvoi à la réalité infralinguistique ou extralinguistique ne constitue pas un élément pertinent pour cerner la nature des informations qui composent la définition, l'hypéronyme gardant sa propriété d'incluant général quel que soit le référent, infralinguistique ou extralinguistique. Ce problème est plus pertinent dans le cadre d'une étude portant sur le métalangage, dans le sens du niveau de métalangage qu'entraîne la phrase métalinguistique issue de la relation entre lemme et information métalinguistique.
Une étude métalinguistique, en effet, porte sur les niveaux de métalangage relevés à partir des phrases issues de la lecture des articles de dictionnaire, notamment la relation existant entre l'entrée et la définition ou, plus généralement, entre l'entrée et l'ensemble des informations qui composent l'article. Nous pouvons affirmer avec [Rey-Debove, 1997] que la relation syntaxique entre l'entrée lexicale et la définition lexicographique est un relation de signifié, où «*X* signifie *Y*» (l'entrée signifie la définition): il s'agit d'une relation entre deux autonymes dont le premier est un autonyme proprement dit et le second un autonyme avec «schize»[2], ce qui place la phrase en question au niveau *n*+1 de langage ou premier niveau de métalangage.
La nature des deux différentes approches, typologique et métalinguistique, justifie et exige deux traitements différents et nous amène à conclure que la question métalinguistique ne peut pas être abordée à l'instar et à l'intérieur de la question typologique.
À partir de cette prémisse, nous allons étendre l'analyse aux phrases métalinguistiques qu'il est possible de dégager de la relation entre l'entré et les informations (autres que la définition) placées entre crochets droits.

3. Le métalangage dans le *T.L.F.* ou la définition en dehors de l'approche typologique

À partir d'un échantillon représentatif de 5% du total des entrées principales du *T.L.F.*, qui nous a servi ailleurs pour une étude beaucoup plus vaste, menée sur toutes les catégories grammaticales et visant à montrer la nécessité de séparer l'étude typologique de la question du métalangage, nous allons présenter, dans les limites de la seule catégorie grammaticale du substantif, qui nous paraît suffisamment représentative, les résultats auxquels nous sommes parvenus au sujet des différents niveaux de métalangage.

[2] Le mot «schize» désigne un autonyme dont on expose l'un des deux fonctifs (expression ou contenu). (cf. Rey-Debove, J. (1997): *Le Métalangage. Étude linguistique du discours sur le langage*, Armand Colin, Paris, pp. 116-118).

L'analyse de l'échantillon nous a amenés à établir les différents types d'informations (contenues dans les *entre-crochets*) qui, en dehors de la définition, composent l'article du *T.L.F.* et que nous avons classés suivant les catégories d'informations qu'ils véhiculent en vue d'établir, pour chaque catégorie, à l'issue des travaux de J. Rey-Debove (et notamment [Rey-Debove, 1997]), le niveau de métalangage impliqué dans la phrase métalinguistique qu'il est possible de dégager de la relation entre l'entrée et l'information dont il est question.

Du point de vue des catégories d'informations que les *entre-crochets* véhiculent, seule la distinction entre *adjuvants démarcatifs* et *adjuvants stylistiques* a été retenue, les *adjuvants rhétoriques* contenant, pour la plupart des cas, une variété très restreinte d'informations.
Les résultats auxquels nous sommes parvenus sont les suivants:

1) Les *adjuvants démarcatifs* comprennent des *entre-crochets* qui contiennent:
   a des informations classématiques (ex: «Troupeau: A. [À propos d'animaux] 1. Ensemble d'animaux domestiques (brebis, moutons, vaches) nourris par l'élevage ou la pâture, faisant partie d'une exploitation agricole et/ou confiés à la garde d'un berger ou d'un vacher»).
   b des informations sémiques (ex: «Mission: 1. *En partic.* [Dans un cadre officiel incluant souvent un déplacement] a) Charge, fonction, mandat donnés à quelqu'un d'accomplir une tâche déterminée»).
   c des informations et classématiques et sémiques (ex: «Arrière: Espace ou partie d'une chose situé(e) dans la direction inverse de celle vers laquelle on regarde, dans laquelle on se déplace. A. [En parlant d'un être ou d'une chose organisée selon un axe de symétrie, la partie qui est dans la direction inverse de la face, de la façade]»).
   d des informations encyclopédiques (ex: «Calice: I.C.1.a) [P. réf. à l'usage [...] de se servir d'une coupe pour boire à la ronde] Portion d'héritage assignée à une personne»).
   e des informations syntaxiques (ex: «Vacance: I. [Suivi d'un comp. déterm.] État de ce qui est vacant»).
   f des informations de type mixte soit syntaxiques soit concernant le rôle agentif associées à des informations sémiques (ex: «Fleuve: D. *Au fig.* [Le compl. déterminatif désigne une chose abstr.] Ce qui suit un cours régulier, paisible; ce qui se développe continûment, avec ampleur.»).
   g des informations circulaires (ex: «Animal[1]: C.1.a) [P. réf. aux sens A et B]»; «Photo: A.1. [Correspond à *photographie* A.1] Photographique»).
2) Les *adjuvants stylistiques* comprennent des *entre-crochets* qui contiennent:
   a des informations exclusivement pragmatiques (ex: «Être[2]: II.B.2. *Cour*. Personne, individu [Pour exprimer une intention affectueuse]»).
   b des informations et pragmatiques et syntaxiques (ex: «Animal[1]: C. *Fam.* [En parlant de l'homme] 1.a.- [Avec un article déf. à valeur emphatique-démonstrative]»).

Les phrases correspondantes que l'on peut déduire permettent ainsi d'établir les différents niveaux du métalangage qui, en fait, ne dépassent jamais le niveau n+1.
Pour ce qui est des *adjuvants démarcatifs*, les *entre-crochets* 1.a, 1.b et 1.c relèvent de la simple description du contenu (ex: *«Troupeau» signifie «Ensemble d'animaux domestiques (brebis, moutons, vaches) nourris par l'élevage ou la pâture, faisant partie d'une exploitation agricole et/ou confiés à la garde d'un berger ou d'un vacher»*, *«Mission» signifie «Charge, fonction, mandat donnés à quelqu'un d'accomplir une tâche déterminée dans un cadre officiel incluant souvent un déplacement»*, *«Arrière» signifie «Espace ou partie d'une chose situé(e)*

dans la direction inverse de celle vers laquelle on regarde, dans laquelle on se déplace, en parlant d'un être ou d'une chose organisée selon un axe de symétrie, la partie qui est dans la direction inverse de la face, de la façade») et, étant de ce fait absorbés par la définition, ils appartiennent au niveau *n*+1 de langage ou premier niveau de métalangage (autonyme avec «schize»).
Les *entre-crochets* 1.d renvoient à des phrases du type *«portion d'héritage assignée à une personne» se réfère à l'usage de se servir d'une coupe pour boire à la ronde*, où l'information véhiculée par l'*entre-crochet* renvoie à la réalité extralinguistique et, par conséquent, au niveau *n* de langage sans entraîner pour autant aucun niveau de métalangage.
Les *entre-crochets* 1.e et 1.f contiennent des informations syntaxiques: dans le cas 1.e il s'agit uniquement de multisignes (*Le complément déterminatif «de» suit «vacance»*) et, de ce fait, du niveau *n*+1 de langage (ou premier niveau de métalangage) alors que dans le cas 1.f (type mixte) les informations syntaxiques renvoient, comme dans le cas 1.e, à des multisignes et donc au niveau *n*+1, tandis que les informations sémiques se référant à la réalité extralinguistique renvoient au niveau *n* de langage (*Le complément déterminatif de «fleuve» désigne une chose abstraite*).
Les informations circulaires entre sens et entre lemmes (1.g) renvoient respectivement au niveau *n*+1 de langage avec «schize» (relation avec «se référer» entre deux autonymes avec «schize») et au niveau *n*+1 de langage (relation avec «correspondre à» entre deux autonymes proprement dits).
Concernant les *adjuvants stylistiques*, les *entre-crochets* 2.a renvoient à la réalité extralinguistique (*L'emploi de «être[2]» exprime une intention affectueuse*) et donc au niveau *n* de langage.
Les *entre-crochets* 2.b, qui entraînent une relation entre deux multisignes (*L'article défini qui accompagne «animal[1]» exprime une valeur emphatique-démonstrative*), renvoient au niveau *n*+1 de langage ou premier niveau de métalangage.

Conclusion

Notre analyse des *entre-crochets* contenus dans les définitions de substantifs du *T.L.F*, a conduit, à l'issue des études sur le métalangage de (et notamment [Rey-Debove, 1997], à une classification à partir des phrases métalinguistiques pour parvenir aux différents niveaux de langage et de métalangage impliqués par l'article de dictionnaire.

L'étude des *entre-crochets* nous a permis, en outre, de mener une analyse sur les informations métalinguistiques, d'en saisir la nature et de proposer la séparation entre typologie et métalangage.

L'intérêt du *T.L.F.* réside non seulement dans l'élargissement considérable de la gamme d'informations proposées par [Putnam, 1975a et 1975b], mais également dans une intuition de la nécessité de séparer la définition proprement dite des différents marqueurs.

Bibliographie

A) *TRESOR DE LA LANGUE FRANÇAISE*

[C.N.R.S., 1971-1994] C.N.R.S. (1971-1994): *Trésor de la langue française. Dictionnaire de la langue du XIX^e^ et du XX^e^ siècle (1789-1960)*, 16 voll., Gallimard, Paris.
C.N.R.S. - A.T.I.L.F.: *Le Trésor de la Langue Française informatisé* <http://atilf.atilf.fr/tlf.htm>.

[Imbs, 1971] Imbs, P. (1971): *Préface* in C.N.R.S. (1971-1994): *Trésor de la langue française. Dictionnaire de la langue du XIX^e^ et du XX^e^ siècle (1789-1960)*, Tome Premier, Gallimard, Paris, pp. IX-XLVII.

B) DOCUMENTS INEDITS

«Normes de rédaction» [4.10.1972].
«Pour un nouveau cahier de normes. Document à discuter les 22, 23 [et 24] février 1979 lors de la "réunion des experts"» [1979].
«Pour un nouveau cahier de normes. Compte rendu: de la réunion des experts tenue les 22 et 23 février 1979; de la réunion des réviseurs et relecteurs tenue le 24 février 1979» [1979].
«Cahier de normes» [s.d.] [1979].

C) ÉTUDES

[Gorcy, 1990] Gorcy, G. (1990): *La polysémie verbale ou le traitement de la polysémie de sens; discussion à partir des normes rédactionnelles du* Trésor de la Langue Française (TLF), «Cahiers de lexicologie» LVI, pp. 109-122.
[Hathout, 1996] Hathout, N. (1996): *Pour la construction d'une base de connaissances lexicologiques à partir du* Trésor de la Langue Française*. Les marqueurs superficiels dans les définitions spécialisées*, «Cahiers de lexicologie» LXVIII, pp. 137-173.
[Lamy, 1980] Lamy, N.M. (1980): *Le dictionnaire et le métalangage*, «Cahiers de lexicologie» XXXVI, pp. 95-110.
[Martin, 1983] Martin, R. (1983): *Pour une logique du sens*, PUF, Paris.
[Pottier, 1965] Pottier, B. (1965): *La définition sémantique dans les dictionnaires*, «Travaux de Linguistique et de Littérature » III, 1, pp. 33-40.
[Putnam, 1975a] Putnam, H. (1975a): *Is Semantics Possible?* in *Mind, Language and Reality. Philosophical Papers II*, Cambridge University Press, Cambridge, pp. 132-152.
[Putnam, 1975b] Putnam, H. (1975b): *The Meaning of "Meaning"* in *Mind, Language and Reality. Philosophical Papers II*, Cambridge University Press, Cambridge, pp. 215-271.
[Radermacher, 2004] Radermacher, R. (2004): «Le Trésor de la Langue Française. Une étude historique et lexicographique», Thèse de doctorat, Strasbourg, Université Marc Bloch.
[Rey-Debove, 1966], Rey-Debove, J. (1966): *La définition lexicographique. Recherches sur l'équation sémique*, «Cahiers de lexicologie» 9, pp. 71-94.
[Rey-Debove, 1967], Rey-Debove, J. (1967): *Autonymie et métalangue*, «Cahiers de Lexicologie» 11-II, pp. 15-27.
[Rey-Debove, 1967], Rey-Debove, J. (1967): *La définition lexicographique; bases d'une typologie formelle*, «Travaux de Linguistique et de Littérature» V, 1, pp. 141-159.
[Rey-Debove, 1969], Rey-Debove, J. (1969): *Les relations entre le signe et la chose dans le discours métalinguistique:* être, s'appeler, désigner, signifier *et* se dire, «Travaux de Linguistique et de Littérature» VII, 1, pp. 113-129.
[Rey-Debove, 1971], Rey-Debove, J. (1971): *Étude linguistique et sémiotique des dictionnaires français contemporains*, Mouton, Paris-La Haye.
[Rey-Debove, 1967], Rey-Debove, J. (1997): *Le Métalangage. Étude linguistique du discours sur le langage*, Armand Colin, Paris.
[Rey-Debove, 1998], Rey-Debove, J. (1998): *La linguistique du signe*, Armand Colin, Paris.

L'intégration de l'information prosodique en lexicographie : nouveaux supports, formats de présentation et techniques de discrimination

Mélanie Petit (1)
melanie.petit@univ-orleans.fr

(1) U.F.R. Lettres, Langues et Sciences Humaines de l'Université d'Orléans

---

Introduction

L'évolution des supports, au niveau du traitement et de l'obtention des données permet de poser la question de l'intégration de l'information prosodique dans le traitement lexicographique. Les formats sous forme imprimée et les traitements de corpus écrits ne permettant pas ou presque de prendre en compte ce type d'information, la question se pose aujourd'hui de savoir s'il est envisageable et profitable de prendre en compte la prosodie au niveau lexical et de faire figurer pour chacune des entrées des éléments informatifs particuliers ayant trait à la prononciation.

1. Discrimination prosodique et représentation lexicographique

La question de décrire ou de rendre compte de la polysémie ou de la polyfonctionnalité des emplois des signes linguistiques est l'une des principales difficultés à laquelle sont confrontés les lexicographes et les lexicologues/sémanticiens. Nous aborderons ici la question de savoir le rôle que peut jouer l'existence d'une discrimination prosodique des emplois à un niveau global puis à un niveau local en traitant plus spécifiquement du cas des connecteurs pragmatiques.

1.1. Evolution des supports lexicographiques

Toute représentation lexicographique est contrainte par la nature et le format de son support. Qualitativement (type d'information) et quantitativement (taille de la description, nombre d'emplois décrits). Or de nouveaux supports permettent aujourd'hui d'envisager une très forte évolution. Les documents maintenant accessibles sur des supports multimédia présentent l'intérêt, contrairement aux ouvrages se présentant sous une forme imprimée, de pouvoir intégrer des fichiers sonores. Le support informatique, pour les dictionnaires électroniques par exemple, offre à l'utilisateur la possibilité d'entendre la réalisation d'une unité lexicale, la difficulté de l'ajout d'un tel élément réside alors dans la mise à jour de liens entre une interprétation et une forme phonologique.
A plus grande échelle, les nouveaux traitements automatiques de grands corpus oraux permettraient d'avoir accès à de nouvelles informations concernant l'intuition des locuteurs.

1.2. Discrimination prosodique : le cas des connecteurs

Très étudiés depuis 30 ans par exemple, y compris du point de vue de leur traitement lexicographique (Dostie, 2004) comme ce sera le cas ici, les connecteurs pragmatiques (ou discourse markers) apparaissent aujourd'hui massivement polyfonctionnels (Fischer, 2006), sorte de paradoxe pour des unités dont la fonction semble être de « guider l'interprétation du discours ». Or pour assurer la désambiguïsation de ces unités, la question de la discrimination phonologique (i.e. prosodique) des emplois est aujourd'hui à l'ordre du jour, aussi bien au niveau descriptif qu'explicatif, dès lors que si discrimination prosodique des emplois il y a, l'ignorer revient pour le linguiste à perdre une information pertinente et présente dès le départ, dont les effets ensuite, rendus imprédictibles par l'effacement de cette information, ne pourront être expliqués que par des moyens artificiels ou des ersatz. Situation qui a conduit récemment à se demander s'il était possible d'établir que l'interprétation d'un terme, présentant théoriquement plusieurs emplois différents, soit très souvent résolue à l'oral par la capacité de l'auditeur à assigner à des contours prosodiques ou à des variations de paramètres acoustiques, un sens particulier parmi tous les choix à sa disposition.

1.2.1. Un exemple

La question se pose par exemple pour des sens opposés d'une même unité. Prenons deux emplois de *enfin*, l'un exprimant le soulagement, l'autre exprimant l'irritation. Et la différence prosodique qui existe entre eux. Comment pouvons-nous intégrer cette discrimination à la représentation lexicographique ? La distinction affective est effectivement représentée dans le TLF, toutefois, il est clair que l'intégration d'une information prosodique serait susceptible de différencier plus nettement ces deux emplois. Elle est indispensable s'il s'avère que les auditeurs, intuitivement, procèdent à cette distinction, et il semble manifeste qu'ils la fassent spontanément dès la première audition d'un discours.

Par ailleurs, dans une optique bilingue, la compréhension d'un article serait rendue plus évidente pour un étranger si la caractéristique d'irritation par exemple apparaissait.

1.2.2. État de la recherche

Certains travaux ayant pour objet la description des connecteurs intègrent déjà la dimension prosodique. C. Chanet [Chanet 2005] s'intéresse par exemple à l'étude des liens entre les réalisations prosodiques de *enfin* et ses rôles dans le discours. D'autre part, elle met en évidence le fait que la prosodie avec laquelle le connecteur *voilà* est réalisé donne des indications sur la structuration engendrée par celui-ci dans le discours.

G. Dostie [Dostie 2004] propose quant à elle un traitement des marqueurs discursifs en envisageant entre autres leurs relations avec la sémantique et la lexicographie. Elle souligne l'intérêt de prendre en compte la prosodie pour repérer les divers sens d'un marqueur discursif, et opte pour une analyse perceptuelle, plutôt que pour une analyse acoustique.

1.3. Problématique

Plus généralement, la question qui se pose est de savoir si nous devons considérer les entrées comme de simples unités graphématiques, moyennant quoi la prosodie est écartée d'emblée comme non pertinente pour définir les entrées. Ou au contraire, s'il faut opter pour une conception qui n'élimine pas a priori la dimension sonore et qui propose de mener en amont une discrimination phonologique, auquel cas la question se pose alors de redéfinir ce qu'est une unité lexicale, et même dans une certaine mesure toute l'organisation lexicographique.

2. Deuxième partie : entre polysémie et homonymie

La prise en compte de la dimension sonore en lexicographie suppose une reconsidération de l'organisation globale de l'entrée, ainsi que de l'élaboration de techniques de discrimination et de formats de présentation.

2.1. Architecture

La hiérarchisation lexicographique suppose d'une part de classer les unités entre elles et de les organiser par rapport à des concepts de monosémie, d'homonymie et de polysémie, et d'autre part de définir quelles seront les informations linguistiques à intégrer dans les articles. En se posant la question d'ajouter une dimension phonologique, comment cela fait-il évoluer la conception lexicographique ? Cette nouvelle dimension change-t-elle la notion d'homonymie et de polysémie ? En refondant l'organisation interne d'une entrée lexicale, c'est-à-dire en procédant à des regroupements et distinctions de sens d'une même unité sur des considérations prosodiques, nous sommes alors en mesure de proposer un objet d'un nouveau type, intermédiaire entre la polysémie et l'homonymie. A partir du moment où il existe plusieurs formes prosodiques d'une même unité, comment choisissons-nous de les traiter?
Une fois établie, la capacité de discrimination montrerait que nous avons bien affaire à des formes prosodiques différentes mais que globalement ce ne sont pas des unités complètement distinctes et qu'il existe bien un lien entre elles (entre la polysémie et l'homonymie). Il serait naïf de nier qu'il existe effectivement de vraies ambiguïtés, non repérables prosodiquement et qui nécessitent simplement un calcul sémantico-pragmatique, mais cela ne doit pas être pour autant considéré comme un obstacle au traitement des autres cas pour lesquels il serait envisageable de s'appuyer sur la discrimination. S'il est possible de mettre en évidence que des variations de formes entraînent des variations de sens, alors il n'y a plus de polysémie mais bien des paires forme/sens qui entretiennent entre elles un autre lien sémantique qui exclut également l'homonymie. Il n'est toutefois pas exclut qu'il puisse exister un autre type de polysémie ou d'homonymie de nature prosodique. Cela signifierait qu'une même forme prosodique correspondrait à plusieurs emplois ou bien qu'à une forme prosodique particulière ne serait associée qu'une seule interprétation possible. La question se poserait alors de définir quelle serait la manière la plus pertinente de décliner les sous-entrées sur cette base.

2.2. Techniques de discrimination

La question qui se pose ensuite est de savoir comment va procéder le lexicographe, si c'est à lui qu'il revient d'effectuer la tâche de discrimination. Est-ce qu'il part de la sémantique en phonologisant les interprétations définies a priori comme a choisi de le faire G. Dostie ou bien de la forme sonore directement? Le choix du signifié ou du signifiant comme point de départ dans la mise en œuvre de la discrimination prosodique est fondamental. Afin de pouvoir mener à bien cette étude, il est nécessaire d'avoir accès aux données sous forme orale et de pouvoir accéder à de grands corpus oraux, ce qui est maintenant possible grâce aux nouveaux supports.
D'autre part, la difficulté du choix et du repérage des emplois va rapidement se poser. Qu'est-ce qui méritera d'être repéré et comment repérerons-nous? Il sera nécessaire de trouver un équilibre entre la prise en considération de critères qualitatifs, qui permettraient de signaler les emplois spéciaux, et de critères quantitatifs. Que choisissons-nous de faire si un cas est rare mais très marqué ? Est-ce que nous ne classons que les formes typées ou bien également les variations aléatoires? Comment délimitons-nous le processus de discrimination ? Si une forme sonore n'apparaît jamais avec un sens précis, il y aurait incompatibilité mais la forme ne marquerait rien de spécifique. Si l'information prosodique se révèle être est trop variable, la solution serait alors d'opter pour la sémantique comme point de départ et rejoindre ainsi le point de vue de G. Dostie en fournissant toutefois des descriptions prosodiques plus précises.
A un niveau plus global, comment serait-il possible d'intégrer la technique de discrimination dans une formation en lexicographie ? L'apprentissage par la mise en œuvre de réseaux de neurones est une possibilité envisageable. Il resterait à définir les éléments indispensables à fournir au système pour que ce dernier soit efficace.

2.3. Formats de présentation

Une fois la discrimination prosodique menée à bien, se pose alors la question du format de présentation de l'entrée. Le statut hiérarchique d'un sous-emploi sera en partie établi en fonction des résultats de la discrimination obtenus mais pas uniquement. Il est possible de conserver une seule entrée nécessaire pour la logique graphématique de consultation et une sous-distinction sur base prosodique. Il sera nécessaire de définir un métalangage approprié (littéraire, acoustique…) afin de caractériser la prosodie. Si toutefois une discrimination prosodique est réalisable, la question de la caractérisation prosodique à fournir devra être réfléchie en fonction des variantes qui seront observées pour un même emploi.
La limite des formats tout comme la taille de la description auront également une influence sur la constitution de l'entrée.

2.4. Le traitement des connecteurs

Bien que la question de savoir si la discrimination est spécifique à certains types de signes est importante, nous avons choisi de traiter plus particulièrement des connecteurs. En effet ces objets étant fortement polysémiques, les emplois variés entraînent une discrimination phonologique importante. Nous avons choisi d'illustrer cette hypothèse sur l'étude particulière du connecteur *enfin*. Dans le TLF, l'emploi de *enfin* synonyme de *voyons* n'est pas présent, ou peut-être est-il sous-entendu dans l'un des autres emplois exprimés, tels que la colère. Une analyse prosodique permettrait de mettre à jour l'existence ou non d'un lien entre ces emplois, de par la similarité de certains paramètres acoustiques. Par ailleurs la distinction proposée pourrait être validée ou infirmée et certains sous-emplois s'avèreraient en fait n'être que des variantes d'un emploi-type, par exemple la reformulation. II B1 *Pour marquer la fin d'une longue attente ou recherche* ne serait-il pas synonyme de I B *Le procès se déroule après un long espace de temps* ? Il existe des entrées très polyfonctionnelles pour lesquelles une description exhaustive serait beaucoup trop importante. C'est par exemple le cas du *mais* explicatif tel que *il a raté son examen mais il était malade* (Nemo 2007). La perspective d'intégrer une information prosodique en lexicographie permettrait de regrouper des emplois sur des bases repérables. Ainsi, la prosodie permettrait de limiter les descriptions en regroupant des sous-entrées en fonction de ce que partagent les sous-emplois.

Conclusion

Conscients du fait que l'analyse d'autres types d'items est possible afin de tenter de mettre à jour une discrimination prosodique pertinente, nous nous proposons toutefois de prendre comme objet étude le cas particulier des connecteurs afin de repenser l'organisation interne d'une entrée lexicographique.

Bibliographie

Béjoint H., Thoiron P. (sous la direction de) (1996), *Les dictionnaires bilingues*, Louvain-la-Neuve, Duculot

Bertrand R., Chanet C. (2005), *Fonctions pragmatiques et prosodie de enfin en français spontané*, *RSP* n°17, p41-68

Corréard M.H. (2002), *Lexicography and Natural Language Processing. A festschrift in honour of B.T.S.*, Atkins, Corréard (ed.), Euralex

Dostie G. (2004), *Pragmaticalisation et marqueurs discursifs, analyse sémantique et traitement lexicographique*, Duculot, Bruxelles

Fischer K. (2006), *Approaches to Discourse Particles*, K. Fischer (ed.), Amsterdam, Benjamins

Nemo F. (2007), *Reconsidering the Discourse Marking Hypothesis. Even, even though, even if, etc. as morpheme/construction pairs* in *Connective as discourse landmark*, Celle, Benjamins

DE LA DESCRIPTION LINGUISTIQUE A LA DESCRIPTION LEXICOGRAPHIQUE : L'EXEMPLE DES ADVERBIAUX DE PHRASE DU TYPE *EN* + LEXEME

Corinne Féron (1)
Corinneferon1@aol.com
Danielle Coltier (1)
coltierdanielle@yahoo.fr

(1) Université du Maine, Le Mans

Comment les rédacteurs d'articles de dictionnaires peuvent-ils tirer parti des travaux consacrés à la description syntaxique, sémantique, énonciative ou pragmatique d'un lexème ou d'une classe de lexèmes ? Telle est la question que nous voudrions examiner aussi concrètement que possible.

Question légitime : d'une part, parce que les linguistes sont parfois critiques à l'égard des descriptions proposées dans les articles de dictionnaires ; d'autre part, parce que certains d'entre eux proposent des analyses susceptibles d'améliorer les descriptions lexicographiques (l'article de Danjou-Flaux 1982 sur *réellement* et *en réalité* fait des suggestions aux lexicographes) ; enfin, parce que le transfert des résultats d'études linguistiques vers le dictionnaire pose de nombreux problèmes, de métalangue notamment (*cf*. Corbin 2002 : 29-35).

Nous examinerons cette question en nous concentrant sur le traitement lexicographique d'adverbiaux de phrase[1].

Notre contribution comprendra trois grandes parties :

1. un répertoire des difficultés que présente le recours aux études linguistiques sur les adverbiaux ;
2. une observation de descriptions d'adverbiaux de phrase proposées par le *Trésor de la langue française* ; nous nous arrêterons aux adverbiaux présentant la structure *en* + lexème ;
3. des essais de traitement lexicographique pour quelques adverbiaux de ce type : *en bref, en clair, en conclusion, en confidence, en gros, en résumé* ; pour ce faire, nous nous appuierons sur des études linguistiques : on pourra ainsi « pointer » quelques problèmes concrets tenant compte d'une part des exigences rédactionnelles d'un dictionnaire tel que le *TLF*, et d'autre part des contraintes auxquelles un dictionnaire informatisé doit obéir.

[1] « Adverbe » renvoie (ordinairement) à une classe de lexèmes, un lexème donné étant susceptible d'avoir deux types de fonctionnement ou plus ; nous préférons parler d'« adverbiaux de phrase » (« complément » étant sous-entendu) plutôt que d'« adverbes de phrase », puisqu'il s'agit de désigner un type de fonctionnement et non une classe de lexèmes.

1. REPERTOIRE DES DIFFICULTES

Les compléments adverbiaux ont suscité depuis les années soixante-dix nombre de travaux dont l'objectif est de proposer des classifications (par ex. Mørdrup 1976, Schlyter 1977, Ducrot 1980, Mélis 1983, Guimier 1996, Molinier et Lévrier 2000[2]). Les « adverbes de phrase » ont tout particulièrement été étudiés, donnant lieu à des essais de typologie (Martin 1974, Nølke 1993[3]), à l'étude de sous-classes (Borillo 1976, Sueur 1978, Rossari 1994, parmi d'autres) ou à des analyses portant sur une ou plusieurs expressions.

Liées au nombre de travaux et à leurs présupposés théoriques, les divergences dans l'analyse des adverbiaux sont nombreuses : divergences dans les critères de classement – avec pour corrélat, des divergences dans la terminologie et dans le classement même de certains compléments adverbiaux –, divergences aussi dans la description du fonctionnement de ces expressions.

1.1 Les critères de classement

Les études fondées exclusivement ou principalement sur des critères syntaxiques distinguent deux grandes classes d'adverbiaux, les adverbiaux de phrase et les adverbiaux de constituant (par ex. Martin 1974, Mørdrup 1976, Molinier et Lévrier 2000) ; en revanche, la notion d'adverbial de phrase n'est plus pertinente dans une classification privilégiant le sens (Guimier 1996 : 5) : s'opposent alors les adverbiaux contribuant à la création du sens référentiel et les adverbiaux véhiculant l'attitude du locuteur, qu'ils affectent la phrase entière ou non (adverbiaux contextuels chez Nølke 1993, exophrastiques chez Guimier 1996).

Des correspondances « terminologiques » sont possibles. Globalement, les fonctionnements regroupés sous l'étiquette « adverbe de phrase » chez Molinier et Lévrier 2000 le sont sous l'étiquette « adverbial contextuel » chez Nølke 1993, « exophrastique » chez Guimier 1996. Mais les classes ne coïncident pas exactement en extension : ainsi *seulement* et *notamment* appartiennent à la même grande classe que *peut-être* ou *pourtant* chez Nølke 1993 et Guimier 1996 (ce sont, pour l'un des adverbiaux contextuels, pour l'autre des exophrastiques), tandis que Molinier et Lévrier 2000 considèrent que ces adverbiaux sont intégrés à la proposition et qu'ils n'appartiennent donc pas à la même classe que *peut-être* et *pourtant*.

Les distinctions introduites dans chacune des grandes classes donnent lieu de même à des terminologies variables et à des sous-classes qui ne se recouvrent pas toujours exactement.

1.2 Description du fonctionnement des adverbiaux

Une autre difficulté concerne la description même de ces expressions, et tient à leur caractère

[2] Les uns disent clairement qu'ils classent des fonctionnements adverbiaux, le classement pouvant se limiter aux fonctionnement des adverbes-mots (Guimier 1996) ou inclure tout type d'expressions : adverbes, locutions adverbiales ou syntagmes libres (Mélis 1983) ; d'autres, tels Mørdrup 1976 ou Molinier et Lévrier 2000, parlent de classements d'adverbes mais, comme ils considèrent qu'un même adverbe-mot peut avoir plusieurs fonctionnements, ce sont en fait ces fonctionnements qu'ils étudient. En ce qui concerne Ducrot 1980, il propose un classement des occurrences d'adverbes (donc des fonctionnements) puis un classement des adverbes-mots.

[3] Si Nølke 1993 ne retient pas le terme d'« adverbe de phrase » dans sa terminologie (il propose *adverbial contextuel*), il travaille pour l'essentiel sur des compléments adverbiaux classés par d'autres comme « adverbes de phrase ».

non référentiel[4] : « les intuitions linguistiques se rapportant à ce genre de segments ne sont pas données », l'occurrence de ces expressions « déclenche chez le sujet des intuitions qui sont souvent difficiles à expliciter, à formuler » (Danjou-Flaux 1980a : 138-139) ; à quoi s'ajoute qu'elles sont parfois polyfonctionnelles. Conséquence de cela : ces intuitions peuvent donner lieu à des constructions fort diverses selon les cadres théoriques adoptés par les linguistes. De plus, la description varie dans une certaine mesure selon le type de données examinées (énoncés construits ou données de corpus, et dans ce dernier cas, corpus écrit ou oral, littéraire ou journalistique, etc. *Cf.* Blumenthal 1996 à propos de la description de *en fait*).

2. OBSERVATION DE DESCRIPTIONS D'ADVERBIAUX DE PHRASE PROPOSEES PAR LE *TLF*

Dans cette partie, nous travaillerons plus précisément sur un ensemble morphologiquement cohérent d'expressions fonctionnant comme adverbiaux de phrases : les expressions formées sur le modèle *EN + X*[5] (*en bref*, *en conclusion, en confidence, en effet*, *en fait*, *en gros*, *en outre*, *en particulier*, *en pratique*, *en principe*, *en réalité*, *en résumé, en revanche, en théorie, en vérité*, etc.).

Pour certains de ces adverbiaux, le *TLF* ne signale que l'emploi « intégré » (adverbial de constituant) ; c'est le cas, par exemple, pour *en conséquence* (« D'une manière logiquement conforme à (telle chose) »), *en clair* et *en substance*. Nous n'avons retenu pour ce qui suit que les descriptions correspondant à des emplois comme adverbiaux de phrase.

Or, les requêtes effectuées dans le *Trésor de la langue française informatisé* sur le site de l'ATILF montrent que les pratiques des rédacteurs concernant ces expressions sont variables – ces différences de traitement n'étant pas toutes, semble-t-il, imputables à la durée de la rédaction du dictionnaire : l'évolution des pratiques rédactionnelles (Martin 1994) et les apports des travaux sur les compléments adverbiaux ne rendent pas compte de toutes les différences relevées (des expressions traitées dans un même volume donnent lieu à des types de description différents).

On donnera ici deux exemples de cette hétérogénéité dans le traitement des adverbiaux de phrase.

2.1 Description du fonctionnement des adverbiaux du type *EN + X*.

Les divergences touchent la précision et la forme des analyses.

- Précision de l'analyse : si la diversité des fonctionnements de *en réalité* est l'objet d'une description précise, inspirée de Danjou-Flaux 1982, pour *en revanche* le *TLF* n'offre qu'une définition synonymique et ne semble donc pas avoir tiré profit de l'étude du même

[4] « Expressions non référentielles » correspond ici à ce que l'on désigne ailleurs par « lexies non descriptives » (Mel'čuk et Iordanskaja 1999) ou encore par « unités non dénominatives » (chez Corbin 2002 : 29, qui nomme ainsi les « mots « grammaticaux », auxiliaires, éléments de formation de mots construits... ») ; cela renvoie aussi aux expressions dont le sens obéit au « modèle instructionnel » (VS « modèle descriptif », Kleiber 1997 : 32-33).
[5] On exclut les expressions comportant plus d'un lexème après *en* : *en toute franchise, en un mot, en fin de compte*...

auteur (Danjou-Flaux 1980b), pourtant citée dans la bibliographie en fin d'article.

- Forme de la description sémantico-fonctionnelle : les rédacteurs ont souvent recours à des définitions synonymiques, éventuellement accompagnées de synonymes ; les définitions métalinguistiques sont rares (deux seulement, pour deux acceptions de *en effet*) ; en revanche, on trouve fréquemment des commentaires métalinguistiques entre crochets (cette hésitation entre définition métalinguistique et commentaire entre crochets n'étant pas exceptionnelle dans le *TLF* ; *cf*. Henry 1996 : 102).

2.2 Métalangue

Une première remarque concerne les indicateurs grammaticaux. L'une des expressions retenues (*en principe*) est insérée comme unité de traitement interne sans statut spécifié. Toutes les autres sont signalées comme « locutions adverbiales », mais cette information peut être fournie dans des champs différents : elle apparait soit avant la séquence, dans la zone des indicateurs grammaticaux (*en effet*, *en plus*, etc.) ou entre crochets (pour *en outre*), soit après la séquence (pour *en réalité* et *en particulier*). Les deux éléments, « loc. » et « adv. » peuvent aussi être disjoints (c'est le cas pour *en conclusion*). De ce fait, en cherchant dans le *TLFI* les locutions adverbiales par le biais de l'objet textuel « indicateur », on obtient une liste qui ne comprend pas toutes les locutions pourtant recensées dans les articles : *en conclusion*, *en outre, en particulier* et *en réalité* n'y figurent pas.

L'hétérogénéité est plus grande encore du côté de l'indication du fonctionnement de l'adverbial. D'une part, cette indication n'est pas systématique ; d'autre part, lorsque les rédacteurs caractérisent le fonctionnement des expressions, ils recourent à des terminologies variables ; trois notions sont ainsi utilisées : « locution adverbiale de phrase », « adverbe d'énonciation », « locution adverbiale de liaison ».

- L'indicateur « adverbe / locution adverbiale de phrase » est utilisé dès le premier volume du *TLF* (par ex. pour les locutions *à coup sûr*, *à propos,* citées dans l'article *à* ; « adv. de phrase » est en outre donné comme code grammatical pour *aussi*[2]) mais de façon apparemment aléatoire : utilisé pour *au contraire* (« loc. adv. de phrase »), il ne l'est pas pour *en revanche*. Pour les adverbiaux du type *EN + X*, cet indicateur n'apparait qu'à propos de *en vérité* (dans un commentaire entre crochets : « Fonctionne comme adv. de phrase »).

- La première occurrence de l'indicateur « adverbe d'énonciation » figure dans le volume 12 (1986) du *TLF* mais il n'est pas utilisé systématiquement, dans ce volume et dans les suivants, pour les expressions qui *a priori* pourraient recevoir cette étiquette : on la trouve pour *en réalité* (dans le champ des indicateurs), mais non pour *en vérité*.

- « Adverbe de liaison » apparait dans le volume 5 (à propos de *cependant*) ; « locution adverbiale de liaison » caractérise quelques-uns des adverbiaux retenus (*en définitive, en effet, en outre, en particulier, en réalité, en revanche, en somme*) cités dans l'article *en.*

3. ESSAIS DE TRAITEMENT LEXICOGRAPHIQUE DE QUELQUES ADVERBIAUX DE PHRASE

Nous proposerons des descriptions lexicographiques de quelques adverbiaux de phrase du type *EN + X* : *en clair*, *en gros, en confidence, en bref, en conclusion, en résumé* – les trois premiers étant classés par Molinier et Lévrier 2000 parmi les disjonctifs de style, les trois autres s'apparentant à la fois aux disjonctifs de style et aux conjonctifs (*cf*. Molinier et Lévrier 2000 : 66 et Mélis 1983 : 155-156).

Au delà de la description de tel adverbial particulier, la question qui se pose est de savoir quelles catégories de données devraient systématiquement figurer dans la description lexicographique de telles expressions. Parmi les informations qui semblent indispensables, nous pouvons citer :

- la position ou les positions de l'expression dans la phrase (entre crochets) ;
- les cotextes dans lesquels elle apparait : il serait souhaitable de spécifier les cotextes qui éliminent les ambigüités ou au moins conduisent à privilégier une interprétation comme adverbial de phrase (les expressions en question pouvant aussi, par ailleurs, fonctionner comme adverbiaux de constituant) ; de telles précisions s'inscrivent dans les réaménagements nécessaires pour faire du dictionnaire informatisé un « dictionnaire automatisé » (Martin 2001 : 68) ;
- la description du rôle de l'adverbial (dans une définition métalinguistique ou entre crochets), incluant, pour les « conjonctifs », la description du rapport qui est établi, par le biais de l'adverbial, avec le cotexte gauche ;
- la caractérisation du type d'adverbial auquel on a affaire.

Ce dernier point parait être le plus délicat. Une indication de ce type exige en effet des choix : choix entre des terminologies diverses et parfois opaques ; choix également du niveau de précision visé : l'article peut indiquer seulement à quelle « grande classe » appartient l'adverbial (« adverbial de phrase » par ex., si l'on décide de se référer à une classification syntaxique) ou mentionner la sous-classe (correspondant au niveau de la distinction « conjonctif », « disjonctif de style » chez Molinier et Lévrier 2000 par exemple).

Des informations de ce type auraient leur place dans un dictionnaire comme le *TLF*, qui cherche à décrire le fonctionnement de la langue de façon systématique : elles permettraient de mettre en évidence l'existence d'une classe d'expressions fréquemment utilisées et qui ont un rôle essentiel, puisque, grâce à elles, le locuteur commente son énonciation ou le produit de celle-ci. Toutefois, l'insertion d'une caractérisation systématique des adverbiaux nécessiterait aussi une mise à jour des articles concernant les notions retenues dans la métalangue : en vertu de la nécessaire « clôture » du dictionnaire, l'article « adverbe » par exemple devrait inclure la définition des différents types d'adverbiaux (la rubrique « commentaire grammatical » de cet article devrait être modifiée) ; par ailleurs, il faudrait déterminer dans quel « champ », dans quel « objet textuel » du dictionnaire informatisé cette caractérisation pourrait trouver sa place.

Une révision du *TLF* – ou la confection d'un dictionnaire des adverbes qui en serait extrait – devrait tirer profit de l'ensemble des travaux publiés, tant pour la description du fonctionnement des expressions que pour la métalangue ; l'utilisation de ces travaux théoriques pour la rédaction des articles de dictionnaires pose donc des problèmes nombreux dont nous voudrions voir si et, le cas échéant, comment, ils sont surmontables.

BIBLIOGRAPHIE

BLUMENTHAL, P. (1996), Le connecteur *en fait*. *In* : Muller, C. éd., *Dépendance et intégration syntaxique*, Tübingen : Niemeyer, 257-269.

BORILLO, A. (1976), Les Adverbes et la modalisation de l'assertion, *Langages*, 43, 74-89.

CORBIN, P. (2002), Lexicographie et linguistique : une articulation difficile. L'exemple du domaine français. *In* : Melka, F. et Augusto, M. C. éds., *De la lexicologie à la lexicographie / From lexicology to lexicography*, Utrecht : University of Utrecht, 9-38.
En ligne : http://www-uilots.let.uu.nl/research/Publications/Melka_Augusto.pdf

DANJOU-FLAUX, N. (1980a), A propos de *de fait, en fait, en effet* et *effectivement, Le Français moderne,* 48, 110-139.

DANJOU-FLAUX, N. (1980b), *Au contraire*, *par contre*, *en revanche* : une évaluation de la synonymie, *Bulletin du centre d'analyse du discours*, 4, 123-146.

DANJOU-FLAUX, N. (1982), *Réellement* et *en réalité* : données lexicographiques et description sémantique. *In* : Gary-Prieur, M.-N. et al. éds., *Adverbes en* -ment*, manière, discours,* Lille : Presses universitaires de Lille, 104-150.

DUCROT, O. (1980), Analyses pragmatiques, *Communications*, 11-60.

GUIMIER, C. (1996), *Les Adverbes du français : le cas des adverbes en* -ment, Gap/Paris : Ophrys.

HENRY, F. (1996), Pour une informatisation du *TLF*. *In* : Piotrowski, D. éd., *Lexicographie et informatique : autour de l'informatisation du Trésor de la Langue Française*, Paris : Didier érudition, 79-139.

KLEIBER, G. (1997), Sens, référence et existence : que faire de l'extra-linguistique ? *Langages*, 127 : 9-37.

MARTIN, R. (1974), La Notion d'adverbe de phrase : essai d'interprétation en grammaire générative. *In :* Rohrer, C. et Ruwet, N. éds., *Actes du Colloque franco-allemand de grammaire transformationnelle, II*, Tübingen : Max Niemeyer Verlag, 66-75.

MARTIN, R. (1994), Présentation, *Le Français moderne*, LXII-2, 129-134

MARTIN, R. (2001), *Sémantique et automate*, Paris : PUF.

MEL'ČUK, I. et IORDANSKAJA, L. (1999), Traitement lexicographique de deux connecteurs textuels du français : *en fait* VS *en réalité*. *In* : *Dictionnaire explicatif et combinatoire du français contemporain*, IV, Montréal : Presses de l'Université de Montréal, 28-41.

MELIS, L. (1983), *Les circonstants et la phrase*, Louvain : Presses universitaires de Louvain.

MOLINIER, C. et LEVRIER, F. (2000), *Grammaire des adverbes : description des formes en* -ment, Genève : Droz.

MØRDRUP, O. (1976), Sur la classification des adverbes en *-ment, Revue romane*, XI.

NØLKE, H. (1993), *Le Regard du locuteur*, Paris : Kimé.

ROSSARI, C. (1994), *Les opérations de reformulation*, Berne : Peter Lang.

SCHLYTER, S. (1977), *La place des adverbes en* –ment *en français*, Constance : thèse.

SUEUR, J.-P. (1978), Adverbes de modalité et verbes modaux épistémiques, *Recherches linguistiques*, Université de Paris-Vincennes, n° 5-6, 235-279.

UN DICTIONNAIRE DES VERBES

Aude GREZKA (1)
Aude.grezka@lli.univ-paris13.fr
Françoise MARTIN-BERTHET (1)
Martin-berthet@wanadoo.fr

(1) Université Paris 13

Ce travail s'insère dans un projet global visant à réaliser des dictionnaires électroniques destinés aux divers systèmes qui opèrent sur des données linguistiques. L'ensemble de ces dictionnaires (portant sur les prédicats adjectivaux, nominaux et verbaux) doit fournir une couverture exhaustive du français, entre autres langues. Il s'agit de décrire le lexique avec des propriétés syntactico-sémantiques standardisées suffisamment explicites pour faire l'objet de procédures informatisées.

Les entrées à traiter sont des emplois verbaux, c'est-à-dire des phrases élémentaires ; on établit les *schémas d'arguments* propres à chaque emploi, en « typant » sémantiquement les noms d'arguments par leurs *classes d'objets* (*Langages* 131, 1998).

Ces entrées sont regroupées en classes sémantiques, offrant une structuration du lexique proche de l'intuition et facilitant par là l'accès à l'information lexicale : par exemple, les verbes de création ; les verbes de parole ; les verbes d'identité ; les verbes de perception ; les verbes de mouvement ; les verbes de sentiment ; les verbes de coups ; etc. On doit chercher le niveau optimum d'homogénéité sémantique et syntaxique.

Sur la base de travaux antérieurs (Levin 1993 ; Dubois et Dubois-Charlier 1997 ; etc.), plusieurs grandes classes ont d'abord été définies, telles que : états élémentaires et modalités ; cognition ; langage et communication ; relations humaines, etc. Chacun de ces groupes se trouve lui-même subdivisé en rubriques plus fines (par exemple, pour langage et communication : parole, écriture/lecture, signes non verbaux, etc.). Cette typologie a été conçue comme un instrument de travail, un point de départ provisoire, ce qui signifie qu'elle doit être révisée, corrigée et complétée progressivement au fur et à mesure de la description. Les classes définitives ne peuvent pas être des classes a priori, elles émaneront du lexique. La classification n'est pas ontologique mais linguistique.

Le regroupement des verbes dans une même classe s'effectue d'abord en fonction de leur signification : les sens des verbes doivent être suffisamment proches pour être appariés. Une

telle approche, bien que prenant appui sur toutes sortes de travaux en lexicographie, est largement intuitive et impose de recourir à des critères définitoires rigoureux pour valider la constitution de classe, en l'occurrence, les propriétés linguistiques.

On doit donc élaborer une grille susceptible de s'appliquer à l'ensemble des classes, faisant appel à l'ensemble des propriétés linguistiques (Buvet et al. 2005 ; Grezka 2006a, 2006b) : propriétés sémantiques (sens de la classe, sens particuliers), structurelles (schémas de base, restructurations), morphologiques (noms et verbes associés), distributionnelles (adverbiaux appropriés). Les différents descripteurs permettent à la fois de dégrouper les emplois et de les regrouper en ensembles sémantiquement homogènes, sans négliger les particularités (constructions singulières, nuances de sens).

Le niveau de classification choisi et l'importance accordée au sens conduit à indiquer éventuellement plusieurs schémas d'arguments pour une classe. Par exemple, les verbes de don entrent dans deux schémas à trois arguments :

1. N0<hum> V N1<objet> à N2<hum>
   Verbes : donner, offrir, céder, fournir, livrer, procurer, etc.

2. N0<hum> V N1<hum> de N2<objet>
   Verbes : doter, gratifier, munir, pourvoir

La restructuration passive topicalise l'objet dans le premier cas, l'humain dans le second cas.

Le regroupement par classes facilite la mise en relation avec les autres catégories morphologiques (noms et adjectifs), en vue de constituer des classes de prédicats transversales. Ainsi, dans la classe des verbes de coups, on rapprochera non seulement les noms morphologiquement associés (exemple : taper et tape), mais aussi les noms prédicatifs autonomes sans correspondance formelle (boxer et coup de poing, se battre et altercation).

Les classes ainsi conçues incluent à part entière les séquences figées (passer à tabac, casser la figure). Les locutions verbales se prêtent, pour l'essentiel, au même type de description que les verbes simples, à quelques adaptations près.

Le dictionnaire des classes est complété par un index des unités (emplois) : chaque emploi est indexé sur la classe dans laquelle il est traité.

Tous les emplois sont illustrés par au moins un exemple forgé ou cité. Les exemples cités viennent notamment à l'appui d'emplois non répertoriés dans les dictionnaires traditionnels.

Ce dictionnaire des verbes est conçu comme un outil central pour le développement de systèmes qui opèrent sur les données linguistiques, aussi bien dans le domaine de la traduction automatique que dans les domaines de l'extraction automatique de l'information, de la recherche d'information ou de la documentation automatique.

Les différents points abordés ci-dessus seront traités dans l'exposé à travers différents exemples rendent compte de l'état du projet. Dans un premier temps, nous présenterons l'ensemble du projet et les différents objectifs. Dans un deuxième temps, nous développerons les différents outils méthodologiques utilisés pour décrire les verbes et leurs emplois. Enfin, dans un troisième temps, nous aborderons les prolongements de ce travail : d'une part, l'exploitation de cette méthode d'analyse pour développer parallèlement le dictionnaire des noms et des adjectifs ; d'autre part, l'extension de l'approche transversale des prédicats pour réaliser un dictionnaire des racines prédicatives (verbes, noms et adjectifs).

Bibliographie

BUVET, P.-A. ; GIRARDIN, C. ; GROSS, G. ; GROUD, C. (2005), « Les prédicats d'<affect> », Revue de linguistique et de didactique des langues, 32, Université de Grenoble, pp. 133-144.

DUBOIS, J. ; DUBOIS-CHARLIER, F. (1997), Les Verbes Français, Larousse-Bordas, Paris.

GREZKA, A. (2006a), Les prédicats de perception. Traitement de la polysémie (Les sens des sens), Thèse de doctorat en Sciences du langage, Paris XIII.

GREZKA, A. (2006b), « Etudes du lexique de la perception : bilan et perspectives », Suvremena Lingvistika, 61, pp. 45-67.

LE PESANT, D. ; MATHIEU-COLAS, M. (eds) (1998), Les classes d'objets, Langages, 131, Larousse, Paris.

LEVIN, B. (1993), English Verb Classes and Alternations - A Preliminary Investigation, The University of Chicago Press, Chicago.

Pour un vrai Trésor du français : Proposition de mise en relation du TLFi et de la BDLP

Claude Poirier (1)
Claude.Poirier@lli.ulaval.ca
collab. de Jean-François Smith
Jean-Francois.Smith@ciral.ulaval.ca

(1) ATILF Nancy Université & CNRS

---

1. Le TLFi et la variation diatopique du français

Empruntant la formulation de Georges Mounin, le *Trésor de la langue française* (TLF) définit le terme *trésor* comme suit : « inventaire des unités lexicales d'une langue visant à l'exhaustivité ». À la lumière de cette définition, on conviendra que l'objectif d'un ouvrage ainsi dénommé ne peut jamais être atteint parfaitement. Il faut toutefois de se demander à partir de quel seuil sa nomenclature peut être considérée comme significative. S'agissant de la représentation des usages du français à travers le monde, on doit se rendre à l'évidence que le contenu du TLF est, en 2008, bien en deçà des promesses d'un trésor. Pourtant, dès la parution du premier tome de ce dictionnaire, ses auteurs avaient pris une avance certaine dans ce domaine en programmant la prise en compte d'un nombre, important pour l'époque, de particularités topolectales et en appliquant à ces emplois le même traitement objectif et philologique que pour ceux du français de référence, donnant ainsi l'exemple à suivre. Mais l'évolution des mentalités et des connaissances sur la variation géographique du français a été telle depuis les années 1970 que le corpus des quelque 2240 emplois que l'utilisateur du *Trésor de la langue française informatisé* (TLFi) peut regrouper en se servant de l'indicateur « régional » n'est plus à la hauteur de l'idéal de départ.

Dans sa présentation du cédérom du TLFi, Jean-Marie Pierrel (2004, p. 1) rappelle que, dans l'esprit de Paul Imbs, le premier objectif de l'entreprise dont il avait la direction était de faire un « dictionnaire du monde francophone ». Or, les attentes dans ce domaine sont devenues très grandes, ce qu'ont bien compris les entreprises dictionnairiques commerciales. Ainsi, *Le Nouveau Petit Robert 2007* a, une nouvelle fois, enrichi et actualisé sa nomenclature des particularités topolectales du français et en a revu le traitement en profondeur. Dans ce contexte, il devient difficile d'accepter que, dans le TLFi, ne soit enregistrée qu'une petite partie des emplois originaux et courants que font du français, en parlant et en écrivant, des millions de locuteurs, en France et hors de France. Il est donc urgent qu'une mise à niveau à cet égard soit inscrite parmi les chantiers prioritaires de l'ATILF, surtout que ce laboratoire a attiré l'attention du monde entier sur le TLFi en en publiant une édition sur cédérom en 2004.

La première solution qui vient à l'esprit est d'améliorer la représentation de la variation diatopique dans le TLFi en y introduisant de nouveaux articles. Cette solution est plus complexe qu'il n'y paraît, surtout pour une équipe déjà investie dans de nombreux travaux de standardisation du produit informatique. Si on l'adopte, il faudrait s'assurer de la

collaboration de spécialistes locaux pour éviter les erreurs d'interprétation qui sont inévitables quand on tente d'expliquer un usage qu'on ne pratique pas soi-même. On pourrait atteindre un résultat plus rapide et, nous en sommes convaincus, beaucoup plus satisfaisant en tirant partie du travail considérable qui a été réalisé depuis trente ans par les équipes d'universitaires qui se sont regroupées, depuis le début des années 1990, au sein du réseau « Étude du français en francophonie » (EFF), de l'Agence universitaire de la francophonie (AUF). Ces équipes ont produit plus d'une vingtaine de dictionnaires ou répertoires dont les contenus ont été ou sont en voie d'être versés dans une base de données qui est en accès libre sur le Web, la Base de données lexicographiques panfrancophone (BDLP). Notre exposé vise donc à mettre en évidence les avantages de mettre en relation le TLFi et la BDLP.

2. La problématique des variétés topolectales en lexicographie

Dans le milieu des lexicographes du français, qu'ils soient de France ou d'ailleurs, les variantes topolectales sont déterminées à partir d'une mise en rapport avec un noyau représentant un modèle que les spécialistes désignent par le terme de *français de référence* (FrR). Ce modèle diffère de celui qui sous-tend la rédaction des dictionnaires de l'anglais où la variété britannique et la variété américaine sont toutes deux prises en considération dans la définition du standard international (McArthur, 2001).

Les variétés maternelles du français dans le monde se répartissent sommairement en deux groupes selon leur genèse (Poirier, 2001) :
a. celles d'Europe, issues d'un modèle parisien qui s'est laissé imprégner jusqu'à un certain point, selon les régions, par les usages sur lesquels il s'est superposé;
b. celles d'Amérique, issues des français populaires de la moitié Nord de la France et sur lesquels le modèle parisien a eu beaucoup moins d'emprise.

À mesure que le français s'implante dans les populations africaines, il se produit une évolution qui est en voie de donner naissance à un troisième ensemble se distinguant du modèle de départ diffusé par l'école (Queffélec, 2000). La façon de parler le français dans l'océan Indien, en Océanie et dans les Antilles présente également, dans chacune de ces situations, des caractéristiques liées à des facteurs locaux et à des conceptions de la langue qui justifieraient peut-être qu'on reconnaisse autant de nouveaux ensembles originaux. C'est pourquoi les lexicographes des VTF estiment de plus en plus que le cadre traditionnel des dictionnaires de l'Hexagone pour le traitement des variantes topolectales est devenu trop réducteur. Il faut pouvoir dégager tous les réseaux de correspondances qui traversent la francophonie, même ceux qui ne passent pas par Paris, si l'on veut comprendre la dynamique historique et contemporaine du français à travers le monde.

De toutes les description du français, celle du TLF est la plus compatible avec la démarche et l'objet d'analyse des lexicographes des VTF. En effet, ce dictionnaire couvre une 'synchronie' qui s'étend sur plus de deux siècles, période au cours de laquelle se sont peu à peu effacés dans le FrR des emplois qui ont pu se maintenir ici et là en France et hors de France. C'est en outre pendant cette période que le français s'est implanté en Afrique et dans l'Océanie, tantôt par l'action de locuteurs venus de diverses régions de France (c'est le cas surtout en Algérie), tantôt à travers les réseaux de l'Administration et de l'école (surtout en Afrique subsaharienne). Cette situation est sans doute à l'origine de nombreuses correspondances que l'on remarque entre des relevés régionaux d'Europe et des usages d'Afrique (voir Frey 2005 qui a tiré ses données européennes du *Dictionnaire des régionalismes de France* de Rézeau, 2000).

Dès la première formulation du projet international de la BDLP (Auger et Poirier, 1996), le TLFi a été identifié comme étant le seul corpus lexicographique qui puisse servir les objectifs poursuivis : rendre compte, en combinant les données des répertoires des VTF avec celles du FrR, de la richesse lexicale de la langue française à travers le monde. Il faut rappeler ici que les deux entreprises ont été profondément influencées par Bernard Quemada envers lequel nous avons tous une dette de reconnaissance. On connaît son investissement personnel considérable dans le TLF et ce, depuis l'étape de la rencontre préliminaire dont le présent colloque souligne le cinquantième anniversaire. La BDLP, quant à elle, est une composante d'un vaste chantier qu'il a planifié dans les années 1980 (Quemada 1990), avant même qu'il n'ait eu mené à terme la révision scrupuleuse de tous les tomes du TLF publiés après 1977. C'est à cette date qu'il remplaça Paul Imbs comme premier responsable de la réalisation du dictionnaire, poste qu'il n'accepta que « [s]ous réserve d'une seule exigence : avoir pour mandat de conduire un ensemble de recherches et de productions lexicographiques dont le dictionnaire [sous-entendu : papier] ne serait que l'une des composantes » (Quemada, 1994). Il ne faut donc pas s'étonner qu'il ait proposé, dès 1981, qu'on informatise le texte du TLF et qu'il ait, quelques années plus tard, constitué une équipe dont la mission était de réfléchir aux meilleurs moyens de réaliser un « Trésor informatisé des vocabulaires francophones ». Le projet de la BDLP est la concrétisation de la première phase de ce programme, c'est-à-dire « la constitution de bases de données lexicographiques nationales ou régionales normalisées et interrogeables selon des logiciels communs » (Quemada, 1990).

Les lexicographes des VTF ont mené des expériences plus ou moins parallèles depuis le début des années 1970, l'Afrique étant la zone où la nécessité de la concertation s'est imposée en premier lieu. Les répertoires ont été conçus en fonction des caractéristiques des variétés étudiées et les attentes des communautés concernées. C'est ce qui explique que les maquettes lexicographiques présentent des différences appréciables selon les zones géographiques et que la terminologie varie. Ainsi, on a recours aux termes *acrolecte*, *mésolecte*, *basilecte* pour la description des français du Sud, qui ne conviennent pas pour ceux du Nord (Gleβgen et Thibaiult, 2005). Le projet de la BDLP a été construit dans le respect de ces différences. La mise en commun a cependant été facilitée par le fait que la méthodologie scientifique était fondée sur les mêmes principes : perspective différentielle, primauté au corpus, description objective, prise en compte de la dimension sociolinguistique (Poirier, 2001, p. 28-31).

3. Compatibilité des deux corpus

Le TLFi et la BDLP sont deux corpus lexicographiques complémentaires en dépit de nombreuses différences relatives à leur taille, à la réputation de leurs auteurs, à la profondeur des analyses, à la puissance de diffusion, etc. La comparaison est généralement en faveur du premier, sauf sur un point : bien qu'aucune de ses composantes ne soit encore achevée, la BDLP, avec ses 14.353 emplois définis et illustrés, donne un meilleur aperçu de la variation topolectale que le TLFi qui ne consacre que 2.838 définitions à cette dimension de la langue. Avec ses 32.658 citations, la BDLP offre en outre un rapport définitions/citations plus avantageux que le TLFi. Ajoutons que la plupart des VTF qui alimentent le corpus de la BDLP ne sont pas prises en compte dans le TLFi, ou ne le sont que de façon ponctuelle; c'est le cas notamment pour la Louisiane, l'Acadie, les pays d'Afrique, ceux du Maghreb, de l'océan Indien et de l'Océanie.

Les deux corpus partagent des caractéristiques communes. Dans les deux cas, il s'agit :

a. de bases intégrant les données d'un grand nombre de descriptions lexicographiques antérieures;
b. de corpus philologiques, donnant une grande place aux exemples, avec des références précises;
c. de corpus permettant des recherches variées, extrêmement utiles pour l'étude et la connaissance du français;
d. de corpus informatisés en accès libre, dont la consultation n'exige que le maniement d'un navigateur.

Les trois principes fondamentaux de la BDLP correspondent ainsi à la philosophie qui a inspiré la rédaction du TLF :
• Les emplois répertoriés doivent être documentés à partir d'un corpus représentatif et correspondre à des usages réels de la variété de français faisant l'objet de la base.
• Toutes les variétés, qu'elles soient nationales ou régionales, parlées par des groupes restreints ou par une population nombreuse, sont considérées sur le même pied.
• Les descriptions visent à dégager tous les emplois lexicaux qui paraissent caractéristiques par rapport à un terme de comparaison commun: le français de référence.

Le TLFi et la BDLP ont été conçus d'après des approches informatiques différentes, mais le fait que le premier soit une base de données textuelles balisée en XML et que la seconde soit une base de données relationnelle n'empêche aucunement leur mise en relation. Le TLFi résulte pour l'essentiel d'une rétroconversion d'un dictionnaire papier dont on a d'ailleurs voulu reproduire l'apparence visuelle. La BDLP est une base de données construite à partir des fichiers informatiques de répertoires lexicographiques, sans préoccupation de conserver la disposition originelle des éléments du texte. La formule de la base de données a été retenue parce qu'elle n'exigeait pas d'uniformiser les textes au préalable, ni pour le forme ni pour le contenu, et qu'elle permettait plus facilement l'affichage d'une série de courts extraits sur un même écran en réponse à une requête.

L'établissement d'une fiche BDLP est réalisé comme suit. Pour un emploi donné, le texte original (définition, citations, commentaire, etc.) est récupéré par rubriques à partir du fichier source, puis versé dans les sections pertinentes de la fiche de saisie. Un même article donne ainsi lieu à plusieurs fiches s'il comporte plus d'une définition. Les caractéristiques typographiques du texte de départ sont conservées si elles sont significatives sur le plan linguistique, mais l'affichage des extraits est standardisé par souci d'uniformité. L'opération de saisie de la fiche BDLP comprend en outre l'introduction d'éléments d'information pour permettre la recherche de contenus de même nature (par ex. classe sémantique, langue d'origine des emprunts, processus néologique, répartition géographique). La programmation exploitera ces deux types de données, autorisant ainsi des recherches à partir du texte même (par ex. repérage d'un mot dans telle rubrique), à partir des informations qui auront été incorporées dans la fiche de saisie (par ex. regroupement d'emplois par classe sémantique, par région, par langue d'origine), ou à partir des deux à la fois. Pour l'utilisateur, la démarche est limpide, l'interface demeurant la même quelle que soit la requête, une distinction étant cependant établie entre la recherche simple (par ordre alphabétique et dans une seule base), et la recherche transversale (à travers les divers champs d'information d'une base ou de plusieurs bases). Ainsi, pour une recherche transversale, l'écran affiche une série de rubriques offrant différents choix qu'il est possible de combiner. Un tel dispositif permet d'intégrer les données de n'importe quel répertoire dans le futur.

4. Avantages de la mise en relation du TLFi et de la BDLP

L'idée de mettre en rapport les deux corpus lexicographiques est une des prémisses du projet de la BDLP. Chacun des répertoires des VTF, quelle que soit sa facture, étant structuré dans sa nomenclature et sa microstructure d'après la comparaison de la variété décrite avec le FrR, la BDLP ne prendra tout son sens que le jour où elle pourra être branchée sur le TLFi, qui représente la description la plus complète et la plus pertinente de la variété de référence.

Le couplage du TLFi et de la BDLP serait, dans l'immédiat, profitable surtout à cette dernière. Outre l'éclairage sémantique et historique que fournirait l'insertion des emplois des VTF dans la structure des articles du TLFi, on pourrait, dans l'exploration de la base panfrancophone, tirer parti des outils puissants qui sont implantés au Centre National de Ressources Textuelles et Lexicales de l'ATILF (dictionnaires de synonymes et d'antonymes, de formes fléchies, etc.). Le rapprochement des deux corpus aurait pour conséquence de favoriser la reconnaissance de la BDLP dans le monde francophone, de favoriser sa fréquentation – le TLFi est l'objet de centaines de milliers de connexions quotidiennes – et ainsi de rendre possible un appui financier plus soutenu de la part d'organismes gouvernementaux et privés.

Le TLFi y gagnerait une mise à jour instantanée et un enrichissement important de la portion topolectale du lexique. Au-delà de la simple augmentation des données traitées, c'est à une nouvelle représentation de la variation géographique de la langue à laquelle auraient accès les utilisateurs du TLFi. La BDLP, étant une base multimédia sur Internet, propose une explication du vocabulaire dans un éclairage culturel. Elle regroupe des bases nationales ou régionales qui peuvent être consultées séparément, par sous-ensembles ou comme un tout, permettant des recherches à partir d'une variété de paramètres. Chaque mot ou expression est accompagné d'une définition, d'exemples authentiques et d'explications portant sur sa répartition géographique, son origine et son histoire. La BDLP comprend en outre des enregistrements sonores (prononciations de mots ou d'expressions, extraits de chansons), des images et des séquences vidéo. Elle est appelée à devenir une véritable encyclopédie linguistique des mots de la francophonie.

Un des atouts de la BDLP réside dans sa capacité de faire apparaître les réseaux qui traversent la francophonie et qui échappent à l'observation dans les autres corpus ou dictionnaires existants. L'aventure de la colonisation aux XVII^{e} et XVIII^{e} siècles a tissé des liens entre les communautés francophones qui se sont établies en Amérique et dans l'océan Indien; les recherches transversales mettent en évidence la parenté qui les unit aux régions de la frange atlantique de la France. De la même façon, des relations privilégiées sont perceptibles entre les français belge, suisse et nord-américains; elles s'expliquent par la maintien dans ces zones d'usages qui sont disparus du FrR et, progressivement, d'une grande partie ou de la totalité du territoire de l'Hexagone. Le lexique du Nord entretient des rapports avec celui du Sud, à travers la filière coloniale des XIX^{e} et XX^{e} siècles, bien sûr, mais parfois aussi entre telle région de France et tel pays d'Afrique, sans que les raisons en soient toujours évidentes (Frey 2005). Mais des oppositions sont également marquées entre le Nord et le Sud, et même entre le Maghreb et l'Afrique subsaharienne, notamment en raison de la provenance différente des emprunts.

Dix-neuf bases sont aujourd'hui en voie d'élaboration, dont quatorze sont en ligne. Il est donc possible d'interroger actuellement les bases suivantes : Acadie, Algérie, Belgique, Burundi, Centrafrique, Congo (Brazzaville), Louisiane, Madagascar, Maroc, Nouvelle-Calédonie,

Québec, La Réunion, Suisse, Tchad. En mars 2008, la BDLP-France sera la 15e à être ouverte à la consultation.

5. Exploration des formules de mise en relation

La conférence explorera trois formules de mise en relation des deux corpus lexicographiques. Un simple renvoi du TLFi vers la BDLP pourrait être réalisé de façon automatique à partir d'un index chaque fois que celle-ci contient des données relatives à un mot figurant dans le TLFi. Des renvois de la BDLP vers le TLFi pourraient également être faits sans trop de difficultés. Ces liens élémentaires apporteraient déjà un éclairage fort intéressant sur la variation topolectale. Mais ce qu'il faudrait viser, à moyen terme, c'est un système de renvois qui permettrait de faire pénétrer l'utilisateur de la BDLP à l'intérieur des articles du TLFi, là justement où s'expliquerait sémantiquement et génétiquement le rapport entre telle variante topolectale et tel emploi du FrR. Inversement, c'est à l'intérieur des articles du TLFi que devraient être implantés les renvois vers la BDLP, pour la même raison. Des projets plus ambitieux pourraient être envisagés, comme celui de tenir compte des variantes topolectales dans les dictionnaires de synonymes, d'antonymes, de formes fléchies, etc. du Centre National de Ressources Textuelles et Lexicales de l'ATILF.

Bibliographie

AUGER, Alain, et POIRIER, Claude, 1996, *Base de données lexicographiques panfrancophone (BDLP). Document de présentation*, Trésor de la langue française au Québec, Université Laval [document mis à jour en 2003, avec la coll. de Nathalie Bacon, Johanna-Pascale Roy et Jean-François Smith; téléchargeable à partir de la page d'accueil de la BDLP : www.bdlp.org].

BAVOUX, Claudine, 2001, « Peut-on appliquer le concept de 'langue polynomique' au français ? », dans *Diversité culturelle et linguistique : quelles normes pour le français?*, IXe Sommet de la Francophonie, Agence universitaire de la Francophonie, p. 74-80.

DENDIEN, Jacques, 2004, « Histoire de l'informatisation du TLF », *Trésor de la langue française informatisé*, Paris, CNRS Éditions, , p. 7-26 [Livre accompagnant le cédérom].

FREY, Claude, 2005, « Régionalismes de France et régionalismes d'Afrique : convergences lexicales et cohérence du français », dans Martin-D. Gleβgen et André Thibault, *La lexicographie différentielle du français et le Dictionnaire des régionalismes de France*, Strasbourg, Presses universitaires de Strasbourg, p. 233-249.

GLEβGEN, Martin, et THIBAULT, André, 2005, « La 'régionalité linguistique' dans la Romania et en français », dans Martin-D. Gleβgen et André Thibault, *La lexicographie différentielle du français et le Dictionnaire des régionalismes de France*, Strasbourg, Presses universitaires de Strasbourg, p. III-XVII.

McARTHUR, Tom, 2001, « World English and world Englishes : Trends, tensions, varieties, and standards », dans *Language Teaching*, vol. 34, no 2, p. 1-20.

PIERREL, Jean-Marie, 2004, « Préface », *Trésor de la langue française informatisé*, Paris, CNRS Éditions, p. 1-5 [Livre accompagnant le cédérom].

PIERREL, Jean-Marie, et PETITJEAN, Étienne, 2007, « Valorisation et exploitation scientifiques de documents numériques pour la recherche en linguistique : l'exemple du CNRTL », Actes de CIDE, Europia, p. 13-25.

POIRIER, Claude, 1983, « L'intrication des mots régionaux et des mots du français général dans le discours québécois », dans *Langues et linguistique*, n° 9, Université Laval, Québec, p. 45-67.

POIRIER, Claude, 2001, « Vers une nouvelle pratique de la lexicographie du français (EFF) », dans *Diversité culturelle et linguistique : quelles normes pour le français?*, IX^e^ Sommet de la Francophonie, Agence universitaire de la Francophonie, p. 19-39.

POIRIER, Claude, 2003, « Variation du français en francophonie et cohérence de la description lexicographique », dans Monique C. Cormier, Aline Francœur et Jean-Claude Boulanger (dir.), *Les dictionnaires Le Robert: genèse et évolution*, Montréal, Les Presses de l'Université de Montréal, p. 189-226.

POIRIER, Claude, 2005, « La dynamique du français à travers l'espace francophone à la lumière de la base de données lexicographiques panfrancophone », dans *Revue de linguistique romane*, Strasbourg, t. 69, n^os^ 275-276, p. 483-516.

QUEFFÉLEC, Ambroise, 2000, « Emprunt ou xénisme : Les apories d'une dichotomie introuvable ? », dans Danièle Latin et Claude Poirier (dir.), avec la coll. de Nathalie Bacon et de Jean Bédard, *Contacts de langues et identités culturelles*, Québec, AUF – Les Presses de l'Université Laval, p. 283-300.

QUEMADA, Bernard, 1990, « Trésor informatisé des vocabulaires francophones », dans André Clas, et Benoît Ouoba (éd.), *Visages du français. Variétés lexicales de l'espace francophone*, Paris, AUPELF et John Libbey Eurotext, p. 141-145.

QUEMADA, Bernard, 1994, « Postface », publié dans le tome 16 du *Trésor de la langue française* et reproduit dans *Trésor de la langue française informatisé*, Paris, CNRS Éditions, p.153-158 [Livre accompagnant le cédérom].

RÉZEAU, Pierre (éd.), 2001, *Dictionnaire des régionalismes de France. Géographie et histoire d'un patrimoine linguistique*, Bruxelles, Duculot.

THIBAULT, André, 2005, « Le traitement des régionalismes dans les rubriques étymologiques du *Trésor de la langue française* : l'exemple du vocabulaire de G. Guèvremont », dans Éva Büchi (éd.), *Actes du Séminaire de méthodologie en étymologie et histoire du lexique (Nancy/ATILF, année universitaire 2005/2006)*, Nancy, ATILF (CNRS/Université Nancy 2/UHP), publication électronique (http://www.atilf.fr/atilf/seminaires/Seminaire_Thibault_2005-10.pdf), 36 p.

L'anglo-normand et le TLF, dans le passé et dans l'avenir

David Trotter (1)
dtt@aber.ac.uk

(1) Aberystwyth University (Royaume-Uni)

Mots-clés : anglo-normand ; ancien français ; moyen français ; lexicographie ; anglicismes

Keywords: Anglo-Norman; Old French; Middle French; lexicography; Anglicisms

Résumé : La communication explique l'importance de l'anglo-normand pour la lexicographie du français, même dans un dictionnaire de la langue moderne (en l'occurrence : le TLF). Trois aspects de la question sont traités à l'aide d'exemples puisés dans le TLF: 1) les facteurs chronologiques (précocité de l'anglo-normand = précocité des attestations) ; 2) mots dont les seules attestations – du moins pour l'instant – sont dans des textes anglo-normands ; 3) l'anglo-normand en tant que français régional du Moyen Âge, voie de transmission de mots anglais (anglo-saxons ou moyen-anglais) en France.

Abstract : The paper discusses the significance of Anglo-Norman for the (historical) lexicography of French, even for the purposes of a dictionary of the modern language like the TLF. Three aspects are examined via examples from the TLF itself : 1) chronological reasons why Anglo-Norman is important (early appearance of Anglo-Norman texts = early attestations) ; 2) words whose only attestations are from Anglo-Norman texts ; 3) Anglo-Norman as a form of medieval regional French, and as a means of transmission of Anglo-Saxon and Middle English words to France.

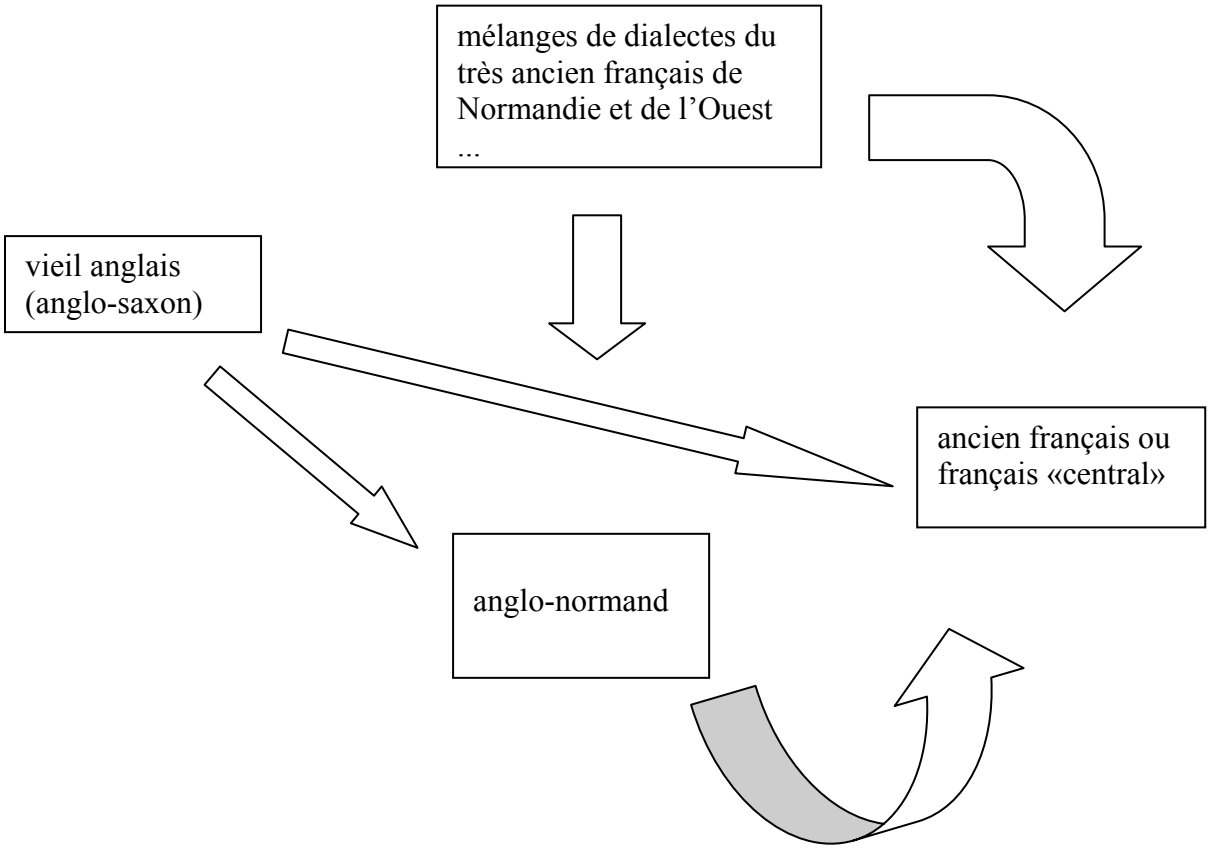
Introduction

Déjà dans son allocution au colloque fondateur du TLF de 1957, Paul Imbs souligna l'importance de l'étymologie : « un dictionnaire ne se conçoit plus sans une série de renseignements précis sur l'étymologie des mots recueillis » [Imbs 1961 : 10]. La tradition se poursuit car une équipe de l'ATILF entreprend actuellement – et de manière approfondie – la révision des étymologies du TLFi. Qui dit étymologie, dit ancien français ; et la communication du père du TLF porte précisément sur « La place du vocabulaire ancien dans un thesaurus de la langue française » [Imbs 1961 : 133-139]. Mais il est frappant aussi que cette discussion de l'ancien français soit suivie de la contribution du regretté Kurt Baldinger, sur « L'importance du vocabulaire dialectal dans un thesaurus de la langue française » (149-163). Ces deux contributions cernent ainsi – à leur insu, sans doute – la problématique de ce qu'on pourrait appeler « L'importance du vocabulaire anglo-normand dans un thesaurus de la

langue française ». D'une part, l'anglo-normand, dialecte de cette pauvre colonie française qui a mal tourné, est de **l'ancien français** ; d'autre part, il s'agit d'**un français régional du Moyen Âge** qui se distingue par son indépendance et (dans une certaine mesure) par son éloignement, voire par son isolement, politique et linguistique [Trotter 2003a ; 2003b : 53 et n.8]. L'intégrer dans un thesaurus de la langue française (donc, dans le TLF) pose des problèmes particuliers mais en même temps, un pareil thesaurus même de la langue moderne ne peut qu'avec difficulté négliger cette variété. Dans cette communication, nous expliquerons pourquoi.

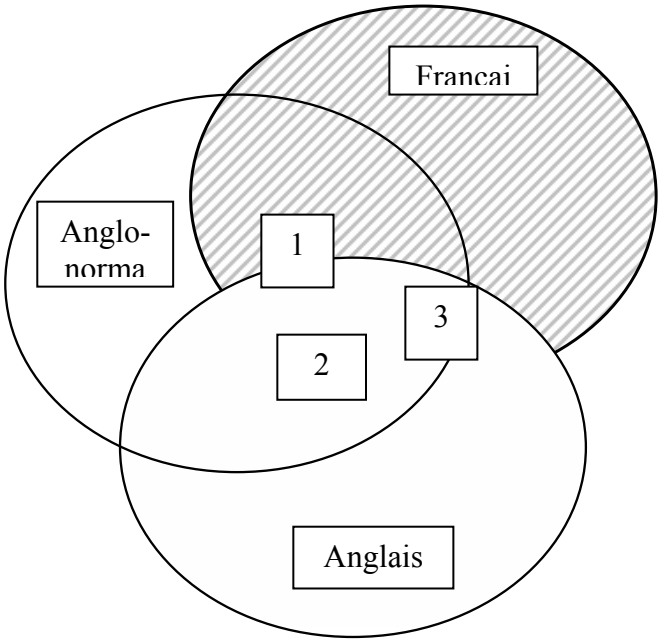
Les rapports entre l'anglo-normand et le français (qui impliquent forcément, pour des raisons qui seront exposées plus loin, l'anglais) peuvent être schématisés de plusieurs façons : en premier lieu, la perspective historico-évolutive, qui met l'accent sur la genèse :

Figure 1 : Rapports historico-évolutifs



Un deuxième schéma, sans doute plus proche de la réalité sociolinguistique de l'époque, présenterait les trois langues synchroniquement et de manière assez différente :

Figure 2 : Rapports linguistiques entre français, anglo-normand, anglais



Zone 1 : contact et influence français ↔ anglo-normand
Zone 2 : contact et influence anglo-normand ↔ anglais
Zone 3 : contact et influence français ↔ anglais

Bien entendu, l'existence de rapports réciproques et de zones de contact n'implique nullement que l'importance de l'influence était la même dans les deux sens. L'on remarquera aussi que nous avons omis dans les deux figures une quatrième langue, le latin, qui bien entendu occupe une place importante, que ce soit dans le schéma « historique » (Fig. 1) ou dans le schéma « synchronique » (Fig. 2). Nous l'avons fait à escient car le latin ne nous concerne pas pour l'analyse qui va suivre.

1. La situation chronologique de l'anglo-normand

L'existence de ces rapports – dus évidemment aux faits historiques – a comme résultat que pour ce qui est de la langue, l'Angleterre toujours sera sœur de la France. Tout dictionnaire qui présente l'historique du français est obligé de tenir compte de l'anglais pour sa partie moderne, et de tenir compte, pour le Moyen Âge, au moins dans une certaine mesure, de l'anglo-normand. Et cela, premièrement, pour des **raisons chrononologiques** dans l'**étymologie** du français. Les habitants de la colonie d'outre-Manche ont eu la mauvaise idée (ou le mauvais goût) de se mettre à écrire en français un peu plus tôt que leurs parents en France (cf. Howlett 1996) avec le résultat inévitable que parmi les premières attestations en ancien français, figurent celles provenant de textes anglo-normands. Que ce soit dans Godefroy, dans TL, ou dans le TLF, cela passe souvent inaperçu, ou n'est pas en tout cas signalé : sub **garant**, **gant**, **adouber, trahison, fermer** du TLF, par exemple, la première attestation dans la notice étymologique dans tous ces cas provient de la *Chanson de Roland*

(éd. Bédier ; manuscrit anglo-normand d'Oxford). L'on ne le dit pas car le texte lui-même est sans doute « français ».[1] Il est légitime peut-être de supposer que les lecteurs du TLF savent que le texte le plus ancien du *Roland* est conservé dans un manuscrit anglo-normand, mais en va-t-il de même pour le *Comput* de Philippe de Thaon, nom aux accents continentaux sinon normands, et source de la plus ancienne attestation de *comput* et de *lunaison*, mais aussi de mots plus banals comme *mercredi* et *novembre* ? Le *Bestiaire* du même auteur (PhThBestWa 3) fournit évidemment la première attestation du mot ... *bestiaire* (TLF **bestiaire**[1]). Dans la rubrique « Étymol. et Hist » de tous ces mots, le TLF n'indique nullement une origine anglo-normande et dans le cas de **comput**, il va jusqu'à préciser : « a.fr. » ; sub **bestiaire** : « seulement en a.fr. ». De même, le *Voyage de Saint-Brendan*, lui aussi texte anglo-normand, fournit la première citation dans la notice étymologique du TLF sub **haler**[1] (rien de plus ancien dans DEAF H 100). Or, le TLF ne dit rien sur l'origine non continentale de ces textes et ainsi, omet de souligner le fait que leur première mise par écrit a eu lieu en Angleterre. Ce phénomène, cependant, montre non seulement la précocité de l'anglo-normand, mais également l'importance de cette variété (souvent méprisée) pour la documentation du plus ancien français. Et d'un certain point de vue, si le TLF ne signale pas l'origine (dialectale) de ses attestations anglo-normandes, ne peut-on pas en conclure que l'anglo-normand a acquis, sans plus, droit de cité au cœur de la lexicographie du français ?

2. L'anglo-normand apporte du nouveau

Deuxième aspect de **la contribution des textes anglo-normands au TLF**, et qui deviendra sans doute plus important dans l'avenir : les **éléments anglo-normands qui manquent** ou surtout, qui ont été découverts après la parution du TLF. Soulignons que l'impression de l'AND date de la période 1977-1992 : ce dictionnaire – le seul à s'occuper de manière différentielle de l'anglo-normand – n'était donc pas disponible pendant la plupart du travail de rédaction du TLF, d'où quelques malheureuses absentes du bouquet du Trésor.[2] Tobler-Lommatzsch, également, n'était pas achevé. Or, pour certains des mots discutés par Paul Imbs lui-même, les connaissances actuelles de l'anglo-normand permettent d'ajouter des renseignements très pertinents : il s'agit d'*arroi* et de *loge*. Sub **arroi** (AND : **arraie**), le TLF fournit dans la partie étymologique – qui reprend sans doute TL 1,541 – deux sens : 1) « manière d'être, organisation » (CleomH, 1285) et 2) « équipage accompagnant une personne » (BastS, s.xiv).[3] Le premier de ces deux sens convient donc à la citation de la *Chanson de Roland* (trad. Bédier ; citée par Imbs 1961 : 134s.) : « Naimes le duc et Jozeran le comte rangent en bel arroi ces deux corps de bataille ». La documentation anglo-normande (en l'occurrence : l'AND) permet d'ajouter une attestation probablement plus ancienne (EchecsCottH 1070, s.xiii) ainsi qu'une quantité importante de citations supplémentaires et de collocations. Pour *loge*, on ajoutera de l'AND la citation de TristBérG 1894 (déjà dans TL), avec un sens qui n'est pas recensé dans la partie historique du TLF (si toutefois la glose de l'AND est correcte : elle est différente de celle fournie par TL 5,575[4]). L'on pourrait ajouter d'autres exemples. GaimarB 6224, texte anglo-normand de 1139, fournit une antédatation

[1] N'entrons pas ici dans la problématique – particulièrement difficile dans le cas de la lexicographie de l'anglo-normand – de la distinction entre texte et manuscrit, pour décider de l'anglo-normandicité ou non d'un texte ...

[2] Pour l'AND, l'on utilisera désormais la version en ligne (gratuite) www.anglo-norman.net. Le dictionnaire électronique comporte 1) une refonte totale de l'AND1 (A-H) et 2) des révisions au texte de la première édition pour la partie I-Z.

[3] Les sigles sont ceux du DEAF.

[4] AND : « porter's lodge » ; TL : «Galerie, Vorhalle, Halle (einer Burg)» ; TLF, sens 2 : « antichambre qui précède la salle principale d'un château », renvoyant à *Couronnement Louis* 1619. Le sens « guérite de portier de ville, de sentinelle» est relevé au XV[e] et celui de «petit réduit servant d'habitation à un portier » depuis 1660 dans le FEW 16,447a.

pour l'**archer** du TLF (actuellement : SThomGuernW[1], ca. 1174) et sub **autoriser** dans TLF, la première attestation : « *Les Loherains,* Richel. 4988, f° 222 r° ds GDF » (c'est-à-dire : AnsMetzS[1], ca. 1300) peut être remplacée par GaimarB 2329. Maria Iliescu signale le cas d'*entrevue*, daté de 1498 dans le TLF mais attesté en 1323-25 dans l'AND, ChaplaisStSardos 20 [Iliescu 2007 : 133]. Comme c'est le cas pour le FEW [Trotter 1998], l'anglo-normand et l'AND permettent de modifier – parfois sensiblement – la datation des premières attestations de mots français par rapport aux données du TLF. Ce n'était pas le but du dictionnaire (qui n'a pas d'ambitions historiques de ce type), mais tant mieux s'il le fait.

3. L'anglo-normand, français régional : le contact avec l'anglais

Troisième volet de l'apport anglo-normand au TLF, peut-être le plus important : son statut assez spécial de **français régional du Moyen Âge**. Dans la mesure où le TLF est surtout un dictionnaire du français « moderne », il est logique qu'il exclue tout ce qui ne fait pas partie de l'ascendance de cette langue ; ainsi, la majeure partie des divagations sémantiques de l'anglo-normand ne s'y trouvera pas. En même temps, si l'on veut un panorama des possibilités historiques de l'ancien français, prédécesseur du français moderne, il y a intérêt à intégrer l'anglo-normand s'il conserve des formes parallèles à l'ancien français et qui aideraient à illustrer l'évolution diachronique de celui-ci. Au fond, l'anglo-normand constitue (sous cet angle) le passé, certes disparu, d'une variété régionale du français qui était (à l'époque) assez étroitement liée à ce qui allait devenir (bien plus tard) la langue « nationale ». Il n'est donc pas surprenant que la variété insulaire ait contribué à la langue continentale, servant notamment de moyen de transfert d'un nombre certes limité mais néanmoins intéressant de mots anglais, en français. Par exemple : le mot **mauve[2]** (TLF ; cf. FEW 16,495b), le seul à notre connaissance à recevoir la balise « langue empruntée » = « anglo-normand » dans le TLF. Il proviendrait de l'anglo-saxon *maew* et serait passé de là en anglo-normand, quitte à traverser la Manche et à s'installer … surtout en Haute-Bretagne. Les dérivés *mouette* et *mauvis* proviennent aussi de cette même origine. Ce sont donc des mots anglo-normands qui font maintenant partie de la langue « mère ». Ils ne sont pas les seuls. D'autres sont très connus : *nord* (du norrois d'après le TLF ; mais de l'anglo-saxon selon FEW 16,603a), *sud*, *est*, et *ouest* semblent provenir de l'anglo-saxon et bien entendu, leurs premières attestations françaises sont dans des textes anglo-normands. L'on retiendra aussi le *haddock* anglais (TLF), devenu anthroponyme sous la plume d'Hergé avec le capitaine Haddock, mot qui a fort bien pu transiter par l'anglo-normand, cf. DEAF H 15 [Trotter 2003a : 53].[5] À côté de l'exemple isolé de l'AND première édition (tiré de OakBookS 2,10), il existe une série d'autres – plus clairement « francisés » – dans LAlbR, et qui tendent à confirmer notre hypothèse d'une transmission en passant par l'anglo-normand. La possibilité de la recherche « langue empruntée » (en sélectionnant bien sûr l'anglais) dans le TLFi permet de retrouver d'autres exemples. Parmi ceux-ci, citons : **beaupré** ; **dogue** ; **drague[1]** ; **haquenée** ; **wharf**. Ce sont tous des mots ayant fait leur entrée en français (c'est-à-dire : en anglo-normand) au Moyen Âge. Le premier (*beaupré*), absent – honni soit qui mal y pense – de l'AND, est passé du bas allemand en anglais, ensuite en anglo-normand, paraissant pour la première fois sur le territoire français dans les comptes du clos des galées de Rouen en 1382 (TLF).[6] Sandahl ne connaît que des attestations latines et (peut-être) anglaises, surtout dans les textes mixtes (SandahlSea 2,24-27). Le MED sub **bou-sprēt(e** a trouvé une attestation anglo-normande reprise par TLF. Voici donc un mot à ajouter à l'AND. Des autres exemples

[5] Baldinger 1960 : 40 le signale aussi en apoit., emprunté directement de l'anglais mais il évoque aussi la possibilité d'un emprunt indirect en passant par l'anglo-normand (n.53).

[6] Nous avons déjà constaté l'importance de ce lieu pour la transmission non seulement des techniques de construction navale, mais aussi du vocabulaire maritime, et qui font de Rouen une véritable «plaque tournante» linguistique (Vidos 1960: 4); cf. Trotter 2003b: 23.

L'anglo-normand et le TLF, dans le passé et dans l'avenir

David Trotter (1)
dtt@aber.ac.uk

(1) Aberystwyth University (Royaume-Uni)

Mots-clés : anglo-normand ; ancien français ; moyen français ; lexicographie ; anglicismes

Keywords: Anglo-Norman; Old French; Middle French; lexicography; Anglicisms

Résumé : La communication explique l'importance de l'anglo-normand pour la lexicographie du français, même dans un dictionnaire de la langue moderne (en l'occurrence : le TLF). Trois aspects de la question sont traités à l'aide d'exemples puisés dans le TLF: 1) les facteurs chronologiques (précocité de l'anglo-normand = précocité des attestations) ; 2) mots dont les seules attestations – du moins pour l'instant – sont dans des textes anglo-normands ; 3) l'anglo-normand en tant que français régional du Moyen Âge, voie de transmission de mots anglais (anglo-saxons ou moyen-anglais) en France.

Abstract : The paper discusses the significance of Anglo-Norman for the (historical) lexicography of French, even for the purposes of a dictionary of the modern language like the TLF. Three aspects are examined via examples from the TLF itself : 1) chronological reasons why Anglo-Norman is important (early appearance of Anglo-Norman texts = early attestations) ; 2) words whose only attestations are from Anglo-Norman texts ; 3) Anglo-Norman as a form of medieval regional French, and as a means of transmission of Anglo-Saxon and Middle English words to France.

Introduction

Déjà dans son allocution au colloque fondateur du TLF de 1957, Paul Imbs souligna l'importance de l'étymologie : « un dictionnaire ne se conçoit plus sans une série de renseignements précis sur l'étymologie des mots recueillis » [Imbs 1961 : 10]. La tradition se poursuit car une équipe de l'ATILF entreprend actuellement – et de manière approfondie – la révision des étymologies du TLFi. Qui dit étymologie, dit ancien français ; et la communication du père du TLF porte précisément sur « La place du vocabulaire ancien dans un thesaurus de la langue française » [Imbs 1961 : 133-139]. Mais il est frappant aussi que cette discussion de l'ancien français soit suivie de la contribution du regretté Kurt Baldinger, sur « L'importance du vocabulaire dialectal dans un thesaurus de la langue française » (149-163). Ces deux contributions cernent ainsi – à leur insu, sans doute – la problématique de ce qu'on pourrait appeler « L'importance du vocabulaire anglo-normand dans un thesaurus de la

Cette entreprise, qui fait partie en premier lieu des tâches qui incombent aux spécialistes de l'anglo-normand, est également importante pour le français lui-même. Évidemment, des collaborations plus étroites entre les différents projets consacrés à la lexicographie du français, en partie grâce à l'apport de l'informatique qui facilite cette coopération, et qui rend possible des liens directs entre différents dictionnaires (Beddow 2007 ; Rothwell / Trotter 2007), joueront un rôle essentiel dans le ré-examen du lexique « anglo-normand ». Pour rendre pleinement compte des sources anglo-normandes dans l'histoire du français, il serait souhaitable que l'AND devenu électronique soit relié au TLFi, ainsi qu'au DEAFi et – le rêve – à un FEWi. Les progrès de l'informatique le permettent, si la volonté lexicographique existe.

Bibliographie

[Baldinger, 1960] Baldinger, Kurt (1960) : *Lexikalische Auswirkungen der englischen Herrschaft in Südwestfrankreich (1152-1453)*, dans Wolfgang Iser (éd.), *Britannica. Festschrift für Hermann M. Flasdieck*, 11-50, Heidelberg, Winter.

[Beddow, 2007] Beddow, Michael (2007) : *L'Anglo-Norman On-line Hub : une présentation technique*, dans David Trotter (éd.), *Actes du XXIV^e Congrès International de Linguistique et de Philologie Romanes, Aberystwyth, 1^er-6 août 2004*, I, 305-310, Tübingen, Niemeyer.

[Bruneau, 1955] Bruneau, Charles (1955) : *Petite histoire de la langue française*, Paris, Colin.

[Howlett, 1996] Howlett, David R. (1996) : *The English origins of Old French literature*, Dublin, Four Courts Press.

[Iliescu, 2007] Iliescu, Maria. (2007) : *Je sème à tout vent*, dans Juhani Härmä, Elina Suomela-Härmä & Olli Välikangas (éds.), *L'Art de la Philologie: Mélanges en l'honeur de Leena Löfstedt*, 131-136, Helsinki, Mémoires de la Société Néophilologique de Helsinki.

[Imbs, 1961] Imbs, Paul (1961) : *Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles. Strasbourg, 12-16 novembre 1957*, Colloques Internationaux du Centre National de la Recherche Scientifique, Paris, Éditions du Centre National de la Recherche Scientifique.

[Roques, 1997] Roques, Gilles (1997) : *Des interférences picardes dans l'Anglo-Norman Dictionary*, dans Stewart Gregory / D.A. Trotter (éds), *De mot en mot : Essays in honour of William Rothwell*, 191-198, Cardiff, MHRA/University of Wales Press.

[Rothwell/Trotter, 2007] Rothwell, Andrew / Trotter, David (2007) : *Évolution et structure de l'Anglo-Norman Dictionary, deuxième édition*, dans David Trotter (éd.), *Actes du XXIV^e Congrès International de Linguistique et de Philologie Romanes, Aberystwyth, 1^er-6 août 2004*, II, 413-421, Tübingen, Niemeyer.

[Trotter, 1998] Trotter, David (1998) : *Les néologismes de l'anglo-français et le FEW*, *Le Moyen Français* 39-41 (1996/1997 [1998]), *Néologie et création verbale : Actes du VIII^e Colloque international sur le moyen français, McGill University, Montréal, Canada, octobre 1996*, 577-635.

[Trotter, 2003a] Trotter, David (2003a) : *Not as eccentric as it looks: Anglo-French and French French*, *Forum for Modern Language Studies* 39, 427-438.

[Trotter, 2003b] Trotter, David (2003b) : *L'anglo-normand : variété insulaire, ou variété isolée ?*, *Médiévales* 45, 43-54.

[Trotter, 2007] Trotter, David (2007) : *Pur meuz acorder en parlance E descorder en variaunce : convergence et divergence dans l'évolution de l'anglo-normand*, dans Sabine Heinemann / Paul Videsott (éds.), Sprachwandel und (Dis-)Kontinuität in *der Romania*, 85-93, Tübingen, Niemeyer.

[Vidos, 1960] Vidos, B.E. (1960) : *Le bilinguisme et le mécanisme de l'emprunt*, *Revue de Linguistique romane* 24, 1-19.

De la préface du *TLF* à l'idéologie francophone : pratiques lexicographiques et description du français en Afrique

Claude Frey (1)
ccfrey@wanadoo.fr

(1) Université de Paris 3

Mots-clés : francophonie, lexicographie, français d'Afrique, variété lexicale, diatopique, IFA, TLFi.

Keywords: french speaking countries, lexicography, french language in Africa, lexical variety, diatopic, IFA, TLFi.

Résumé : Le *TLF* aujourd'hui informatisé (*TLFi*), a pour vocation théorique une description lexicale couvrant la totalité de l'espace francophone dans la combinaison rupture et continuité. Or sa consultation met en évidence que, si la couverture est bien réalisée sur le plan diachronique, elle l'est moins sur le plan diatopique. La communication cherche à montrer, en s'appuyant sur des exemples tirés de différents dictionnaires, que les variétés lexicales africaines compléteraient utilement les perspectives du *TLF* en respectant les idéaux francophones affichés dans la Préface originale de Paul Imbs en 1971. Cinquante ans après la naissance du *TLF*, la riche documentation par l'*IFA* et les productions lexicographiques ultérieures, combinées aux potentialités de l'informatique, devraient permettre d'aller dans cette direction.

Abstract : *TLF*, to-day computerized as *TLFi*, has as a theoretical vocation, a description of French lexicon covering the whole range of French speaking countries, combining differences and identities. However, consulting *TLFi* points out a correct covering from a chronological point of view, but some gaps from a geographical point of view. My purpose here is, with some examples out of different dictionaries, to show that African varieties of French lexicon would profitably complete *TLF* outlooks, as proposed in the original Preface by Paul Imbs, 1971. Fifty years after *TLF* birth, the important data by *IFA* and other lexicographic outputs, combined with computer possibilities, should allow to go further in that direction.

Introduction

Au début des années 1960 se concrétise le projet lexicographique du *TLF*. Le corpus de textes écrits est balisé par deux dates symboliques : 1789 et 1960. Si l'une et l'autre renvoient à des événements de dimension internationale, je retiendrai surtout 1960, la décolonisation, qui amène « *dans son sillage la contestation universelle des valeurs transmises par voie d'autorité* » [Imbs, Préface du TLF, 1971]. Décolonisation, contestation, valeurs, autorité :

autant de termes qui renvoient aux rapports des pays africains avec l'ancienne métropole, et qui ont des répercussions sur le lexique. L'indépendance politique implique en effet une forme d'indépendance linguistique, qui va osciller entre l'assimilation à la métropole et l'affirmation de l'identité africaine, entre le respect de la norme exogène, française, et le besoin d'affirmer les réalités socioculturelles endogènes africaines avec une langue étrangère devenue langue seconde, et parfois revendiquée comme langue africaine. Le lexique est une mesure de la distance entre ces deux pôles. La lexicographie peut en rendre compte.

Dans la même décennie se répandent deux mots : *francophonie* : « ensemble des pays de langue française » (*TLFi*), proposé 80 ans plus tôt par Onésime Reclus, dans *France, Algérie et colonies* ; et *francophone* : « [En parlant d'une collectivité] Dont la langue officielle ou dominante est le français » (*TLFi*). Le terme, qui apparaît avec ce sens à plusieurs reprises dans la préface du *TLF* (par exemple : « *langue commune d'extension nationale et francophone* » [Imbs, 1971]), est rappelé dans celle du *TLFi* : « *Un dictionnaire du monde francophone* » [Pierrel, 2004]. On peut donc considérer qu'il s'agit d'une constante.

Par ailleurs, en 1983, paraît l'*Inventaire des particularités lexicales du français en Afrique noire* (*IFA*) à partir de travaux initiés plusieurs années auparavant, et dont la genèse remonte, là aussi, au début des années 1960. L'*IFA* est complété aujourd'hui par de nombreuses autres descriptions lexicographiques, et reprises, partiellement pour l'instant, dans la *Base de données lexicales panfrancophone* (*BDLP*), née avec ce millénaire, et qui s'affiche dès sa dénomination comme une entreprise francophone. Elle a, entre autres avantages, celui de donner, avec les nombreuses contributions lexicographiques parues à ce jour, un matériau lexical auquel le *TLF* n'avait guère accès auparavant.

Mon propos consistera à évaluer la dimension francophone qu'a prise ou que peut prendre le *TLFi*.

1. Les contenus lexicographiques

L'usage du terme « francophonie » pose la question des contenus du *TLFi* par rapport à sa couverture géographique et, au-delà, de ses perspectives francophones. Par « contenus » j'entendrai ici :

1.1 La nomenclature

C'est l'*« une des caractéristiques les plus révélatrices de la nature d'un dictionnaire »* [Imbs, 1971]. La nomenclature du *TLFi* propose 20 entrées ou sous-entrées issues de l'*IFA* :

margouillat, subst. masc.
moqueur, *-euse*, adj. et subst.
palabre, subst.
parent, -ente, subst. et adj.
parenté, subst. fém.
paresseux, *-euse*, adj. et subst.
pater[2], subst. masc.
pileur, euse, dans article *piler* verbe
pili-pili, *pilipili*, subst. masc.
pistache, subst. fém.
piste, subst. fém.
pistard, dans article *piste*
potto, subst. masc.
ressortissant, *-ante*, part. prés., adj. et subst.
singe, subst. masc.
solde[1], subst. fém.
taxi-brousse, subst. masc. dans article *taxi*
teint, subst. masc.
torcher, verbe trans.
viande, subst. fém.

Le *TLFi* propose également, puisées à d'autres sources, *azobé, bambara, bantou, boubou, boy, canari, cauri, cola, élaeis, filao, franc CFA, makoré, marigot, safari, sorgho,* etc. : l'Afrique est ainsi mentionnée dans 343 définitions (ou 338 « textes de définitions »), dont la plupart renvoient à des référents africains plus qu'à des créations endogènes d'usage courant sur le continent.
Mais beaucoup de mots sont absents, et c'est en vain qu'on chercherait, parmi de nombreuses autres créations courantes en Afrique et de bonne facture : *abacost, essencerie, mamba, nacco/nacot, noix de palme, primature, table-banc, taximan, tradipraticien, (pagne) wax* ; des sens supplémentaires : *broussard, gouvernance, pisteur, maraboutisme, sous-région* ; ou encore des formes particulières comme *du n'importe quoi* « n'importe quoi », *de toutes les façons* « de toutes façons », *accoucher un enfant* « donner naissance à un enfant », *téléphoner qqn* « téléphoner à qqn », etc., dont certaines d'ailleurs sont entendues en France, selon les cas par nécessité référentielle ou en raison d'une évolution de l'usage.

1.2 Les définitions

« *Du fait qu'elles ne visent pas à saisir la réalité, mais des vues sur la réalité ou des créations imaginaires, les définitions linguistiques sont toujours sujettes à variation, liées qu'elles sont à la situation contingente des sujets parlants qui ont élaboré ces vues ou ces créations imaginaires* » [Imbs, 1971]. Je mentionnerai ici deux exemples, illustrant deux aspects. Le premier est celui de *mangue*, ainsi définie :
- par le *TLFi* :

> *BOT.* Drupe du manguier, de la grosseur d'une poire, couleur jaune ou orange, à gros noyau, à la chair très savoureuse.

- par le *Petit Robert électronique*

> 1996 : Fruit du manguier, de la taille d'une grosse pêche, à peau lisse, à chair jaune très parfumée et savoureuse, à odeur de térébenthine.
> 2007 : Fruit du manguier, à peau lisse, à chair jaune orangé très parfumée et savoureuse, à odeur de térébenthine.

Qu'il soit pêche ou poire, le comparant relève bien d'une vue ethnocentrée de la réalité, corrigée par le *Petit Robert* 2007, qui conserve la discutable « odeur de térébenthine ».

L'autre exemple est celui de *case*, qui connaît en Afrique une acception supplémentaire :
- par le *TLFi* :

> Habitation rudimentaire, en particulier en Afrique noire *(cf. cabane, cahute, hutte, paillotte)*
> ─*P. ext., vieilli, fam.* Maison de style rustique
> ─*Région.* (fr. d'Afrique). ,,Maison de construction légère`` (J. DAVID, *Dict. du fr. fondamental pour l'Afrique,* Paris, Didier, 1974).

- par l'*IFA* :

> 1° Habitation traditionnelle en paille (ou tout autre matériau traditionnel).
> 2° *par ext., parfois plais.* N'importe quelle maison particulière, y compris la villa de type européen.

Le caractère « rudimentaire » ou « léger » apparaît systématiquement dans l'imaginaire occidental, et donc dans la définition lexicographique *vs* la possibilité d'une architecture élaborée dans l'usage extra hexagonal. On notera dans cet ordre d'idée, les renvois à *cabane, cahute, hutte, paillotte* dans le *TLFi*, mais, dans l'*IFA* et divers inventaires africains, à *case de passage* « maison réservée aux visiteurs », *case à étages* « immeuble », *case en dur* « en béton », *case de santé* « dispensaire », *case à palabres, grande case,* etc.

1.3. Les exemples

« *On les a voulus […] authentiquement historiques (ou « culturels »), c'est-à-dire démonstratifs d'usages datés et liés à des conditions de milieu* » [Imbs, 1971]. Pourtant la réalisation révèle également ici une histoire et une culture francocentrée.

D'abord en ce qui concerne les « usages datés », puisque le *TLFi* donne deux définitions de *camisole* qui renvoient au passé : « Vêtement court ou long et à manches, qui se port**ait**[1] sur la chemise » et **«** *En partic., **vx.***[2] Court vêtement de nuit porté par les femmes sur la chemise de jour », illustrées respectivement par des citations de 1870 et 1859 :

> - *Pourtant à la fin tout le monde se calma. Le père s'essuya la figure; il mit sa camisole, son bonnet des dimanches* (ERCKMANN-CHATRIAN, *Histoire d'un paysan,* t. 1, 1870, p. 385).
> - *Cette horrible femme vêtue d'une camisole de nuit, (...), et portant par-dessus sa camisole un châle tartan à carreaux verts* (PONSON DU TERRAIL, *Rocambole,* t. 1, L'Héritage mystérieux, 1859, p. 210).

Pourtant, des sources africaines fournissent des exemples attestant un usage actuel en Afrique :

> - *Sa camisole et son pagne d'un vert foncé faisaient ressortir la clarté de son teint.* (Sow Fall, *Le revenant* (1976), cité par *IFA*).
> - *Christiane me parut laminée, sans défense, essoufflée, la camisole mouillée par la transpiration. »* (Alain Mabanckou, *Les petits-fils nègres de Vercingétorix* (2002), p. 98).
> - *Une porte s'ouvre, [...] une manche de camisole, un bras, des doigts vernis. Une femme en lunettes apparaît* (*Ivoir'Soir*, 16-12-1997) [Lafage, 2002-2003].

Ensuite, les remarques vont dans le même sens en ce qui concerne les « conditions de milieu », illustrées ci-dessous par la rubrique *case*, déjà évoquée :

> 1. Allez aux pays des Noirs, gîtés en des cases *de boue*; aux pays des Arabes blancs, abrités sous une toile brune qui flotte au vent, ... MAUPASSANT, *Contes et nouvelles,* t. 2, Fou, 1885, p. 1012.
> 2. ... je souris à ma maison, car il n'en est pas de plus mienne que cette grande case de granit gris, persiennes dépeintes et ouvertes, nuit et jour, sur des fenêtres sans défiance. COLETTE, *Claudine en ménage,* 1902, p. 274.

L'exemple pour le français d'Afrique, certes bienvenu, ne remet cependant pas en cause la vision française de la *case* :

> ─*Région.* (fr. d'Afrique). „Maison de construction légère`` (J. DAVID, *Dict. du fr. fondamental pour l'Afrique,* Paris, Didier, 1974) : *il habite la plus grande case du village* (J. DAVID, *Dict. du fr. fondamental pour l'Afrique,* Paris, Didier, 1974).

La comparaison avec les exemples suivants est révélatrice[3] :

> - *Il a fait construire une case moderne avec tout le confort à Lomé pour la louer à une ambassade.* (exemple *IFA*).
> - *Il a une belle case avec piscine aux Deux Plateaux* (Ingénieur, Abidjan) [Lafage, 2002-2003].
> - *La case des Sainte-Rose était une imposante villa coloniale aux toits de tôle à double pente avec d'innombrables portes et fenêtres à persiennes.* (Henri Lopes, *Le lys et le flamboyant*, 1997).

[1] Je souligne.
[2] Je souligne.
[3] Les photos de différentes « cases », que l'on peut consulter sur la *BDLP*, à partir de l'inventaire de la Réunion, sont instructives.

On admettra que ces illustrations sont représentatives de l'usage propre lié aux « conditions de milieu » des auteurs sélectionnés par les lexicographes.

1.4 Les auteurs

Il s'agit d'auteurs « *usant de la langue sans préoccupation linguistique directe et donc non suspects de gauchir les matériaux de la preuve dans le sens de la thèse à prouver* » [Imbs, 1971]. Force est de constater que les auteurs africains sont peu représentés dans le *TLFi*. On trouve parmi les élus :

- de l'Afrique subsaharienne
  - Léopold Sedar Senghor : 1 fois, pour *francité*
  - Henri Lopes : 1 fois, pour *tontine*
- du Maghreb
  - Albert Memmi : 1 fois, pour *tunisois*
- des Antilles françaises, un peu mieux représentées grâce à René Maran
  - Aimé Césaire : 2 fois, pour *gestapo* et *négritude*
  - Frantz Fanon : 1 fois, pour *djellaba*
  - René Maran : 149 fois (dont 148 de *Batouala*, prix Goncourt, 1921)

Ce sont par contre des « auteurs-voyageurs » français qui évoquent l'Afrique (même si la plupart de leurs exemples ne concernent pas ce continent) :

- Maurice Genevoix : 1585 fois (dont 27 *Fatou Cissé*)
- André Gide : 6835 fois (dont 80 *Voyage au Congo* et 67 *Retour du Tchad*)
- Louis-Ferdinand Céline : 2652 fois (dont 218 *Voyage au bout de la nuit*)
- Joseph Kessel : 3 fois (dont 1 *Le lion*)
- Paul Morand : 2565 fois (dont 44 *Magie noire* et 49 *Paris-Tombouctou*),

auxquels on ajoutera, dans un registre particulier

- Maurice Houis : 2 fois, pour *bambara*

La place manque ici pour citer tous les absents : Ahmadou Kourouma, Ferdinand Oyono, Mongo Beti pour l'Afrique noire ; Rachid Boudjedra, Tahar Ben Jelloun, Mouloud Feraoun pour le Maghreb ; etc. Mais on perçoit, entre l'intention des propos de Paul Imbs et la réalisation lexicographique, les écarts qui se présentent selon que le point de vue, et par suite la perspective de lecture didactique, est hexagonal ou francophone :

- dans le premier cas la description est ethnocentrée, pluriculturelle dans le second ;
- concernant le public d'utilisateurs, souvent mentionné dans la Préface de 1971, l'ouvrage terminal sera ou un instrument assimilateur, avec un point de vue français, ou un vecteur francophone, décrivant l'ensemble des usages.

2. Les publics

Cela nous amène au public. Même sur un continent, l'Afrique en l'occurrence, où les nouvelles technologies sont moins accessibles qu'en France, le potentiel de consultation du support informatique est plus élevé que celui du support papier. Le public est donc plus large : les consultations quotidiennes passent de 50 en 2002 à 160000 en 2004 [cf. Pierrel, 2004], et toutes, vraisemblablement, ne sont pas le fait des compatriotes de Voltaire. Mais ce public qui n'est pas que français doit aussi se reconnaître dans un français qui n'est plus seulement « de France », et dont la description lexicographique doit respecter l'identité.

Dans l'autre sens, les usagers français (et particulièrement le public cultivé auquel le *TLF* s'adresse) ne peuvent ignorer totalement les usages lexicaux africains. La vocation didactique du dictionnaire est ici concernée : « *le dictionnaire de langue est d'abord un dictionnaire de*

lecture, d'interprétation, *de décodage, et il répond à la question : si je rencontre tel mot, qu'est-ce qu'il veut ou peut dire ?* » [Imbs, 1971]. On ne peut alors passer sous silence la littérature africaine d'expression française qui prend aujourd'hui de l'ampleur, comme en témoignent les prix littéraires[4] et les rayons des librairies, pas plus qu'on ne le peut pour les médias qui emploient, fût-ce de façon ponctuelle, des termes souvent incontournables appropriés aux realias africains.

3. Rupture e continuité : quelles solutions ?

3.1 Les « fertilisations croisées »

Il apparaît essentiel dans ce contexte de présenter et de définir des mots du français d'Afrique dont certains sont en usage sur toute la partie francophone du continent et réalisent ce que Paul Imbs définissait en 1971 comme un équilibre entre la rupture et la continuité. Il faut, dit-il, « *attester les permanences sous les ruptures* » ; et Willy Bal [1983, p. XIX] ne dit pas autre chose dans l'introduction de l'*IFA* : *« la diversité interne d'une langue n'a rien que de normal et ne constitue pas un cas pathologique ou exceptionnel. Elle n'est d'ailleurs nullement incompatible avec l'unité. »* Seulement, si le *TLFi* prend en compte essentiellement la dimension diachronique en couvrant une période (1789-1960), l'*IFA* gère quant à lui la dimension diatopique en couvrant un espace (l'Afrique francophone). Or, la vie de la langue, telle qu'elle peut être représentée dans un ouvrage lexicographique, doit prendre en considération l'une et l'autre afin de *« relever ce défi de l'Un et du Multiple »* [Bal, 1983]. Il est intéressant de constater, dans les présentations des deux entreprises, l'identité des points de vues : *« la confrontation de la vision plongeante des faits en diachronie et de leur vue étendue en synchronie pourrait procurer un accroissement qualitatif de connaissance, résultat d'une de ces « fertilisations croisées » »* [Imbs, 1971].
Avec donc aujourd'hui, d'une part le matériau lexical disponible, et d'autre part la ressource informatique, le choix d'inscrire le lexique du français hors de France (que je limite ici à l'Afrique) dans une description lexicographique n'est plus une question de contraintes matérielles ou éditoriales évoquant des volumes gigantesques (les 16 volumes du *TLF* sont, en effet, déjà énormes), mais une question de choix idéologique entre ethnocentrisme et ouverture francophone multiculturelle. Dès lors que ce dernier cas est privilégié, quelle forme lexicographique peut être envisagée ? Le concept et la cohérence de l'ensemble dépend du choix idéologique premier, et de là découle sa conception formelle. Je perçois en gros trois formes possibles.

3.2 Trois réalisations possibles

Deux formes extrêmes :

- deux descriptions indépendantes concernant l'une la France et l'autre la francophonie. Dans cette forme la description du français de France ignore celle du français hors de France. Cette option n'a de sens francophone qu'en termes de locuteurs français parlant le français dit « de France », ou de locuteurs étrangers parlant le français langue étrangère. Mais on s'aperçoit vite des intersections entre les deux descriptions, un certain nombre d'items figurant dans l'une comme dans l'autre : même les critères de sélection les plus rigoureux ne peuvent permettre de frontière toujours nette, et la rupture théorique ou idéologique se concrétise dans une continuité effective. On peut à ce sujet consulter entre

[4] En 2006 : le Renaudot au Congolais Alain Mabanckou (*Mémoires d'un porc-épic*), le Goncourt à l'Américain Jonathan Little (*Les bienveillantes*), le Femina à la Canadienne Nancy Huston (*Lignes de faille*), et le Goncourt des lycéens à la Camerounaise Lénora Miano (*Contours du jour qui vient*).

autres *Le Lexique français de Côte d'Ivoire* [Lafage, 2002-2003], où des entrées comme *éléphant, masque* ou *ventre* offrent des perspectives différentes à partir de mots identiques.

- une description synthétique réunissant l'ensemble des variétés diatopiques : cette option véritablement francophone apparaît comme un idéal, que je proposais lors des Journées de l'AUF à Ouagadougou [Frey, 2004]. Mais elle implique d'une manière générale, et à certains lieux lexicographiques en particulier, une forte restructuration de l'ensemble lexicographique, ce qui peut paraître déstabilisant pour qui ne s'intéresse pas au français hors de France, utopique pour les concepteurs car difficile à mettre en place (encore que les apports de l'informatique autorisent une souplesse de conception et de consultation que ne permet guère le support papier), ou franchement iconoclaste pour les puristes : en Afrique comme en France, le débat n'est pas clos concernant l'avilissement ou l'enrichissement de la langue française par les néologismes africains, rejoignant selon les cas les idéologies assimilationnistes ou identitaires, le néocolonialisme ou l'indépendance linguistique.

Une troisième forme est possible, intermédiaire, avec deux descriptions parallèles et complémentaires, dans une cohabitation proposant des renvois et pour qui le souhaite une consultation non linéaire permise par les supports informatiques. Cette forme implique des intersections choisies, qui apparaissent d'ailleurs dans le *TLFi* puisqu'il n'est pas opposé dans le principe à cette ouverture, tout en ne la réalisant qu'à ses marges :

le *TLFi* renvoie à l'*IFA*, avec 20 mentions dans le *TLFi* en tant que source de définitions, d'exemples ou d'auteurs) ; le nombre de ces renvois peut augmenter dans des proportions considérables, d'autant plus que depuis 1983, l'*IFA* a été complété par de nombreuses productions lexicographiques précises et actualisées ;

le *TLFi* renvoie au français d'Afrique (343 mentions de l'Afrique dans les définitions du *TLFi*) ;

Ces mentions, de même que les sources littéraires africaines, constituent une proportion infime en regard des 90000 entrées du *TLFi* et, dans une perspective francophone, ne suffisent par à donner une image complète de la mosaïque francophone.

3.3 Langue, cultures et créativité

Ces intersections, existantes mais insuffisantes, émanent des identités et des complémentarités liées :

- à la nécessité de désigner les realias africains, par exemple la faune : *colobe, tisserin*, et la flore : *filao, okoumé*. Si ces termes sont forcément reçus d'un point de vue normatif, leur présence dans les inventaires nationaux africains est parfois discutée en raison de leur présence même dans les dictionnaires généraux ;
- à l'histoire des mots dans le cadre de la conquête coloniale et de leurs conséquences post-coloniales : *concession, parcelle, case, gouvernance* ;
- à des similitudes socioculturelles qui apparaissent en Afrique après la France (du fait de l'influence de l'Occident), et en France après l'Afrique : *quartiers, cités, micro-crédit* ;
- aux contacts des cultures, du fait de la colonisation, et après 1960 surtout, de l'immigration, des voyages, des médias : *sous-région* (présent dans le *TLFi*, sa définition occulte le sens africain et son usage dans les médias français) ;
- à des habitus linguistiques tels que *du n'importe quoi* « n'importe quoi », *à plus* « à plus tard », *de toutes les façons* « de toutes façons » : introduits depuis plusieurs années dans la plupart des inventaires africains, l'usage de ces formes se répand en France même ;
- et sur le plan linguistique, aux principes mêmes de la formation néologique (avec ses ressorts lexicaux et sémantiques) : en se limitant au suffixe *–er*, dans la mesure où *torcher* « éclairer avec une torche » est mentionné comme régionalisme africain dans le *TLFi*, il

faudrait alors également en mentionner de nombreux autres : *saboter* « poser un sabot à une voiture », *enceinter* « mettre enceinte », *insolencer* « traiter avec mépris, se montrer insolent envers », *réfectionner* « effectuer des travaux de réfection », *dévierger* « dépuceler », etc.
Il apparaît bien ici que *« la créativité d'une langue se mesure, entre autres paramètres, à sa capacité d'invention de mots nouveaux par le procédé de la dérivation préfixale ou suffixale »* [Imbs, 1971]. Nous sommes bien ici à l'articulation entre la rupture et la continuité, sur le plan de la création lexicale comme sur celui de la description lexicographique. Bien sûr n'entre pas dans le dictionnaire qui veut, et il faudrait mettre en place un système de sélection des candidats à l'entrée. Cette question ne sera pas abordée ici. Mais cette solution de renvois systématiques et raisonnés est compatible avec l'esprit du *TLF*, si nous prenons à la lettre ce propos de P. Imbs : *« au lieu d'un dictionnaire unique où chaque mot serait l'objet d'une monographie qui le suivrait tout le long de son histoire à l'intérieur de la langue, il y aurait une suite de dictionnaires comprenant chacun le vocabulaire d'une des cinq ou six couches relativement homogènes en quoi se laisse découper l'histoire du vocabulaire français. »* Il y aurait alors possibilité de jouer sur les liens entre « *un dictionnaire de base ou dictionnaire général* » et « *une série de lexiques spéciaux* » que, « l*e cas échéant, des encyclopédies compléteraient* » [Imbs, 1971].

Conclusion

Si la préface originale revient fréquemment sur la dimension diachronique du lexique, il paraît nécessaire aujourd'hui d'inviter plus largement la dimension diatopique. Ceci ne paraît pas contraire à l'esprit de cette entreprise née dans les années 1960, mais n'ayant pu tenir compte de ces aspects parce que :

- le matériau lexical francophone n'était pas présent ou incertain ;
- l'outil informatique était insuffisamment performant ;
- l'idéologie n'était pas la même : il y a eu évolution des cultures et des conceptions depuis 1971, ce qu'avait bien exprimé P. Imbs dans sa préface : « *le dictionnaire reflète une* culture *ou une succession de cultures* », et plus loin : « *notre méthode s'est constituée au fur et à mesure de l'évolution doctrinale de la linguistique moderne* ».

Dans cette perspective francophone, on peut envisager, selon des proportions, des modalités et surtout des principes qui doivent être définis :

- intégrer dans la nomenclature ou la microstructure de plus nombreux termes de « français d'Afrique » ;
- compléter certaines définitions correspondant à des usages non hexagonaux ;
- ajouter des exemples francophones en tant que *« témoins et preuves »* [Imbs, 1971] ;
- et pour cela élargir le corpus aux auteurs francophones ;
- ménager dans le *TLFi* et des renvois vers les autres descriptions afin de couvrir la diatopie francophone.

A ce titre, la *BDLP*, prolongement très élargi et informatisé de l'*IFA*, présente le double avantage de réunir les occurrences francophones et de renvoyer en son sein à chacune des bases de données nationales. Cette complémentarité permettrait d'étendre encore le potentiel descriptif et la vocation didactique du *TLFi*, de gérer le compromis entre le « principe d'identité » et le « principe d'innovation » tout en conservant au *TLFi* sa cohérence qui, comme le formule P. Imbs « *ne peut jamais être que celle d'un point de vue.* » Si ce point de vue est francophone, l'élargissement est nécessaire. Et je laisserai au franco-congolais Henri Lopes le mot de la fin, tiré d'un ouvrage dont le titre même est intéressant :

> « Mots de France ou mots d'Afrique, j'écris pour courir après eux, pour les éplucher et disséquer leur chair, pour tenter de percer leurs mystères. Chaque fois que je joue avec

eux et que je les palpe, je suis comme l'aveugle qui tâche de reconstituer la forme de l'objet entre ses doigts et d'en imaginer la couleur. » (Henri Lopes, *Ma grand-mère bantoue et mes ancêtres les Gaulois*, Gallimard, Paris, 2003, p. 113).

Je pense qu'aujourd'hui, le *TLFi* a les moyens de favoriser cette reconstitution et cette imagination.

Bibliographie

[Bal, 1983] Bal, W. (1983) : « Introduction. Genèse et travaux de base », Equipe IFA, *Inventaire des particularités lexicales du français en Afrique noire*, EDICEF – AUPELF, Vanves.

[Frey, 2004] Frey, C. (2004) : « Les structures lexicographiques dans les dictionnaires francophones, une rencontre symbolique des mots et des cultures », *Penser la francophonie. Concepts, actions et outils linguistiques*, EAC – AUF, Paris, pp. 197-210.

[Imbs, 1971] Imbs, P. (1971) : « La Préface originale du *TLF*. L'œuvre et ses ouvriers », *Le Trésor de la langue française*, http://www.atilf.fr

[Lafage, 2002-2003] Lafage, S. *Le lexique français de Côte d'Ivoire, Appropriation et créativité*, Tomes 1 et 2, *Le français en Afrique*, Revue du Réseau des Observatoires du Français Contemporain en Afrique Noire, n° 16 et n° 17, Institut de Linguistique française , Nice .

[Pierrel, 2004] Pierrel, J.-M. (2004) : « La Préface du *TLFi* », *TLFi : Le Trésor de la langue française informatisé*, http://www.atilf.fr

Le traitement des emprunts au portugais dans le TLF(i)

Myriam Benarroch (1)
myriam.benarroch@atilf.fr

(1) ATILF Nancy Université & CNRS

La relative indigence de la lexicographie portugaise contemporaine et le retard considérable pris par le portugais au regard des autres langues romanes quant au dépouillement lexical des textes médiévaux et classiques expliquent certainement, en partie tout au moins, que cette langue n'ait pas été suffisamment prise en compte dans l'étymologie de ses langues sœurs. La langue portugaise a fourni une petite quantité de lexèmes au français, depuis le 15e siècle, et tout particulièrement au cours des 16e et 17e siècles, époque correspondant aux voyages des Portugais en Afrique, en Asie et au Brésil. Nous en avons, pour le moment, répertorié une centaine dans le TLF(i). Nous nous proposons dans le cadre de ce colloque à l'occasion du 50e anniversaire du projet du lancement du *Trésor de la Langue Française* de faire le point sur la question. Nous analyserons la manière dont sont traités les emprunts au portugais dans les notices étymologiques du TLF(i), ainsi que les problèmes soulevés par ce traitement, auquel nous tenterons d'apporter des améliorations substantielles. Enfin, nous essayerons d'établir une typologie des emprunts au portugais attestés dans ce dictionnaire.

1. Repérage des emprunts au portugais

La première question qui se pose est de repérer dans le vaste océan lexical des entrées du TLF(i) le petit ruisseau constitué par les mots empruntés à la langue portugaise. Dans un premier temps, les ressources électroniques du dictionnaire permettent, à travers la recherche de l'objet « langue empruntée », de relever la liste des lexèmes pour lesquels la notice étymologique mentionne un emprunt au portugais, signalé par l'abréviation « empr. au port. ». Nous relevons ainsi 31 lexèmes. Il nous est apparu très vite que cette liste n'était pas exhaustive, ne serait-ce que parce que manquaient à l'appel des emprunts avérés tels *fétiche*, *piranha*. Nous constatons déjà que cette recherche électronique a ses limites car même des mots comportant ce critère de recherche « empr. au port. » n'ont pas été repérés (*abricot*, *alastrim*, *alcatraz*, *anthroponymie*, *balise* et de nombreux autres), ou, à l'inverse, parce que ce critère n'est mentionné que pour être écarté (*albacore*, *albinos*).

Nous avons ensuite prospecté dans un corpus beaucoup plus étendu, en recherchant, toujours par voie électronique, les mentions « port. » et « portug ». Naturellement, nous avons dû commencer par éliminer un grand nombre de « parasites ». Dans le premier cas, en particulier parce que souvent le portugais est cité, dans les notices étymologiques consacrées aux lexèmes hérités, parmi d'autres langues romanes, dans la rubrique des correspondants romans « corresp. rom. » ; dans le second, on l'imagine aisément, parce que la référence au Portugal ou aux Portugais dans un texte, n'établit pas nécessairement un lien avec la langue portugaise. Cette double recherche (« port. » et « portug ») nous a néanmoins permis de relever un nombre non négligeable de lexèmes mentionnés comme emprunts au portugais, enrichissant ainsi la liste de départ réalisée à

partir du seul critère « empr. au port. ». Il conviendra, bien entendu, de vérifier la pertinence des étymologies attribuées à ces lexèmes.

D'autre part, certains lexèmes sont considérés comme des xénismes et portent la mention « mot port. » (*fado*) ou « port. » (*samba*), alors qu'ils sont passés en français dans la langue courante.

2. *Etimologia prossima* et *etimologia remota*

L'esprit de la refonte des notices étymologiques du TLF(i) (*cf.* http://www.atilf.fr/tlf-etym; Buchi *et alii* 2005) privilégie l'*etimologia prossima* et « exclut ainsi clairement de son champ d'étude l'*etimologia remota* (l'étymologie des étymons) qui relève, pour ce qui est du lexique héréditaire, de la linguistique latine et indoeuropéenne, pour ce qui est des anglicismes, de la linguistique anglaise et germanique, etc. » (Buchi 2005). C'est là un changement considérable par rapport à la rédaction de la partie étymologique des articles du TLF(i). Ce choix permettra d'éviter la confusion régnant dans la manière dont sont présentés les emprunts dans ce dictionnaire. Pour ce qui est du portugais, nous relevons, en particulier, une grande incohérence dans la diversité présidant à la présentation des étymons du français. À côté des articles où le lexème français est présenté clairement comme un emprunt au portugais (« empr. au port. »), y compris si celui-ci est donné comme issu d'une autre langue (*s. v. matchiche* : « Empr. au port. *maxixe* '*id.*' (1890 ds MACH.), lui-même issu d'un mot indigène du Brésil (v. *FEW* t. 20, p.70a) », nous constatons que souvent, le mot français est considéré comme emprunt par l'intermédiaire du portugais à une autre langue (c'est là une des raisons pour lesquelles le mot ne figure pas dans la liste électronique des emprunts au portugais). Donnons-en quelques exemples :

• *datura* : « Empr., par l'intermédiaire du port. (attesté dep. 1563, Garcia da Orta ds DALG.), au skr. *dhattūra* » ;

• *mangoustan* : « Empr., par l'intermédiaire du port. *mangostae* «fruit du mangoustan» (XVI^{e} s. *mangostae* ds MACH. 1977), au malais *manggoestan* '*id.*'» ;

• *mangue* : « Empr., par l'intermédiaire du port. *manga* fém. '*id.*'» (XVI^{e} s. ds DALG. ET MACH. 1977), au tamoul *mān-gay* ou *mān-kay* '*id.*'» ;

• *nabab* : « Empr. (par l'intermédiaire du port. *nababo* [1600 ds DALG.] pour l'attest. de 1614) à l'hindoustani *nawwāb, nabbāb* 'vice-roi, gouverneur', lui-même empr. à l'ar. *nuwwāb,* plur. de *nā'ib* 'lieutenant, représentant, remplaçant', part. actif de *nāba* 'prendre la place de (quelqu'un), représenter, remplacer' » ;

• *piranha* : « Empr., par l'intermédiaire du port. *piranha* '*id.*' (1587 ds FRIED.; aussi *Piray*, comme mot indigène, en 1555, *ibid.*), au tupi *piranha*, var. de *piraya* (v. FRIED. et MACH.[3]) ».

D'autre part, certains lexèmes sont présentés comme emprunts au portugais mais à travers des traductions de textes rédigés dans d'autres langues (*igname*, *jonc*, *négondo*). Ils n'ont donc pas été pris en compte dans la recherche « empr. au port. ».

Si l'origine de l'étymon portugais présente un intérêt dans une recherche consacrée à la typologie des emprunts du français à la langue portugaise (nous y reviendrons plus loin), il convient avant tout de mettre de l'ordre, pour la rédaction des notices du TLF(i), dans la présentation des étymons du français.

3. Dans quelle mesure le portugais est-il pris en compte dans la recherche de l'étymon ?

3.1 La datation : critère déterminant ?

La date de première attestation d'un lexème dans une langue est un critère permettant souvent de déterminer, dans le cas où il traduit une antériorité par rapport à d'autres langues, que c'est à cette langue que les autres l'ont emprunté. Ce critère n'est toutefois pas suffisant, et encore moins dans le cas des emprunts au portugais, pour lesquels il convient d'être extrêmement prudent. En effet, la faible quantité de documents médiévaux et classiques dépouillés et l'absence de dictionnaires historiques du portugais font que, très souvent, la date donnée comme première attestation est très éloignée de la date où le mot est entré dans la langue, et même de celle où il est attesté dans un texte. Le dictionnaire étymologique de référence du portugais, le DELP de José Pedro Machado (1977[3]), abondamment cité dans le TLF(i) comme source de datation et aussi pour l'étymologie du portugais, en est un bon exemple. Des progrès considérables ont été réalisés avec la publication du *Dicionário Houaiss da Língua portuguesa* (2001), qui antédate un nombre très important de lexèmes du DELP, grâce, notamment, au fichier sur le vocabulaire médiéval réalisé par Antônio Geraldo da Cunha, conservé à la Fundação Casa de Rui Barbosa à Rio de Janeiro et qui a, depuis la publication du *Houaiss*, été édité sous forme de CD-Rom (VHCPM 2002). Toutefois, le *Houaiss*, considéré aujourd'hui comme le meilleur dictionnaire étymologique du portugais (Monjour 2004), est encore très contestable du point de vue de la datation. D'une part, les ouvrages mentionnés dans la bibliographie n'ont pas été consultés systématiquement ; d'autre part, les sources principales pour la datation des lexèmes sont essentiellement des dictionnaires et non des textes permettant de situer le lexème dans un contexte vivant. Lors d'un précédent travail, nous avons pu antédater de plusieurs siècles certains lexèmes, uniquement à partir de la consultation des dictionnaires de Jerónimo Cardoso, publiés entre 1551 et 1570 et pourtant cités dans la bibliographie du *Houaiss*. En voici les exemples les plus spectaculaires : *alardeadeiro* : 1986 > 1562 (- 424 ans) ; *chancarona* : 1899 >1562 (- 337 ans) ; *dandão* : 1958 > 1562 (- 396 ans).

Dans le TLF(i), il arrive ainsi qu'un étymon portugais soit rejeté pour la seule raison que sa première attestation est tardive. C'est, par exemple, le cas de *anhinga*, dont l'origine tupi laisse penser à un très probable passage par le portugais du Brésil, et donc à un étymon portugais, mais pour lequel on peut lire dans la notice étymologique qui lui est consacrée : « Empr. au tupi *anhinga* ornith. (*cf.* FRIED. 1960); le port. n'a pu servir d'intermédiaire, ne datant que de 1871 (MACH. t. 1 1967) ».

3.2 L'espagnol privilégié dans la recherche de l'étymon

Le rejet d'un étymon portugais se fait, en particulier, au profit de l'espagnol, pour lequel les travaux lexicographiques et étymologiques sont beaucoup plus avancés. L'antériorité d'attestation d'un lexème en espagnol par rapport à son équivalent portugais peut ne pas signifier, pour un lexème français, qu'il est passé de l'espagnol au français mais simplement qu'il n'a pas encore été répertorié dans les textes d'ancien portugais.

Parfois, tout simplement, l'antériorité du lexème en portugais n'est pas prise en compte dans le choix de l'étymon. Ainsi, dans le TLF(i), *démarcation*, attesté en 1700 dans un contexte où il est question du traité de Tordesilhas qui délimite en 1494 les possessions portugaises et espagnoles, est donné comme un emprunt à l'espagnol *demarcación* (1609), alors que *demarcação* est attesté en portugais dès 1451 (*Houaiss*, mais déjà Machado donnait la date de 1473).

Soulignons encore que les critères phonologiques et morphologiques ne sont pas toujours suffisamment pris en compte dans la recherche de l'étymon. Ainsi *autodafé*, attesté , pour le TLF(i), en 1714 sous la forme *auto-da-fé*, et en 1759 pour la forme actuelle, est donné comme un emprunt à l'espagnol *auto de fé* « ([…] en raison du texte d'où est tirée la 1re attest., influencé par plusieurs romans esp.) croisé avec le port. *auto da fe* '*id.*' ». La lexie composée du portugais

porte, dans le *Houaiss* (s. v. *auto-de-fé*), la date de 1544 et la forme *auto da fee*. Elle est absente du DELP et du dictionnaire de Nascentes. Quelle date pour l'espagnol ? Pas de trace de la lexie chez Corominas (DCECH) et, par conséquent, aucune mention de la date de première attestation en espagnol dans le TLF(i). Or l'élément *da*, contraction de la préposition *de* et de l'article défini féminin *a*, est exclusive du portugais (l'équivalent espagnol serait **auto de la fe*) et c'est sous cette unique forme que la lexie est attestée en français au 18e siècle. Dès lors, comment ne pas penser à un emprunt direct au portugais ?

3.3 Le portugais comme intermédiaire

Le portugais n'est pas suffisamment pris en compte comme intermédiaire possible entre une langue non romane et le français. Ceci est vrai en particulier pour les emprunts du français à l'arabe, où la recherche d'un éventuel intermédiaire roman s'oriente surtout vers l'italien et l'espagnol, comme en témoignent ces lignes de Frankwalt Möhren, à propos du lexème *jarde* du TLF(i) :

> La proposition étymologique « Emprunté à l'it. *giarda*… 13e s. » correspond à une pétition de principe selon laquelle « un emprunt arabe du français est passé par l'italien ou par l'espagnol, chose certaine quand les attestations sont plus anciennes dans ces deux langues ». Or une telle affirmation ne saurait être valable en général : à moins d'entreprendre des recherches poussées sur les traditions discursives des vétérinaires à l'échelle européenne, il faut rester très prudent. C'est donc l'ancienneté relative des attestations italiennes qui a dicté l'étymologie (Möhren 2006 : 3).

Dans le cas de *jarde*, Möhren, qui dénonce l'a priori privilégiant l'italien et l'espagnol, postule pour un emprunt au moyen-français. Mais il existe des lexèmes pour lesquels c'est le portugais qui a servi d'intermédiaire entre l'arabe et le français, et dont l'étymon direct est donc bien portugais (*anil*, *marabout*, *tincal*). En outre, pour certains lexèmes, le TLF(i) suggère, sans hiérarchiser les hypothèses, un étymon espagnol, portugais ou latin, alors que le contexte historique plaide pour une origine portugaise. C'est le cas, par exemple, de *négondo* (« Empr., par l'intermédiaire du port., de l'esp. et du lat. *negundo (cf. supra),* au concani [lang. du territoire de Goa] »), attesté en 1602 : la référence à Goa, capitale de l'Etat portugais de l'Inde dès 1510 privilégie l'hypothèse du portugais comme langue prêteuse.

Il conviendra donc de reconsidérer systématiquement tout lexème donné comme d'origine « espagnole ou portugaise » afin de déterminer à laquelle des deux langues ibériques appartient l'étymon.

4. Typologie des emprunts au portugais

Pour être menée à bien, cette partie de notre recherche exige un travail approfondi que nous n'avons pas encore mené. Nous nous proposons de l'aborder sous différents aspects :

1. Graphie et phonétique : adaptation des lexèmes portugais aux systèmes graphique et phonétique du français ; emprunts de formes (*colles* pour *coolie*) ;
2. Morphosyntaxe : en particulier, la classe grammaticale la plus représentée (les substantifs) et le changement de genre dans le passage d'une langue à l'autre (*samba*, masc. en port. et fém. en fr.) ;
3. Sémantique : évolution sémantique et champs sémantiques les plus représentés ; emprunts d'acceptions (*alcatraz* 'albatros' ; *banian*[1] 'habitant de l'Inde, de religion brahmanique et s'adonnant au commerce' ; *dom*[2] 'titre de courtoisie donné au Portugal' ; *paillotte* 'hutte de paille dans certains pays d'outre-mer') ;
4. Emprunts à l'onomastique portugaise ou brésilienne (*andradite*, *martinisme*, pour les anthroponymes, *porto*, *acunien* pour les toponymes).
5. *Etimologia remota*.

Quelques mots sur ce dernier point. En dépit des réserves émises par les rédacteurs du projet de révision des notices étymologiques du TLF(i), il nous semble utile, dans le cas du portugais, de prendre en considération les langues dont sont issus les étymons portugais du français, lorsque ces derniers sont d'origine autre que latine ou romane. Ces langues présentent le double intérêt de retracer l'histoire culturelle des lexèmes et de permettre de corroborer ou d'infirmer certaines étymologies présentées dans le TLF(i). Nous distinguons d'ores et déjà trois aires géographiques significatives qui témoignent des contacts linguistiques qu'ont eus les Portugais avec d'autres peuples au cours de leur histoire :

1. les langues africaines, en particulier les langues bantoues ;
2. les langues asiatiques, dravidiennes et indo-européennes, en particulier ;
3. les langues du Brésil : le tupi, d'abord, mais aussi les africanismes ainsi que les lexèmes d'origine latine propres au portugais du Brésil.

À ces trois aires géographiques productives en matières d'étymons du portugais, il nous faut ajouter une quatrième catégorie qui transgresse les continents et traverse l'histoire de la langue portugaise : il s'agit de l'arabe, dans ses variétés hispanique, maghrébine et orientale, grand pourvoyeur de lexèmes portugais dont certains sont souvent omis dans le TLF(i) en tant qu'étymons directs du français.

Bibliographie

Buchi, Eva *et al.* (16 décembre 2005) : *Projet TLF-Étym : mise à jour des notices étymologiques du* Trésor de la Langue française informatisé. *Dossier de présentation*, Nancy, ATILF (CNRS/Université Nancy 2/UHP).

DELP[3] = Machado, José Pedro (1977[3] [1952[1]]) : *Dicionario etimológico da língua portuguesa com a mais antiga documentação escrita e conhecida de muitos dos vocábulos estudados*, 5 volumes, Lisboa, Livros Horizonte.

Houaiss = Houaiss, Antônio ; Villar, Mauro de Salles ; Franco, Francisco Manoel de Mello (2001) : *Dicionário Houaiss da língua portuguesa*, Rio de Janeiro, Objetiva.

Möhren, Frankwalt (2006) : L'importance de la critique des sources en étymologie. In : Buchi (Éva) (éd.) : *Actes du Séminaire de méthodologie en étymologie et histoire du lexique (Nancy/ATILF, année universitaire 2005/2006)*, Nancy, ATILF (CNRS/Université Nancy 2/UHP), publication électronique (http://www.atilf.fr/atilf/seminaires/ Seminaire_Möhren_2005 -11.pdf), 17 pages.

Monjour, Alf (2004) : « El diccionário *Houaiss* y la etimología portuguesa », in *Novi te ex nomine*. *Estudos filolóxicos ofrecidos ao Prof. Dieter Kremer*, Ana Isabel Boullón Agrelo, A Coruña, Fundación Pedro Barrié de la Maza.

Nascentes, Antenor (1932): *Dicionário Etimológico da Língua Portuguesa*, com prefácio de W. Meyer Lübke, Rio de Janeiro.

VHCPM = Cunha, Antônio Geraldo da (dir.) (2002) : *Vocabulário Histórico-Cronológico do Português Medieval* (CD-Rom), Rio de Janeiro, Fundação Casa de Rui Barbosa, Ministério da Cultura.

Quoi de neuf du côté de la lexicographie étymologique ?
La méthodologie utilisée dans le cadre du projet TLF-Étym pour distinguer les emprunts au latin de l'Antiquité de ceux faits au latin médiéval

Nadine Steinfeld (1)
nadine.steinfeld@atilf.fr
Marta Andronache (1)
marta.andronache@atilf.fr

(1) ATILF Nancy Université & CNRS

Mots-clés : voie d'emprunt, latin de l'Antiquité, latin médiéval, domaine français, lexicographie, lexicologie

Keywords: loan ways, Classical Latin, Medieval Latin, French vocabulary, Lexicography, Lexicology

Résumé : L'évolution de la pensée lexicographique et la refonte de certaines notices étymologiques a suscité des réflexions qui enrichissent et précisent la voie/les voies d'emprunt au latin. Pour l'heure il n'y a pas d'étude dans le domaine français consacrée à ce problème d'étymologisation.
Nous proposons une réflexion concernant ce problème dans le domaine de la lexicographie française autour de quelques exemples des notices étymologiques du projet TLF-Étym. Elle s'inscrit à la fois dans la thématique « les projets lexicographiques dans le sillage du TLF » et dans celle des « grands chantiers lexicographiques actuels et leurs méthodologies » et se présente comme un bilan et surtout comme une mise en perspective concernant la problématique des voies d'emprunt du latin.

Abstract : Today, the development of the lexicology imposes more precision of the methodology and the lexicographical terminology. We propose an approach for the loans from the Latin in the French vocabulary. Our conference focuses concomitantly the "Lexicographical projects inspired by the TLF" and the "Major research areas and methodologies in current lexicographical practices" fields.
The loans from the Latin are defined by the opposition with the hereditary vocabulary, but it is necessary to distinguish precisely the differences between different types of loans from the Latin. In the etymological dictionaries, we find a multitude of expressions to define the French loans from the Latin; we took our examples from the *Trésor de la Langue Française* (TLF) and we found miscellaneous terminologies: loans from scientific Latin, late Latin", imperial Latin, Christian Latin, ecclesiastical Latin, etc. We propose a discussion to harmonise and redefine the terminology of the loans from the Latin.
Our topic repose to the recent theories of the etymological Lexicography developed by the project TLF-Étym (Trésor de la Langue Française – Etymologie) of the linguistic laboratory ATILF (Analyse et traitement informatisé de la langue française) / CNRS (France).

We find three typical cases in which the lexicographer is confronted in the lexicographical practice : 1) Loans from Latin of the Antiquity; 2) Loans from the Medieval Latin; 3) Unclear cases.The analysis of any case is based on the relevant examples of the etymological notices reworked for the TLF-Etym.
In the sight of our work, we incite the scientific community to work on these problematic unclear cases in order to precisely define and harmonise the terminology and the work methodology.

Introduction

Le projet TLF-Étym[1] (*cf.* Buchi 2005 ; Steinfeld 2006) se propose de réviser les notices étymologiques et historiques du TLFi en y injectant des datations et des étymologies nouvelles. Ayant répertorié dans les notices en question du TLFi non seulement des insuffisances mais aussi des retards par rapport aux développements récents de la recherche et désireuses de combler certaines déficiences — sans toutefois refaire la totalité des notices — et surtout animées de la volonté de mettre à jour la méthodologie étymologique à partir de cas cruciaux, nous avons saisi l'opportunité de ce colloque international pour proposer une présentation d'un remodelage de la catégorie : emprunt au latin. Celle-ci se définit par opposition avec mot héréditaire, etc., mais il faut faire des distinctions à l'intérieur de cette catégorie qui sont déjà faites en partie (mais, comme on va le montrer, de façon souvent impropre) par le TLF. La tradition du TLF fait que les emprunts français au latin sont traités comme : emprunt au latin scientifique, au « bas latin », au latin impérial, au latin chrétien, au « bas latin » médical, au latin ecclésiastique, etc. Il convient de reconsidérer ces divisions et de les redéfinir.

0. PRELIMINAIRES

La pratique quotidienne et l'évolution de la pensée lexicographique font qu'actuellement les choses ne se présentent plus de manière aussi hétéroclite : la refonte de certaines notices étymologiques a suscité des réflexions qui enrichissent et précisent la voie/les voies d'emprunt au latin. Pour l'heure il n'y a pas d'étude dans le domaine français consacrée à ce problème d'étymologisation.

En effet, même le manuel de référence en étymologie romane (Pfister/Lupis 2001) n'évoque pas ce problème ; ce qui vaut pareillement pour Wunderli 2001. C'est aussi le cas de Bork 2006, qui n'aborde pas du tout les cas indécidables sur les voies d'emprunt du latin – et pourtant, tous les étymologistes y sont régulièrement confrontés.

Par contre, dans le domaine italien Jan Reinhardt (Reinhardt 2004) pose le problème des emprunts au latin médiéval. Dans son ouvrage sur la lexicographie historique italienne et le latin médiéval, il aborde le lexique italien pour lequel il est difficile de décider avec précision la voie d'emprunt au latin de l'Antiquité ou au latin médiéval (Reinhardt 2004 : 27-33).

C'est pourquoi notre réflexion concernant ce problème dans le domaine de la lexicographie française se cristallise autour de quelques exemples des notices étymologiques du projet TLF-Étym. Elle s'inscrit à la fois dans la thématique « les projets lexicographiques dans le sillage du TLF » et dans celle des « grands chantiers lexicographiques actuels et leurs méthodologies » et se présente comme un bilan et surtout comme une mise en perspective concernant la problématique des voies d'emprunt du latin.

[1] http://www.atilf.fr/tlf-etym/

PRÉLIMINAIRES TERMINOLOGIQUES

Le latin de l'Antiquité et le latin médiéval sont deux variétés du diasystème latin, qui se distinguent :

- sociolinguistiquement (langue vernaculaire *versus* langue véhiculaire) ;
- diachroniquement (conceptuellement : avant/après la naissance des langues romanes ; par convention : avant 600 *vs* après 600) ;
- diaphasiquement (totalité de l'éventail diaphasique *versus* situations de communication « hautes »).

Nous sommes parties dans notre démarche de trois cas typiques auxquels nous sommes susceptibles d'être confrontées dans le cadre de la rédaction de l'étymologie d'un emprunt au latin :

1. **Clairement emprunt au latin de l'Antiquité**
2. **Clairement emprunt au latin médiéval**
3. **Cas indécidables**

1. CLAIREMENT EMPRUNT AU LATIN DE L'ANTIQUITÉ

Il s'agit clairement d'un emprunt au latin de l'Antiquité (appelé *latin* tout court dans la métalangue de TLF-Étym)si la ou les première(s) attestation(s) incite(nt) à penser que l'emprunt a été effectué à travers la traduction d'un texte antique.

différer

La notice historique et étymologique révisée du projet TLF-Étym[2] propose *ca* 1355 comme première attestation absolue de *différer*, verbe transitif direct, relevé avec le sémantisme « remettre à plus tard » dans une traduction d'un texte latin. Il s'agit en l'occurrence de la traduction du 14e siècle des *Décades* de Tite Live (59 avant J.-C. — 17 après J.-C.) : « et leur dist que pour certain Turnez avoit ier a lui et a eulz la mort appareillé, afin que il peust tous seulz l'empire des Latins avoir et maintenir, mes que la chose **avoit esté differee** pour l'absence de lui, lequel principalment il desirroit a occire » (BERSUIRE, *Tite-Live* 1 : 85, § 51.3).

Cet exemple illustre donc notre cas premier puisque l'emprunt s'est fait par voie directe du latin de l'Antiquité, *differre* verbe transitif « remettre à plus tard » attesté depuis Cicéron dans ce sens précis (*Thesaurus Linguae Latinae* 1904 : 5/1, 1071-1072), avec changement de conjugaison.

2. CLAIREMENT EMPRUNT AU LATIN MÉDIÉVAL

Il s'agit clairement d'un emprunt au latin médiéval :

- **2.1.** si l'étymon a été créé en latin médiéval (dérivé, emprunt à une langue vernaculaire) ;
- **2.2.** si l'étymon a développé un sémantisme secondaire (inconnu du latin de l'Antiquité) en latin médiéval ;
- **2.3.** si l'étymon présente en latin médiéval une particularité morpho-syntaxique inconnue du latin de l'Antiquité ;

[2] Voir *Annexe*, *infra*., p. 5

Dans notre résumé nous ne présentons que deux exemples illustrant les cas 2.3. et 2.4., les autres cas sont destinés à être traités lors de la communication.

Étude de cas : *obole* (illustre le cas 2.3.)

La révision de la notice étymologique et historique de l'article *obole* montre que le mot a été emprunté à deux reprises et à deux époques différentes : tout d'abord au Moyen Âge et plus précisément vers 1200, puis de nouveau à la Renaissance. En ce qui concerne le substantif féminin *obole*, qui désigne une pièce de monnaie en usage en France, sous l'Ancien Régime, il s'agit d'un emprunt au latin médiéval, qui connaît les trois genres, masculin, féminin, neutre, contrairement au latin de l'Antiquité, qui n'a que le masculin pour ce terme. Le genre féminin qu'a le mot actuellement en français s'explique en effet par l'emprunt au latin médiéval, où *obola* (subst. fém.) alternait avec *obolus* (subst. masc.) et *obolum* (subst. neutre) dans les livres de comptes.

Cet exemple illustre donc notre deuxième cas puisque l'emprunt s'est fait par la voie du latin médiéval *obolus, -a, -um* subst. masc., fém. ou neutre « petite pièce de monnaie » (attesté dès 1086 [masc. et fém.] et dès 1218 [neutre], LATHAM 1965, *Word-list*), variante tardive du latin classique *obolus* subst. masc. « *id*. ».

3. CAS INDÉCIDABLES

Cela laisse pourtant de la place aux cas pour lesquels il est difficile de décider entre la voie d'emprunt au latin de l'Antiquité ou au latin médiéval. Il convient de faire état de ces doutes, sans généraliser le « principe du doute » de façon mécanique. Ce qui peut orienter quand même :

3.1. il s'agira plutôt d'emprunt au latin de l'Antiquité si l'emprunt semble appartenir au français (littéraire) général ;

3.2. il s'agira plutôt d'emprunt au latin médiéval si l'emprunt reste confiné, au moins au début, à des genres textuels savants, représentants de sociolectes comme celui des juristes, celui des médecins, etc. : on peut supposer que la communication à l'intérieur de ces milieux passait par des textes contemporains rédigés en latin (= latin médiéval).

Étude de cas : *défectif* (illustre le cas 3.2.)

La notice historique et étymologique révisée du projet TLF-Étym propose fin du 13e siècle comme première attestation absolue de *défectif* adj. « défectueux (en parlant d'un inanimé) » : « Li point qe comande qe la grande chartre soit tenue en touz ces poinz est defective par defaute de adicion de peine» (*The Mirror of justices*[3], page 183 = AND2 *s.v. addicion*).

Le sens grammatical est attesté depuis la 2e moitié du 14e siècle dans un glossaire latin-français : « cedo. cedite : verbe **defectif** » (*Aalma*[4], volume 2, page 55, n° 1561 = STÄDTLER 1988 : 196).

La révision de la notice étymologique et historique de *défectif* montre la difficulté de décider entre les deux voies d'emprunt : latin de l'Antiquité *defectivus* ou latin médiéval *defectivus*.

[3] Whittaker (William Joseph) (éd.), 1895. *The Mirror of justices*, Londres, Quaritch (Selden Society : 7). Date du texte : fin 13e siècle. Date du manuscrit : 1er tiers 14e siècle. Région d'origine du texte : Angleterre (anglo-normand).

[4] Roques (Mario), 1938. *Aalma, Lexiques alphabétiques*, *in* : *Recueil général des lexiques français du moyen âge (XIIe-XVe siècle)*, édité par Mario Roques, volume 2, Paris, Champion (Bibliothèque de l'École des Hautes Études : 269).

L'emprunt au latin *defectivus* adj. « imparfait, vicieux » (attesté depuis *ca* 400 [saint Augustin ; peut-être déjà Tertullien], TLL 5/1, 290), « (mot) dont la conjugaison ou la déclinaison ne possède pas toutes les formes (terme de grammaire) » (attesté depuis le 4e siècle [Charisius], TLL 5/1, 290) s'est fait soit à travers des auteurs tardifs comme Saint Augustin ou Saint Grégoire pour le sens général et Donat pour le sens grammatical, soit à travers des emprunts techniques dans des contextes savants en latin médiéval (cf. NIERMEYER[2]).

Cet exemple illustre notre troisième cas puisque l'analyse des matériaux n'a pas permis de trancher pour l'un ou l'autre des canaux d'emprunt.

Ces cas restent heureusement rares et concernent, à en juger d'après les exemples que nous avons pu réunir, des termes techniques (en grammaire, en médecine, en droit).

Conclusion

Dans la pratique quotidienne du lexicographe le premier et le deuxième cas ne posent pas de problème [mais il faut déjà les établir, ce qui n'est pas si simple que ça !] étant donné qu'il existe des critères clairement établis pour déterminer la voie d'emprunt.

Pour le cas premier, si le mot est entré en français par l'intermédiaire d'une traduction d'un texte antique, la voie d'emprunt ne fait aucun doute. Nombreuses sont les études portant sur la traduction au Moyen Âge, par exemple la recherche que mène Bertrand (2004) pour évaluer les différents procédés de formation du vocabulaire français de la science politique durant la période du règne de Charles V. Ce chercheur constate :

> Le XIVe siècle français est une période de l'histoire de la langue qui n'a de cesse de traduire les œuvres qui appartiennent à la conscience collective et représentent d'une certaine manière le patrimoine intellectuel de l'humanité jusque là réservé à une élite savante (Bertrand 2004 : 23)

Pour le deuxième cas si on dégage un des quatre critères indiqués ci-dessus (*cf*. 2) ou la combinaison de plusieurs d'entre eux, nous sommes en présence d'un emprunt au latin médiéval.

Par contre, en ce qui concerne le cas troisième, on ne peut que déplorer qu'il n'existe aucun effort de théorisation globale qui rend compte des cas indécidables concernant les voies d'emprunt du latin. Pour l'heure, ce troisième cas ne semble pas avoir été traité dans les ouvrages théoriques de référence pour la linguistique historique. Du point de vue méthodologique, il faudrait de toute évidence procéder au dépouillement exhaustif des textes médiévaux et des articles de terminologie technique.

L'idée de débattre de ce troisième cas où deux voies d'emprunt peuvent être envisagées a surgi de notre pratique quotidienne — dans le cadre de la rédaction du TLF-Étym — et du constat de la pénurie de travaux d'ensemble concernant cette problématique. Ce serait aussi l'occasion de répondre à l'idée émise par Giovanni Rovere 2007 qui déplore dans le domaine italien – mais la situation est la même pour le français – l'absence d'un filon d'études théoriques, alors que les projets lexicographiques sont florissants :

> È noto, infatti, che in ambito lessicografico la produzione italiana appare nel confronto europeo molto ricca, e che d'altra parte a questa ricchezza non corrisponde una altrettanto consistente ricerca teorica, paradosso su cui aveva già attirato l'attenzione Aldo Duro, quando nel 1971 lamentava l'assenza in Italia di cattedre di lessicologia e lessicografia. Esistono sì manuali, descrizioni metodologiche, ricerche storiche, ma non un filone di studi teorici. (Rovere 2007 : 3).

La visée de notre communication est d'inciter la communauté scientifique à se pencher sur cette problématique pour théoriser ce troisième cas possible d'emprunt et définir une méthodologie de travail.

Bibliographie

AalmaR = Roques (Mario), 1938. *Aalma, Lexiques alphabétiques*, *in* : *Recueil général des lexiques français du moyen âge (XII^e-XV^e siècle)*, édité par Mario Roques, volume 2, (Bibliothèque de l'École des Hautes Études : 269), Paris : Champion.

AND = Rothwell (William)/Gregory (Stewart)/Trotter (D. A.) (dir.), 2005—[2] [1977—1992[1]]. *Anglo-Norman Dictionary*, Londres, Maney Publishing/Modern Humanities Research Association.

Bailly = Bailly (Anatole), 2000 [1894[1]]. *Dictionnaire grec-français*, édité par Émile Egger, Paris : Hachette.

Bertrand, Olivier (2004) : *Du vocabulaire religieux à la théorie politique en France au XIVe siècle. Les néologismes chez les rédacteurs de Charles V*, Paris : Connaissance et Savoir.

Bork, Hans Dieter (2006) "Sprachkontakte : Latein und Galloromania", *in* Ernst (Gerhard) et *al*. (éd.), *Romanische Sprachgeschichte* [titre parallèle*: Histoire linguistique de la Romania*], 2, Berlin/New-York : Walter de Gruyter : 1582-1590.

Buchi, Éva (2005) : Le projet TLF-Étym (projet de révision sélective des notices étymologiques du Trésor de la langue française informatisé). In : *Estudis romànics* 27, 569 – 571.

FEW = Wartburg (Walther von) *et al.*, 1922—2002. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*, 25 volumes, Bonn/Heidelberg/Leipzig-Berlin/Bâle : Klopp/Winter/Teubner/Zbinden.

Latham = Latham (R. E.), 1965. *Revised Medieval Latin Word-list from British and Irish sources*, Londres : Oxford University Press.

Liddell-Scott = Liddell (Henry George)/Scott (Robert), 1996 [1843[1]]. *A Greek-English Lexicon with a revised supplement*, édité par Henry Stuart Jones, Oxford : Clarendon.

MirJustW = Whittaker (William Joseph) (éd.), 1895. *The Mirror of justices*. Date du texte : fin 13^e siècle. Date du manuscrit : 1^er tiers 14^e siècle. Région d'origine du texte : Angleterre (anglo-normand), Londres : Quaritch (Selden Society : 7).

Niermeyer = Niermeyer (Jan Frederik) *et al.*, 2002[2] [1954—1976[1]]. *Mediae Latinitatis lexicon minus* : *lexique latin médiéval*, 2 volumes, Leiden, Brill.

Pfister, Max et Lupis, Antonio (2001), *Introduzione all'etimologia romanza*, Soveria Manenelli (Catanzaro) : Rubbettino Editore.

Reinhardt, Jan (2004) : *Mittellatein und italienische historische Lexikographie*, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien : Peter Lang.

Rovere, Giovanni (2007), "Per una teoria generale della lessicografia : l'impostazione di Herbert Ernst Wiegand", [Introduzione], *in Studi italiani di linguistica teorica e applicata* 1, Ospedaletto (Pisa) : Pacini editore : 3-6.

Städtler, Thomas (1988) : *Zu den Anfängen der französischen Grammatiksprache. Textausgaben und Wortschatzstudien*, Tübingen : Niemeyer (Beihefte zur Zeitschrift für romanische Philologie : 223).

Steinfeld, Nadine (2006a) : "Observations méthodologiques sur la traque des premières attestations en étymologie et histoire du lexique (domaine français)", *in* : Buchi (Éva) (éd.), *Actes du Séminaire de méthodologie en étymologie et histoire du lexique (Nancy/ATILF, année universitaire 2005/2006)*, Nancy, ATILF (CNRS/Université Nancy 2/UHP), site internet : http://www.atilf.fr/atilf/seminaires/Seminaire_Steinfeld_2005-11.pdf.

Steinfeld, Nadine (2006b) : "L'apport du roman de *Perceforest* pour la mise à jour des notices étymologiques du TLF(i)", *in* : Buchi (Éva) (éd.), *Actes de la Journée d'étude « Lexicographie historique française : autour de la mise à jour des notices étymologiques du* Trésor de la langue française informatisé » *(Nancy/ATILF, 4 novembre 2005)*, Nancy, ATILF (CNRS/Université Nancy 2/UHP), site internet : http://www.atilf.fr/atilf/evenement/JourneeEtude/LHF2005/Steinfeld.pdf.

TLF = Imbs (Paul)/Quemada (Bernard) (dir.), 1971—1994. *Trésor de la Langue Française. Dictionnaire de la langue du XIX^e et du XX^e siècle (1789—1960)*, 16 volumes, Paris, Éditions du CNRS/Gallimard. (Volume 1, *A — affiner* : 1971 ; volume 2, *affinerie — anfractuosité* : 1973 ; volume 3, *ange — badin* : 1974 ; volume 4, *badinage — cage* : 1975 ; volume 5, *cageot — constat* : 1977 ; volume 6, *constatation — désobliger* : 1978 ; volume 7, *désobstruer — épicurisme* : 1979 ; volume 8, *épicycle — fuyard* : 1980 ; volume 9, *G — incarner* : 1981 ; volume 10, *incartade — losangique* : 1983 ; volume 11, *lot — natalité* : 1985 ; volume 12, *natation — pénétrer* : 1986 ; volume 13, *pénible — ptarmigan* : 1988 ; volume 14, *-ptère — salaud* : 1990 ; volume 15, *sale — teindre* : 1992 ; volume 16, teint — zzz... : 1994).

TLL = 1900—. Thesavrvs lingvae latinae, Leipzig : B. G. Teubner.

Vignay (Jean de), 1373. [Miroir historial]. Manuscrit BnF, fonds français 316. Date du texte : ca 1328.

Wunderli, Peter (2001) : "La philologie romane de Diez aux néogrammairiens", *in* Holtus (Günter), Metzeltin (Michael), Schmitt (Christian) (éd.) : *Lexikon der romanistischen Linguistik* 1, Tübingen : Niemeyer : 121-175.

ANNEXES

1. EXEMPLE DE NOTICE ÉTYMOLOGIQUE DU PROJET TLF-ÉTYM

***différer*[2]**, verbe trans.

ÉTYMOLOGIE

Histoire :

A. 1. Transitif direct : « remettre à plus tard ». Attesté depuis *ca* 1355 [dans une traduction d'un texte latin] (BERS., *Tite-Live* T. I, 1, page 85, paragraphe 51.3 : et leur dist que pour certain Turnez avoit ier a lui et a eulz la mort appareillé, afin que il peust tous seulz l'empire des Latins avoir et maintenir, mes que la chose **avoit esté differee** pour l'absence de lui, lequel principalment il desirroit a occire). Remarque : *Trad. Ovide Remède d'Amour*, donné comme première attestation par TLF, est à dater de *ca* 1370/1380 et non de la fin du 13e siècle. -

A. 2. Transitif indirect : *différer de/à* + infinitif « tarder à ». Attesté depuis *ca* 1370 [dans une adaptation d'un texte latin] (OresmeEthM, page 451 = DMF2 : Et se l'un de eulz estoit mauvais, l'autre devroit fuir ou **differer a** lui ministrer du sien [« des subsides »]). -

B. Absolu : « tarder, temporiser ». Attesté depuis *ca* 1453/1457 (*Aff. Jacques Cœur* M, tome 1, page 255 = DocDMF : mon entencion n'estoit point de **différer**, ainçois estoit de procéder contre eulx et les contraindre à paier). *Cf.* la locution *différer le temps* « attendre », qui préfigure l'émergence de l'emploi absolu, attestée dès 1369 (GuillMachPriseM, page 202, vers 6647 = DMF2 : Et en Rodes s'en vuet aler. Là vuet il **le temps differer** Pour veoir que ce devenra Et se son Tricoplier venra). -

Origine :

Transfert linguistique : emprunt au latin *differre* verbe trans. « remettre à plus tard » (attesté depuis Plaute [depuis Cicéron dans ce sens précis], TLL 5/1, 1071-1072), avec changement de conjugaison. *Cf.* VON WARTBURG *in* FEW 3, 73b, DIFFERRE 2. Le terme semble être entré en français à travers une traduction de Tite-Live (*cf.* ci-dessus) .

Rédaction TLF 1979 : Équipe diachronique du TLF. - Mise à jour 2005 : Nadine Steinfeld. - Relecture mise à jour 2005 : Stephen Dörr ; Frédéric Duval ; Éva Buchi.

2. Exemple de notice étymologique du projet TLF-ÉTYM — notice présentant la structure lexicographico-informatique (DTD) avec balises XML délimitant des champs et des sous-champs

Extraction de collocations à partir du champ syntagme du *TLFi* : application aux noms transdisciplinaires des écrits scientifiques

Veronika Lux-Pogodalla (1)
veronika.lux@inist.fr
Agnès Tutin (2)
agnes.tutin@u-grenoble3.fr

(1) Inist-CNRS, 2, allée de Brabois, 54514 Vandoeuvre-lès-Nancy
(2) LIDILEM, Université Grenoble 3, UFR des sciences du langage, BP25, 380440 Grenoble cedex 9

Mots-clés : lexique scientifique général, collocations, extraction de collocations, aide à la rédaction

Keywords: general scientific lexicon, collocations, collocation extraction, writing help

Résumé : Le *TLFi* est une source de données lexicales extrêmement riche et relativement peu exploitée. Dans cet article, nous souhaitons évaluer dans quelle mesure il est possible d'extraire semi-automatiquement un sous-ensemble de collocations (champ « syntagme » du *TLFi*) liées au lexique général des écrits scientifiques, ce lexique transdisciplinaire qui renvoie à la description, au processus et au raisonnement des activités scientifiques. Nous voudrions déterminer comment et dans quelle mesure on peut tirer parti de cette ressource électronique en la combinant avec des informations automatiquement extraites de corpus à l'aide d'un analyseur syntaxique.

Abstract : The *TLFi* is a rich and underexploited source of lexical data. Here, we want to evaluate if a list of collocations can be semi-automatically extracted (using the "syntagme" field of the TFLi), that are related to the general lexicon of scientific writing. Items in this lexicon are used for the descriptions, the processes and the reasonings of scientific activities. We want to determine how and to which extend this electronic resource can be used for our aim, combining data extracted from the *TLFi* with data automatically extracted from text corpora.

Problématique

Le *TLFi* est une source de données lexicales extrêmement riche qui, pour certain de ses champs, a de façon surprenante été assez peu exploitée. Le champ syntagme, qui nous intéresse particulièrement, a, à notre connaissance, peu fait l'objet d'études systématiques,

hormis celles de [Hausmann, 1995]. Dans cet article, nous souhaitons évaluer dans quelle mesure il est possible d'extraire semi-automatiquement un sous-ensemble de collocations liées au lexique général des écrits scientifiques, ce lexique transdisciplinaire qui renvoie à la description, au processus et au raisonnement des activités scientifiques. Nous voudrions déterminer comment et dans quelle mesure on peut tirer parti de cette ressource électronique en la combinant avec des informations extraites de corpus à l'aide d'un analyseur syntaxique. Nous examinerons ainsi les collocations de type V-N (*faire l'hypothèse, vérifier l'hypothèse, mener une étude*) et Adj-N (*une hypothèse pertinente, des résultats encourageants*) extraites semi-automatiquement du champ syntagme du *TLFi* et les extractions effectuées entièrement automatiquement à partir d'un corpus d'écrits scientifiques de 2 millions de mots analysé à l'aide du logiciel Syntex [Bourigault, 2007]. La façon dont on peut combiner ces deux ressources sera évaluée par des linguistes s'intéressant à la problématique du lexique des écrits scientifiques.

1. Les collocations de langue scientifique générale

Dans le cadre de cette expérimentation, nous nous intéressons aux collocations transdisciplinaires des écrits scientifiques. Les collocations étant une notion à géométrie variable, il convient bien entendu d'en préciser les contours. Notre approche des collocations se situe dans la lignée de Mel'čuk et Hausmann (par exemple, [Hausmann, 1989], [Mel'čuk, 1998]). Ce sont pour nous des expressions linguistiques composées de deux éléments linguistiques, apparaissant fréquemment en cooccurrence et entretenant une relation syntaxique, et dont l'un des éléments, la base conserve son sens habituel, alors que le collocatif apparaît moins prédictible. Dans *résultats encourageants,* le mot *résultats* aurait ainsi la fonction de base, alors que *encourageants* ferait office de collocatif, étant conditionné par le mot *résultats*. Nous nous intéressons ici particulièrement aux collocations apparaissant dans les écrits scientifiques et qui sont emblématiques de ce genre, renvoyant à la description de l'activité scientifique, aux résultats, aux évaluations, au raisonnement mis en jeu dans les écrits de ce type. Ce lexique, qui transcende en grande partie les disciplines (il ne renvoie pas à la terminologie du domaine), comporte des expressions comme *faire une hypothèse, rejeter une hypothèse, approche traditionnelle, étude théorique, thèse classique* … . Notre objectif à plus long terme est de proposer un inventaire exhaustif et une description sémantique de ce lexique pour des applications d'aide à la rédaction en langue étrangère. Notre expérimentation vise à évaluer l'emploi du *TLFi* comme source possible pour l'acquisition de ce lexique.

2. Extraction des collocations du lexique transdisciplinaire du *TLFi*

Le *TLF* est souvent considéré comme un dictionnaire de langue littéraire. La langue scientifique et technique n'y a cependant pas été négligée et nous pensons que le *TLFi* peut en partie servir de base pour la constitution d'un dictionnaire des collocations de la langue scientifique générale, même si l'état de langue décrit y est un peu ancien et le lexique extrait doit être en partie actualisé.

Un avantage considérable du *TLF* est la possibilité d'extraire des éléments de champs ciblés à partir de la version informatisée balisée, procédure déjà adoptée pour des dictionnaires bilingues (Cf. par exemple [Fontenelle, 1997]). En outre, par rapport à d'autres dictionnaires de langue comme le *Petit Robert*, le *TLF* apparaît très bien structuré dans le traitement des collocations. Tout d'abord, le concept de « syntagme » (à peu près équivalent à notre notion de collocation) est différencié de la notion de locution et fait en principe l'objet d'un

traitement spécifique[1], même si dans les faits, quelques incohérences demeurent [Henry, 1995]. L'informatisation du dictionnaire permet d'avoir accès à la « richesse de la face cachée » du *TLF*, comme le décrit [Haussmann, 1995] (p. 38), en accédant aux collocations à partir de certains champs de l'article : (a) les éléments retenus dans la rubrique « SYNT. » (pour « syntagme »), une rubrique spécifiquement dédiée aux informations de cooccurrences (Cf. (1) ci-dessous),

> (1) *TLFI* s.v. *conclusion* II.A[2] → le champ « SYNT »
> SYNT. (très fréq.). *Conclusion audacieuse, bonne, erronée, hardie, juste, nulle, prématurée; les conclusions qui se dégagent de ces travaux; exposer la conclusion de ses réflexions; aboutir, arriver, être conduit à certaines conclusions, aux conclusions suivantes; affirmer, confirmer, infirmer une conclusion; établir, formuler une conclusion inattaquable; amener qqn à ses conclusions.*

(b) les « collocations enchaînées » qui suivent la définition, mais ne sont pas glosées ou définies, (c) les « collocations définies », qui ne sont introduites par aucun indicateur, (d) les fausses locutions, qui sont des sous-vedettes[3], introduites par l'indicateur « Loc. », qui sont en réalité des collocations. Haussmann signale également d'autres champs où les collocations apparaissent également de façon plus diffuse, et où l'extraction automatique n'apparaît pas envisageable en l'état : (e) les collocations dans les citations, détachées ou enchaînées ; (f) les collocations dans les définitions, où la collocation doit être reconstruite ; (g) les collocations synonymiques ou antonymiques.
Dans le *TLFi*, on peut extraire les collocations du champ « syntagmes»[4] qui correspond à peu près aux champs (a)-(d) du *TLF* « papier ». Dans cette version électronique, on pourra également effectuer une recherche sur tous les articles à partir de la recherche assistée ou de la recherche complexe, en extrayant tous les articles qui contiennent un mot donné dans le champ syntagme. Il est même possible de préciser dans la recherche complexe dans quel type de champ on veut extraire le syntagme : paragraphe syntagme (rubrique dédiée), syntagme défini, ou syntagme enchaîné (que l'indicateur « Loc. » apparaisse ou non). Sur l'interface en ligne, il est ensuite possible d'afficher tous les champs contenant les syntagmes, sans passer par l'affichage de l'article complet.

Pour cette communication, nous avons obtenu la liste des syntagmes comprenant 90 noms considérés comme relevant du lexique général (*hypothèse, cas, données, thèse, approche* …)[5] sous forme de texte balisé XML[6] dont nous avons extrait un sous-ensemble de syntagmes sous les vedettes verbales et adjectivales. Par exemple, sous la vedette verbale MENER, on repère des syntagmes comportant les noms transdisciplinaires *analyse, étude* et *recherches*, comme on peut le voir dans la Figure 1.

```
<article>
   <vedette>MENER, verbe trans.</vedette>
```

[1] [Henry, 1995] mentionne (p.107) l'existence de définitions précises pour les notions de « syntagme », « locution » et « phrase figée ayant valeur de vérité générale » dans le *Cahier des normes* utilisé en interne pour la rédaction, tout en signalant quelques incohérences dans le traitement des articles.
[2] La définition principale donnée pour cette acception est la suivante : « Proposition tirée des données de l'observation ou d'un raisonnement. »
[3] Dans le *TLFi*, le terme sous-vedette a une extension plus étroite. Il semble surtout renvoyer à des mots composés très figés, souvent réunis par des traits d'union.
[4] Dans le *TLF*, on peut postuler les degrés de figement suivants pour les unités polylexicales (du plus figé au moins figé) : vedettes (ex : *pied à terre*), sous-vedette (ex : *maison-témoin*), syntagme défini avec marqueur « Loc. » (ex : *casser la tête de qqun*), syntagme défini (ex : *maison seigneuriale*), les syntagmes enchaînés (ex : *crainte, peur irraisonnée*).
[5] Voir liste en annexe.
[6] Un grand merci à Etienne Petitjean de nous avoir fourni cette liste.

```
    <occurrences>
      <occurrence>
        <mot>analyse</mot>
        <syntagme>Mener une action, une analyse, une étude,
        une politique, des recherches.</syntagme>
       </occurrence>
      <occurrence>
        <mot>étude</mot>
        <syntagme>Mener une action, une analyse, une étude,
        une politique, des recherches.</syntagme>
       </occurrence>
      <occurrence>
        <mot>recherches</mot>
        <syntagme>Mener une action, une analyse, une étude,
        une politique, des recherches.</syntagme>
       </occurrence>
    </occurrences>
  </article>
```

Figure 1 : Extrait du *TLF* de syntagmes comprenant *mener* (c'est nous qui mettons en gras les informations pertinentes pour l'extraction des collocations)

Ces informations structurées peuvent ensuite être filtrées pour extraire une liste pertinente de collocations transdisciplinaires. Une première expérimentation en ce sens a été réalisée avec succès. Un important filtrage manuel doit néanmoins être effectué puisqu'aucune désambiguïsation automatique ne peut être effectuée – en tout cas facilement – à partir du dictionnaire. On relève ainsi un très grand nombre de cooccurrences non pertinentes comme *allouer à qqn un traitement* ou *méthode aratoire* qu'il faut bien entendu exclure de la liste désirée.

Le filtrage manuel permet néanmoins d'extraire assez rapidement un sous-ensemble d'éléments *a priori* pertinents pour notre projet. A titre d'exemple, nous listons dans le Tableau 1 un sous-ensemble de collocations *recherche(s)* -Adj extraites automatiquement et filtrées manuellement.

	recherche	appliquée
	recherche	bibliographique
	recherches	épistémologiques
	recherche	fondamentale
	recherche	inaccessible
	recherche	infructueuse
	recherche	prévisionnelle
	recherche	pure
	recherche	scientifique
	recherche	spéculative
	recherche	statistique
	recherche	stérile
	recherche	tâtonnante
	recherche	technique
	recherche	théorique
premières	recherches	

	recherches	récentes

Tableau 1 : Ensemble des collocations *recherche(s)* + Adjectif extraites automatiquement et filtrées manuellement

3. Extraction des collocations transdisciplinaires à partir de corpus, combinaison et évaluation des données extraites

Notre objectif final est de constituer des ressources lexicales utilisables. Nous souhaitons ainsi compléter les données extraites du *TLF* à l'aide de données extraites automatiquement de corpus d'écrits scientifiques, en utilisant l'analyseur syntaxique Syntex développé par Didier Bourigault (2007). Nous pensons que ces deux types de ressources sont complémentaires : les données du *TLFi* peuvent renvoyer à des collocations rares mais utiles (qui, du fait de leur faible fréquence seraient écartées lors d'une extraction automatique sur corpus exploitant des critères de fréquence et de répartition) alors que les données extraites des corpus correspondront à des données plus récentes, mais peut-être moins nombreuses. Dans le cadre de la méthode automatique, les collocations du lexique transdisciplinaire sont obtenues à partir d'un corpus d'écrits scientifiques diversifiés de 2 millions de mots[7], auquel l'analyseur Syntex a été appliqué. Les noms retenus sont les noms les plus fréquents qui apparaissent à la fois dans les trois disciplines du corpus (économie, linguistique et médecine). Cette liste de noms est identique à celle qui a été exploitée pour l'extraction semi-automatique des collocations du *TLF*. Les collocations de type N-Adjectif et V-N sont simplement obtenues en extrayant les relations syntaxiques de dépendance de type épithète et objet direct apparaissant au moins trois fois dans deux des trois disciplines. Pour les relations de type V-N, les verbes *être* et *avoir* ont été exclus. Le Tableau 2 ci-dessous présente un extrait de collocations de type N-Adjectif obtenues automatiquement et dont on a enlevé les adjectifs qu'on peut considérer comme des mots grammaticaux (i.e. *autre, même, différent,* les adjectifs ordinaux et cardinaux).

Collocation N-Adjectif	Fréquence
étude récente	54
études empiriques	50
présente étude	31
approches théoriques	30
analyse statistique	24
analyse factorielle	20
études antérieures	18
étude précédente	15
étude préliminaire	15
argument supplémentaire	14
études complémentaires	13
analyse fine	12

[7] Le corpus comprend le corpus d'articles scientifiques KIAP de l'équipe de Kjersti Fløttum, augmenté par nos soins de rapports de recherche et thèses. Le corpus de 2 millions de mots se répartit équitablement entre les domaines de la médecine, de la linguistique et de l'économie.

analyse automatique	11
étude comparative	10
étude spécifique	9

Tableau 2 : Collocations transdisciplinaires de type N-Adj extraites automatiquement

Sur un premier test effectué sur 8 noms transdisciplinaires[8], les données extraites du *TLFi* semblent bien plus nombreuses que celles qui sont extraites du corpus. Elles comportent aussi beaucoup moins d'erreurs que les données automatiquement extraites de corpus où les erreurs d'analyse syntaxique introduisent un bruit non négligeable. Il apparaît en revanche difficile de porter un jugement clair sur certaines collocations extraites du *TLFi* (en tout cas, sans plus de contexte). Par exemple, l'expression *recherche infructueuse* qui relève clairement de la langue courante (Ex : *ma recherche d'appartement s'est révélée infructueuse*), apparaît-elle dans les écrits scientifiques ? A-t-elle sa place dans un lexique scientifique transdisciplinaire ?

Afin d'établir une liste de collocations de ce champ sémantique, nous proposons de constituer d'abord une liste de collocations candidates issues des deux ressources, corpus et *TLFi*, que nous avons exploitées. Y figureront en tête les collocations trouvées dans les deux ressources puis celles qui ne sont présentes que dans le *TLFi* puis celles qui ne sont présentes que dans le corpus (triées selon leur fréquence). Cette première liste sera évaluée par des linguistes experts du domaine. Notre hypothèse actuelle est que les collocations communes aux deux méthodes d'extraction seront les mieux évaluées ; l'analyse des résultats permettra de mieux cerner ce qu'apporte chaque ressource et, éventuellement, de se faire de nouvelles idées sur de meilleures stratégies pour les combiner.

Lors de l'évaluation, nous demanderons aussi aux linguistes d'établir si, selon eux, l'expression relève du lexique scientifique transdisciplinaire, s'ils l'emploieraient, s'ils la relèveraient pour une application d'aide à la rédaction et aussi, si l'expression relève de la langue générale ou de la terminologie du domaine.

Bibliographie

[Bourigault, 2007] Bourigault, D. (2007). *Syntex, analyseur syntaxique opérationnel*. Thèse d'Habilitation à Diriger des Recherches. Université Toulouse le Mirail, juin 2007.

[Dendien&Pierrel, 2003] Dendien, J. & Pierrel, J.-M. (2003). « Le Trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence », *TAL*, vol. 44 - n°2, 11-39.

[Fontenelle, 1997] Fontenelle Th. (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. (Lexicographica /Series maior). Tübingen, Niemeyer Verlag.

[Hausmann, 1989] Hausmann F. J. (1989). Le dictionnaire de collocations. In Hausmann F.J., Reichmann O., Wiegand H.E., Zgusta L. (eds), *Wörterbücher : ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires*. Berlin/New-York, De Gruyter, 1010-1019.

[Hausmann, 1996] Hausmann F. (1996). La syntagmatique dans le *TLF* informatisé, in *Autour de l'informatisation du TLF*, Actes du Colloque International de Nancy (29-31 mai 1995), D. Piotrowski (ed.). Paris, Didier, 51-77.

[8] Les noms : *analyse, approche, argument, concept, démarche, étude, idée, recherche.*

[Henry, 1996] Henry F. (1996). Pour une informatisation du *TLF,* , in *Autour de l'informatisation du TLF*, Actes du Colloque International de Nancy (29-31 mai 1995), D. Piotrowski (ed.). Paris, Didier, 79-139.

[Mel'čuk, 1998] Mel'čuk I. (1998). Collocations and Lexical Functions. In A. P. Cowie (ed.), *Phraseology. Theory, Analysis and Applications*. Oxford, Clarendon Press, 23-53.

La lexicographie explicative et combinatoire à l'épreuve de l'informatisation

Alain Polguère (1)
alain.polguere@umontreal.ca

(1) OLST — Département de linguistique et de traduction, Université de Montréal
C.P. 6128, succ. Centre-ville Montréal (Québec) H3C 3J7 Canada

Mots-clés : lexicographie informatisée, Lexicologie Explicative et Combinatoire (LEC), Dictionnaire Explicatif et Combinatoire (DEC), base lexicale DiCo, *Lexique actif du français* (LAF)

Keywords: computational lexicography, Explanatory Combinatorial Lexicology (ECL), Explanatory Combinatorial Dictionary (ECD), DiCo lexical database, *Lexique actif du français* (LAF)

Résumé : Nous présentons un bilan de la lexicographie explicative et combinatoire du point de vue de son rapport avec l'informatisation du travail lexicographique. Nous traitons de l'évolution de l'approche lexicologique qui sous-tend cette lexicographie, en nous fondant sur l'expérience accumulée dans le cadre de la construction de la base lexicale DiCo des dérivations sémantiques et collocations du français. Plusieurs produits dérivés de cette base sont également considérés, dont le dictionnaire pédagogique *Lexique actif du français*. Nous concluons par une présentation d'un ensemble de tâches essentielles qu'il reste à accomplir.

Abstract : We present an assessment of explanatory combinatorial lexicography from the perspective of the computerization of lexicographic work. We examine the evolution of the theoretical approach that underlies this lexicography, using insights gained while working on the DiCo database of French semantic derivations and collocations. Several byproducts of this database are also considered, among which the learner's dictionary *Lexique actif du français*. We conclude with a presentation of core tasks that remain to be accomplished.

Introduction

Le terme *lexicographie explicative et combinatoire* désigne la lexicographie pratiquée depuis plus de trente ans dans le cadre théorique de la Lexicologie Explicative et Combinatoire. Nous présentons ici un bilan de l'évolution de cette approche du point de vue de son rapport avec l'informatisation du travail lexicographique. Il ne s'agit pas d'une revue complète des travaux effectués et, faute de place, nous nous voyons contraint de présenter ici un exposé dépouillé de données lexicographiques (exemples d'articles de dictionnaires ou d'enregistrements de bases de données lexicales). Pour remédier à cela, nous donnons un nombre important de références bibliographiques et de sites web où le lecteur trouvera de quoi rassasier son appétit linguistique.

Nous procéderons en trois étapes. La Section 1 examine les *Dictionnaires explicatifs et combinatoires* en tant que produits d'une lexicographie dite « théorique ». La Section 2 traite

de la notion d'informatisation de la Lexicologie Explicative et Combinatoire, par opposition à une simple informatisation des DEC eux-mêmes, en se concentrant sur le cas de la base lexicale DiCo des dérivations sémantiques et collocations du français. La Section 3 traite des nouvelles avenues ouvertes par l'approche DiCo à travers la présentation de trois produits lexicographiques dérivés de cette base. Finalement, nous concluons en présentant l'évolution future que nous anticipons pour la lexicographie explicative et combinatoire.

1. Les Dictionnaires Explicatifs et Combinatoires en tant que produits d'une lexicographie théorique

Les *Dictionnaires Explicatifs et Combinatoires* ou DEC [Mel'čuk et autres, 1984, 1988, 1992, 1999 ; Mel'čuk et Zholkovsky, 1984] se démarquent des dictionnaires commerciaux, et autres dictionnaires grand public, en ce qu'ils sont les produits directs d'une lexicographie théorique. Ce dernier terme est ambigu. Il peut dénoter des projets lexicographiques s'appuyant sur une théorie lexicologique et un ensemble de notions théoriques bien définies. Il peut aussi dénoter des projets lexicographiques ayant une visée plus théorique que descriptive : à travers la modélisation (partielle) des lexiques, on vise avant tout l'exploration et la meilleure compréhension de l'organisation et du fonctionnement lexical des langues naturelles.

L'entreprise lexicographique derrière la production des DEC est une lexicographie théorique aux deux sens du terme. En effet, le travail de production des DEC ne se conçoit que par rapport à l'approche lexicologique qui la sous-tend : la Lexicologie Explicative et Combinatoire [Mel'čuk et autres, 1995], qui est la « branche lexicale » de la théorie linguistique Sens-Texte [Mel'čuk, 1997 ; Polguère 1998]. Ce travail lexicographique est pratiqué comme un outil de recherche sémantique : étude des structures sémantiques lexicales et interconnexion des phénomènes sémantiques avec les niveaux syntaxiques, morphologiques et phonologiques de fonctionnement de la langue. Cependant, la Lexicologie Explicative et Combinatoire est orientée vers la description lexicographique et propose à la fois des notions lexicologiques de base et une méthodologie d'application de ces notions dans un contexte de construction de modèles formels des lexiques. On peut dire que la Lexicologie Explicative et Combinatoire — dorénavant, LEC — a comme raison d'être l'activité lexicographique, car le but premier d'une étude lexicologique d'une langue menée dans ce cadre théorique est très clair : construire un dictionnaire du type DEC de cette langue [Mel'čuk, 2006].

Cette double orientation, qui à notre avis fait la force de la LEC, a eu comme conséquence perverse de favoriser une description des lexiques menée en profondeur d'abord, plutôt qu'en largeur. On reproche ainsi fréquemment à la lexicographie explicative et combinatoire de ne pas être « professionnelle », car elle ne mène qu'à des descriptions partielles des lexiques, à des échantillons de dictionnaires : il n'existe pas de DEC suffisamment complet pour pouvoir se mesurer à un dictionnaire plus classique (non théorique) du type *Petit Robert*. Pour ce qui est des ressources informatiques exploitables en TAL, il n'existe pas de bases de données lexicales fondées sur la LEC dont la nomenclature pourrait se mesurer à celle de la base WordNet de l'anglais [Fellbaum, 1998].

Ce reproche fait aux DEC et à la LEC est justifié, dans le sens où il s'appuie sur des faits exacts. Il est clair que les produits lexicographiques de la LEC pour le français n'offrent qu'une couverture très partielle de la nomenclature minimale exigée pour, par exemple, un véritable dictionnaire d'apprentissage. Cependant, ce qui importe vraiment, c'est de savoir si cette faiblesse au niveau de la production lexicographique relève d'un problème méthodologique ou d'un vice de forme de l'approche elle-même. Dans ce qui suit, nous allons montrer, à travers l'examen de changements récents qu'a connus la LEC au niveau de

l'informatisation de ses descriptions, que celle-ci est entrée dans une nouvelle étape de son évolution qui la rend compatible avec des entreprises lexicographiques grandeur nature.

2. Informatisation des DEC ou de la LEC ?

2.1 Construire un DEC, c'est pas humain

Un des principaux problèmes posés à la lexicographie explicative et combinatoire est la richesse de l'information à gérer. Cette richesse se manifeste à deux niveaux.

Tout d'abord, il y a la quantité même d'informations qui doit être incluse dans chaque article de dictionnaire, puisque chaque unité lexicale doit recevoir une description minutieuse et, en théorie, exhaustive de son sens — définition —, de sa combinatoire grammaticale — morphologie, régime syntaxique, etc. —, de sa combinatoire lexicale — collocations qu'elle contrôle — et des liens paradigmatiques — appelés dérivations sémantiques — qui l'unissent à d'autres unités lexicales de la langue. La description de tout vocable un moindrement polysémique prend, dans un DEC, plusieurs pages. Même en changeant le mode de présentation et en le compactant au maximum, la taille du texte constituant une entrée d'un vocable dans le DEC dépassera toujours de loin celle du texte consacré au même vocable dans les dictionnaires standard.

Ensuite, il y a la nature de l'information lexicographique, puisque toute information sur les propriétés des unités lexicales doit être, dans la LEC, explicitement encodée de façon formelle. Il va de soi que plus on cherche à être explicite, plus on a de chance de faire des erreurs. Par exemple, un dictionnaire qui se contente d'énumérer des collocations dans ses articles[1], a beaucoup moins de chance de « faillir » qu'un dictionnaire qui se donne pour tâche d'encoder explicitement le lien que le mot-vedette entretient avec ses collocatifs en indiquant précisément le rôle sémantico-syntaxique que les collocatifs jouent auprès de leur base dans les collocations. Il en va bien sûr de même avec la description des liens lexicaux paradigmatiques ou dérivations sémantiques. Ajoutons à cela que le contenu de tout dictionnaire est, du fait de la structure même des lexiques des langues naturelles, de nature extrêmement relationnelle. La formalisation des DEC impose alors au lexicographe de gérer de façon explicite et cohérente un gigantesque treillis de connexions interlexicales, que le cerveau humain seul ne peut parcourir pour le consulter ou le valider.

Il est clair que la lexicographie fondée sur la LEC ne peut se faire « grandeur nature » sans le recours à l'outil informatique (à moins, bien sûr, de disposer d'une armée de lexicographes). L'humain n'y suffit pas et le problème de l'informatisation du travail lexicographique s'est posé de façon de plus en plus aiguë au fur et à mesure que la lexicographie explicative et combinatoire prenait de la maturité.

Deux approches ont été envisagées pour utiliser l'outil informatique afin de supporter le travail lexicographique fait dans le cadre de la LEC : 1) informatiser les DEC, notamment au moyen d'un éditeur dédié à cette tâche et 2) informatiser la LEC, en visant en premier lieu la construction de bases de données lexicographiques génériques — destinées aussi bien à servir de ressources pour le traitement automatique de la langue (TAL) qu'à être à la source de la production automatique de descriptions dictionnairiques. Nous allons examiner tour à tour ces deux approches, en nous attardant surtout sur la seconde, qui est celle que nous poursuivons.

2.2 Premières tentatives : informatisation des DEC

Les volumes des DEC publiés pour le français et pour le russe ont tous été élaborés sans aucun support informatique spécifique, autre qu'un traitement de texte. La méthode de travail,

[1]Nous parlons ici de collocations au sens de [Hausmann, 1979] et [Mel'čuk, 2003].

pour ce qui est de la construction et de la gestion des descriptions, n'était en fin de compte pas très différente de ce qui se pratique depuis que les dictionnaires existent, si ce n'est que le support papier avait été remplacé par les fichiers de textes électroniques. Plusieurs expérimentations ont été menées pour informatiser la rédaction des DEC grâce à un éditeur « intelligent » ; pour le DEC français : [Décary et Lapalme, 1990] et [Sérasset, 1997].

Sans entrer dans le détail de ce qui a été réalisé alors, notons qu'une caractéristique première de ces projets était de chercher à respecter au maximum la structure et le format des DEC publiés. Il s'agissait de construire des outils au service des lexicographes et lexicologues, en collant au plus près à leur façon de faire, sans rien remettre en question de celle-ci et en leur donnant un outil ne bridant en rien leur travail. Une conséquence de cette approche est qu'il s'est avéré impossible de dépasser le stade du prototype. En effet, le DEC étant en fait semi-formalisé et les conventions d'encodage évoluant au fur et à mesure que de nouvelles nécessités descriptives apparaissaient, il aurait fallu entretenir en permanence un travail de mise à jour d'un éditeur d'une très grande complexité informatique : chose que seuls les auteurs d'un tel programme seraient à même de faire. Faute d'une équipe de recherche incluant un ou plusieurs informaticiens « en résidence », l'équipe du DEC n'est jamais parvenue au stade de production lexicographique utilisant des éditeurs dédiés.

Partant du principe qu'il fallait absolument déconnecter la LEC des DEC, qui n'en sont que la manifestation, une nouvelle approche a été envisagée au début des années 90, puis graduellement mise en place par I. Mel'čuk et nous-même : celle de la base DiCo des dérivations sémantiques et collocations du français.

2.3 L'approche DiCo : vers une informatisation de la LEC

L'idée centrale derrière le projet DiCo [Polguère, 2000a ; Polguère, 2000b] était qu'il fallait reprendre presque à zéro la formalisation des descriptions lexicographiques faites dans le cadre de la LEC, en imposant des contraintes formelles strictes à la modélisation, quitte à en limiter provisoirement la portée. Toute description faite dans cette optique devait répondre aux deux critères suivants :

- aucune utilisation de formats ou polices spéciales comme mode de formalisation linguistique, afin de rendre la description aussi indépendante que possible d'un support informatique donné et d'éviter ce que nous considérions à l'époque comme une débauche de formatage dans les DEC publiés ;
- aucune ambiguïté syntaxique dans la description : tout élément d'information doit être suffisamment balisé pour rendre possible une analyse automatique du contenu des descriptions lexicographiques (afin, notamment, de compiler ces descriptions dans diverses autres structures de données).

De plus, nous avons décidé de concentrer la description sur de l'information correspondant *grosso modo* à l'état de l'art en TAL : nous avons ainsi choisi de n'inclure dans le DiCo que les données dont nous savions qu'elles pourraient éventuellement être exploitables par des systèmes de TAL, avec leur degré de sophistication de l'époque (qui, notons-le, n'as pas beaucoup évolué en 15 ans).

La définition lexicographique est l'élément de description le plus important qui a temporairement été laissé de côté. Celle-ci a été remplacée par une description de la structure actancielle des unités prédicatives (avec typage sémantique des actants) combinée à un étiquetage sémantique élaboré, conçu strictement dans une perspective d'analyse sémantique de type Sens-Texte — pour une description du système d'étiquetage sémantique du DiCo, voir [Polguère, 2003].

La description du régime syntaxique des unités lexicales a été conservée, mais nettement épurée. Les schémas de régime, qui décrivent comment s'expriment aux niveaux syntaxiques

profond et de surface les actants des mots-vedettes, sont effet très utiles en TAL (notamment, en analyse automatique) et ils font définitivement partie de l'information minimale à encoder dans une base telle que le DiCo.

La modélisation des liens lexicaux paradigmatiques — dérivations sémantiques — et syntagmatiques — collocations — contrôlés par les mots-vedettes est au centre du projet lexicographique DiCo [Mel'čuk et Polguère, 2006]. Cette description s'appuie sur les fonctions lexicales de la théorie Sens-Texte [Mel'čuk, 1996] et il est justifié de dire que le DiCo et tous ses produits informatiques dérivés ont été notamment conçus comme des modèles des liens de fonctions lexicales qui tissent le lexique du français.

Il est important de noter que toutes les informations retenues pour le modèle lexicographique DiCo (étiquette sémantique, structure actancielle, régime syntaxique, liens paradigmatiques et syntagmatiques) sont en étroite relation dans la langue, et aussi bien sûr dans la description de type LEC. La méthodologie lexicographique fondée sur la LEC (telle que décrite dans [Mel'čuk et autres, 1995]) exploite cette interconnexion et fait d'une fiche DiCo une structure de données où tout se tient. Une fiche DiCo est, pour des raisons pratiques, segmentée en champs distincts. Cependant, tous les champs de la fiche communiquent conceptuellement.

Le DiCo s'est graduellement développé comme un produit de la LEC à part entière, distinct des DEC, tout en étant compatible à 100 % avec le type de modélisation qu'offrent ces derniers ; il a fait école et a inspiré notamment le projet DICE, pour l'espagnol [Alonso Ramos, 2004][2]. De plus, le DiCo, en tant que base lexicale générique, a jusqu'à présent donné naissance à trois produits dérivés lexicographiques : le *Lexique actif du français*, l'interface DiCouèbe d'accès aux données du DiCo stockées sous forme de tables SQL et l'environnement DiCoPop de navigation dans le réseau lexical du DiCo.

3. Produits dérivés du DiCo

3.1 Le *Lexique actif du français*

Le projet du *Lexique actif du français* ou LAF [Mel'čuk et Polguère, 2007] s'est développé en parallèle et en interaction avec le DiCo. En fait, le DiCo visait dès son origine, en plus de la compilation automatique des données sous des formes exploitables par des programmes informatiques « clients » (voir Section 3.2 ci-dessous), l'extraction manuelle d'un dictionnaire grand public à vocation pédagogique. La formalisation traditionnelle des DEC a donc été rationalisée, épurée et, dans le cas des fonctions lexicales, vulgarisée afin de la rendre plus compatible avec la production d'un dictionnaire pédagogique des dérivations sémantiques et collocations du français : le LAF[3].

Le fait de travailler simultanément sur le DiCo et sur sa traduction en format grand public LAF a bien entendu considérablement retardé l'accomplissement de la tâche lexicographique. Cependant, cela a permis de faire avancer la compréhension que nous avions de l'interaction entre les différents paramètres la description lexicale faite dans le cadre de la LEC. Nous avons notamment isolé un ensemble minimal de notions descriptives de la LEC dont il est, selon nous, impossible de faire l'économie dans le cadre d'un dictionnaire pédagogique des dérivations sémantiques et des collocations, même simplifié au maximum. Les notions en question sont, pour chaque mot-vedette : son noyau sémantique (étiquette sémantique), sa structure actancielle, son schéma de régime (décrivant sa valence active), la valeur sémantico-syntaxique des liens paradigmatiques et syntagmatiques qu'il contrôle (encodée dans la LEC au moyen des fonctions lexicales) et, enfin, le régime de ses collocatifs (***faire*** *fortune* [sans

[2] http://www.dicesp.com/

[3] Le LAF publié possède un site web d'accompagnement : http://olst.ling.umontreal.ca/laf/.

déterminant] vs ***amasser*** *une fortune*). La prise en charge de toutes ces notions s'est avérée nécessaire si l'on veut offrir une description lexicographique qui soit à la fois nécessaire et suffisante (bien qu'incomplète) dans un contexte d'enseignement et d'apprentissage du vocabulaire. Notons que, par *prise en charge*, nous n'entendons pas uniquement la nécessité de refléter dans l'encodage lexicographique la manifestation linguistique (règles lexicales) de ces notions, mais aussi la nécessité de fournir aux utilisateurs (enseignants et élèves) d'outils tels que le LAF une explication de ces notions qui leur soit accessible (voir la première partie du LAF, pages 11 à 79). On trouvera dans [Polguère, 2007] un bilan du projet LAF et de son apport au travail sur le DiCo, notamment pour ce qui est de la caractérisation sémantique des unités lexicales et de l'encodage métalinguistique, sous forme de formules de vulgarisation, des liens de fonctions lexicales.

3.2 Le DiCo-SQL et l'interface de consultation DiCouèbe

La production du LAF à partir du DiCo s'est faite manuellement, car la compilation automatique des données originelles du DiCo vers une structure qui pourrait être directement exploitable par un programme de génération automatique d'articles de dictionnaire formatés n'a été réalisée que tardivement, dans le cadre d'une coopération avec Sylvain Kahane et l'équipe Talana de Paris 7 [Steinlin et autres, 2005].

Le problème de la compilation du DiCo n'était pas trivial. En effet, les données originelles du DiCo, bien que formalisées de façon beaucoup plus rigoureuse que celles du DEC, restent de nature « textuelle », sans recours à un balisage de type XML : un article du DiCo correspond à un enregistrement (*record*) dans un fichier FileMaker, dont chaque champ est un bloc de texte encodant une zone donnée d'un article de type DEC (étiquetage sémantique, forme propositionnelle, schéma de régime, liens de fonctions lexicales, etc.). Il s'agissait donc d'analyser le texte de chaque champ pour en extraire les éléments informationnels élémentaires et stocker ceux-ci sous une forme atomisée : dans ce cas précis, dans une cellule de table SQL. Le processus d'élaboration du compilateur de DiCo sous la forme d'un DiCo-SQL a été l'occasion d'améliorer notre formalisation et de mettre au jour des problèmes posés par l'encodage traditionnel des DEC, qui n'avaient pas encore été élucidés dans le cadre du projet DiCo. Finalement, avec le compilateur, une interface web d'accès aux données du DiCo-SQL a été réalisée par l'équipe de Paris 7 : le DiCouèbe [Jousse et Polguère, 2005][4].

La mise en ligne du DiCo-SQL et du DiCouèbe a été un palier important dans nos travaux sur l'informatisation de la LEC, puisqu'elle nous permettait pour la première fois de rendre disponibles nos descriptions de façon immédiate et constamment à jour, sans devoir passer par l'étape fastidieuse et frustrante de la publication traditionnelle. Pour rendre justice à nos collègues, notons que cette mise en ligne s'est effectuée après qu'au moins trois autres produits lexicographiques ayant une filiation directe avec la LEC ont été rendus accessibles sur le web : la version électronique du *Dictionnaire d'apprentissage du français des affaires* DAFA [Binon et autres, 2000], son corrélat de langue générale le *Dictionnaire d'apprentissage du français langue étrangère* DAFLES [Selva et autres, 2003][5] et le *Diccionario de colocaciones del español* DICE (déjà mentionné à la Section 2.3).

3.3 L'environnement DiCoPop de navigation dans les données du DiCo

Le DiCouèbe est une interface très utile aux données du DiCo puisqu'elle permet de générer des tables de données lexicographiques au moyen de requêtes SQL formulées indirectement par un « remplissage de formulaire » en ligne. Cependant, si elle ne nécessite pour son utilisation aucune connaissance de la structure de données elle-même et du langage de requête

[4] http://olst.ling.umontreal.ca/dicouebe/index.php.

[5] Le DAFA et le DAFLES sont accessibles à l'adresse suivante : https://www.kuleuven.be/ilt/blf/.

SQL, elle n'est pleinement exploitable que par un usager connaissant bien l'encodage pratiqué dans le DiCo. L'étude d'une documentation d'une quarantaine de pages [Jousse et Polguère, 2005] est ainsi requise pour une bonne utilisation du DiCouèbe. Cela n'est pas satisfaisant pour les usagers non-spécialistes, intéressés strictement par les données lexicologiques. Suite à la publication récente du LAF, nous avons poursuivi le travail sur l'accès aux données du DiCo-SQL, ce qui nous a amené à mettre en place l'interface DiCoPop[6].

Le DiCoPop est une ressource en ligne dérivée de deux sources : le DiCo-SQL (pour ce qui est des données lexicographiques proprement dites) et la hiérarchie d'étiquettes sémantiques du DiCo (exportée en format XML à partir de sa représentation originelle construite au moyen de l'éditeur d'ontologies Protégé). Le DiCoPop a une double finalité.

Premièrement, il s'agit d'offrir une interface de navigation grand public dans les données du DiCo-SQL. La navigation proposée est multiple : 1) de type dictionnaire — à travers la nomenclature (alphabétiquement présentée) de la base, 2) sémantique — à travers le graphe formé par la hiérarchie des étiquettes sémantiques du DiCo, 3) textuelle — à travers la recherche de chaînes de caractères dans les principaux champs de la description lexicographique.

Deuxièmement, le DiCoPop génère automatiquement, à la volée, des articles de dictionnaire de type LAF pour toute unité lexicale possédant un article de statut « 0 » (le statut d'avancement maximal de la description) dans le DiCo. L'utilisation du DiCoPop permet d'avoir accès à un LAF virtuel, régulièrement corrigé et en progression constante pour ce qui est de sa couverture. Il nous permet ainsi d'atteindre un but que nous avions dès le début du projet DiCo : la production automatique d'articles lexicographiques grand public à partir des données formelles d'une base de données explicative et combinatoire. Le DiCo combiné au DiCoPop relève de ce que [Selva et autres, 2003] appellent la *deuxième génération de dictionnaires électroniques* : des dictionnaires dynamiquement produits à partir d'une structure de données lexicographiques indépendante de la structure textuelle dictionnairique.

4. Ce qu'il reste à faire

Nous voudrions brièvement conclure en mentionnant quels sont les quatre buts principaux qu'il nous semble approprié de fixer à la LEC, maintenant que sont franchies les premières étapes d'une informatisation véritable.

Premièrement, et cela est nécessaire pour la mise en place d'une production lexicographique à large couverture, il faut pouvoir stocker les descriptions lexicographiques du DiCo dans une structure de données beaucoup plus souple que celle que nous utilisons actuellement dans le DiCo-SQL. Nous visons ici une structure de graphe particulière, que nous appelons *système lexical*. Nous avons déjà effectué des expérimentations sur la production et la manipulation d'un système lexical à partir du DiCo-SQL (voir [Polguère, 2006]), mais il y a encore loin de la coupe aux lèvres et, pour l'instant, nous n'avons pas dépassé le stade du prototypage.

Deuxièmement, il est vital de construire un éditeur de DiCo ayant les propriétés suivantes :

1. il doit fonctionner en mode textuel, mais extraire et stocker l'information lexicographique directement dans une structure de données de type *système lexical* ;
2. il doit permettre d'effectuer automatiquement des vérifications quant à la cohérence et à la complétude des informations lexicographiques ;

[6] Le DiCoPop est un projet que nous avons mené avec Sébastien Cabot, qui en a assuré la programmation. Son adresse web est : http://olst.ling.umontreal.ca/dicopop.

3. il doit finalement interagir avec un module de génération automatique d'ébauches de descriptions lexicographiques exploitant, d'une part, l'information encodée dans le DiCo et, d'autre part, des règles générales d'inférence fondées sur l'extraction de généralisations (associées aux étiquettes sémantiques, aux configurations de régime,s aux divers liens de fonctions lexicales, etc.).

Troisièment, il faudra exploiter ces nouvelles ressources pour dépasser le stade du maquettage de dictionnaires. Pour cela, la description massive du lexique devra passer par la saturation de vocabulaires « fondamentaux » bien circonscrits, isolés dans une perspective d'enseignement de la langue.

Finalement, il faudra adjoindre aux bases de type DiCo la composante « reine » de la description lexicographique : la définition lexicale. Cette dernière, présente sous forme textuelle dans les DEC, doit être entièrement formalisée et connectée au reste de l'information lexicographique suivant les principes établis, notamment, dans le cadre du travail sur la base de définitions Bdéf [Altman et Polguère, 2003].

Remerciements

Les recherches en lexicologie et lexicographie menées à l'Observatoire de linguistique Sens-Texte (OLST) de l'Université de Montréal sont financées par le Fonds québécois de recherche sur la société et la culture (FQRSC) et le Conseil de recherches en sciences humaines du Canada (CRSH).

Bibliographie

[Alonso Ramos, 2004] Alonso Ramos, M. (2004) : Elaboración del Diccionario de colocaciones del español y sus aplicaciones. In M. P. Bataner et J. DeCesaris (réd.) : *De lexicografia. Actes del I Symposium Internacional de Lexicografía*, Barcelone, Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra.

[Altman et Polguère, 2003] Altman, J. et Polguère, A. (2003) : La BDéf : base de définitions dérivée du Dictionnaire explicatif et combinatoire, *Actes de la première conférence internationale de théorie Sens-Texte* (MTT 2003), Paris, 43-54.

[Binon et autres, 2000] Binon, J., Verlinde, C., Van Dyck, J. et Bertels, A. (2000) : *Dictionnaire d'apprentissage du français des affaires*, Paris, Didier.

[Décary et Lapalme, 1990] Décary, M et Lapalme, G. (1990) : An Editor for the Explanatory Dictionary of Contemporary French (DECFC), *Computational Linguistics*, 16:3, 145-154.

[Fellbaum, 1998] Fellbaum, C. (1998) : *WordNet: An Etronic Lexical Database*, Cambridge MA, MIT Press.

[Hausmann, 1979] Hausmann, F. J. (1979) : Un dictionnaire des collocations est-il possible ?, *Travaux de littérature et de linguistique de l'Université de Strasbourg*, XVII:1, 187-195.

[Jousse et Polguère, 2005] Jousse, A.-L. et Polguère, A. (2005) : *Le DiCo et sa version DiCouèbe*. Document descriptif et manuel d'utilisation, Département de linguistique et de traduction, Université de Montréal. [http://olst.ling.umontreal.ca/dicouebe/DiCoDOC.pdf]

[Mel'čuk, 1996] Mel'čuk, I. (1996) : Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner (réd.): *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia, Benjamins, 37-102.

[Mel'čuk, 1997] Mel'čuk, I. (1997) : *Vers une linguistique Sens*-Texte. Leçon inaugurale (faite le Vendredi 10 janvier 1997), Collège de France, Chaire internationale, Paris.

[Mel'čuk, 2003] Mel'čuk, I. (2003) : Collocations dans le dictionnaire. In T. Szende (réd.) : *Les écarts culturels dans les dictionnaires bilingues*, Paris, Champion, 19-64.

[Mel'čuk, 2006] Mel'čuk, I. (2006) : Explanatory Combinatorial Dictionary. In G. Sica (réd.) : *Open Problems in Linguistics and Lexicography*, Monza, Polimetrica, 225-355.

[Mel'čuk et autres, 1984, 1988, 1992, 1999] Mel'čuk, I. et autres (1984, 1988, 1992, 1999) : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques*, vol. I-IV, Montréal, Les Presses de l'Université de Montréal.

[Mel'čuk et autres, 1995] Mel'čuk, I., Clas, A. et Polguère, A. (1995) : *Introduction à la lexicologie explicative et combinatoire*, Paris/Louvain-la-Neuve, Duculot.

[Mel'čuk et Polguère, 2006] Mel'čuk, I. et Polguère, A. (2006) : Dérivations sémantiques et collocations dans le DiCo/LAF, *Langue française*, 150, 66-83.

[Mel'čuk et Polguère, 2007] Mel'čuk, I. et Polguère, A. (2007) : *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*, Bruxelles, De Boeck & Larcier.

[Mel'čuk et Zholkovsky, 1984] Mel'čuk, I. et Zholkovsky, A. (1984) : *Explanatory Combinatorial Dictionary of Modern Russian*, Vienne, Wiener Slawistischer Almanach.

[Polguère, 1998] Polguère, A. (1998) : La théorie Sens-Texte, *Dialangue*, 8-9, Université du Québec à Chicoutimi, 9-30.

[Polguère, 2000a] Polguère, A. (2000) : Une base de données lexicale du français et ses applications possibles en didactique, *Revue de Linguistique et de Didactique des Langues* (LIDIL), 21, 75-97.

[Polguère, 2000b] Polguère, A. (2000) Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French, *Proceedings of EURALEX 2000*, Stuttgart, 517-527.

[Polguère, 2003] Polguère, A. (2003) : Étiquetage sémantique des lexies dans la base de données DiCo, *T.a.l.*, 44:2, 39-68.

[Polguère, 2006] Polguère, A. (2006) : Structural properties of Lexical Systems: Monolingual and Multilingual Perspectives, *Proceedings of the Workshop on Multilingual Language Resources and Interoperability* (COLING/ACL 2006), Sydney, 50-59.

[Polguère, 2007] Polguère, 7. (2007) : Lessons from the *Lexique actif du français*, *Proceedings of the Third Meaning-Text Conference* (MTT'07), Klagenfurt, *Wiener Slawistischer Almanach*, Sonderband 69.

[Selva et autres, 2003] Selva, T., Verlinde, S. et Binon, J. (2003) : Vers une deuxième génération de dictionnaires électroniques, *T.a.l.*, 44:2, 177-197.

[Sérasset, 1997] Sérasset, G. (1997) : Le projet NADIA-DEC : vers un dictionnaire explicatif et combinatoire informatisé ? In A. Clas, S. Mejri et T. Baccouche (réd.) : *La mémoire des mots*, actes des cinquièmes Journées scientifiques du Réseau « Lexicologie, Terminologie, Traduction » de l'AUF, Tunis, 25-27 septembre 1997, 149-159.

[Steinlin et autres, 2005] Steinlin, J., Kahane, S. et Polguère, A. (2005) : Compiling a "classical" explanatory combinatorial lexicographic description into a relational database, *Proceedings of the Second International Conference on the Meaning Text Theory* (MTT'07), Moscou, 477-485.

Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français

Benoît Sagot (1)
benoit.sagot@inria.fr
Laurence Danlos (1)
danlos@linguist.jussieu.fr

(1) Projet Alpage-INRIA Rocquencourt et Université Paris 7

Mots-clés : Lexiques syntaxique, lexique de référence, validation manuelle, validation sur corpus

Keywords: Syntactic lexicon, reference lexicon, manual validation, corpus validation

Résumé : Cet article présente une méthodologie lexicographique originale pour la constitution d'un lexique syntaxique de référence pour le français. Cette méthodologie se propose de fusionner les informations lexicales issues des principales ressources existant à l'heure actuelle et ensuite de les valider manuellement et sur corpus, dans la continuité de travaux récents déjà engagés sur trois d'entre elles. L'ensemble du processus fait donc appel à une double validation - à l'aide d'interfaces adaptées - reposant à la fois sur des travaux de lexicographie effectués par des linguistes et sur l'exploitation des résultats d'analyseurs de surface et d'analyseurs syntaxiques profonds traitant des corpus volumineux, textuels ou dictionnairiques.

Abstract : This paper introduces an novel lexicographic methodology whose goal is the development of a reference syntactic lexicon for French. This methodology proposes to gather and validate both manually and on corpora lexical information coming from the main existing resources, following recent works on three of them. The proposed process relies on a double validation with appropriate user interfaces: on the one hand thanks to lexicographic work by linguists, and on the other hand by the exploitation of automatically parsed textual and dictionary corpora.

Introduction

Qu'il s'agisse de travaux en linguistique ou en traitement automatique des langues (TAL), l'analyse et l'exploitation manuelle ou automatique de corpus de textes rédigés en français souffrent de la non-existence d'un lexique électronique couvrant pour le français courant,

disposant d'informations linguistiques riches et qui soit librement distribué. Un certain nombre de ressources lexicales pour le français existent pourtant, mais elles ne sont que partiellement satisfaisantes. Certaines se limitent aux informations morphologiques (e.g Multex, Morphalou, DELAS) ou à des informations syntaxiques limitées (TLFi). D'autres (présentées dans les Sections 1 et 2) contiennent des informations syntaxiques plus riches qui ont été soit compilées par des linguistes sur la base de l'introspection soit acquises (semi-)automatiquement à partir de corpus. Les ressources issues de travaux linguistiques présentent les défauts inhérents à la méthode introspective (couverture et tolérance relativement arbitraires) et ne sont souvent pas directement exploitables dans des applications automatiques. À l'inverse, les ressources développées à partir de techniques d'apprentissage sur corpus pâtissent d'un manque de formalisation linguistique et de validation manuelle.
Nous proposons ici une méthodologie lexicographique pour la constitution d'un lexique morphologique et syntaxique de référence du français qui cherche à palier aux limites des deux types d'approches rappelées ci-dessus. L'idée est de fusionner dans un modèle lexical consensuel des informations extraites des principales ressources déjà existantes, et de confronter le résultat à la fois à une validation linguistique manuelle et à une validation sur corpus.
Nous montrons l'intérêt de cette démarche en décrivant (Section 1) des travaux déjà effectués sur la conception d'un modèle lexical et sur son utilisation pour rassembler au sein du Le*fff* (Lexique des Formes Fléchies du Français) des informations provenant d'autres ressources. Nous décrivons ensuite la façon dont nous envisageons d'étendre et de compléter cette démarche à un éventail plus large de ressources (Section 2). Nous montrons dans la Section 3 la façon dont interagiront des travaux linguistiques de validation manuelle et des procédures de validation grâce à des systèmes automatiques d'analyse de corpus textuels et dictionnairiques.

1. Modèle lexical et enrichissement du Le*fff*

La méthodologie lexicographique que nous proposons repose sur un modèle lexical unique et consensuel dans lequel on peut représenter de façon linguistiquement satisfaisante et automatiquement exploitable les informations jugées nécessaires. Nous décrivons ici une version préliminaire de ce modèle, qui est utilisée dans le Le*fff*. Nous présentons ensuite des travaux effectués sur la comparaison et la fusion d'informations lexicales en vue de l'enrichissement du Le*fff*, travaux qui préfigurent la méthodologie proposée dans cet article.

1.1 Modèle lexical

Le modèle lexical utilisé dans la version actuelle du Le*fff*, lexique électronique du français courant (520 000 entrées) librement disponible (Sagot et al. 2006, et www.lefff.net), est une version préliminaire du modèle lexical que nous souhaitons exploiter pour mettre en œuvre la méthodologie décrite dans cet article. Il est issu pour partie de travaux réalisés au sein du projet LexSynt, mené au sein de l'ILF (Institut de Linguistique Française) et dirigé par Sylvain Kahane de 2005 à 2007. L'objectif de LexSynt était de commencer à faire coopérer plusieurs équipes de recherche francophones spécialistes de lexicologie, de modélisation d'informations linguistiques (tant pour les lexiques que pour les formalismes grammaticaux) et de TAL, ces derniers cherchant à coupler dans un même programme informatique lexiques et grammaires.
Parmi les avancées de LexSynt, et grâce à une comparaison entre diverses ressources lexicales [Danlos et Sagot 2007, Sagot et Danlos 2007], il a été mis à jour un consensus sur les informations syntaxiques caractérisant un emploi de verbe simple et plein du français courant.

Ces informations syntaxiques donnent le cadre de sous-catégorisation (la valence, la rection) d'un emploi de verbe plein sous forme de liste de fonctions syntaxiques (Suj, Obj, Objà, etc.)[1] assorties de leurs diverses réalisations de surface (réalisations sous forme de groupe nominal, de groupe adjectival, de complétive, d'interrogative indirecte, d'infinitive ou de clitique, chacune de ces réalisations pouvant être précédées d'une préposition). Soulignons bien que ce modèle lexical se veut indépendant des choix théoriques de ses utilisateurs, et en particulier des théories syntaxiques. C'est ce qui permet au Le*fff* d'être utilisé dans divers outils de TAL (e.g. analyseurs syntaxiques reposant sur TAG [Thomasset et de La Clergerie, 2005] ou sur LFG [Boullier et Sagot, 2005]).
À titre d'illustration, voici deux entrées extraites du Le*fff* pour le verbe *apprendre* dans une construction active standard, ainsi qu'une entrée pronominale (dite « à agent fantôme ») :

Luc apprend la conduite / à conduire à Marie
apprend v [pred='apprendre_teach_1<**Suj**:sn|cln, **Obj**:(sn|cla|à-sinf|scompl|qcompl), **Objà**:(à-sn|cld)>',
cat=v, @personnel, @CtrlObjàObj, @CompInd, @P3s]
Marie apprend la conduite / à conduire
apprend v [pred='apprendre_learn_2<**Suj**:sn|cln, **Obj**:(sn|cla|à-sinf|scompl|qcompl)>',
cat=v, @personnel, @CtrlObjàObj, @CompInd, @P3s]
La conduite / conduire s'apprend facilement
apprend v [pred='apprendre_learn_2<**Suj**:sn|cln|sinf>',
cat=v, @personnel, @pronominal, @P3s]

On y distingue un identifiant sémantique de l'entrée « intensionnelle » correspondante (i.e. l'entrée factorisée de niveau lemme à partir de laquelle sont générées automatiquement les entrées pour chaque forme fléchie du lemme), la liste (entre chevrons) de fonctions syntaxiques assorties de leurs réalisations, et un ensemble de couples attributs-valeurs — souvent factorisés par le biais d'une « macro » préfixée par @ — pour décrire les autres informations morphosyntaxiques (catégorie syntaxique, pronominalité, impersonnalité, contrôles, attributifs, traits morphologiques, etc.).
Ce modèle consensuel issu du projet LexSynt est limité à plusieurs titres. En particulier, la modélisation des entrées autres que les verbes pleins et simples n'a pas été étudiée en détail. Toutefois, c'est en partie grâce à ce modèle que le Le*fff* est utilisé dans divers outils de TAL, et qu'il a pu être enrichi comme décrit ci-dessous.

1.2 Enrichissement du Le*fff* à l'aide d'autres ressources

Des travaux récents sur le Le*fff* ont initié le projet décrit ici. À ce jour, ces travaux ont couvert les constructions impersonnelles [Sagot et Danlos 2007] suite aux travaux de [Danlos 2005], certaines expressions verbales figées (de façon préliminaire) [Danlos *et al.* 2006], et les adverbes en *–ment* [Sagot et Fort 2007]. Ils se sont penchés sur l'extraction, manuelle ou automatique, d'informations du lexique-grammaire (qui est décrit brièvement dans la Section 2). L'intégration de ces informations dans le Le*fff* a été validé par comparaison avec Dicovalence (voir Section 2) en ce qui concerne les constructions impersonnelles [Danlos et Sagot 2007].
Ces travaux, dont les résultats sont encourageants, ont montré qu'il est nécessaire, malgré la richesse du lexique-grammaire, de procéder à un double travail de linguistique et de modélisation, afin d'exploiter son contenu dans un lexique destiné au TAL. Ces travaux préfigurent les phases de fusion et de validation manuelle décrites ci-dessous. Mais il leur

[1] L'inventaire des fonctions syntaxiques ainsi que leurs critères définitoires ont de nombreux points communs avec les « paradigmes » de Dicovalence, ressource décrite dans la Section 2.

manque à la fois une plus grande variété de ressources de départ, une modélisation plus fine des phénomènes, une validation manuelle plus poussée, et une validation sur corpus plus systématique.

2. Intégration de ressources lexicales

2.1 Principales ressources existantes disponibles

La méthodologie décrite dans cet article repose sur la disponibilité (c'est-à-dire l'existence et la libre diffusion) d'un certain nombre de ressources lexicales pour le français, qui, bien que de nature, d'origine, de couverture et de qualité variées, constituent toutes ensemble un corpus de données morphologiques et syntaxiques très important. Parmi les principales ressources lexicales disponibles pour le français, outre le Le*fff* décrit ci-dessus, on peut citer :

- Dicovalence : Ce dictionnaire [van den Eynde & Mertens 2006] est une ressource informatique qui répertorie les cadres de valence de plus de 3 700 verbes simples du français et qui contient plus de 8 000 entrées. Sa particularité réside dans le fait que les informations valencielles sont définies selon les principes de « l'Approche Pronominale » [van den Eynde & Blanche-Benveniste 1978]. Pour chaque place de valence (appelée « paradigme »), Dicovalence précise le paradigme de pronoms qui y est associé et qui couvre « en intention » les lexicalisations possibles. Ensuite, la délimitation d'un cadre de valence, appelé « formulation », repose non seulement sur la configuration (nombre, nature, caractère facultatif, composition) de ces paradigmes pronominaux, mais aussi sur les autres propriétés de construction associées à cette configuration, comme les « reformulations » passives.
- SynLex : Cette ressource [Gardent *et al.* 2006] est issue d'une conversion automatique du sous-ensemble du lexique-grammaire des verbes qui est disponible[2] vers un format mieux approprié pour le TAL. Le lexique Synlex contient 28 000 cadres de sous-catégorisation pour 4 100 verbes.
- DiCo: Le Dictionnaire de Combinatoire [Polguère 2003] est une base de données lexicale du français, développée à l'OLST (Observatoire de Linguistique Sens-Texte de l'Université de Montréal) par Igor Meľčuk et Alain Polguère. La finalité première de cette ressource est de décrire chaque entrée (« lexie ») selon deux axes : les dérivations sémantiques (relations sémantiques fortes) qui la lient à d'autres lexies de la langue et les collocations (expressions semi-idiomatiques) qu'elle contrôle. Cette description s'accompagne d'une modélisation des structures syntaxiques régies par la lexie et d'une modélisation de son sens, sous forme d'étiquetage sémantique. Actuellement, le DiCo inclut 1075 lexies (acceptions) et 25 540 liens lexicaux.
- DicoLPL : Ce lexique du LPL (Laboratoire Parole et Langage) [van Rullen *et al.* 2005] contient 580 000 formes fléchies. Il décrit pour chaque entrée ses traits morphologiques, sa forme phonétisée, sa fréquence et le lemme sous-jacent. Les verbes contiennent quant à eux quelques informations concernant la sous-catégorisation. DicoLPL a été constitué sur la base d'un lexique interne au LPL et complété par croisement de ressources existantes et vérification sur corpus.

[2] Le lexique-grammaire a été développé au LADL (Laboratoire d'Automatique Documentaire et Linguistique) sous la direction de Maurice Gross. Il contient des données électroniques extensives sur les propriétés morphosyntaxiques des foncteurs du français (verbes, noms, adjectifs, adverbes). Il est actuellement maintenu et développé à l'IGM (Institut Gaspard Monge) sous la direction d'Eric Laporte. Seules 60% des données du lexique-grammaire sont actuellement diffusées librement.

2.2 Fusion des informations lexicales

L'intégration des ressources que nous venons de décrire pour former un point de départ au lexique syntaxique de référence se déroule en deux étapes successives : l'adaptation de ces ressources au modèle lexical décrit à la section 1.1 et la fusion de ces ressources dérivées en une ressource préliminaire unique.
L'étape de conversion est délicate, car elle nécessite une interprétation des données présentes dans chaque ressource. Ainsi, notre modèle lexical fait appel à la notion de fonction syntaxique alors que la plupart des ressources lexicales citées ne font pas directement appel à cette notion. Il faut donc « traduire » les informations codées dans ces ressources en termes de fonctions syntaxiques.
Une fois les conversions effectuées, il faut fusionner les ressources dérivées en un seul lexique qui constitue une version préliminaire du lexique de référence. La difficulté majeure de cette étape est la gestion des conflits. Certains devraient pouvoir se résoudre assez facilement. Par exemple, si un actant d'un verbe est marqué comme obligatoire dans une ressource et facultatif dans une autre, il est aisé d'indiquer que le statut obligatoire ou facultatif de cet actant est non déterminé dans cette version préliminaire. D'autres conflits vont demander plus de réflexion. Nous pensons entre autres à la séparation des entrées pour un même lemme. Si un lemme a un certain nombre d'entrées dans une des ressources et un autre nombre d'entrées dans une autre, l'identification des entrées similaires est aisée, mais que faire des autres ? À partir de quel moment va-t-on considérer que deux entrées sont à fusionner, ou, à l'inverse, à distinguer ? Une différence telle que le statut obligatoire ou facultatif d'un actant ne devrait pas compter (voir ci-dessus) ; à l'inverse, une différence de fonctions syntaxiques devrait compter, mais qu'en est-il d'une différence de réalisations de surface des fonctions syntaxiques ? En résumé, cette phase de fusion demande un gros travail collaboratif de lexicologie permettant de mettre au point des heuristiques de résolution automatique des conflits les plus courants.

3. Double validation

Un des objectifs majeurs de la méthodologie proposée est de construire une ressource lexicale dont la pertinence linguistique soit ancrée à la fois dans l'introspection linguistique *et* dans la validation sur corpus. Nous considérons en effet que les avantages et inconvénients de ces deux types d'approches sont complémentaires.
La validation par introspection est confiée à des linguistes qui corrigent les entrées de la version préliminaire du lexique de référence, par exemple en établissant des critères pour gérer les conflits mal traités automatiquement lors de la phase de fusion et en appliquant ces critères de façon systématique. Ce travail linguistique sera guidé par des informations extraites de corpus étiquetés ou analysés superficiellement, ainsi que par des recherches dans des dictionnaires électroniques comme le TLFi. L'objectif de cette validation manuelle est de savoir si telle ou telle entrée est bien attestée en corpus ou, au contraire, de s'apercevoir qu'une entrée est manquante.
L'objectif de la seconde validation est de vérifier et de probabiliser les entrées du lexique de référence par son utilisation dans différents analyseurs syntaxiques. En effet, les recherches actuelles en analyse syntaxique, en particulier dans les pays anglo-saxons, montrent qu'il est indispensable que les descriptions linguistiques utilisées en TAL soient ancrées dans la réalité de corpus concrets, notamment à des fins de validation, de probabilisation et d'évolution dynamique permanente. Dans un premier temps, il est raisonnable de se restreindre aux deux premiers de ces aspects. Pour chaque analyseur syntaxique concerné, il faut en premier lieu adapter l'analyseur afin qu'il puisse fonctionner avec un lexique au format de la ressource. Il

faut ensuite procéder à l'analyse syntaxique de corpus variés et volumineux (plusieurs dizaines de millions de mots) pour :

- identifier les entrées lexicales qui semblent superflues, car n'entrant (presque) jamais dans l'analyse d'aucune phrase par (presque) aucun analyseur ;
- identifier les entrées lexicales probablement erronées ou incomplètes, à l'aide de techniques de fouille d'erreurs dans les résultats d'analyseurs syntaxique telles que celles de [Sagot et de La Clergerie, 2006] ;
- probabiliser les entrées lexicales et les structures syntaxiques qu'elles comportent (fréquence de réalisation des fonctions syntaxiques, fréquences de chaque type de réalisation, etc.) ;
- fournir (si besoin est) des exemples en corpus pour illustrer les entrées lexicales.

La double validation décrite ci-dessus ne peut se faire de façon précise et efficace qu'à l'aide d'un environnement de lexicographie permettant de visualiser aisément les entrées d'un lexique, de les corriger selon des protocoles bien établis, et de connaître leur degré et type de validation. Il faut en effet permettre l'accès à toutes les informations permettant d'accélérer la validation manuelle tout en visualisant ou intégrant les résultats des analyseurs syntaxiques. De plus, dans un contexte de travail collaboratif, il est important de garder une trace des modifications effectuées, de leurs auteurs, voire de commentaires associés pour justifier une décision.

Conclusion

Nous avons présenté dans cet article une méthodologie lexicographique originale, qui fait suite à des travaux prometteurs, et dont le but est d'exploiter au mieux l'existant afin de faire franchir un cap décisif aux ressources disponibles pour le français. Les travaux lexicographique sur le français sont en effet nombreux, mais leur exploitation est en retard par rapport à ce qui se passe pour d'autres langues — et pas seulement pour l'anglais — en raison de l'éparpillement des informations lexicales entre différentes ressources et, parfois, de leur manque de formalisation, en particulier dans une optique de TAL. La méthodologie proposée sera mise en œuvre dans les prochaines années et devrait déboucher sur une vraie ressource syntaxique de référence pour le français, qui sera précise, couvrante, linguistiquement pertinente et directement exploitable et exploitée, tant par des linguistes que dans des outils automatiques.

Bibliographie

[Boullier et Sagot, 2005] Boullier, P. et Sagot, B. (2005) : *Analyse syntaxique profonde à grande échelle : SXLFG,* Traitement Automatique des Langues n°46/2.

[Danlos *et al.*, 2006] Danlos, L., Sagot, B. et Salmon-Alt, S. (2006) : *French frozen verbal expressions: from lexicon-grammar tables to NLP applications,* Actes du Colloque Lexique Grammaire 2006, Palerme, Italie.

[Danlos et Sagot, 2007] Danlos, L. et Sagot, B. (2007) : *Comparaison du Lexique-grammaire et de Dicovalence : vers une intégration dans le Lefff,* Actes de TALN 2007, Toulouse, France.

[Danlos, 2005] Danlos, L. (2005) : *ILIMP : Outil pour reprérer les occurrences du pronom impersonnel* il*,* Actes de TALN 2005, Dourdan, France.

[Gardent *et al.*, 2006] Gardent, G., Guillaume, B., Perrier, G. et Falk, I. (2006) : *Extraction d'information de sous-catégorisation à partir des tables du LADL,* Actes de TALN 2006, Louvain, Belgique.

[Polguère, 2003] Polguère, A. (2003) : *Étiquetage sémantique des lexies dans la base de données DiCo,* Traitement Automatique des Langues n°44/2.

[Sagot *et al.*, 2006] Sagot, B., Clément, L., de La Clergerie, É et Boullier, P. (2006) : *The Lefff 2 syntactic lexicon for French : architecture, acquisition, use,* Actes de LREC 2006, Gênes, Italie.

[Sagot et Danlos, 2007] Sagot, B. et Danlos, L. (2007) : *Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire. Constructions impersonnelles,* Cahiers du Cental, Louvain, Belgique.

[Sagot et de La Clergerie, 2006] Sagot, B. et de La Clergerie, É. (2006) : *Error mining in parsing results,* Actes de ACL-CoLing 2006, Sydney, Australie.

[Sagot et Fort, 2007] Sagot, B. et Fort, K. (2007) : *Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire. Adverbes en* -ment*,* Actes du Colloque Lexique Grammaire 2007, Bonifacio, France (à paraître).

[Thomasset et de La Clergerie, 2005] Thomasset, F. et de La Clergerie, É. (2005) : *Comment obtenir plus des méta-grammaires,* Actes de TALN 2005, Dourdan, France.

[van den Eynde et Blanche-Benveniste, 1978] van den Eynde, K. et Blanche-Benveniste, C. (1978) : *Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale,* Cahiers de Lexicologie n°32 (pages 3–27).

[van den Eynde et Mertens, 2006] van den Eynde, K. et Mertens, P. (2007) : *Le dictionnaire de valence Dicovalence : manuel d'utilisation (version 1.2),* en ligne à l'adresse http://bach.arts.kuleuven.be/dicovalence/.

[van Rullen *et al.*, 2005] van Rullen, T., Blache, P., Portes, C., Rauzy, S., Maeyheux, J.-F., Guénot, M.-L., Balfourier, J.-M. et Bellengier, E. (2005) : *Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales,* Actes de TALN 2005, Dourdan, France.

Prolexbase : Une base de données lexicale de noms propres pour le Tal

Denis Maurel (1)
denis.maurel@univ-tours.fr

(1) *Université François Rabelais Tours - Laboratoire d'informatique*

Mots-clés : nom propre, base de données lexicales, multilingue, traduction, Prolexbase.

Keywords: proper name, lexical database, multilingual, translation, Prolexbase

Résumé : Cet article présente un modèle de description des noms propres dans une base de données lexicale multilingue, Prolexbase, qui permet d'envisager le traitement automatique de cette classe de mot, même dans des situations complexes (translations multilingues, anaphores lexicales…). Il existe, d'une langue à l'autre, un nombre différent de lemmes (et de formes) directement en lien avec un même référent ; ceux-ci sont rassemblés, dans Prolexbase, autour d'un concept multilingue, le *nom propre conceptuel*, qui se projette sur chaque langue en un *prolexème*.

Abstract: This paper presents a model of description of proper names into a multilingual lexical database, Prolexbase that allows proper name processing, even in a complex situation (multilingual translation, lexical anaphora…). From two different languages, the number of lemmas (and forms), from the same referent, is different; into Prolexbase, they are put together a multilingual concept, the *conceptual proper name*, that is projected unto each language as a *prolexeme*.

Introduction

Les noms propres ne figurent que rarement dans les bases de données lexicales utilisées pour le traitement automatique des langues. Or leur importance n'est pas à démontrer : [Coates-Stephens, 1993] estime leur présence dans les textes journalistiques à 10 % et, si nous prenons un exemple littéraire, nous pouvons remarquer que les quatre noms les plus fréquents du *Tour du monde en quatre-vingts jours* de Jules Verne sont des noms propres (*Fogg*, *Passepartout*, *Phileas* et *Fix*), le premier nom commun, cinquième de la liste, étant le mot *heures*. Même en lemmatisant, le lemme *heure* n'arrive qu'en troisième position derrière les deux premiers noms propres.
D'autre part, les noms propres sont souvent la base de dérivations morphologiques, plus ou moins importantes suivant les langues. En français déjà, la plupart des toponymes sont associés à un gentilé, parfois imprévisible (*Tours* - *Tourangeau*) [Eggert, 2005] et un grand nombre de célébrités politiques, scientifiques ou littéraires, ont des "partisans" ou des

"disciples" (*De Gaulle* - *Gaulliste* - *Gaullien*). Ces noms sont souvent absents des dictionnaires de langue, comme le signale par exemple [Rey, 1977].
Si nous nous plaçons dans un contexte multilingue, certaines langues, comme le serbe, ont une productivité morphologique très abondante, d'abord par un système casuel, mais aussi par un système dérivationnel très important : par exemple, à partir d'un nom de personne, on peut créer régulièrement un nom relationnel, un adjectif relationnel, un adjectif possessif, etc. Reprenons l'exemple du roman de Jules Verne : le nom *Passepartout* se retrouve dans la traduction en serbe sous quatre formes (quatre cas) du nom, mais aussi sous six formes de l'adjectif possessif.
Aussi, si certains signalent que l'utilisation en Tal de dictionnaires de nom propre est quasiment inutile pour l'extraction d'information [Mikheev *et al.*, 1999], grâce à l'utilisation d'outils de reconnaissance d'entités nommés [Chinchor, 1997], cela semble exact dans une langue morphologiquement pauvre comme l'anglais et pour ce genre de tâches. Cela est certainement bien différent dans une langue morphologiquement riche [Maurel et al., 2007] et sur des tâches comme la veille multilingue ou l'aide à la traduction, ou même la correction orthographique.
Pour traiter efficacement cette question de l'existence de lemmes et de formes en nombre différent d'une langue à une autre, la base de données lexicale Prolexbase est construite autour d'un concept multilingue, le *nom propre conceptuel*, qui se projette sur chaque langue en un ensemble de lemmes (et de formes) directement en lien avec le nom propre en question, le *prolexème*. Ainsi, le nom propre conceptuel *Passepartout* dans le roman de Jules Verne correspondrait aux ensembles de lemmes :

- Prolexème-Fra$_{\text{Passepartout}}$ = {Passepartout.N}
- Prolexème-Srp$_{\text{Paspartu}}$ = {Paspartu.N, Paspartuov.AP}

Des relations entre les noms propres conceptuels permettent aussi de traiter des anaphores lexicales ou de gloser des translations d'une langue à une autre.
Dans une première partie nous décrirons la structure de Prolexbase, puis, dans une deuxième, la présentation sous un format XML de cette base actuellement disponible sur le site du CNRTL[1]. Puis nous conclurons sur les prochaines perspectives du projet.

1. La structure de Prolexbase

Comme cela vient d'être brièvement présenté en introduction, Prolexbase comporte deux parties bien distinctes, le niveau conceptuel, indépendant de la langue, dont l'élément principal est le nom propre conceptuel, et le niveau linguistique, qui le projette, pour une langue donnée, sur un ensemble de lemmes, le prolexème.

1.1 Le niveau conceptuel

Un nom commun se définit par son (ou ses) sens, alors qu'un nom propre désigne son référent. Les sens ne sont pas universels, les concepts qu'ils représentent peuvent être raffinés dans une langue et non dans une autre (par exemple *rivière* et *fleuve* en français versus *river* en anglais, voir [Mangeot, 2000]). Ils dépendent en fait des langues considérées. Par contre, le référent, lui, ne dépend pas de la langue, pas plus que les relations qu'il entretient avec d'autres référents. Pour cela, nous avons décidé de créer dans Prolexbase un niveau conceptuel, indépendant de la langue considérée, pour y factoriser un maximum d'informations.

[1] http://www.cnrtl.fr/lexiques/prolex/

Pourtant, la notion de référent ne nous a pas paru la mieux à même de répondre à l'attente du traducteur lorsqu'un même référent correspond à plusieurs noms propres. Donnons quelques exemples ; si le terme *Cité phocéenne* est supposé suffisamment clair pour un lecteur en langue cible, on pourra le traduire (par exemple en anglais par son équivalent *Phocean city*), sinon, il faudra le remplacer par son synonyme *Marseille*[2] ; de même, des constructions spécifiques à une langue comme l'*ex-République populaire de Pologne* doivent être adaptées… Pour cela, le nom propre conceptuel qui est le centre de notre modèle n'est pas le référent, mais un point de vue sur celui-ci[3]. Les différents points de vue sur un même référent sont considérés dans Prolexbase comme des synonymes, marqués en suivant la diasystématique de [Coseriu, 1998] :

- diachronique (variété dans le temps), *République populaire de Pologne* et *République de Pologne* ;
- diastratique (variété relative à la stratification socioculturelle), *Paul-Alain Leclerc* et *Julien Clerc* ;
- diaphasique (variété concernant les finalités de l'emploi), la *Cité phocéenne* et *Marseille*.

Cependant, on peut considérer que le référent d'un nom propre est quand même modélisé dans Prolexbase par l'ensemble de tous les points de vue le concernant ; il s'agit donc d'un ensemble de synonymes, une sorte de *synset* pour employer le terme utilisé par le projet Wordnet [Miller *et al.*, 1990].

Quatre autres relations sont placées à ce niveau :

- Une relation partitive, sorte de méronymie étendue, *Tours* est en *Indre-et-Loire*, la *guerre de Corée* est à l'*Époque contemporaine*, le *Deutéronome* est un livre du *Pentateuque*…
- Une relation associative, qui permet l'accessibilité [Ariel, 1990] du référent, *Salammbô* est un roman de *Gustave Flaubert*, l'*Abbé Pierre* est le fondateur des *Compagnons d'Emmaüs*, *Peugeot* est la firme *sochalienne*…
- Deux relations génériques, l'une vers une typologie hiérarchique de trente types et neuf supertypes (présentés sur la Figure 1), l'autre vers un paradigme d'existence comportant trois instances (historique, fictif et religieux).

<table>
<tr><td colspan="6">Nom propre</td></tr>
<tr><td colspan="3">Anthroponyme</td><td colspan="2">Toponyme</td><td>Ergonyme</td><td>Pragmonyme</td></tr>
<tr><td>Individuel</td><td colspan="2">Collectif</td><td colspan="2"></td><td></td><td></td></tr>
<tr><td></td><td></td><td>Groupement</td><td></td><td>Territoire</td><td></td><td></td></tr>
<tr><td>Célébrité
Patronyme
Prénom
Pseudo Anthroponyme</td><td>Dynastie
Ethnonyme</td><td>Association
Ensemble
Entreprise
Institution
Organisation</td><td>Astronyme
Edifice
Géonyme
Hydronyme
Ville
Voie</td><td>Pays
Région
Supranational</td><td>Objet
Œuvre
Pensée
Produit
Vaisseau</td><td>Catastrophe
Fête
Histoire
Manifestation
Météorologie</td></tr>
</table>

Figure 1. Les types de Prolexbase

Le niveau conceptuel de Prolexbase constitue une sorte d'ontologie des noms propres, présentée Figure 2.

[2] Et peut-être même aussi ajouter une glose déduite des relations de méronymie et d'expansion classifiante (voir plus loin), comme *la ville française de Marseille.*

[3] Il est représenté dans la base par un identifiant (un numéro de pivot).

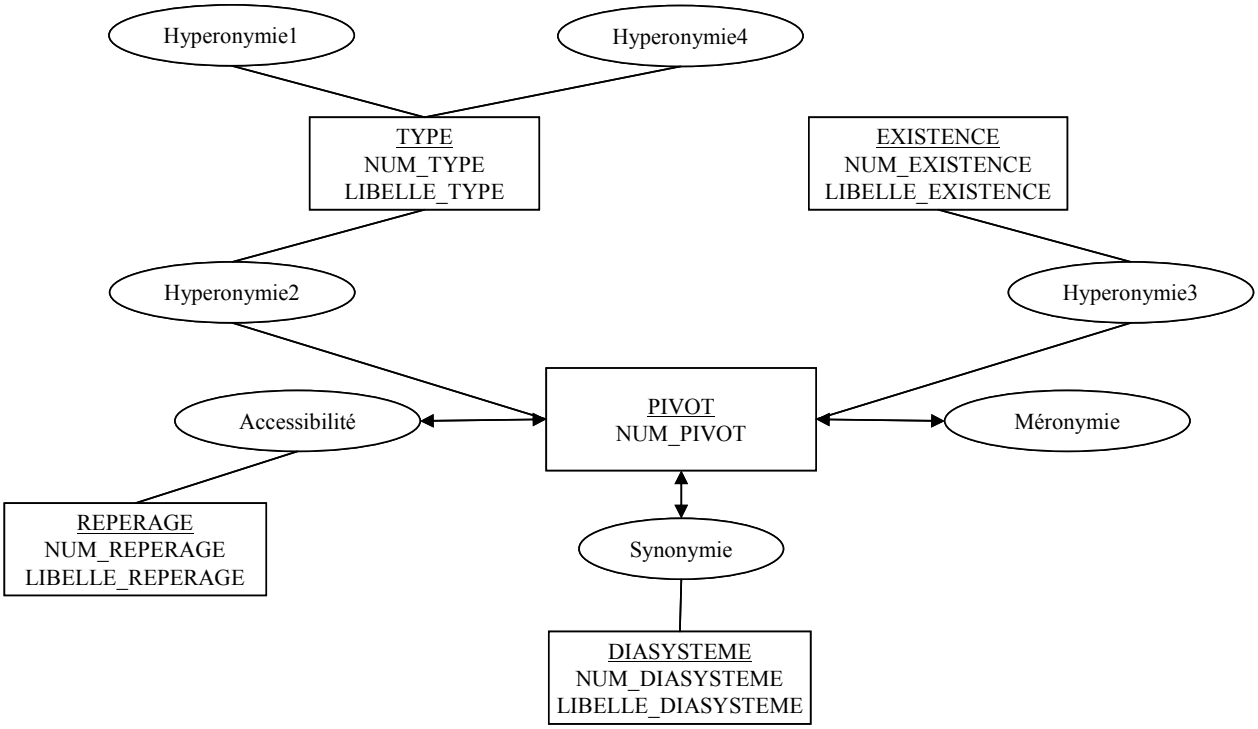


Figure 2. L'ontologie des noms propres de Prolexbase

1.2 Le niveau linguistique

Pour une langue donnée, le niveau linguistique est principalement un ensemble de lemmes, le prolexème, considéré comme la projection d'un unique nom propre conceptuel. Ce prolexème comprend tout d'abord différents alias du nom propre, forme intégrale, nom usuel, variante, abréviation, forme courte, sigle, acronyme, forme translittérée, transcrite ou romanisée, quasi synonyme… Il comprend aussi les dérivés dont le sens se déduit de celui du nom propre. Par exemple, Finlandais, qui est le nom relationnel associé à Finlande, sera dans son prolexème, mais, le verbe finlandiser, non, car son sens ne se déduit pas du nom Finlande, bien que cette dérivation soit productive morphologiquement (balkaniser, libaniser… [Tolédano, Candel, 2002]). Pour simplifier l'accès au nom propre, la forme intégrale a été choisie comme forme principale et est parfois abusivement appelée aussi prolexème. Donnons deux exemples :

– Prolexème-Fra$_{\text{Finlande}}$ = {Finlande.N, Finlandais.NR, finlandais.AR}

– Prolexème-Fra$_{\text{Organisation des Nations unies}}$ = {Organisation des Nations unies.N, Nations unies.N, Onu.N, Onusien.NR, onusien.AR}

Quatre relations (qui dépendent de la langue) sont associées au prolexème :

– Trois relations syntagmatiques, la cooccurrence (pour l'instant, en français, on y note la présence ou l'absence de déterminant), l'expansion classifiante (collocation libre), le *pape* Jean-Paul II, la *ville* de Tours… et l'éponymie (lexicalisation et figement), qui concerne l'antonomase du nom propre (un *bic*), la terminologie (la *maladie de Creutzfeldt-Jakob*) et les idiomes (*tous les chemins mènent à Rome*).

– Une relation de notoriété (usage rare, peu fréquent ou fréquent).

Enfin, des règles de flexion (mono- et polylexicales), utilisant le système Multiflex [Savary, 2005], permettent la génération de toutes les formes associées (Figure 3).

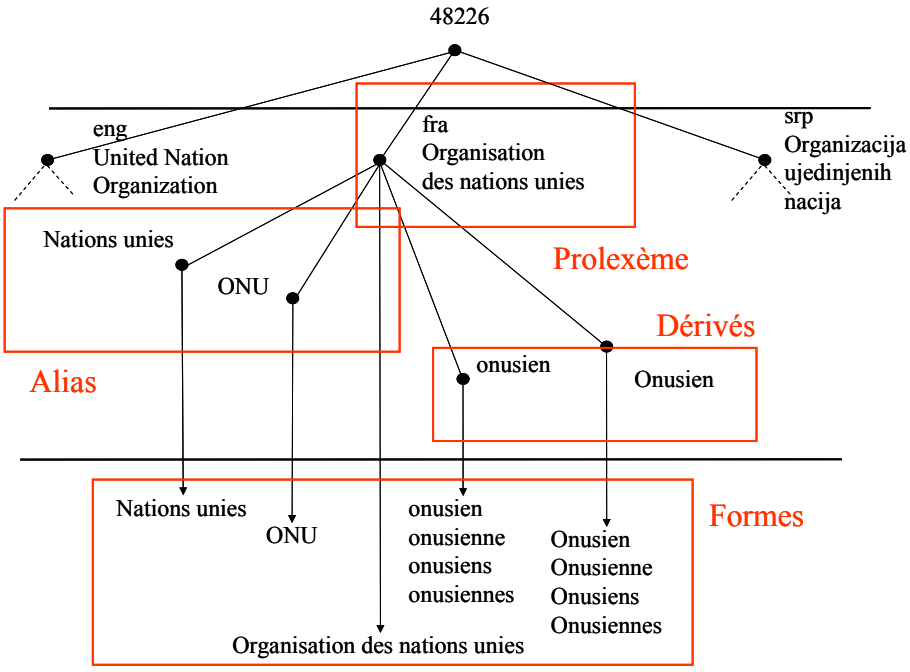


Figure 3. Le prolexème et les formes associés au nom propre *Organisation des Nations unies*

2. Le format XML actuel

La partie française de Prolexbase est actuellement disponible sur le site du CNRTL, sous une licence libre (LGPL-LR), dans un format XML inspiré de la norme ISO 16642[4], mais sans utiliser les termes préconisés par la norme ISO 12620[5].
Comme dans Prolexbase, l'entrée est le nom propre conceptuel (Figure 4), ou pivot interlingue, suivi de son type, de son existence et des prolexèmes propres à chaque langue. A l'intérieur du prolexème, on trouve ses alias et dérivés. Chaque lemme est suivi de la liste de ses formes (Figure 5).
C'est dans la partie interlingue, juste après le pivot, que sont placées les relations de synonymie, de méronymie et d'accessibilité (Figure 6).

[4] La norme TMF (*Terminological Markup Framework*) précise l'organisation des bases terminologiques.

[5] Cette norme s'intitule *Aides informatiques en terminologie - Catégories de données* et offre un vocabulaire normalisé pour la définition des bases de données terminologiques.

```
<struct type="Prolex">
  <struct type="pivot">
        <feat type="type">Organisation</feat>
        <feat type="existence">Historique</feat>
        <feat type="identifier">48226</feat>
…
        <struct type="prolexeme">
              <feat type="language">fr</feat>
              <feat type="lemma">Organisation des nations unies</feat>
              <feat type="pos">name</feat>
              <feat type="category">proper name</feat>
…
              <struct type="alias">
                    <feat type="lemma">ONU</feat>
                    <feat type="pos">name</feat>
                    <feat type="category">Acronyme ou sigle</feat>
…
              </struct>
              <struct type="alias">
                    <feat type="lemma">Nations unies</feat>
                    <feat type="pos">name</feat>
                    <feat type="category">Abréviation</feat>
…
              </struct>
              <struct type="derivative">
                    <feat type="lemma">Onusien</feat>
                    <feat type="pos">nom</feat>
                    <feat type="category">Nom relationnel</feat>
…
              </struct>
              <struct type="derivative">
                    <feat type="lemma">onusien</feat>
                    <feat type="pos">adj</feat>
                    <feat type="category">Adjectif relationnel</feat>
…
              </struct>
        </struct>
  </struct>
</struct>
```

Figure 4. Le nom propre conceptuel et le prolexème français associés au nom propre *Organisation des Nations unies*

```
              <struct type="derivative">
                    <feat type="lemma">Onusien</feat>
…
                    <struct type="inflection">
                          <feat type="form">Onusiens</feat>
                          <feat type="gender">masculine</feat>
                          <feat type="number">plural</feat>
                    </struct>
…
              </struct>
```

Figure 5. La forme *Onusiens*

```
<struct                                                type="relationship">
        <feat                                  type="relation">accessibility</feat>
        <feat                                  type="identifier">44712[6]</feat>
        <feat                                    type="argument">arg1</feat>
        <feat                                    type="context">Siège</feat>
</struct>
```

Figure 6. Une relation associée au nom propre *Organisation des Nations unies*

Conclusion

Nous venons de présenter un modèle de description des noms propres dans une base de données lexicale multilingue, Prolexbase, qui permet d'envisager le traitement automatique de cette classe de mot, même dans des situations complexes (translations multilingues, anaphores lexicales…).
Prolexbase est disponible sous une licence libre sur le site du CNRTL, dans un format XML inspiré de TMF. Plus de détails sur ce projet sont donnés dans [Tran, Maurel, 2006]. Nous comptons améliorer prochainement ce format en utilisant les catégories de données de la norme ISO 12620. Nous travaillons aussi sur la définition d'un format XML compatible avec la future norme LMF[7], afin d'extraire de Prolexbase un dictionnaire de noms propres et de leurs dérivés.

Bibliographie

[Ariel, 1990] Ariel, M. (1990), *Accessing Noun Phrases Antecedents*, Routledge, London.

[Chinchor, 1997] Chinchor, N. (1997), *Muc-7 Named Entity Task Definition*, consultable sur le site http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html.

[Coates-Stephens, 1993] Coates-Stephens, S. (1993), *The Analysis and Acquisition of Proper Names for the Understanding of Free Text*, Kluwer Academic Publishers, Hingham, MA.

[Coseriu, 1998] Coseriu, E. (1998), Le double problème des unités dia-s, *Les Cahiers δια. Etudes sur la diachronie et la variation linguistique* 1:9-16.

[Eggert, 2005] Eggert, E. (2005), *Bisontins ou besançonnais? À la recherche des règles pour la formation des gentilés pour une application au traitement automatique*, Tübinger Beiträge zur Linguistik: Band 480, Softcover, Germany.

[Mangeot, 2000] Mangeot M. (2000), Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links, *WAINS'7, 7th Workshop on Advanced Information Network and System*, Kasetsart University, Bangkok, Thailand.

[Maurel et al., 2007] Maurel D., Vitas D., Krstev S., Koeva S. (2007), Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian, Bulag, 32 (à paraître).

[Mikheev *et al.*, 1999] Mikheev, A., Moens, M., Grover, C. (1999), Named entity Recognition without Gazetteers, *EACL'99*:1-8.

[Miller *et al.*, 1990] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. (1990), Introduction to WordNet : an on-line lexical database, *International Journal of Lexicography*, n°3, p. 235-244.

[Rey, 1977] Rey, A. (1977), *Le lexique : images et modèles. Du dictionnaire à la lexicologie*, Paris, Armand Colin.

[6] Il s'agit du pivot correspondant au nom propre *Genève*.

[7] La norme LMF (*Language resource management - Lexical markup framework*) décrit une organisation générale des dictionnaires et des informations morphologiques, syntaxiques ou sémantiques liées aux entrées.

[Savary, 2005] Savary, A. (2005), A formalism for the computational morphology of multi-word units, *Archives of Control Sciences, 15(LI)*, Silesian University of Technology.

[Tolédano, Candel, 2002] Tolédano, V., Candel, D. (2002), Dérivation suffixale de toponymes :étude d'un terrain propice à la création lexicale, *Meta*, 47-1:105-124.

[Tran, Maurel, 2006] Tran M., Maurel D. (2006), Prolexbase : Un dictionnaire relationnel multilingue de noms propres, *Traitement automatique des langues*, Vol. 47-3, (à paraître).

Le «Trésor» de la langue roumaine

Monica Busuioc (1)
monica.busuioc@yahoo.com
Florin Vasilescu (1)
florin_vasilescu2000@yahoo.com

(1) Institut de Linguistique « Iorgu Iordan – Al. Rosetti » près l'Académie Roumaine, Bucarest

---

La rédaction du grand *Dictionnaire de la langue roumaine* touche à sa fin. Il est connu sous le nom de *Dictionnaire de l'Académie,* car il avait représenté l'un des desiderata, dès sa fondation, de la Société Académique Roumaine, devenue par la suite l'Académie Roumaine. En 2008, une fois publiées les trois lettres qui sont encore en voie de parution ou de mise au point, s'achèvera l'élaboration du plus ample ouvrage de la lexicographie roumaine.

L'histoire du *Dictionnaire,* qui a commencé en 1906 et a continué pendant un siècle, a été marquée par de nombreuses vicissitudes et par certaines interruptions pendant et après les deux guerres mondiales. Sous la direction du grand linguiste roumain Sextil Puşcariu ont été publiées les lettres *A – J*, qui forment la première partie du dit *Dictionnaire,* connue sous le sigle DA (la lettre *E* n'a pas été rédigée, et les lettres *D* et *L* ont paru partiellement).

L'élaboration du *Dictionnaire* a été reprise en 1959 et a continué sans interruption jusqu'à présent dans le cadre des départements de lexicographie et de lexicologie des instituts spécialisés de Bucarest, Cluj-Napoca et Iaşi près l'Académie Roumaine. Au cours de cette période est parue la deuxième partie du *Dictionnaire,* portant le sous-titre *Nouvelle série* et connue sous le sigle DLR. Elaborée sous la direction de Iorgu Iordan, Alexandru Graur et Ion Coteanu et, à partir de l'an 2000, Marius Sala et Gheorghe Mihăilă, tous membres de l'Académie Roumaine, elle comprend les lettres *M – Z.*

Les lettres *D, E* et *L* seront achevées sous peu. Le *Dictionnaire* comprendra 13 tomes en 36 volumes, totalisant plus de 20.000 pages et plus de 175.000 entrées, y compris les variantes.

C'est un dictionnaire historique, descriptif et normatif, dont les principes sont inspirés de la grande tradition lexicographique française pour ce qui est du type de dictionnaire, sa structure, la typologie des définitions et de la filiation des sens, l'étymologie des mots, etc. À l'instar du *Trésor de la langue française,* notre *Dictionnaire* est souvent dénommé à son tour *trésor,* en vertu de sa grande richesse lexicale, illustrée à l'aide d'amples citations appartenant à toutes les époques, depuis les plus anciens textes jusqu'à nos jours. La structure des deux séries du *Dictionnaire de la langue roumaine* est restée la même pour l'essentiel: nature de l'inventaire, choix des mots vedettes, organisation des articles et principes de leur rédaction, ce qui confère au dictionnaire une unité remarquable. Les quelques différences ont trait notamment aux modalités concrètes de réalisation des différentes composantes du *Dictionnaire*; ainsi, on a malheureusement renoncé à traduire en français les entrées et les définitions, considérant qu'il s'agit d'un dictionnaire monolingue et que la langue roumaine est maintenant plus connue qu'au début du XX^{e} siècle, et d'autre part, on a renoncé à regrouper autour d'un terme vedette placé en entrée les dérivés et les composés qui s'y rattachent par leur sens, méthode de travail usitée à l'époque.

Le *Dictionnaire de la langue roumaine* une fois achevé, la nécessité s'impose d'en préparer une nouvelle édition, plus moderne par rapport surtout à sa première édition. Le problème qui se pose est celui de la manière dont cette nouvelle version devra être réalisée. Une seule chose est certaine à ce moment : dans notre époque dominée par l'informatisation, la nouvelle édition du *Dictionnaire de l'Académie* ne saurait être qu'informatisée, suivant le modèle des grands dictionnaires, dont le *Trésor de la Langue Française,* ce qui situera la lexicographie roumaine au niveau mondial, contribuant au développement de la collaboration internationale dans le domaine de l'étude comparée des langues et des cultures.

Le caractère extrêmement laborieux et de longue durée du travail lexicographique, d'une part, et l'essor de l'informatique en Roumanie après 1989, d'autre part, ont amené la possibilité, voire la nécessité, d'informatiser notre dictionnaire, processus qui a connu deux directions.

La première consiste en la création d'une archive électronique de textes comprenant toutes les sources dépouillées en vue de l'élaboration du *Dictionnaire.* C'est à cette opération que se consacrent les lexicographes de l'Institut de Linguistique « Iorgu Iordan – Al. Rosetti » de Bucarest.

La deuxième direction consiste en la conversion des volumes du *Dictionnaire* par l'annotation en format *XML*[1] de toutes les entrées. Cette opération a été assumée par l'équipe de l'Institut de Philologie Roumaine « Al. Philippide » de Iaşi et par un groupe d'informaticiens de l'Université « A. I. Cuza » de Iaşi, en coopération avec les linguistes de Bucarest et de Cluj-Napoca.

Ces deux directions, qui forment une démarche cohérente, répondent à la nécessité de développer des instruments susceptibles de faciliter et d'accélérer le travail en lexicographie. Ce processus a été favorisé par la création de logiciels adaptés aux particularités de l'orthographe du roumain. Les tentatives, remontant à plus de dix ans, de balayer des textes roumains en format électronique n'ont pas eu au début le succès escompté de par l'absence de logiciels adaptés aux polices roumaines. Le taux de transfert en format texte des logiciels de reconnaissance optique des caractères (type OCR – Optical Character Recognition) était assez faible; c'est pourquoi, jusqu'à la création de versions appropriées de ces logiciels, on n'a pas pu obtenir de résultats significatifs.

Vu que l'une des opérations essentielles de la rédaction d'un dictionnaire est le dépouillement des sources, c'est dans cette direction que l'équipe de Bucarest a dirigé son attention. L'importance d'une base de textes en format électronique, qui soit accessible à tous les chercheurs – et non seulement à ceux qui travaillent au *Dictionnaire* – est indiscutable. Elle est l'une des conditions préliminaires et indispensables de l'informatisation du travail lexicographique. Le dépouillement des textes est une opération extrêmement difficile, lente et coûteuse. Le transfert des sources du *Dictionnaire* en format électronique est devenu possible grâce au développement des outils informatiques et des équipements périphériques. Ce processus a tenu compte des aspects propres à l'activité lexicographique ayant pour objet la langue roumaine – limites, mentionnées ci-dessus, des logiciels et difficultés de la transcription des textes écrits en alphabet cyrillique notamment –, ainsi que des questions concrètes concernant l'organisation du travail, son financement, etc. Il nous faut mentionner qu'au début, une partie des textes ont été saisis en format Word Perfect ou Word par des personnes ayant une formation philologique insuffisante (étudiants ou opérateurs), ce qui a imposé par la suite un corrigé laborieux et soigné.

Aussi à l'Institut de Linguistique de Bucarest est-on en train d'élaborer, dans le cadre d'un programme financé par le Conseil National de la Recherche Scientifique de l'Enseignement Supérieur, un projet visant à réaliser une archive électronique de textes et un

[1] XML («eXtended Markup Language») est un langage de balisage pour la mise en forme, le partage et la conversion des ressources textuelles et terminologiques.

corpus de référence de la langue roumaine. La base de textes devra comprendre, au terme de sa réalisation, les plus de mille ouvrages de la bibliographie du *Dictionnaire,* auxquels s'ajouteront de nouvelles sources capables d'offrir une image fidèle de la circulation actuelle des mots en roumain. Une attention spéciale devra être accordée à la sélection des sources pour la nouvelle édition du *Dictionnaire*, afin de maintenir uniquement les dernières éditions critiques pour des œuvres parues le long du temps et mentionnées dans la bibliographie des éditions successives. Toutefois, à la différence du *Trésor de la langue française,* nous pensons que le Dictionnaire devra comprendre toutes les périodes de la langue roumaine écrite. Ceci à cause de la précarité des grands dictionnaires du roumain en ce qui concerne l'inventaire des mots et leurs premières attestations.

Comme nous venons de le préciser, pour commencer, les textes en format électronique ont été saisis à l'ordinateur. À présent, grâce à l'acquisition de l'infrastructure technique nécessaire, on a fait appel au scanner. L'archive électronique de textes et le corpus de référence formeront une base de données représentative pour la langue roumaine, qui servira aux futurs projets de notre institut non seulement dans le domaine de la lexicographie, mais aussi dans celui de l'histoire du roumain ou de la linguistique appliquée, etc. L'emploi de ces instruments de travail permet de traiter plus rapidement les données, d'identifier les premières attestations des mots, d'établir la filiation des sens, d'enrichir la structure sémantique des articles, d'élucider certaines étymologies controversées, etc.

La réalisation de l'archive électronique de textes s'effectue en base de principes précis et se fonde sur un grand nombre de textes d'époques différentes, depuis le XVI^{e} siècle (date des premiers textes attestés en roumain) jusqu'à nos jours et appartenant à toutes les variétés fonctionnelles. L'un des objectifs de ce projet vise également à résoudre le problème des périodes de «vide d'attestation», lorsqu'un mot semble avoir disparu de la circulation, pour réapparaître après un certain intervalle.

C'est pourquoi, pour créer cette archive, nous avons commencé par enregistrer en format électronique les sources les plus anciennes. L'archive, enrichie continuellement, devra assurer par la suite une répartition uniforme des textes au cours de toutes les époques, jusqu'au XXI^{e} siècle. En plus de l'attention accordée aux sources présentes dans la bibliographie du *Dictionnaire de la langue roumaine*, choisies sur l'avis des spécialistes en matière d'histoire de la langue et de la littérature, les chercheurs travaillant à la constitution du corpus font preuve d'une ouverture particulière à la langue roumaine actuelle. L'archive et le corpus pourront devenir ainsi une importante ressource pour les linguistes, visant la découverte rapide de certains mots absents jusqu'à présent de l'archive de l'Institut, mais surtout d'attestations plus anciennes ou de sens non attestés des mots existants dans le fichier manuel de l'institut. Bien que la base actuelle de textes ne contienne qu'une petite partie des sources du *Dictionnaire,* l'avantage immédiat du travail sur ordinateur a déjà pu être vérifié lors de la rédaction de la lettre *D* de l'édition actuelle (dont les trois parties sont déjà parues en 2006 et 2007). Nous étudions maintenant la possibilité d'utiliser un moteur de recherche pour transférer des textes utiles pour notre entreprise à partir de l'Internet.

Il est certain que la grande complexité des contextes et des situations mis en évidence par l'utilisation d'un tel instrument conduira à la réévaluation des modalités traditionnelles de travail en lexicographie et diminuera dans une mesure significative l'effort humain et la durée de la documentation et de la rédaction de la prochaine édition du *Dictionnaire de l'Académie.*

La constitution de l'archive et du corpus tient compte de la variété des textes choisis et de leurs dimensions. Nous sommes en train d'effectuer l'unification typographique et de la graphie des textes et l'insertion des signes diacritiques. Nous envisageons une utilisation du corpus aussi large que possible, grâce à l'annotation au niveau du mot d'informations concernant le mot de base et sa description morphosyntaxique. Les textes utilisés à présent sont annotés et corrigés en vue de lever les ambiguïtés. Nous utilisons les balises

recommandées par la norme EAGLES pour l'annotation morphosyntaxique, ainsi que le marqueur automatique TnT (Thorsten Brants 'TnT – A Statistical Part of Speech Tagger, 2000), ce qui permettra d'améliorer les normes d'annotation du corpus en les adaptant aux particularités du roumain.

Chaque texte contient des informations bibliographiques (auteur, date, source) et concernant le genre littéraire (poésie, prose, etc.), le type de texte (artistique, informatif, scientifique), le domaine (vie quotidienne, politique, législation, etc.). Nos textes contiennent des annotations au niveau de la page de l'édition utilisée comme source, page qui doit être indiquée lorsque l'on emploie des citations tirées de ces textes pour illustrer les définitions du *Dictionnaire.* L'annotation des textes se réalise en format XML, conformément à la norme XCES[2], par l'élaboration de la définition correspondant au type de document, dénommée DTD (Document Type Definition), qui spécifie la structure de l'information annotée par des éléments et des attributs.

Au corpus de textes constitué par des linguistes s'ajoute un instrument de travail conçu par des informaticiens en coopération avec des linguistes. Il s'agit d'un logiciel de concordances construisant un index des mots contenus dans les textes, fournissant les contextes dans lesquels ces mots apparaissent et permettant des recherches selon divers attributs XML du texte annoté. La capacité actuelle du logiciel est d'environ 100 Mb. Il est amélioré en permanence afin de permettre le traitement d'un volume croissant de textes. Son optimisation se fonde sur la variante Lucene du moteur de recherche Java.

Le logiciel a été adapté aux modifications d'annotation des textes par la généralisation des options, évitant ainsi la nécessité de connaître les éléments et les attributs XML utilisés dans le corpus, extraits automatiquement par le logiciel de concordances. Cette généralisation rend possible l'utilisation du logiciel, de manière indépendante, par n'importe quel utilisateur. De ce point de vue, la caractéristique essentielle du format XML est de décrire la structure logique d'un document indépendamment de sa présentation. L'existence du corpus de textes s'avèrera utile en corrélation avec d'autres instruments d'étude du roumain et avec les recherches linguistiques assistées par l'ordinateur dans le domaine de la dialectologie, qui sont en cours à Iaşi et à Cluj-Napoca.

La deuxième direction mentionnée, consistant en la conversion en format XML de toutes les entrées du dictionnaire, a en vue, dans une première phase, le balayage et la reconnaissance optique des caractères de tous les tomes de l'ouvrage. Ces opérations sont effectuées par des informaticiens de Iaşi, tout le matériel devant être collationné ensuite par les linguistes de Bucarest, Iaşi et Cluj-Napoca. Cette démarche vise la réalisation d'une version informatisée du dictionnaire actuel, sans aucune modification, susceptible d'être consultée sur Internet et de servir de base à une future édition du dictionnaire.

Nous espérons qu'à l'aide des instruments de travail informatiques, la nouvelle édition du *«Trésor» de la langue roumaine* pourra être élaborée beaucoup plus rapidement et plus facilement et qu'elle sera moins sujette aux omissions et aux différences de documentation et de rédaction, inévitables jusqu'à l'introduction de ces facilités modernes dans le travail lexicographique. Le nouveau *Dictionnaire* en format électronique sera ainsi accessible à un plus grand nombre d'utilisateurs.

Ces démarches vers l'informatisation ne sont que le commencement. Une fois achevée la base de textes, comprenant toutes les sources bibliographiques et la version informatisée du dictionnaire actuel, le travail en vue d'une nouvelle édition pourra enfin commencer. Il faut encore réfléchir à la manière dont elle sera élaborée. Toutes proportions gardées, le problème est similaire à celui posé à l'occasion du lancement du projet du *Trésor de la langue française*, à savoir : faut-il rééditer tel quel le dictionnaire existant, le soumettre à une révision attentive ou le remplacer par un dictionnaire entièrement nouveau ?

[2] XCES, application de la norme XML, utilisée pour l'annotation des corpus textuels (Corpus Encoding Standard – CES).

Nous estimons que, disposant de la version informatisée du dictionnaire actuel, il ne faudrait pas le rééditer tel quel. On pourrait toutefois le réimprimer sur papier, afin de mettre à la disposition des spécialistes un outil qui n'existe aujourd'hui que dans un nombre très limité d'exemplaires.

L'idée de soumettre le dictionnaire à une révision attentive vise surtout la première partie, vraiment datée de pas sa parution au cours de la première moitié du XX^{e} siècle. Refaire la première partie (lettres A – J) sur le modèle de la deuxième est une solution de compromis. Car il y a des problèmes qui exigent réflexion, telles la nomenclature (d'un côté la préférence pour les termes du vocabulaire ancien et historique de même que pour les termes régionaux et dialectaux mentionnés sous toutes les variantes et formes possibles et avec toutes les attestations et de l'autre la réserve vis-à-vis des néologismes appartenant au vocabulaire scientifique et technique actuel) ; la définition (qui manque quelquefois de concision) ; les exemples (la multitude, voire l'abondance pour les mots polysémantiques), etc.

Autant de raisons qui nous déterminent de croire que la nouvelle édition devrait être en fait un dictionnaire nouveau et, à ce titre, l'exemple du *Trésor de la langue française* pourrait être pour nous un modèle. Une collaboration souhaitable avec les auteurs du *Trésor de la langue française* sera également bénéfique.

The Dictionary of the Danish Language Online: From Book to Screen – and Beyond

Henrik Lorentzen (1)
hl@dsl.dk
Lars Trap-Jensen (
ltj@dsl.dk

(1) Det Danske Sprog- og Litteraturselskab (Society for Danish Language and Literature), Christians Brygge 1, DK-1219 Copenhagen

---

Keywords: historical dictionary, digitisation, XML mark-up, supplementary volumes, dating information, advanced search, corpus, WordNet

Abstract : The Dictionary of the Danish Language was digitised and published online for the first time in 2005. The paper first describes the digitisation method (the so-called double-keying) and then presents and discusses some of the challenges for the future, e.g. the integration of supplementary volumes, a refined mark-up and better dating information. Finally, the perspectives involved in the integration of corpora, dictionaries and other language resources are briefly introduced.

Introduction

This paper is concerned with a large-scale monolingual dictionary, the Danish counterpart to le Trésor de la Langue Française, to the Oxford English Dictionary, to the Woordenboek der Nederlandsche Taal and others, i.e. historical dictionaries in many volumes that aspire to be comprehensive. The Dictionary of the Danish Language (Ordbog over det danske Sprog, henceforth ODS) was published from 1919 to 1956 in 28 volumes. It covers the Danish language from about 1700 until about 1950. During the compilation period, editorial guidelines changed considerably, and as the staff continued to collect material, the inevitable consequence was that the last part of the alphabet was covered more comprehensively than the first. Soon after the last volume was published it was decided to start gathering material for a supplement. With only a small staff, it took some decades to complete, but from 1992 to 2005 five supplementary volumes were published, meaning that the complete work comprises 33 volumes.

The original dictionary was conceived by the Danish linguist Verner Dahlerup as a much smaller project, originally designed to be carried out by one man, himself, but even before the first volume was published, he realised that the job far exceeded his working capability, and it was decided that the project should come under the auspices of the Society for Danish Language and Literature (Det Danske Sprog- og Litteraturselskab, DSL). It is also within the framework of this institution that the work with the online version of the dictionary takes place.

The goals of the current project, which runs from 2004 until 2010, are threefold:

1. to provide public online access to the dictionary
2. to integrate the supplementary volumes with the original dictionary
3. to integrate the new supplemented dictionary with other linguistic resources developed by DSL, in particular The Danish Dictionary (Den Danske Ordbog, DDO, a medium-size dictionary of modern Danish) and corpora of modern Danish

Digitisation

Since the original was only available as a printed book, the first step was raw digitisation. Different methods have been explored by other projects:

1. keying and proofreading (OED)
2. scanning and proofreading (SAOB, the large Swedish dictionary)
3. double-keying without proofreading (Grimm's Deutsches Wörterbuch)

In this project it was decided to adopt the model used for the German dictionary founded by the Grimm brothers, Deutsches Wörterbuch, the so-called double-keying method. The printed dictionary data is keyed twice, by two independent typing teams, and afterwards the two versions are compared electronically. In this way, the number of typing errors is reduced to almost nothing, and thus the labour-intensive process of proofreading that is normally applied can be avoided.
This part of the project was carried out in collaboration with the University of Trier, and monitored by the same department which was responsible for the Grimm project (http://germazope.uni-trier.de/Projects/KoZe2/). The actual typing of the dictionary took place in Nanjing, China, by the company TQY DoubleKey which has specialised in this type of assignment and can deal with different sorts of typefaces including black-letters and even hand-written manuscripts. After the keying process, the two versions were automatically compared in Trier and a list of discrepancies generated. The list was subsequently processed semi-automatically and manually. All the difficult and dubious cases had been specially marked by the keyboarders, and about 2,000 instances had to be solved by the editors in Copenhagen because native-speaker competence was required. In many cases the answer was straightforward for a native speaker, but in other cases it was necessary to go back to the original slip to resolve the issue. The percentage of genuine errors has not yet been calculated, but spot checks carried out by the Grimm project yielded a result as low as 1 error in 33,000 characters – hardly surprising when you think of it, as it is highly unlikely that two keyboarders would make identical mistakes at exactly the same place. In the ODS, the rate may be expected to be even lower, an estimated 1 in 100,000 characters, because of a clearer structure and typography in the printed text.
The ultimate aim is to establish a fine-grained XML mark-up of the dictionary text, but it cannot be done in a single round. As a first step, a crude mark-up has been implemented in which only the headword, homograph number (if any) and the part-of-speech are identified. The rest of the entry is treated as one chunk. This mark-up allowed us to release a first preliminary version in November 2005 where the only possible search was for headwords. A second version was launched in April 2006 where wildcards and parts-of-speech were

introduced as search criteria. The new features have improved the online version, no doubt, but there is still a lot of work to be done.
The files from Trier are in a format close to the TUSTEP[1] standard where every typographic detail from the book is rendered by means of codes: font size, bold, italics, spacing; special characters like the Danish *æ*, *ø* and *å*, and symbols used as labels (e.g. anchor = nautical language; book = literary; note = music). The files also give exact information about the page and the line in the dictionary, information that will prove useful when cross-references are to going be processed.
In the current online version the dictionary contains more than 180,000 headwords to which number may be added about 70,000 sub-headwords (words that do not have their own entry, but are nested within another entry) and supplementary articles which will take the total number to 250,000. The number of definitions, citations and multi-word units is still unknown but the improved mark-up will reveal it in a couple of years.

Future work

1. Integrating the supplement

An important task is to integrate the five supplementary volumes into the original dictionary, which is far from being a trivial issue. In the online version of the OED (the so-called third version), the editors have included the Additions to the Second Edition (printed in three volumes 1993-1997) as well as additions that have never been published in print. The additions are always presented in a section of their own, after the original entry or as an entirely new entry if that is the case. At least as a preliminary measure for a work in progress, this seems a reasonable approach and it is likely that a similar procedure will be adopted for the integration of the ODS and its supplementary volumes.
A major challenge lies in the fact that the supplementary volumes were meant to be used as printed books in connection with the original printed volumes; therefore a thorough knowledge of the structure of the original dictionary is required in order to benefit from the additions and corrections proposed by the supplement entries. An elaborate system of markings is used to indicate how a supplement entry should be interpreted.
Additions are marked by the plus sign (+). A plus sign before a headword means that it is a completely new entry. This is a rather straightforward situation since the new entry can be treated on a par with the existing entries and the headword can be added to the general list of headwords. Examples of this are for instance loan words that were excluded from the original dictionary due to a somewhat purist approach on the part of the editorial staff: *caddie, café au lait, cafeteria, cancer* (cf. [Hjorth, 1990]). Another category of new headwords is compounds that were not included in the original dictionary, for instance due to low or no frequency in the collection of dictionary slips; examples of compounds with the word *dag* ‘day’ are *dagsaktuel* ‘topical’, *dagsdosis* ‘daily dosis’, *dagsprogram* ‘programme or plan for the day’.
A third major category is new words that have entered the language during the period covered, but after publication of the relevant volume, e.g. *a-bombe* ‘A-bomb’, *bilradio* ‘car radio’, *dobbeltmoral* ‘double standard’.
Things turn more complicated when the plus sign occurs in front of other information types. In principle any part of the microstructure can be affected by the additions. Thus the plus sign may precede a new main sense, a new sub-sense, a new citation including a new date and a new author, a new cross-reference etc. In contrast to the added headwords this type of

[1]TUSTEP (= Tübinger System von Textverarbeitungsprogrammen) is a text processing and layout system especially used for philological text editing.

addition is hard to handle automatically and requires a substantial amount of manual effort because the editors of the supplementary volumes put confidence in the human user's capability to interpret the information correctly and locate the right place in the original entry.

2. Refining the mark-up

Another important and necessary task is to refine the mark-up of the entries. The information that is encoded in the typographical setting as well as in the serial ordering of information reveals to a high degree which microstructural element is involved.
The headword, the homograph number and the part-of-speech have already been identified. The next information categories we want to identify are the definitions, the citations and the citation sources. The printed book has some typographical pointers as to the interpretation of the data. The definition for instance is in italics and occurs after the etymology which is in turn placed within bold brackets. In contrast to many other dictionaries the citations are in ordinary typeface, not in italics, whereas the sources for the citations are in italics and relatively well documented in a list of sources in volume 28. Multi-word units are another subtype which is fairly easy to identify as many of them are in spaced typeface and often preceded by formulae like "in the expression", "in the phrase" etc.
It is possible to identify the etymological information by means of the bold brackets, but the actual contents are often pretty hard to interpret and structure. However, we hope to benefit from the experience of the modern Danish dictionary, DDO, which is stored in XML, and possibly also from the work on the etymologies of the TLF conducted by Salmon-Alt [Salmon-Alt, 2006].

3. Dating information

Unfortunately, the ODS does not provide explicit information about the first occurrence of a word form in the language, but the information can to a certain extent be deduced from either the earlier word forms given in the etymological section or from the dates assigned to the citations in the source list. As a rule, the ODS brings the oldest occurrence as the first citation, meaning that the first edition of the source text could be used as basis for dating. However, this fact is obscured by an editorial practice of using collections of texts as a source rather than first editions, e.g. complete editions of a writer's work. The year given for a particular citation is therefore potentially misleading for dating purposes if the citation is taken from a collection of works. In that case, the year would refer to the publication of the collection and not to the first edition of the text. By way of example, the fairy tales of Hans Christian Andersen are cited from an edition that was published in 1919, 44 years after Andersen's death. This may perhaps lead some users to conclude that The Ugly Duckling was published in 1919 and not in 1843 which is the actual year of first publication. At the moment work is being undertaken to improve the dating information by assigning the first year of publication to all texts being cited from collected works. Thereby we will, on the one hand, be able to provide explicit information about the first year of publication for all citations and, on the other, always be in a position to use the first edition of a text when using a citation as evidence for dating a word.

4. Cross-references

As with all large dictionaries, the ODS is full of cross-references, and there is a lot of work to be done in detecting and structuring them. Two major distinctions can be made:

1. between internal references (the target is in the same entry as the source) and external references (the target is outside the entry, i.e. in another entry, in the source list or even outside the dictionary)
2. between complete references (all information is to be found in the target) and references that only provide supplementary information (some information is already given in the source entry)

The original dictionary text provides many clues as to the interpretation of cross-references, e.g. the complete references are generally marked "see <target>" and the supplementary references are marked "cf. <target>" or "as opposed to <target>". In principle, all target entries should be marked as hyperlinks and thus made clickable in order to facilitate the user's navigation within the dictionary.

Perspectives

The ultimate goal is to create a state-of-the-art dictionary base, thereby reviving a project that is not only the largest one undertaken in Danish lexicography, but also one of the country's finest academic achievements in the field, and granting it the public attention it deserves. A fine-grained mark-up in combination with advanced search facilities will enable its users to broaden the scope of their inquiries as they will be able to make queries and find answers to questions such as: "Which words were borrowed from Arabic in the 18th century?", "Which adjectives occur in citations from authors like Hans Christian Andersen or Søren Kierkegaard?" or "How do the entries and the citations reflect the Danes' view of the Jews during the first half of the 20th century?" This means that the dictionary may be used in new ways and also for new purposes, for example cultural or literary studies, because the great wealth of information becomes accessible in ways which were not immediately available in the printed medium.
As a simple example of this you can compare the following entry for *Jødeskole* ('Jewish school') which is presented in full in the screen version (figure 1). Figure 2 shows a possible XML tagging of a relevant extract of the same entry. The citation in figure 2 runs like this in English translation: 'I command everybody else to be quiet; this place will soon be like a Jewish school'; apparently a Jewish school was seen as a place full of noise and shouting.

Jødeskole, en. *egl.: synagoge; ogs.: skole for jødiske børn. Moth.J109. VSO. MO. StSprO.Nr.113.11*. || *nu især (dagl.) i udtr. for stærk støjen af mennesker, raaben i munden paa hinanden olgn.* (*jf.* -kirke). Her er en Støi som i en Jødeskole. *VSO. Mau.I.496*. Uden at fornærme nogen kunde man jo tro, at man var i en Jødeskole. *Sven Clausen. Forensiske Skuespil.(1920).25*. *(jeg)* kommanderer . . alle andre til at tie stille imens; her er jo snart som i en Jødeskole! *Borregaard.VL.III.363.*

Figure 1. Screen version of an entry.

```
<Lemma>Jødeskole</Lemma>
    <POS>en</POS>
    <Citation>(jeg) kommanderer . . alle andre til at tie stille imens;
    her er jo snart som i en Jødeskole!</Citation>
        <Source>Borregaard.VL.III.363.</Source>
        <Author>Einar Borregaard</Author>
        <Title>Viktor Løwe, I-III</Title>
        <PublYear>1924-26</PublYear>
```

Figure 2. Extract of the entry in XML format.

The XML format allows queries for particular words in citations from a particular period, in this case for instance compound words with *jøde* ('Jew') and the period 1900 to 1950. Such a query leads to a number of words, not all equally flattering, such as *jødepris* ('Jew's price, exorbitant price'), *jøderente* ('Jew's interest, usury interest'), *jødesmovs* ('yid') and *jødesnabel* ('Jew's conk', literally 'Jew's trunk').

Another goal is to integrate the ODS with other linguistic resources, both dictionaries and corpora. The ODS covers the period from about 1700 to about 1950, the DDO covers the period from 1950 until today, so together the two dictionary resources provide coverage of the last three centuries of the Danish language. A dictionary of the Old Danish language, i.e. from 1100 to 1500, is being compiled at the DSL, unfortunately by a very small staff so the number of actual dictionary entries is not very high. There are plans, however, to digitise the dictionary slips and publish them on the Internet so that the raw material can be made available to scholars and other people who take an interest in the earlier stages of the language.

The DSL already has corpora of modern Danish which were used for the compilation of the DDO and have been made publicly available [Andersen et al., 2002]. The idea is to link the corpora and the dictionary so that the user can navigate freely between the two resources. It is not very difficult to do since both resources are marked up in XML. A corpus of the language of the 18th and 19th centuries, however, is not yet available but it can be developed on the basis of the texts gathered in the Archive of Danish Literature (www.adl.dk) which covers the essential parts of the canonical Danish literature. A major obstacle is the large number of spelling variants but fortunately the ODS accounts for many of them. The long-term goal is to create the same integration and easy navigation in the older dictionary and corpora as in the modern ones.

A third perspective is the possible integration of wordnet resources. A wordnet for contemporary Danish following the format of Princeton WordNet, GermaNet, EuroWordNet and others is being prepared on the basis of several resources, among others the DDO [Pedersen et al., 2006]. It would be a useful improvement of the wordnet if it could link to the more comprehensive older dictionary, but it would also be a quite labour-intensive task to adapt the material to the wordnet standard and no concrete plans exist at the moment.

Bibliographie

[Andersen et al., 2002] Andersen, Mette Skovgaard; Asmussen, Helle; Asmussen, Jørg (2002): The Project of Korpus 2000 Going Public. In: Braasch & Povlsen (eds.): *Proceedings of the Tenth EURALEX International Congress.* Copenhagen, 291-299.

The Dictionary of the Danish Language online: http://ordnet.dk/ods

[Hjorth, 1990] Hjorth, Poul Lindegård (1990): Danish Lexicography. In: *Wörterbücher, Dictionaries, Dictionnaires – Handbücher zur Sprach- und Kommunikationswissenschaft,* Band 5.2. Berlin, New York: Walter de Gruyter, 1913-1922.

[Pedersen et al., 2006] Pedersen, Nimb, Asmussen, Sørensen, Trap-Jensen, Lorentzen (2006): DanNet – a wordnet for Danish. In: *Proceedings from Third International Conference on Global Wordnets*. Jeju, South Korea.

[Salmon-Alt, 2006] Salmon-Alt, Susanne (2006): Data Structures for Etymology: towards an Etymological Lexical Network. In: Corino, Marello & Oensti (eds.): *Proceedings XII EURALEX International Congress.* Torino, 79-87.

Le dictionnaire électronique du coréen contemporain : le Dic *Sejong*

– ses caractéristiques et son intérêt –

Seong Heon Lee (1)
lsh0717@snu.ac.kr

Chai-Song Hong (1)
cshong@snu.ac.kr

(1) Université Nationale de Séoul (Corée du Sud), Dépt. de Langue et Littérature françaises

---

Mots-clés : dictionnaire électronique, le Dic *Sejong*, sous-dictionnaires, le Dic détaillé, le Dic élémentaire, le Dic intégral, extraction des informations, le TAL

Résumé : Cette étude présente la méthode avec laquelle nous avons construit le Dic *Sejong*, mais aussi les caractéristiques et l'intérêt de celui-ci. Pour ce faire, nous abordons notre sujet sous trois angles distincts. D'abord, nous traitons des objectifs, de l'organisation et de l'état actuel de nos travaux. Puis, nous parlons de ce qui caractérise la macrostructure et la microstructure du Dic *Sejong*. Finalement, nous montrons toute l'efficacité de son fonctionnement pour extraire des informations demandées lors du traitement automatique de la langue.

Introduction

Cette étude se propose de présenter les caractéristiques du dictionnaire électronique *Sejong* et de montrer son intérêt tant dans ce qui relève des études lexicographiques ou linguistiques que dans le traitement automatique de la langue. Ce faisant, nous essaierons également de présenter la méthode que nous avons élaborée en vue de la construction du Dic *Sejong*, et de justifier notre choix. Et cela pour apporter notre contribution à des travaux de ce type qui sont à venir ou déjà en cours.

A cet effet, nous nous attacherons dans cette étude à présenter le Dic *Sejong* sous trois angles : les généralités des travaux de sa construction, ses caractéristiques et son intérêt. En ce qui concerne le premier point, nous parlerons des objectifs de ces travaux, de leur organisation ainsi que de leur état des lieux. Pour ce qui est du deuxième point, nous présenterons la macrostructure et la microstructure du Dic *Sejong*, et cela en parlant de l'intérêt des informations qu'il véhicule dans les différents domaines tant dans le traitement automatique que dans les recherches linguistiques. Et finalement, nous montrerons son fonctionnement efficace pour extraire des informations demandées lors du traitement automatique de la langue.

1. Le Dic *Sejong*

Depuis 9 ans, nous nous consacrons à la construction d'un dictionnaire électronique du coréen contemporain de grande dimension et à usages multiples. Il s'agit de travaux à échelle nationale dont la durée s'étale sur 10 ans (1998-2007) et qui sont subventionnés par l'Etat. Ces travaux s'effectuent en effet dans le cadre du « Projet *Sejong* pour le 21ème siècle », projet qui vise à informatiser tous les types de données sur le coréen et à bâtir une infrastructure favorisant le développement tant de l'industrie langagière que des sciences concernées. Ainsi conçu, le Dic *Sejong* s'est fixé de concilier les caractéristiques suivantes :

– une synthèse (conciliation) de différentes théories

– une recherche minutieuse, ainsi qu'une exactitude et une richesse du contenu dans la description des mots

– un système de présentation standardisé et explicite : codé en XML

– un dictionnaire de référence avec assez de flexibilité pour de possibles changements allant de concert avec le futur environnement technologique (le TAL).

Comme nous pouvons le noter en fonction de ce qui vient d'être dit, dans la construction du Dic *Sejong*, nous avons tout aussi bien mis l'accent sur la validité technique pour le TAL que sur la qualité de la description lexicographique, c'est-à-dire de son contenu. L'objectif principal du Dic *Sejong* réside dans la prise en compte de l'efficacité lexicographique et informatique qui peut être établie ainsi :

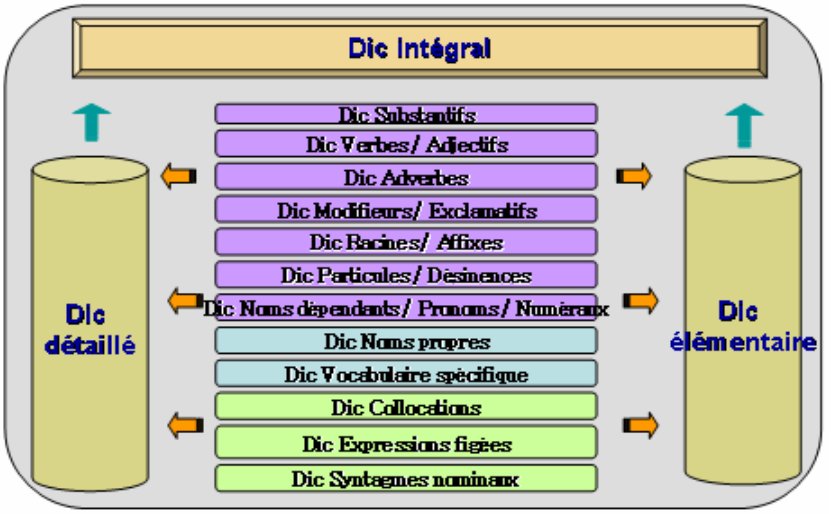
– décrire autant de mots que possible (600 000 entrées) avec la plus grande exactitude et la plus grande minutie

– construire une méthode de présentation efficace des données lexicographiques de telle sorte que ces données puissent être mises en pratique dans le TAL.

2. Les caractéristiques du Dic *Sejong*

Pour atteindre son objectif de la manière la plus efficace, le Dic *Sejong* est construit selon les procédés suivants : les sous-dictionnaires des différentes parties du discours sont établis séparément et intégrés à l'ensemble par la suite[1]. Chacun de ces sous-dictionnaires se compose de deux types de dictionnaire : le Dic *détaillé* qui offre une description complète et le Dic *élémentaire* qui offre une description réduite.

[1] Au total, 17 catégories grammaticales sont réparties en 12 sous-dictionnaires.



Il faut noter ici que, du point de vue de sa macrostructure et de son fonctionnement, le Dic *Sejong* se caractérise en particulier par l'intégration des dictionnaires des collocations, des expressions figées et des syntagmes nominaux dont les entrées sont des unités syntagmatiques, ce qui permet d'optimiser son fonctionnement lors de la reconnaissance et de la génération des phrases ; mais aussi par l'intégration d'un dictionnaire des adverbes disposant d'informations abondantes et détaillées sur leurs combinaisons avec les verbes et les adjectifs ainsi que leurs combinatoires spécifiques, ce qui permet d'optimiser l'exactitude dans la désambiguïsation et la génération des phrases ; par ailleurs, l'intégration d'un dictionnaire du vocabulaire spécifique comme le vocabulaire de l'actualité, les emprunts et mots étrangers, les abréviations ou acronymes et les mots en chiffres permet d'optimiser le traitement des mots non enregistrés dans le dictionnaire. Cela dit, ces différents sous-dictionnaires sont basés sur les mêmes principes :

– l'examen critique des problématiques linguistiques courantes concernant les questions de lexicographie de la partie du discours en question et la synthèse des divers résultats.

– la reconnaissance et la définition de ce type de données lexico-syntaxiques nécessaires à la description du vocabulaire toujours dans la partie du discours concernée.

– l'établissement d'une structure logique qui va permettre une présentation systématique, explicite et exhaustive de chaque type de données.

– la description des données pour chaque entrée sous les rubriques d'informations.

Les sous-dictionnaires ainsi construits disposent de rubriques d'informations dont on peut voir ici le nombre :

Sous-Dics	Nombre des rubriques d'infos	Sous-Dics	Nombre des rubriques d'infos
Substantifs	**87**	**Racines/Affixes**	**25**
Verbes/Adjectifs	**60**	**Collocations**	**44**
Adverbes	**62**	**Expressions figées**	**48**
Particules/ Désinences	**36**	**Noms propres**	**15**
Modifieurs	**32**	**Vocabulaire spécifique**	**18**
Exclamatifs	**32**	**Syntagmes nominaux**	**25**

A titre d'exemple, voici la microstructure du Dic Substantifs :

Exemple
de la microstructure du Dic (1) : Dic Substantifs

```
표제항정보구획
<superEntry>
  <orth></orth>
  <entry pos="X">
관리정보구획
<mnt_grp>
   <cre>
      <cre_date></cre_date>
      <cre_writer></cre_writer>
      <cre_note></cre_note>
   </cre>
   <mod>
      <mod_date></mod_date>
      <mod_writer></mod_writer>
      <mod_note></mod_note>
   </mod>
   <add></add>
</mnt_grp>
<see></see>
형태정보구획
<morph_grp>
   <var type="etc"></var>
   <str type="X"></str>
   <org lg="X"></org>
</morph_grp>
관용표현구획
<idm_grp>
   <idm type="etc"></idm>
</idm_grp>
<sense>
의미정보구획
<sem_grp>
   <eg></eg>
   <trans></trans>
   <domain></domain>
   <sem></sem>
   <lr>
      <syn type="X"></syn>
      <ant type="X"></ant>
   </lr>
</sem_grp>
통사정보구획
<syn_grp>
   <n_aj_grp>
      <n_aj type="etc"></n_aj>
   </n_aj_grp>
   <n_n_grp>
      <n_n type="etc"></n_n>
   </n_n_grp>
   <n_v_grp>
      <n_v type="X">
      <form vcompound="X"></form>
      <frame></frame>
      </n_v>
   </n_v_grp>
      <max_n></max_n>
      <sel_rst arg="X"></sel_rst>
   <s_n></s_n>
   <cl type=""></cl>
   <prt></prt>
   <av></av>
   </syn_grp>
   </sense>
</entry>
</superEntry>
```

Cette microstructure permet d'établir une description conforme à la distinction des noms prédicatifs et des noms argumentaux. Par exemple, les rubriques d'informations <frame>, <max_n>, <sel_res arg>, <n_v type= "x">[2] concernent les noms prédicatifs, tandis que les rubriques <s_n> et <prt>[3] concernent les noms argumentaux. De plus, le Dic Substantifs contient des informations diverses et minutieuses représentant les propriétés de la catégorie grammaticale des substantifs telles que les restrictions sur leurs combinaisons avec les particules (surtout casuelles) et leurs emplois adverbiaux.

A l'heure actuelle, nous avons pratiquement terminé la construction du Dic *Sejong* et nous sommes en train de le parfaire pour optimiser son fonctionnement dans le TAL. Voici l'état actuel du développement du Dic *Sejong* :

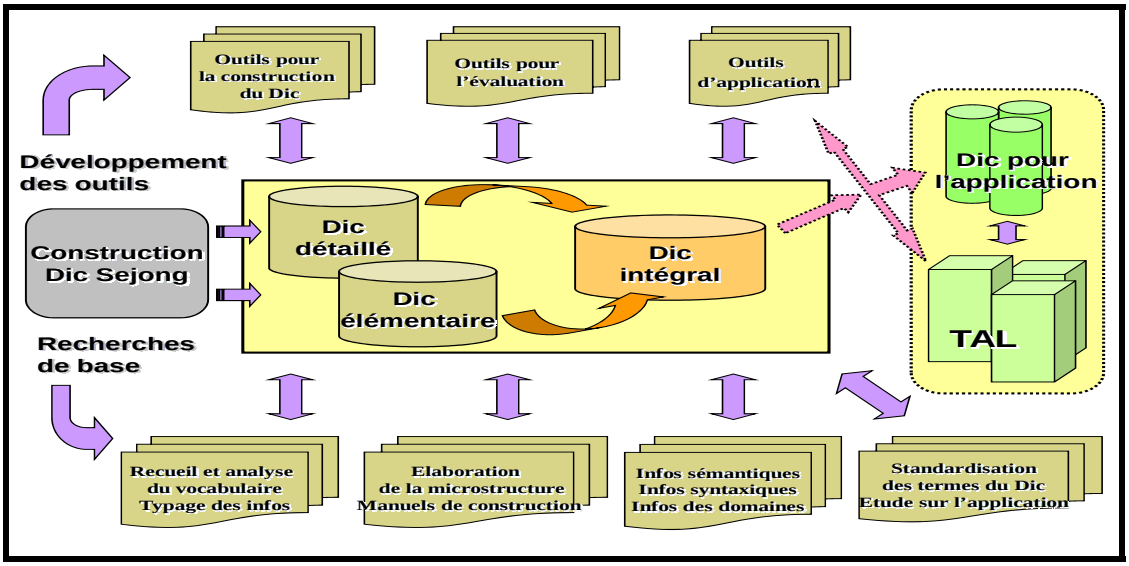
Sous-Dics	Nombre des entrées		Sous-Dics	Nombre des entrées	
	Dic détaillé	Dic élémentaire		Dic détaillé	Dic élémentaire
Substantifs	35 100	136 000	Adverbes	4 500	11 000
Verbes/Adjectifs	27 700	50 000	Collocations	9 000	9 000
Racines/Affixes	1 700	1 700	Expressions figées	5 000	5 000
Pronoms/Numéraux Noms dépendants	1 200	1 200	Vocabulaire spécifique	35 000	60 000
Noms propres	22 000	158 600	Syntagmes nominaux	5 800	15 200
Particules/Désinences/ Modifieurs/Exclamatifs	2 650	2 650	Total	149 650	450 350
				600 000	

Il est à noter ici que la plupart des informations, qu'elles soient morphologiques, syntaxiques ou sémantiques, sont décrites dans le Dic *Sejong* à l'aide de codes dont le système a été préalablement construit. Par exemple, pour décrire le sens d'une entrée lexicale ou les restrictions imposées sur ses arguments, nous disposons des classes sémantiques dont le nom sert de code pour représenter une portée sémantique. Il en est de même pour la description de la structure de la phrase. Le Dic *Sejong* dispose de la typologie des constructions syntaxiques.

[2] Ces rubriques représentent respectivement les informations sur la structure de la phrase, la structure maximale des arguments, les restrictions sur la sélection des arguments et les verbes supports compatibles.

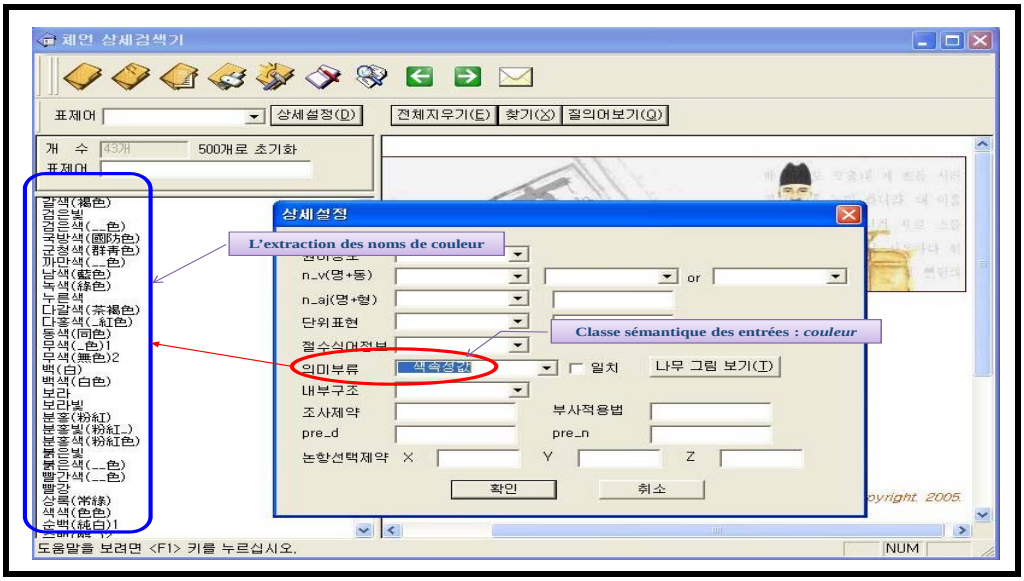
[3] Ces rubriques contiennent respectivement les informations sur les modifieurs propositionnels et les classifieurs.

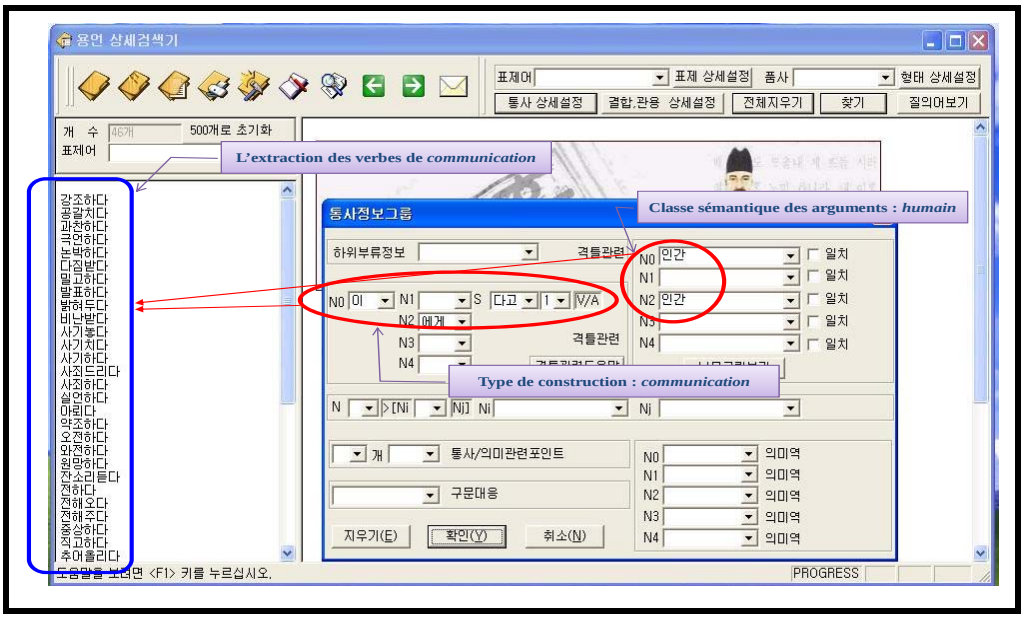
C'est ainsi que la construction du Dic *Sejong* ne se limite pas à la description lexicographique et qu'elle implique divers grands chantiers comme des études de base pour le développement du Dic *Sejong* et des travaux pour le développement des différents outils indispensables à sa construction et à son application. Les premières contiennent l'établissement de la liste du vocabulaire général du coréen contemporain, la redéfinition des catégories grammaticales du coréen, l'étude et construction des classes sémantiques des substantifs coréens, l'étude et établissement de la typologie des constructions syntaxiques en coréen, la construction du dictionnaire des termes utilisés dans ces travaux et l'étude et développement des différentes applications du Dic *Sejong*. Quant aux seconds, ils concernent le développement des outils pour la construction du Dic *Sejong* (les éditeurs pour les sous-dictionnaires, l'éditeur intégral du Dic *Sejong*, son recenseur, son gestionnaire) et des outils pour son application (l'analyseur des morphèmes, le parseur des constructions, l'analyseur des syntagmes nominaux, etc.). Voici l'organisation de notre travail :



3. L'intérêt du Dic *Sejong*

Le Dic *Sejong* ainsi construit montre toute l'efficacité de son fonctionnement, en particulier dans l'extraction des informations demandées lors du traitement automatique. En voici quelques exemples :





Comme on le voit dans les exemples ci-dessus, le Dic *Sejong* est susceptible de permettre d'extraire toutes les informations demandées. Ce qui revient à dire qu'il intègre toutes les informations sur les entrées lexicales qui sont nécessaires au traitement automatique de la langue, et cela de manière formelle et systématique, de sorte que l'ordinateur puisse à lui seul le consulter lors de son fonctionnement.

Conclusion

Le Dic *Sejong* ainsi construit représente les valeurs suivantes :

(a) Une valeur scientifique en lexicographie et en linguistique : il offre une base de données complète sur le lexique coréen contemporain et permet sa gestion permanente,

(b) Une valeur scientifique en informatique : il permet de développer de différents systèmes pour le traitement automatique du coréen comme le recensement des informations, la classification et les résumés automatiques des dossiers, la traduction automatique (assistée par ordinateur), le système de questions-réponses automatique ; il offre une base pour le développement d'un système d'ontologie.

(c) Des valeurs stratégiques industrielles : il permet d'établir une infrastructure pour l'informatisation des connaissances.

Bibliographie

Candel, D. *et al.*, (1990) Aspects de la documentation scientifique et technique dans un grand dictionnaire de langue in *Autour d'un dictionnaire : le « Trésor de la langue française ». témoignage d'atelier et voies nouvelles*, ed. B. Quemada, Didier édition, Paris.

Corréar, M.-H. ed., (2002), *Lexicography and Natural Language Processing* : A Festchrift in Honor of B.T.S. Atkins. EURALEX. Info: http://www.ims.uni-stuttgart.de/euralex/.

Dirven, R. eds., (1995), *Current approaches to the lexicon*, Duisburg Papers on Research in Language and Culture, Peter Lang.

Fellbaum, C. ed., (1998), *Wordnet. An Electronic Lexical Database*, The MIT Press.

Fontenelle, T., (1997), *Turning a Bilingual Dictionary into a Lexical-Semantic Database,* Max Niemeyer Verlag.

_____, (1998), *ACTES EURALEX'98 PROCEEDINGS,* Communications soumises à EURALEX'98 (Huitième Congrès International de Lexicographie) à Liège, Belgique / Papers submitted to the Eighth EURALEX, International Congress on Lexicography in Liège, Belgium, English and Dutch Departments, University of Liège.

Gross, M., (1975), *Méthodes en syntaxe,* Hermann.

Gross. G., (1992), *Formes d'un dictionnaire électronique, L'environnement traductionnel,* Silley, Presses de l'Université du Québec.

Gross, G., Vivès, R., (1986), Les constructions nominales et l'élaboration d'un lexique-grammaire, *Langue Française* 69, Larousse, Paris.

Cruse, A. *et al.* eds., (2002), *Lexicology I*, Berlin, New York, Walter de Gruyter.

Guillet, A., Leclère, C., (1981), Restructuration du groupe nominal, Formes syntaxiques et prédicats sémantiques, *Langages* no.63, Larousse, Paris.

_____, (1992), *La structure des phrases simples en français. Constructions transitives locatives*, Droz.

Heid, U., (2000), *Proceedings of the Ninth, EURALEX International Congress.* EURALEX 2000. Stuttgart. Germany.

Mel'čuk, *et al*., (1984, 1988, 1992, 1999), *Dictionnaire explicatif et combinatoire du français contemporain I Ⅰ-IV*, Les presses de l'université de Montréal.

_____, (1995), *Introduction à la lexicologie explicative et combinatoire*, Editions Duculot.

Hong, C.-S (1998-2006), *Développement du dictionnaire électronique* Sejong, rapports techniques, Ministère coréen de la Culture et du Tourisme / Institut national de la langue coréenne, Séoul (en coréen).

Hong, C.-S., Lee, S.-H., (2003) Representation of Lexico-Syntactic Information for the Description of Predicate Nouns in the Sejong Electronic Dictionary, *Proceedings of ICKL-TU Berlin International Conference on Korean/Corpus Linguistics,* ICKL-TU Berlin.

Knowles, F. E., (1990), The Computer in Lexicography in F. J. Hausmann *et al*. eds. 1989-91:1645-72.

Pustejovsky, J., (1995), *The Generative Lexicon*, MIT Press.

Pustejovsky, J., Bergler, S. eds., (1992), *Lexical Semantics and Knowledge Representation,* Springer Verlag.

Sinclair, J., (1991), *Corpus, Concordance, Collocation,* Oxford University Press.

Vosen P. ed., (1998), *EuroWordNet,* Dordrecht : Keuwer Academic Publishers.

Walter A., Cook, S. J., (1989), *Case Grammar Theory*, Georgetown University Press.

Weigand, E., (1998), *Contrastive Lexical Semantics,* University of Munster.

L'exploitation statistique des bases lexicographiques

Etienne Brunet (1)
brunet@unice.fr

(1) Université de Nice

Charles Muller, le maître de la statistique linguistique, assistait au colloque de Strasbourg, à côté de son ami Paul Imbs. On lui prêterait à tort, en cette occasion, un rôle de promoteur de ce qui allait devenir la lexicométrie. Sa conversion n'intervint qu'un peu plus tard, sur le chemin de Besançon, auprès de Quemada, Evrard et Moreau. Mais en 1957, Pierre Guiraud, qui allait publier l'année suivante « Problèmes et méthodes de la statistique linguistique » se trouvait aussi à Stasbourg parmi les intervenants. Certes la statistique ne fut guère évoquée dans les débats (on parlait plutôt de relevés et de dénombrements) et Guiraud choisit de parler d'un sujet un peu moins provocateur : l'argot. Mais les participants au colloque avaient pour la plupart une idée plus confuse encore de l'informatique (le mot n'existait pas alors), que certains confondaient alors – ce n'était pas le cas de Quemada ou de Wagner - avec la mécanographie.

Mais dans les années qui ont suivi le Colloque de Strasbourg, la statistique, présumée un peu plus accessible que l'informatique aux esprits littéraires, bénéficia soudain d'un essor remarquable auquel contribuèrent le Centre d'étude du vocabulaire français de Besançon, l'entreprise du « français fondamental », les travaux d'Evrard en Belgique et du Père Busa en Italie, les publications de Guiraud et, bien sûr, la thèse et le manuel de Muller.

Cela est si vrai que la première publication du TLF a été de nature statistique. C'est en 1971, l'année même où le premier tome du Trésor allait être livré aux lecteurs, que paraissent, avec la signature de Robert Martin, les quatre volumes du *Dictionnaire des fréquences*.

Cinquante ans après on doit reconnaître que la veine statistique dans la mine de Nancy n'a pas été exploitée comme on pouvait l'espérer. Le filon qui devait conduire au trésor n'a pas suscité les vocations attendues, malgré les efforts des directeurs du TLF. Inversement l'informatique, un peu timide les premières années, a pris un essor spectaculaire dans l'entreprise nancéenne, quand un informaticien de grand talent, Jacques Dendien, a rejoint l'INaLF.

Il n'y a pas concurrence entre l'informatique et la statistique. Les deux sont associées dans les moteurs de recherche, le data mining et la plupart des industries de la langue. Et si *l'ATILF* rend des services inégalés en matière documentaire, les informations statistiques qu'il distribue sont d'un grand intérêt et donnent à tout le moins les données de base pour les études quantitatives. On en donnera quelques exemples tirés de *Frantext*, du *TLFI*, du *BHVF*…

On souhaiterait toutefois qu'elles ne soient pas restreintes aux graphies et qu'elles s'étendent aux lemmes, puisqu'aussi bien les textes ont été, dans leur majorité, étiquetés et lemmatisés. On aimerait aussi qu'un véritable atelier statistique soit ouvert à Nancy et que le serveur puisse offrir des traitements statistiques évolués, et pas seulement des fréquences et des

pourcentages. On réclame enfin que dans le domaine des données chiffrées le copyright des éditeurs soit abandonné et que l'abonnement ne soit plus nécessaire, puisqu'à aucun moment le texte ne serait communiqué.
Ce sont là des projets à court terme dans le développement des bases constituées à Nancy. Ils n'excluent pas des projets plus ambitieux qu'on tentera d'évoquer.

La lexicographie au service de l'apprentissage/enseignement des combinaisons de mots

Serge Verlinde (1)
serge.verlinde@ilt.kuleuven.be
Jean Binon (1)
jean .binon@ilt.kuleuven.be
Ann Bertels (1)
ann.bertels@ilt.kuleuven.be

(1) Groupe de recherche en lexicographie pédagogique (GRELEP), ILT, K.U. Leuven (Belgique)

Mots-clés : lexicographie, enseignement assisté par ordinateur (EAO), combinatoire des mots, base de données, environnement d'apprentissage

Keywords: lexicography, Computer Assisted Language Learning and Teaching (CALLT), word combinations, database, learning environment

Résumé : Dans cette contribution, nous démontrons comment un dictionnaire peut être transformé en réel environnement d'apprentissage en ligne pour apprenants de français langue étrangère ou seconde : la *Base lexicale du français* (BLF). La BLF se compose d'un dictionnaire interactif ainsi que d'un générateur d'exercices intégré. Nous nous concentrons sur le phénomène de la combinatoire des mots, qui constitue l'un des aspects du lexique les plus difficiles à maîtriser par les apprenants. En conclusion, nous démontrons comment des analyses de corpus plus poussées et l'intégration d'outils de TAL peuvent encore améliorer la qualité des dictionnaires d'apprentissage.

Abstract : In our article, we illustrate how a dictionary could be turned into a powerful online learning environment pour learners of French as a foreign or second language: the *Base lexicale du français* (BLF). The BLF consists of an interactive dictionary and an integrated exercise generator. We focus on the topic of word combinations, which seems to be one of the most difficult issues to master in foreign language learning. To conclude we make some suggestions for further research on improving the integration of corpus analysis results and NLP in the field of pedagogical lexicography.

Introduction

Depuis toujours, le dictionnaire occupe une place importante dans le processus d'apprentissage d'une langue étrangère. Le plus souvent, en classe, on a recours au dictionnaire bilingue. Celui-ci n'est toutefois pas exempt de critiques [Bogaards, 2006]. En outre, on a assisté depuis une bonne dizaine d'années à l'avènement de nouveaux concepts de

dictionnaires d'apprentissage pour l'anglais, monolingues et semi-bilingues [Herbst et Popp, 1999], Et, enfin, on ne peut ignorer les avantages qu'offre le dictionnaire électronique, sur CD ou en ligne [de Schryver, 2003].
Pour le français, on dispose de dictionnaires bilingues de bonne qualité, mais de conception somme toute assez traditionnelle, à quelques exceptions près [Back, 2004]. Du côté des monolingues, les dictionnaires d'apprentissage récents sont essentiellement destinés à des apprenants de langue maternelle, quoi que puissent affirmer les auteurs [Verlinde e.a., à paraître b] et les idées lancées par la lexicographie pédagogique anglaise n'ont pas vraiment fait école [Bogaards, 2001]. Quant aux versions électroniques des dictionnaires papier, elles restent provisoirement assez proches de leur version d'origine (Petit Robert) ou présentent des exploitations très pointues qui sont moins intéressantes pour des apprenants de français langue étrangère (TLFi, atilf.atilf.fr/).
Comme professeurs de français langue étrangère, nous ne disposons pas toujours des ouvrages répondant à nos besoins. C'est pourquoi nous avons développé à l'Institut des langues vivantes (ILT) de la K.U.Leuven (Belgique) des outils lexicographiques qui servent d'outils de référence pour nos étudiants, d'abord dans le domaine du français des affaires [Binon e.a., 2000] et plus récemment pour le français général. Ces outils ont été rassemblés sur le site de la Base lexicale du français (www.kuleuven.be/ilt/blf), qui est en accès libre.
Dans ce qui suit, nous présentons la méthodologie qui sous-tend une part de ce travail de développement ainsi que les dispositifs mis en oeuvre (2.). Nous nous concentrons à cette occasion sur le problème de la combinatoire des mots en français général (1.). Ce choix s'explique par le fait qu'il est omniprésent et qu'il pose de nombreux problèmes, aussi bien lors du processus d'apprentissage/enseignement [Granger, 1998 ; Lewis, 2000] que lorsque l'apprenant se livre à l'encodage ou au décodage de messages. Ou, comme le formulent Hausmann et Blumenthal [2006:4], c'est le sentiment de 'détresse' que ressent le locuteur (natif aussi parfois d'ailleurs) lorsqu'il est à la recherche du collocatif approprié. Nous signalons également les aspects de l'outil qui sont susceptibles d'être améliorés (3.).

1. La combinatoire des mots dans les dictionnaires

On observe un regain d'intérêt pour la description de la combinatoire des mots, illustré par la publication récente de deux gros dictionnaires [Le Fur, 2007 ; Mel'čuk et Polguère, 2007] ainsi que de deux numéros de revue thématiques consacrés à un segment particulier de la combinatoire des mots: les collocations [Blumenthal et Hausmann, 2006 ; Grossmann et Tutin, 2003]. De même, sur le web, Bourigault (www.irit.fr:8080/voisinsdelemonde/) propose un outil qui permet d'explorer la façon dont les mots se combinent entre eux. On notera également que la dernière version du correcteur orthographique Antidote offre un inventaire extrêmement fourni de combinaisons de mots tirées d'un corpus de 500 millions de mots composé d'œuvres littéraires classiques et contemporaines et de textes recueillis sur le web. Ces outils viennent compléter les informations que l'on trouvait déjà dans des dictionnaires de référence tels que le GR, le PR et le TLF ou d'autres inventaires de combinaisons de mots [voir Verlinde e.a., 2006 pour plus de détails].
Traditionnellement, les dictionnaires, qu'ils soient monolingues ou bilingues, mentionnent toutes les combinaisons de mots dans le corps des articles, rattachées au sens de l'une de leurs composantes.[1] Elles sont accompagnées dans un certain nombre de cas d'une indication de sens sous la forme d'une définition ou d'un synonyme. Cette présentation minimale sert essentiellement au décodage.

[1] Le classement des combinaisons de mots n'est pas toujours transparent : certaines, comme par exemple *pomme de terre*, font l'objet d'une entrée, d'autres non (*pomme d'Adam*).

Aussi fournis soient-ils, ces inventaires connaissent un certain nombre de faiblesses : manque de cohérence lors de la sélection des combinaisons, difficulté de repérage de ces combinaisons dans des articles plus longs, manque de critères stricts lors de l'attribution de la combinaison à l'un ou l'autre article. Avec l'avènement des dictionnaires électroniques et leurs puissantes fonctionnalités de recherche, les problèmes de repérage sont en grande partie résolus. Le PR électronique permet par exemple d'effectuer une recherche directe sur un ensemble de combinaisons de mots sorties des articles. Le TLFi a récemment adopté une palette de couleurs pour faire ressortir davantage certaines rubriques d'un article, dont les inventaires de combinaisons de mots.

Dans Le Fur [2007], dont le modèle de présentation des collocations n'est pas sans rappeler celui adopté dans le DAFA [Binon e.a., 2000], les collocations sont explicitement classées par catégorie grammaticale et implicitement par sens, avec une référence explicite aux fonctions lexicales de Mel'čuk. Les combinaisons sont en effet classées par 'grands groupes de sens correspondant à des notions universelles' [Le Fur, 2007 :IV], telles l'intensité et l'atténuation, le début, le déroulement et la fin d'un processus, etc. Le fait que ces indications de sens ne soient pas mentionnées explicitement dans le corps des articles oblige quand même l'utilisateur à parcourir l'ensemble de l'article, ce qui rend la classification moins fonctionnelle. Comme il n'y a presque pas d'indications sur le sens des mots mentionnés et donc sur le sens des collocations, cet ouvrage semble principalement réservé à des locuteurs natifs.

Le *Lexique actif du français* [LAF, Mel'čuk et Polguère, 2007] suit une démarche inverse, puisque chaque collocation ou groupe de collocations est accompagné d'une formule décrivant la 'relation particulière de contrôle fonctionnel entre les deux lexies qui la constituent: la base contrôle le choix du collocatif' [Mel'čuk et Polguère, 2007:22]. L'apprenant y trouve un instrument intéressant, tant pour le décodage que pour l'encodage, même si le formalisme utilisé nécessite une initiation préalable et que la nomenclature couverte est très réduite. Malheureusement, ni l'ouvrage de Le Fur, ni le LAF ne sont disponibles en version électronique, ce qui, compte tenu de la densité des informations, rend la consultation ardue.

Parallèlement à ce travail descriptif, les travaux de Mel'čuk [1984-1999] et de Blumenthal [2006] ont démontré que le lexique, et plus particulièrement les combinaisons de mots, obéissent à certaines règles, même si celles-ci ne sont pas nécessairement absolues ni transparentes au premier abord.

Si le locuteur natif marie les mots (presque) sans en être conscient, il n'en est pas de même pour les apprenants, qui ne disposent pas de cette intuition linguistique. D'où la nécessité de décrire ce phénomène de façon très détaillée et conviviale dans un dictionnaire destiné à ce public cible.

2. La combinatoire des mots dans un dictionnaire d'apprentissage

Pour réaliser l'inventaire des combinaisons de mots présenté dans la BLF, nous avons eu recours à un corpus de textes journalistique de 50 millions de mots et à l'utilisation d'une mesure statistique permettant de calculer la pertinence statistique des mots apparaissant dans un contexte de quatre mots à gauche et à droite autour d'un mot-pivot, le Z-score.

Comme un apprenant connaît différents types de besoins, la BLF, qui est une base de données lexicographique interrogeable en ligne, tente de s'adapter aux besoins variés de l'utilisateur et propose les informations consignées dans la base sous différentes formes, dont nous donnons un aperçu dans ce qui suit.

2.1 Décoder les combinaisons de mots

Le décodage d'un message à l'aide d'un dictionnaire (papier ou électronique) peut s'avérer assez simple, si l'apprenant a uniquement besoin de la traduction d'un mot isolé ou de son sens. Le décodage s'avère nettement plus compliqué lorsqu'il s'agit d'une séquence de mots. L'intérêt des sites de traduction est indéniable dans ce contexte, même s'ils sont loin de donner des traductions parfaites.[2] Idéalement, ce genre de site devrait être relié à un dictionnaire de sorte que l'utilisateur puisse contrôler éventuellement la traduction fournie ou le sens d'un mot. On trouve une telle application dans la BLF. L'apprenant peut soumettre une séquence de mots limitée qui est analysée en ligne. L'application identifie les mots et, jusqu'à un certain degré, les combinaisons de mots. Cette application n'est qu'un prototype, qu'il conviendrait de perfectionner, puisque l'analyse génère du bruit, surtout lors de l'identification des combinaisons de mots (voir 3.).

Dans une démarche plus classique, l'utilisateur d'un dictionnaire se reportera à l'article du mot qu'il ne connaît pas. Comme les problèmes de compréhension sont souvent des problèmes de compréhension de séquences de mots, la version électronique du DAFA, le dictionnaire du français des affaires intégré à la BLF, attire l'attention de l'apprenant sur ce phénomène en juxtaposant sur le premier écran de l'article le mot simple et toutes les combinaisons de mots dans lesquelles il entre.

Pour faciliter la compréhension des combinaisons de mots mentionnées dans chaque entrée, celles-ci sont classées par 'groupes de sens', comme dans Le Fur [2007]. Dans la BLF, l'apprenant peut accéder aux combinaisons de mots à partir des articles de chacune de leurs composantes. En plus, en cliquant sur une combinaison de mots, il peut obtenir les définitions des mots qui la composent. Dans les cas où le sens n'est pas compositionnel, une véritable définition est donnée. Enfin, des exemples authentiques de corpus peuvent être appelés. Il s'agit d'exemples de notre corpus journalistique et d'autres corpus disponibles sur le web.

2.2 Encoder les combinaisons de mots

La BLF propose plusieurs aides à l'encodage d'un message. L'utilisateur peut soumettre un mot dans sa langue maternelle.[3] Comme dans un dictionnaire semi-bilingue, où la langue maternelle constitue une voie d'accès à un dictionnaire monolingue en langue étrangère, il accède à une multitude d'informations qui peuvent servir à l'encodage. Il obtient ainsi non seulement des suggestions de traduction (*high = élevé, haut, gros, fort,* …), mais également les définitions des mots suggérés, des informations sur les constructions syntaxiques dans lesquelles ils entrent, comme par exemple la place de l'adjectif, une phrase exemple et des indications sur les cooccurrents de ces mots. Pour l'instant, seules sont fournies les traductions de mots de la nomenclature. Nous prévoyons aussi l'ajout des traductions pour les combinaisons de mots.

La description lexicographique, aussi fournie soit-elle, ne rend toutefois jamais compte de toutes les possibilités combinatoires de la langue. Si la combinaison n'est pas mentionnée explicitement, l'apprenant trouvera difficilement quel verbe support (*pratiquer, effectuer, réaliser,* ... + nom) ou quel adjectif adéquat (*haut, élevé grand, fort, gros + nom*) choisir. Afin de résoudre ce problème, la BLF donne directement accès au corpus de textes journalistiques: l'apprenant peut ainsi contrôler l'existence de certaines combinaisons de mots, voire en comparer la fréquence d'occurrence. A chaque fois, il a également la possibilité d'appeler les exemples du corpus.

---

[2] On distingue principalement deux types de sites de traduction. Aux sites de traduction automatique viennent s'ajouter des sites qui proposent des corpus parallèles alignés (urd.let.rug.nl/tiedeman/OPUS/). L'avantage de ces sites est qu'ils proposent des traductions en contexte et non hors contexte, comme dans les dictionnaires de traduction classiques.

[3] Le néerlandais et en moindre mesure l'anglais pour l'instant.

Une dernière aide se fonde sur une analyse des combinaisons de mots statistiquement pertinentes dans le corpus. Nous avons établi le profil combinatoire [Blumenthal, 2006] de 13.000 mots et ajouté à chaque mot apparaissant dans un profil la catégorie grammaticale. Si un apprenant est à la recherche d'un adjectif qui pourrait traduire l'idée de 'très bon' ou 'très mauvais' auprès d'un nom, il lui suffit d'appeler la liste des adjectifs qui se combinent de façon statistiquement pertinente avec celui-ci et de la parcourir. Ici aussi, chaque adjectif est relié à son article de dictionnaire et on peut faire apparaître des exemples de chaque combinaison.

2.3 Encoder les combinaisons de mots

Dans l'enseignement, on n'a pas l'habitude d'utiliser le dictionnaire comme manuel d'apprentissage du vocabulaire. Pourtant, si la description lexicographique est bien structurée, cet ouvrage de référence contient toutes les informations nécessaires à cet effet. Ainsi, dans la BLF, des requêtes transversales peuvent être exécutées sur toutes les données consignées dans la base. Pour les combinaisons de mots, la recherche est paramétrable en fonction du lien sémantique exprimé (intensification, début, ...), de la catégorie grammaticale du mot dominant de la combinaison (nom, adjectif, verbe, ...), du domaine auquel appartient le mot dominant (la base des collocations: économie, informatique, politique, ...), de la fréquence, de la compositionnalité du sens et du type de combinaison de mots (locution, expression impersonnelle, proverbe). A l'aide de cette fonctionnalité, un professeur ou un apprenant peut extraire des contenus d'apprentissage sous forme de listes de combinaisons spécifiques. Par le biais des hyperliens, les combinaisons extraites restent reliées aux articles dans lesquels elles apparaissent et auxquelles l'apprenant accède par un simple clique.
En complément à la description lexicographique, la BLF propose également un générateur d'exercices (semi-)automatisé. Parmi les exercices offerts, ceux portant sur les combinaisons de mots occupent une place importante. Des exercices sont ainsi disponibles pour les verbes support (‘*poser’ une question, ‘procéder’ à une vérification*), les verbes aspectuels (‘*entamer’ des négociations*), les locutions (*en ‘face’ de, en ‘fin’ de*), les expressions (‘*mieux’ que rien, joindre les deux ‘bouts’*) et l'expression de l'intensification (‘*gravement’ malade*), de la multiplicité (*une ‘troupe’ d'acteurs*) et du partitionnement (*une ‘phase’ de jeu*). Tous ces exercices sont construits sur le même modèle et selon le même principe. Dans un premier temps, une série de phrases illustrant l'emploi des combinaisons de mots a été extraite de notre corpus journalistique pour être stockée ensuite dans la base de données. Au lancement de l'exercice, un nombre restreint de phrases est extrait au hasard de cette base et proposé à l'apprenant. Celui-ci complète ensuite la partie manquante de la combinaison de mots dans la phrase. Après correction, l'application conserve en mémoire les erreurs commises, ainsi que les items que l'apprenant a sélectionnés pour réapparaître éventuellement dans un exercice ultérieur. Sur la page de correction apparaît aussi un retour d'information tiré de la base lexicale.
Le générateur se décline en trois versions: outre la version de base, il est possible de confronter l'apprenant aux erreurs qu'il a commises dans les exercices précédents pour contrôler le degré de rétention. Il est également possible de préparer des tests composés des mêmes phrases pour tout un groupe d'apprenants.
Cette composante de la BLF est également en accès libre. Toutefois, certains exercices sur les combinaisons de mots et certaines fonctionnalités sont réservés aux étudiants de notre université.

3. Perspectives

L'analyse de corpus est devenu une pratique courante dans la lexicographie moderne. Toutefois, on constate que pour le français, les corpus utilisés à l'heure actuelle ne sont pas vraiment représentatifs: il s'agit dans la plupart des cas de corpus journalistiques. Il manque au français un corpus bien équilibré bâti sur le modèle du BNC par exemple. Il devrait se composer d'un large échantillon de textes diversifiés, ainsi que des transcriptions d'enregistrements oraux afin de capter les particularités du français parlé, qui connaît des usages différents du français écrit. Ce constat, Blanche-Benveniste l'avait déjà fait en 1996. Il reste donc toujours d'actualité.

Blumenthal [2006] a ouvert une piste de recherche intéressante en démontrant que la combinatoire des mots est (partiellement) régie par une série de règles sous-jacentes qu'il serait utile de mettre à nu et de communiquer aux apprenants. Une application de la BLF permet à l'heure actuelle déjà de comparer les profils de deux mots. Pour comparer plus de deux profils, il est impossible de recourir à des techniques rudimentaires. Dans ce cas, on doit recourir à des approches statistiques plus sophistiquées, comme par exemple le positionnement multidimensionnel. Grâce à ces approches, on devrait être en mesure par exemple de mieux cerner les propriétés combinatoires des verbes support sémantiquement vides (*commettre*, *exercer*, *effectuer*, *opérer*, *perpétrer*, *pratiquer*, *procéder*, *réaliser*, ..., et *faire*) [Verlinde e.a., à paraître a].

Pour terminer, le traitement des corpus à l'aide d'outils de traitement automatique des langues (TAL) devrait permettre de pallier certaines faiblesses des outils actuels [Antoniadis e.a., 2005, 2007]. Ainsi, dans la BLF, la sélection d'exemples d'emploi de collocations se fait sur base d'une simple correspondance entre mots. Si les exemples de corpus avaient fait l'objet d'une analyse syntaxique, il aurait été possible de varier davantage ces exemples et de réduire le taux d'exemples inadéquats. Cela vaut également pour les phrases utilisées dans les exercices.

De même, l'outil d'aide au décodage qui analyse des séquences de mots devrait pouvoir être combiné à un analyseur syntaxique en ligne, tel Synthia (vm.socher.org/syn/main-cgi.lsp).

Conclusion

Vingt ans après que l'idée a été lancée en France par Galisson [1987] et avec l'aide de l'informatique, le dictionnaire de dépannage s'est bel et bien transformé en un véritable dictionnaire d'apprentissage. Il offre aujourd'hui un large éventail de fonctionnalités qui permettent entre autres d'accorder au problème de la combinatoire des mots la place qu'il mérite dans l'enseignement des langues.

Bibliographie

[Antoniadis e.a., 2005] Antoniadis G., S. Echinard, O. Kraif, T. Lebarbé et C. Ponton. 2005. Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO. Apprentissage des langues et systèmes d'information et de communication (*ALSIC*). vol. 8, n° 1. 65-79.

[Antoniadis e.a., 2007] Antoniadis G., C. Ponton et V. Zampa. 2007. De la nécessité du TAL dans les EIAH en langues - Les cas EXXELANT et MIRTO. In *Actes EIAH 2007*. Lausanne.

[Back, 2004] Back, M. 2004. A New Bilingual Learner's Dictionary Format: the Junior Bilingue. In Williams, G. et S. Vessier (éds.). *Actes EURALEX 2004 Proceedings*. Lorient, vol. II, 451-455.

[Binon, 2000] Binon, J., S. Verlinde, J. Van Dyck et A. Bertels. 2000. *Dictionnaire d'apprentissage du français des affaires*. Paris : Didier.

[Blanche-Benveniste, 1996] Blanche-Benveniste, Cl. 1996. De l'utilité du corpus linguistique. *Revue française de linguistique appliquée* I.2, 25-42.

[Blumenthal et Hausmann, 2006] Blumenthal, P. et F.J. Hausmann. 2006. Collocations, corpus, dictionnaires. *Langue française* 150.

[Bogaards , 2001] Bogaards, P. 2001. Compte rendu du Dictionnaire du français. *International Journal of Lexicography* 14.4, 319-324.

[Bogaards , 2006] Bogaards, P. 2006. Produire en L2 au moyen d'un dictionnaire bilingue. in Szende, T. (éd.). *Le français dans les dictionnaires bilingues*. Paris, Champion. 23-34.

[Cowie, 1998] Cowie A. (éd.) 1998. *Phraseology: theory, analysis and applications*. Oxford: Clarendon Press. 145-160.

[de Schryver, 2003] de Schryver, G.-M. 2003. Lexicographers' Dreams in the Electronic-dictionary Age. *International Journal of Lexicography* 16.2. 143-199.

[Galisson, 1987] Galisson, R. 1987. De la lexicographie de dépannage à la lexicographie d'apprentissage. *Cahiers de lexicologie* 51.2, 95-117.

[Granger, 1998] Granger, S. 1998a. Prefabricated patterns in advanced EFL writing: collocations and formulae. In Cowie, A. (éd.). Phraseology: theory, analysis and applications. Oxford: Oxford University Press. 145-160.

[Grossmann et Tutin, 2003] Grossmann, F. et A. Tutin. 2003. Les collocations. Analyse et traitement. *Travaux et recherches en linguistique appliquée*, série E, No 1.

[Hausmann et Blumenthal, 2006] Hausmann, F.J. et P. Blumenthal. 2006. Présentation: collocations, corpus, dictionnaires. *Langue française* 150. 3-13.

[Herbst et Popp, 1999] Herbst, T. et K. Popp. 1999. *The Perfect Learners' Dictionary (?).* Tübingen : Max Niemeyer Verlag.

[Le Fur, 2007] Le Fur, D. 2007. *Dictionnaire des combinaisons de mots*. Paris : Editions Le Robert.

[Lewis, 2000] Lewis, M. (éd.) 2000. *Teaching Collocation: Further Developments in the Lexical Approach*. Hove: Language Teaching Publications.

[Mel'čuk, 1984-1999] Mel'čuk, I. e.a. 1984, 1988, 1992 et 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I-IV*. Montréal, Les Presses de l'Université de Montréal.

[Mel'čuk et Polguère, 2007] Mel'čuk, I. et A. Polguère. 2007. *Lexique actif du français*. Bruxelles : De Boeck Université.

[Verlinde e.a., 2006] Verlinde S., J. Binon et T. Selva. 2006. Corpus, collocations et dictionnaires d'apprentissage. In: Blumenthal, P. et F.J. Hausmann. 2006.

[Verlinde e.a., à paraître a] Verlinde, S., J. Binon, S. Ostyn et A. Bertels. (à paraître a). La Base lexicale du français (BLF): un portail pour l'apprentissage du lexique français. *Cahiers de lexicologie*.

[Verlinde e.a., à paraître b] Verlinde, S., T. Selva et J. Binon. (à paraître b). La Base lexicale du français (BLF): de la lexicographie d'apprentissage à l'environnement d'apprentissage. *Lexicographica*.

Utilisation pédagogique du TLFi - Ce que des étudiants peuvent apprendre d'un dictionnaire informatisé

Cécile Fabre(1)
cfabre@univ-tlse2.fr

(1) Université de Toulouse-Le Mirail (Département de Sciences du langage)

Mots-clés : dictionnaires informatisés, enseignement, interrogation du *TLFi*

Keywords: electronic dictionaries, teaching, *TLFi* search

Résumé : Nous proposons un bilan de l'utilisation du *Trésor de la Langue Française Informatisé* dans un enseignement destiné à des étudiants de 2ème année en sciences du langage. L'objet d'un tel enseignement est à la fois de faire découvrir aux étudiants l'ampleur et la complexité de ce dictionnaire et de leur apprendre à l'utiliser pour constituer par la suite des données linguistiques dans leurs travaux personnels. Nous présentons les atouts qu'offre le TLFi pour développer ces compétences diversifiées, de nature à la fois méthodologique, lexicographique et linguistique. Nous faisons également le point sur les difficultés que nous rencontrons, qui sont la contrepartie de la richesse et de la complexité du fonds dans lequel l'interface permet de puiser.

Abstract : This paper presents a teaching experiment based on the use of the *Trésor de la Langue Française Informatisé* by second year students in linguistics. The aim of this course is twofold : it is meant to help students get into this large complex dictionary and lead them to use it as a ressource from which they will be able to draw linguistic data for various tasks. We present the advantages of the TLFi as a means to develop methodological, lexicographical and linguistic skills. We also assess the difficulties we have to cope with in return, due to the richness and complexity of the lexical data made available by the interface.

1. Pouvoir d'attraction du dictionnaire informatisé

Avant d'évoquer les possibilités qu'offre le *Trésor de la Langue Informatisé* (désormais TLFi), il faut insister sur les mérites des dictionnaires informatisés en général pour qui veut susciter l'intérêt des étudiants vis-à-vis des dictionnaires, et initier de nouvelles démarches pour exploiter les ressources qu'ils contiennent. Le support informatique change le regard porté sur les dictionnaires, et facilite la découverte. L'accès à de nombreux dictionnaires numérisés via le site de l'ATILF[1] ou celui du projet ARTFL[2] qui met en ligne des dictionnaires de différentes époques,

[1] http://www.atilf.fr/atilf/res_ling_info.htm

[2] http://www.lib.uchicago.edu/efts/ARTFL/projects/dicos/

rend possibles des travaux pratiques sur machine et crée les conditions d'un travail différent d'observation. Une séance de travaux dirigés combinant la découverte de ces deux sites permet de mesurer avec précision et dans toutes ses dimensions la « révolution électronique » dont parle Jean Pruvost [2006 :153] :

- l'expression de la requête est facilitée par un ensemble de fonctionnalités (troncature, lemmatisation, approximation orthographique, calcul phonétique),
- le balisage de différents champs de description multiplie des modes d'accès au contenu - le TLFi allant évidemment le plus loin dans cette direction,
- des cheminements nouveaux sont suscités par des liens hypertextuels internes (parcours analogiques) ou externes (accès à des ressources complémentaires),
- les modes de visualisation se diversifient (listage des entrées, concordances…).

Ces environnements créent les conditions d'un renouveau de l'intérêt pour les dictionnaires et offrent des pistes nouvelles pour leur exploitation. Les étudiants sont alors prêts à un apprentissage technique relativement rébarbatif parce qu'ils peuvent aussitôt le mettre en pratique et en saisir les bénéfices pour leur formation en linguistique. La difficulté consiste sans doute à faire en sorte que ce type d'apprentissage ne fasse pas obstacle à l'essentiel : la découverte de la richesse lexicographique de ces dictionnaires, et l'accès à des ressources lexicales uniques. Le TLFi a le mérite de rendre compatibles ces différents niveaux d'apprentissage.

2. Une utilisation avancée du TLFi

L'utilisation avancée du TLFi consiste à utiliser pleinement les fonctionnalités du mode de recherche complexe [Dendien et Pierrel 2003] qui comprend :

- la possibilité de formuler des requêtes portant sur plusieurs objets textuels entretenant entre eux des liens de dépendance hiérarchique (lorsqu'un objet est situé dans la portée d'un autre objet) ou d'inclusion (dans le cas d'objets composites comme l'entrée ou l'exemple),
- la formulation de contenus textuels spécifiques permettant d'exprimer des contraintes sur le positionnement des mots dans l'objet, d'intégrer des lemmes, d'exclure certains types d'éléments, etc.
- la gestion de listes constituées manuellement par l'usager ou extraites automatiquement à partir du dictionnaire grâce à des critères graphiques.

Ces fonctionnalités requièrent un apprentissage assez exigeant pour des étudiants de licence. En 2003, Jacques Dendien et Jean-Marie Pierrel déploraient le fait que les usagers dans leur très grande majorité s'en tenaient à une utilisation élémentaire du TLFi et se contentaient de faire une recherche simple de mot [Dendien et Pierrel 2003 :27]. Cette situation n'a probablement guère évolué aujourd'hui, même si un mode de recherche assistée a été mis en place. L'accès à la puissance de requêtage qu'offre l'interface de consultation en mode complexe passe, pour des usagers novices, par un apprentissage technique assez exigeant. Il est néanmoins rapidement payant. A l'issue d'un apprentissage de quelques heures, les étudiants sont à même de mener de petites études linguistiques portant sur des données qu'ils auront eux-même construites. Ils peuvent par exemple étudier la valeur sémantique des suffixes (en observant par exemple les indicateurs associés aux adjectifs en *–eux*, les domaines couverts par les noms en *–ose*...), les différentes dimensions de variation du vocabulaire (vocabulaire argotique, familier, usages régionaux...), l'organisation sémantique du vocabulaire (on peut construire par le biais de différentes requêtes le vocabulaire qui concerne la prison, ou s'intéresser à la sémantique de tous les verbes qui ont trait à l'activité de manger...).

3. Objectifs et compétences sollicitées

Un enseignement articulé autour de la découverte du TLFi[3] sollicite des compétences diverses et facilite plusieurs types d'initiation :

- compétences techniques : il s'agit d'apprendre à formuler les requêtes, ce qui passe par l'apprentissage d'une syntaxe particulière dans la formulation des contenus et surtout par la maîtrise de l'expression des liens logiques entre objets textuels [Pierrel 2003], qui permet de manipuler très utilement la notion d'arborescence.
- compétences lexicographiques : la maîtrise des aspects techniques repose nécessairement sur une connaissance approfondie de la microstructure du TLF. Il s'agit de connaître les différents objets qui la structurent (indicateur, construction syntaxique, syntagme, définition...) et de savoir par quels types d'objets une information donnée est codée. Un contact préalable avec la version papier du dictionnaire s'avère indispensable.
- compétences d'ordre méthodologique : le recours à l'outil informatique, désormais banalisé chez les étudiants, est mis en perspective et questionné de diverses manières par ce type d'utilisation. Ils doivent apprendre à ne pas tout attendre de l'outil, à combiner requêtes automatiques et filtrage manuel pour se conformer à un objectif de recherche. Une réflexion concrète sur les notions de précision et de rappel des requêtes est très bénéfique. Comment évaluer la performance d'une requête ? Comment s'assurer qu'elle ne passe pas à côté de résultats pertinents ? Avec le TLFi, les étudiants découvrent qu'il ne suffit pas qu'une requête donne un résultat pour qu'on puisse s'en satisfaire, mais qu'il faut observer de très près les résultats, réitérer la recherche. C'est un excellent moyen d'aiguiser le regard sur les données linguistiques et d'acquérir du recul par rapport aux utilisations de l'informatique.
- compétences linguistiques : l'objectif premier de cet enseignement est d'amener l'étudiant à avoir plus tard le réflexe de puiser dans le TLFi certaines des données dont il a besoin. Cette démarche est encouragée en prenant appui sur les enseignements donnés dans des modules de linguistique générale, en particulier de lexicologie. L'utilisation du TLFi les amène à constituer des données pour illustrer les types de variation langagière, pour travailler sur la néologie, la polysémie, etc.
- découverte des traitements informatisés : le TLFi est un point de départ utile pour initier les étudiants à la compréhension de certains traitements de base qui seront utiles pour ceux qui souhaitent poursuivre une formation en TAL ou en linguistique de corpus : lemmatisation, création de listes de mots à partir de critères proches des expressions régulières, compréhension des limites de l'interrogation sur texte libre et sensibilisation aux techniques de balisage de textes structurés…
- sensibilisation à des objectifs de recherche concrets : il n'est pas possible de mener à bien avec des étudiants peu avancés des recherches à grande échelle. Mais l'utilisation du TLFi met sur la piste d'expériences de recherches lexicales réalistes, utiles à ceux qui travaillent sur la langue. Voici un exemple (issu d'une expérience réelle) : dans le cadre d'une activité d'expression écrite, un professeur des écoles souhaite constituer un lexique des mots qui riment avec *bulle*. Ce lexique doit permettre d'aider ses élèves de CE1 à écrire de petits textes en puisant dans un vocabulaire proposé. Les étudiants ont pu réaliser cette activité dans le cadre d'un travail dirigé : une liste nommée *ule* est extraite automatiquement du dictionnaire selon le critère `/.*[^aeo]ull?e?r?/` et permet de récupérer des mots vedettes que l'on peut ensuite classer par catégorie grammaticale pour obtenir différentes listes. On trouve par exemple une centaine de verbes en *uler*. On peut également les classer selon les mots contenus dans les définitions, ou le domaine dont ils relèvent, de manière à tenter des regroupements sémantiques. La requête présentée dans le tableau 1 (et faisant appel à la liste *ule* grâce au symbole &l) permet ainsi de

[3] Ce bilan s'appuie sur une expérience d'enseignement assuré dans un département de sciences du langage. D'abord intégré dans un module méthodologique de découverte des bases lexicales et textuelles informatisées, cet enseignement de 16 heures fait désormais partie d'une UE de lexicologie et lexicographie qui accorde une place importante aux dictionnaires informatisés.

trouver 41 noms en *ule* ayant trait à la zoologie. On peut y puiser un vocabulaire propre à stimuler l'imagination d'enfants de 7 ans (*antennule*, *mandibule*, *pustule*, *tentacule*)…

n° d'objet	type de l'objet	lien	contenu
1	Mot vedette	Inclus dans l'objet 4	&lule
2	Code grammatical	Inclus dans l'objet 4	substantif
3	Domaine technique	Dépendant de l'objet 4	zoologie
4	entrée		

Tableau 1: les substantifs rimant avec bulle et relevant du domaine de la zoologie

Le TLFi est indéniablement un bon support d'activités pédagogiques du fait de la variété des compétences qu'il sollicite. On rencontre néanmoins un certain nombre d'écueils, qui sont la contrepartie de la richesse de l'outil.

4. Difficultés

Les difficultés posées par l'utilisation du TLFi sont d'abord liées à des choix d'implémentation. On aimerait que les étudiants puissent réinvestir dans d'autres contextes les connaissances techniques acquises, mais la syntaxe imposée par le TLFi pour la formulation des contenus textuels n'est pas normalisée – un même symbole (|) a par exemple deux sémantiques différentes dans le TLFi et dans Frantext, ce qui est difficile à accepter, et encore plus pour un étudiant novice. Une limitation est plus problématique encore : elle tient à l'impossibilité d'exporter aisément les résultats d'une recherche, pour les retraiter à l'aide d'autres outils – les mettre en forme, effectuer des comptages, par exemple avec un outil bureautique comme Excel. Il y aurait beaucoup à gagner à une mise à plat de ces aspects techniques.
C'est cependant à une difficulté plus radicale, déjà signalée par [Corbin et al. 1995] ou [Martin 2001] que les étudiants sont confrontés en utilisant le TLFi. Evoquant l'exemple de l'étude des constructions impersonnelles du français, Corbin et ses collègues illustraient les problèmes de cohérence rencontrés par l'utilisateur, l'information étant codée de différentes manières (dans l'entrée du verbe, mais aussi via les crochets, l'indicateur, la construction syntaxique) ou pas codée du tout (lorsqu'on trouvait au hasard d'une lecture d'exemple une illustration de construction impersonnelle non prise en compte dans l'article). Ces problèmes ont été détaillés par R. Martin, mesurant l'ampleur du travail d'«aménagement des données » [Martin 2001:101] qui serait requis pour faire du TLFi un dictionnaire automatisé, c'est-à-dire exploitable par des traitements automatiques. Dans le cas d'une consultation humaine, le besoin d'aménagement est moindre ; néanmoins, il reste conséquent. Et les étudiants se heurtent fréquemment au problème des « variantes notationnelles » que décrit Martin [2001:102]. Comme le signale également J. Pruvost [2000:80] : « Si dans le dictionnaire papier la méthodologie peut différer d'un article à l'autre sans que le lecteur s'en aperçoive et en soit même gêné, il en va tout autrement avec le dictionnaire électronique : on a métamorphosé en effet le lecteur du dictionnaire en formulateur de requêtes et indirectement renforcé ses exigences. » L'utilisateur a en effet du mal à accepter de ne pas pouvoir s'appuyer sur des libellés stables pour interroger le TLFi comme il le fait d'autres bases de données. On peut néanmoins accorder à ces écueils une vertu, celle d'obliger les étudiants à passer par une recherche préalable minutieuse des sources de variation. Ainsi par exemple pour la recherche des onomatopées dans le dictionnaire. Une observation un peu minutieuse de quelques exemples montre que l'information apparaît 57 fois dans le code grammatical, 31 fois dans la définition, 25 fois dans les crochets, 23 fois dans l'indicateur. Mais deux problèmes se posent alors : 1) on n'est jamais tout à fait sûr que la recherche soit close. On peut simplement considérer qu'il est raisonnable de penser que le lexicographe n'a pas décidé d'indiquer cette information ailleurs, ou sous une autre forme. 2) Il est impossible de faire un

décompte immédiat des onomatopées, dans la mesure où on ne peut pas se contenter de faire l'union des ensembles d'unités lexicales récupérés. Dans la plupart des cas, la mention apparaît dans plusieurs objets. Par exemple, lorsque la mention apparaît dans l'indicateur, elle apparaît toujours également dans le code grammatical. Alors que lorsqu'elle apparaît en tête de définition, cette information n'est la plupart du temps pas reprise dans le code grammatical (dont la valeur est plutôt *interjection*). On est amené à élaborer des recettes, plus ou moins satisfaisantes, pour s'approcher au mieux d'un besoin de recherche. C'est particulièrement le cas de recherches sémantiques, du type de celles que [Corbin et al. 1995] évoquaient à propos de l'extraction des noms désignant des statuts sociaux liés à une activité. A défaut de disposer d'étiquettes sémantiques qui seraient extrêmement précieuses, on est amené à exploiter les premiers mots de la définition – lorsque celle-ci est de type définition par inclusion. Nous avons cherché par exemple à recenser les noms d'animaux employés de manière métaphorique. C'est le cas du mot *hyène* :

> HYÈNE, subst. fém.
> Animal nocturne d'Afrique et d'Asie, carnassier très vorace, […]
> *P. compar.* (ou *p. métaph.*). [Symbole de la laideur, de la lâcheté, de la cruauté] *Lui n'a jamais adouci pour moi son œil d'hyène* (DUMAS père, *Cte Hermann,* 1849, IV, 10, p. 303).

La requête présentée dans le tableau 2 pêche surtout par un mauvais taux de rappel, et ce pour au moins deux raisons : le mot *animal* (dont on veut qu'il apparaisse au début de la définition – &d2 signifie que 2 mots au maximum le séparent du début de l'objet) ne figure pas dans la définition de tous les noms d'animaux ; l'usage métaphorique peut être codé par d'autres indicateurs (*p. anal.*, *p. compar.*, *fig.*). Corriger ce silence passe par la constitution de listes manuelles, liste restreinte dans le cas des valeurs de l'indicateur, ouverte dans le cas des noms d'animaux. On apprend alors à l'étudiant à mesurer le coût du travail à accomplir, à se contenter d'un certain état de la recherche, tout en l'encourageant à persévérer et à ne pas s'arrêter trop tôt.

n° d'objet	type de l'objet	lien	contenu
1	Code grammatical		subst
2	définition	Dépendant de l'objet 1	&d2 animal
3	indicateur	Dépendant de l'objet 2	métaphore

Tableau 2 : les noms d'animaux ayant un emploi métaphorique

Ces problèmes font le quotidien du travail linguistique. Ils constituent des énigmes qu'un linguiste a plaisir à élucider. Mais l'effet de ces tâtonnements peut être redoutable chez les étudiants, qui en retirent l'impression d'un à peu près, d'un manque de rigueur qu'on a beau jeu ensuite de leur reprocher. C'est donc un défi à relever, qui passe par un cadrage très strict des activités que l'on propose aux étudiants (celles dont on sait qu'elles peuvent donner lieu à des recherches fécondes sur le dictionnaire, et qui ne passeront pas par des détours trop laborieux), ce qui risque parfois de brider la spontanéité de la découverte.

Conclusion

Ce bilan d'une expérience de nature pédagogique ne rend pas compte de tous les usages qui peuvent être faits du TLFi par des linguistes aguerris. Notre objectif était en effet de réfléchir aux possibilités qu'offre cet outil en tant que support d'une initiation à la collecte de données linguistiques, au cours des premières années de formation en linguistique. Les illustrations proposées correspondent à des objets d'étude très modestes, qui sont à la portée d'étudiants débutants. Nous avons cependant insisté sur la richesse des activités que le TLFi rend possibles, et sur sa capacité à encourager les étudiants à manipuler des données lexicales et à devenir

autonomes dans la formulation d'un objectif de recherche et dans la collecte des données correspondantes. C'est en cela un outil très stimulant. En contrepartie, son utilisation est limitée par les inévitables incohérences d'un dictionnaire volumineux qui n'a pas été conçu à l'origine pour une interrogation extensive. D'autres bases de données et d'autres outils doivent prendre le relais lorsqu'on veut envisager des études lexicales à grande échelle. C'est l'objet d'autres apprentissages.

Remerciements
Ce travail a bénéficié de discussions avec Michelle Lecolle (CELTED), Marie-Paule Péry-Woodley et Josette Rebeyrolle (CLLE-ERSS).

Bibliographie

[Corbin et al., 1995] Corbin, D., Corbin, P., Tutin, A., Aliquot, S. (1995) : « Ce que des linguistes peuvent attendre d'un dictionnaire informatisé », in D. Piotrowski (éd.), *Lexicographie et informatique. Autour de l'informatisation du Trésor de la langue française*, Actes du colloque international de Nancy (29, 30, 31 mai 1995), Paris, Didier Erudition, p. 51-77.

[Dendien et Pierrel, 2003] Dendien, J., Pierrel, J.-M. (2003) : « Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence » (2003), *TAL (Traitement Automatique des Langues)*, numéro sur les dictionnaires électroniques, Hermes Sciences Edition , vol. 44, n° 2, p. 11-38.

[Martin R., 2001] Martin, R. (2001) : *Sémantique et automate*, *l'apport du dictionnaire informatisé*, PUF, Écritures électroniques, Paris.

[Pierrel J.-M., 2003] Pierrel, J.-M. (2003) : « Un ensemble de ressources de référence pour l'étude du français : TLFi, Frantext et le logiciel Stella », revue québécoise de linguistique, numéro sur TALN, Web et corpus, vol.32, n°1, p. 155-176.

[Pruvost J., 2000] Pruvost, J. (2000) : *Dictionnaires et nouvelles technologies*, PUF, Écritures électroniques.

[Pruvost J., 2006] Pruvost, J. (2006) : *Les dictionnaires français outils d'une langue et d'une culture,* Ophrys, L'Essentiel français.

Automatisation du langage, premiers corpus informatisés et lexicographie dans les années 1950-60 : étude comparée

Jacqueline Léon (1)
jleon@linguist.jussieu.fr

(1) Laboratoire d'Histoire des Théories Linguistiques, CNRS, Université Paris 7 Denis Diderot

---

Lors de l'automatisation du langage apparue dans les années 1950-1960[1], le lexique a constitué un axe d'investigation privilégié, suscitant parfois un regain d'intérêt pour le « mot », unité alors très controversée par les linguistes structuralistes. Toutefois, bien que le phénomène d'automatisation ait été plus ou moins simultané dans les différents pays, cet intérêt a été très divers et d'importance variable selon les traditions linguistiques et les positions théoriques, méthodologiques ou pratiques diversement adoptées. Dans notre intervention, nous proposons une étude historique et épistémologique de l'automatisation du lexique, en insistant sur les traitements effectués à l'aide des premiers corpus informatisés. En particulier, nous examinerons un certain nombre de questions que suscitent ces travaux pionniers.

Le remaniement des conceptions du 'mot' et du lexique occasionné par l'informatisation a-t-il été le même pour les différentes approches, empiristes et/ou structuralistes? En particulier, les questions que pose l'informatisation sont-elles les mêmes selon que l'on s'intéresse à la structure des mots ou à leur sens ? Peut-on parler de positions structuralistes ou bien empiristes homogènes, indépendantes des traditions linguistiques concernées? Autrement dit, peut-on limiter les différents traitements à deux courants "analyse structurelle" et "analyse statistique" comme le proposent Habert et Jacquemin (1993) ou bien faut-il affiner les distinctions? Quel rôle ont joué les objectifs pratiques - traduction automatique, construction de dictionnaire, enseignement des langues … - dans les projets d'automatisation du lexique et quelles différences d'approches du traitement des mots ont-ils occasionné ? Lorsque les linguistes ont eu recours à un corpus, ont-ils toujours jugé pertinent de distinguer corpus et texte, ou bien encore corpus, texte et contexte pour l'étude du lexique ? L'utilisation de

[1] Par automatisation du langage, nous entendons les processus de modification des théories et des méthodologies des sciences du langage, dès lors que le traitement des langues par ordinateur a pu être envisagé. Ce processus a été engagé dès 1948 avec les premières expériences de traduction automatique qui furent les premières appplications non-numériques des ordinateurs.

méthodes statistiques du lexique a-t-elle conduit nécessairement à poser la question de la nature probabiliste du langage ?

A l'examen de ces questions, plusieurs lignes de partage se dessinent. Si l'on considère tout d'abord les objectifs pratiques, la prise en considération du lexique est radicalement différente selon que l'on se soit donné pour tâche la traduction automatique de textes, qui comme on le sait a concentré la majeure partie des efforts pionniers de traitement automatique du langage, ou au contraire qu'on ait eu pour objectif l'automatisation des comptages de vocabulaire déjà florissants dans les années 1950 à des fins lexicographiques ou d'enseignement des langues.
Les premiers travaux en traduction automatique ont suscité, en France, un renouveau de réflexion sur le statut du 'mot' chez les linguistes structuralistes (Léon 2001, 2004). L'objectif consistant à vouloir traduire des textes par une machine électronique, même s'il n'a donné lieu qu'à peu de réalisations concrètes, les a conduits à s'interroger de façon originale sur le statut des unités traitées. Ils ont été ainsi amenés à définir des unités syntaxiques supérieures au mot et à les appréhender non seulement du point de vue de leur mode de construction interne mais aussi de leurs rapports avec le reste de l'énoncé. Les termes forgés à l'occasion de ces réflexions, *lexies* chez Bernard Pottier, *synapsies* chez Emile Benveniste et *synthèmes* chez André Martinet, traduisent des interrogations inédites sur les critères d'identification, de construction et de classement. On voit émerger au travers de questionnements sur l'ancrage morphologique et/ou syntaxique des mots composés, et sur leur statut discursif, un renouveau de la lexicologie
A noter toutefois que, dans le contexte de la linguistique structuraliste des années 1960, l'objectif pratique de traduction automatique n'a été souvent qu'un prétexte pour développer des considérations essentiellement théoriques. De plus, même si l'objectif de traduction a contraint les linguistes à s'intéresser au texte, celui-ci n'est conçu que comme contexte phrastique ou extra-phrastique pour trouver des règles de repérage des unités. Nul besoin dans ce cas de corpus ou de contexte.
La tradition structuraliste n'est pas homogène. Alors que les structuralistes français s'intéressent au « mot » et à sa structuration interne ou à celle des groupes de mots, les structuralistes américains, qui ne distinguent pas la morphologie de la syntaxe, ne lui accordent aucune importance[2]. Au départ, Harris ne s'intéresse pas aux mots. Même si une approche empiriste et inductive le conduit à utiliser des méthodes probabilistes, celles-ci servent à repérer des morphèmes dans un énoncé et non à étudier le lexique en tant que tel (Harris 1955). Lorsqu'il va s'intéresser aux cooccurrences de mots, Harris les conçoit comme classes de mots dans une perspective distributionnelle, à partir de restrictions de sélection lexicales, et les critères sont syntaxiques et non sémantiques. Texte, corpus et contexte (ou discours) n'ont pas de définition autonome ; ils sont complètement déterminés pour un sous-langage donné, pour lequel il s'agit de donner une représentation structurale et au sein duquel les classes de mots sont fermées (Harris 1968). Il en ira autrement pour les héritiers de Harris au sein du structuralisme français qui ont appliqué l'analyse distributionnelle à l'étude des groupes de mots. Pour Jean Dubois (1960), pionnier de l'analyse du discours française, le corpus est un ensemble de textes, défini selon des critères socio-historiques. Les groupes de mots sont des unités sémantiques ayant une fonction discursive pour un corpus donné. Pour Maurice Gross (1982), en revanche, les classes de mots restent déterminées de façon morpho-syntaxique au sein de son lexique-grammaire.
La prise en considération du 'mot' est radicalement différente lorsqu'il s'agit de mener à bien des entreprises pratiques, lexicographiques ou pour l'enseignement des langues. Deux aspects

[2] Par ailleurs, au sein des différents courants de grammaires formelles, Bar-Hillel (1955), confronté au problème posé par la traduction automatique des unités lexicales complexes, est le seul à s'intéresser aux 'idioms'.

deviennent alors cruciaux : la définition du corpus et la discussion sur les propriétés statistiques du vocabulaire et plus largement du langage. Toutefois traditions françaises et britanniques s'opposent ici bien que des points de convergence apparaissent nettement.
En France, à partir des travaux de Mario Roques et de son *Inventaire Général de la Langue Française* (1936), on observe un intérêt continu pour l'étude du vocabulaire. La stylistique et l'étude du vocabulaire français, préoccupations centrales de nombre de linguistes à l'époque, président à l'organisation en 1957 du colloque de Strasbourg *Lexicologie et lexicographie françaises et romanes* ayant abouti à la création du TLF en 1960. Y est évoqué l'apport prometteur des machines mécanographiques et électroniques dans l'accélération des dépouillements et des classements du lexique. Cet essor de la lexicologie par son automatisation est également marqué par la création, en 1959 à Besançon, du Laboratoire d'analyse lexicologique et des deux revues, *Les Cahiers de Lexicologie* et *Les Etudes de Linguistique Appliquée*, tous trois sous la direction de Bernard Quemada, suivie par un colloque international sur la mécanisation des recherches lexicologiques qui a eu lieu en 1961, toujours à Besançon, et d'un second colloque à Strasbourg en 1964 intitulé *Statistiques et analyse linguistique* auquel ont participéde nombreux linguistes, G.Bourquin, E.Coseriu, J. Dubois, F. François, G.Gougenheim, A.J. Greimas, P.Guiraud, R.Martin, H.Mitterand, G.Moignet, R. Moreau, Ch. Muller, B.Pottier, B. Quemada, G.Straka. Les communications portent sur l'application de méthodes statistiques à la stylistique, à la philologie, à la dialectologie et à l'enseignement des langues. La position de Charles Muller (1968) qui avait pour ambition de promouvoir la statistique lexicale comme partie intégrante de la linguistique, domine les travaux de cette époque. Ceux-ci trouveront un essor particulier avec la création du Laboratoire de Lexicologie politique de l'ENS de St Cloud à la fin des années 1960.
Dans la tradition empiriste de la linguistique britannique, en particulier au sein de la London School dont un chef de file est J.R. Firth, la linguistique descriptive (ou empirique), centrée sur l'étude de l'usage, occupe une place essentielle. La linguistique est une science appliquée, orientée vers la pratique: enseignement des langues, traduction, confection de grammaires et de dictionnaires, vulgarisation et enseignement de la linguistique, traitement automatique des langues, etc. Au sein d'une conception polysystémique du langage, opposée à la conception monosystémique du structuralisme excluant l'étude du sens, J.R. Firth (1951) défend une approche du sens des mots par collocation « meaning by collocation » selon laquelle le sens lexical réside dans l'usage des mots en contexte, sur l'attente mutuelle ('mutual expectancy' ou 'collocability') qu'un mot fait porter sur un autre, et non dans une sémantique du mot *a priori*, conceptualiste, logique ou psychologique. La notion de texte est centrale pour Firth et l'attestation des collocations dans des textes authentiques est essentielle pour l'étude du sens lexical (Léon, sous presse). Dans les années 1960, MAK Halliday, élève de Firth, propose la notion de « lexicalness », faisant pendant à la notion chomskyenne de « grammaticalness », afin de rendre compte de l'idée d'un continuum entre lexique et grammaire, déjà présente dans la tradition britannique, et cohérente avec la conception polysystémique de Firth. Lexique et grammaire participent alors d'un même niveau, le «*lexicogrammar*»[3]. En proposant la recherche de patterns de collocations dans des textes à l'aide de chaînes de Markov, il commence à mettre en oeuvre sa conception probabiliste du langage (Halliday 1966). Il soutient et surpervise le projet OSTI (UK Government Office for Scientific and Technical Information) entrepris en 1963 par John Sinclair, alors jeune chercheur à l'Université d'Edimbourg, et destiné à l'étude par ordinateur des *patterns* de collocations dans

[3] La proximité terminologique entre « lexique-grammaire » et « lexicogrammar » n'est pas uniquement une coïncidence. Bien qu'issues d'approches théoriques différentes, les deux notions mettent en cause les positions chomskyennes sur la créativité et le rôle de la mémoire dans l'apprentissage.

des enregistrements d'anglais oral et écrit. La notion de corpus se pose alors. Pour Sinclair, comme pour Firth, le corpus doit être constitué de textes authentiques et intégraux. De plus pour Sinclair (1965, 1966), l'idée qu'un texte intégral constitue à lui seul un échantillon de langage a pour conséquence de concevoir le corpus comme potentiellement infini. Cette conception du corpus qui s'oppose en tous points à celle de Randolph Quirk (1960), autre élève de Firth et auteur du SEU (Survey of English Language), corpus pré-informatisé, qu'on peut considérer comme le précurseur du Corpus Brown (Kucera et Francis 1967). Pour Quirk, un corpus est l'aboutissement d'une construction systématique par le linguiste, constitué par des échantillons selon les genres, obtenus à partir de textes authentiques mais aussi à partir de tests expérimentaux (Léon 2005). On notera toutefois que le Corpus Brown comme le SEU, conçus essentiellement dans le contexte du structuralisme américain, serviront plus à l'étude de structures grammaticales que lexicales[4].

Il est important de souligner que les contacts entre néo-firthiens et lexicologues français sont nombreux. P.J. Wexler, qui deviendra président du comité de pilotage du projet OSTI, assiste en 1957 au colloque où fut décidée la création du TLF. Halliday et Sinclair visitent le Laboratoire d'analyse lexicologique de Besançon en 1963. Ils publieront respectivement dans les *Etudes de Linguistique Appliquée* et dans les *Cahiers de Lexicologie*. Enfin la notion statistique de disponibilité du lexique du *Français Elémentaire* (Gougenheim et al. 1954), citée par Halliday, n'est pas sans rapport avec la notion probabiliste de « lexicalness ».

Contrairement à la situation française des années 1960, éclatée entre structuralisme saussurien et statistiques lexicales encore peu théorisées, la cohérence des linguistes britanniques est tout à fait significative. Unis par une vision empiriste commune - conception de la linguistique comme science appliquée, rôle crucial de l'usage et idée d'un continuum entre lexique et grammaire - ne les empêchent pas de discuter les thèses chomskyennes alors en plein développement. Cette cohérence n'a sans doute pas été sans conséquence sur la prééminence des linguistes britanniques dans ce qu'on appelle actuellement la « Corpus linguistics » alors que, pourtant, les Français étaient, grâce au TLF, parmi les pionniers dans la constitution de corpus informatisés.

Bibliographie

Bar-Hillel Yehoshua. 1955. "Idioms", *Machine Translation of Languages*, 14 essays, William N. Locke et Andrew D. Booth, éds., New York, MIT et John Wiley, p.183-193.

Dubois Jean. 1960. "Les notions d'unité sémantique complexe et de neutralisation dans le lexique", *Cahiers de lexicologie*, n°2, p.62-66.

Firth John Rupert, 1957 [1951], "Modes of Meaning", *Papers in Linguistics (1934-1951)*, Oxford, Oxford University Press, p.190-215.

Gougenheim Georges, Robert Michea, Paul Rivenc et Aurélien Sauvageot, 1954, *L'élaboration du français élémentaire*, Paris, Didier.

Gross Maurice, 1982, "Une classification des phrases figées du français", *Revue québécoise de linguistique*, n°11-2, p.151-185.

Guiraud Pierre, 1954, *Les Caractères statistiques du vocabulaire* Paris : PUF

[4] Quirk, bien que considéré comme faisant partie de la London School, a fait un PhD en syntaxe et a complété sa formation aux Etats-Unis au début des années 1950.

Habert Benoît et Christian Jacquemin, 1993, "Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques", *TAL,* n°34-2, p.5-43.

Halliday, M.A.K., 1966 ,'Lexis as a Linguistic Level'. In *In memory of J.R. Firth*, C.E. Bazell, J.C. Catford, M.A.K Halliday, R.H. Robins, (eds.) Longmans, p.148-162.

Harris, Z.S. 1955. From phoneme to morpheme. *Language* 31, p.190-222.

Harris, Z.S. 1968. *Mathematical Structures of Language*. New York: John Wiley & Sons.

Kucera, H. & N. Francis. 1967. *Computational Analysis of Present Day American English*. Providence : Brown University Press.

Léon J. 2001. "Conceptions du mot et débuts de la traduction automatique", *Histoire Epistémologie Langage*, n°23-1, p.81-106.

Léon J. 2004. "*Lexies, synapsies, synthèmes*: le renouveau des études lexicales en France au début des années 1960 " *History of Linguistics in Texts and Concepts ˜ Geschichte der Sprachwissenschaft in Texten und Konzeptionen,* G. Hassler (ed.) Münster: Nodus Publikationen, p.405-418.

Léon, J. 2005. « Claimed and unclaimed sources of Corpus Linguistics ». *The Henry Sweet Society Bulletin of History of Linguistics* n°44, p.34-48. Reprint in 2007, *Corpus Linguistics: Critical Concepts in Linguistics* 6 volumes (Ramesh Krishnamurthy & Wolfgang Teubert eds.) vol. 1, London & New York : Routledge, p.326-341.

Léon, J. (sous presse). " Meaning by collocation. The Firthian filiation of Corpus Linguistics » Proceedings of *ICHoLS X, 10th International Conference on the History of Language Sciences*, (D. Kibbee ed.), John Benjamins Publishing Company.

Muller, Charles. 1968. *Initiation à la linguistique statistique*, Paris, Larousse.

Quirk, Randolph. 1960. « Towards a description of English Usage », *Transactions of the Philological Society*, p.40-61.

Sinclair, John. 1965. «When is a poem like a sunset ?» *A Review of English Literature* 6-2, p.76-91.

Sinclair John. 1966, "Beginning the Study of Lexis" *In Memory of J. R. Firth,* Bazell Charles E., John C. Catford, Michael A.K. Halliday, Robert H. Robins éds., Londres, Longmans, p.410-30

Annotation sémantique : profilage textuel et lexical

Mick Grzesitchak (1)

mickgrz@gmail.com

Evelyne Jacquey (1)

ejacquey@atilf.fr

Fabienne Baider (2)

(1) *ATILF - Nancy Université & CNRS*

(2)*Université de Chypre*

Mots-clés : sémantique textuelle, recherche d'informations, annotation sémantique, statistiques textuelles, TAL

Keywords: textual semantics, information retrievial, semantic annotation, textometry, NLP

Résumé : Cet article présente la plateforme d'annotation sémantique développée au sein du projet DIXEM. Celle-ci s'appuie sur un premier lexique extrait automatiquement depuis le TLFi. Différents aspects de cet outil sont abordés, son cadre théorique, les modes de représentation choisis, ses capacités actuelles ou encore les perspectives et les objectifs que nous suivons. Enfin, nous décrirons ses utilisations afin d'observer, sous un angle nouveau, deux corpus préparés à l'origine pour une étude linguistique, respectivement de la féminisation dans le vocabulaire français et du discours journalistique à propos de l'immigration en France.

Abstract : This article presents a semantic annotation system which has been built in the framework of the DIXEM project. It starts with a first automatically-extracted lexicon from the TLFi dictionary. Various aspects of this system are described: its theoritic framework, its interface, its present possibilities and the goals we would like to reach. Two experimentations are described: one about the feminisation in the french vocabulary and the second about immigration in France.

Introduction

La plateforme a été réalisée dans le but de pouvoir analyser des corpus de textes, les annoter sémantiquement et d'essayer d'en extraire des données sémantiquement informatives sur la base de procédures statistiques.

Nous optons pour une approche complémentaire en croisant sémantique interprétative et statistiques, et souhaitons offrir à la linguistique ainsi qu'au TAL un nouvel objet d'étude. Nous suivons en cela plusieurs travaux actuels [Rossignol & Sébillot 2006], [Enjalbert & Victorri 2005 : p. 82-83] et [Caillet, Pessiot, Amini & Gallinari 2004]. Au final, nous voulons isoler des informations sémantiques (et/ou thématiques) en étudiant les isotopies textuelles d'un texte par des méthodes statistiques.

La plateforme d'annotation a été réalisée en Python, langage de prototypage reconnu, qui permet d'écrire des programmes informatiques rapidement en respectant néanmoins la plupart

des contraintes et normes des langages robustes tel Java. Elle se veut être également un composant Python facilement réutilisable (et modifiable ou étendable) en se présentant sous la forme d'un paquetage indépendant et transportable.

L'outil s'utilise en ligne de commande (via des scripts python) ainsi qu'en manipulant un fichier de configuration. Dans un premier temps, aucune interface graphique ne vient accompagner le logiciel. Cela n'exclut pas la possibilité d'en développer par la suite. Il est fréquent que des applications graphiques soient simplement des « front-end » d'applications en ligne de commande, c'est-à-dire des interfaces graphiques pour des programmes non graphiques.

Enfin la première version de la plateforme souhaite, malgré sa jeunesse, constituer un socle logiciel cohérent, pérenne, efficace, avec le moins d'erreurs et le plus de contrôles possibles. Le projet cherche actuellement à établir des bases solides pour les enrichissements futurs. C'est dans ce but également qu'un manuel d'utilisation a été réalisé.

1. Présentation de la plateforme d'annotation « Sémy »

La plateforme s'appuie à la base sur les entrées du TLFi[1] dont voici un exemple :

VERTEX, subst. masc.

A. 1. ANAT., ANTHROPOL. Point le plus élevé de la voûte crânienne.
2. ENTOMOL. „Région de l'épicrâne située immédiatement derrière le front entre les yeux composés`` (SÉGUY 1967).
B. 1. ASTRON. „Point représentatif, sur la sphère céleste, de la direction du vecteur vitesse d'un ensemble d'étoiles`` (Astron. 1980).
2. GÉOD. Point de latitude maximale d'une courbe qui est tracée sur une surface de révolution.

Tous les mots sémantiquement pleins (verbes, noms, adjectifs, adverbes) de chacune des définitions sont considérés par hypothèse comme ses traits sémantiques. Un lexique de traits sémantiques, dont la structure est détaillé dans la partie 1.1, est donc crée de la sorte et nous permet d'annoter sémantiquement des textes.

1.1 Représentation du lexique

Comme l'illustre la figure 1, le lexique est structuré par association d'éléments. Ainsi, il contient une entrée 'vertex', qui contient une seule entrée 'commonNoun' (puisque l'entrée 'vertex' n'a qu'une seule forme lemmatique) à laquelle sont associées toutes les informations qui sont nécessaires : le nombre de définitions de l'entrée, les traits sémantiques extraits, les informations de domaines ou de crochets. Toutes ces données sont représentées de façon compacte mais sans perdre d'information.

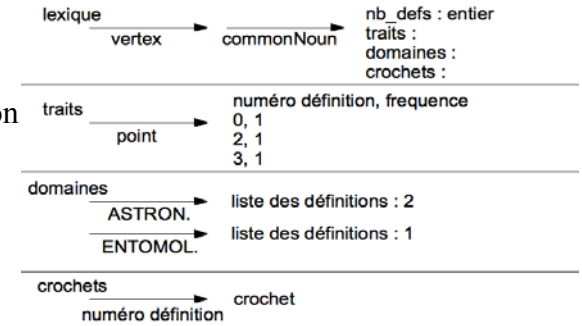


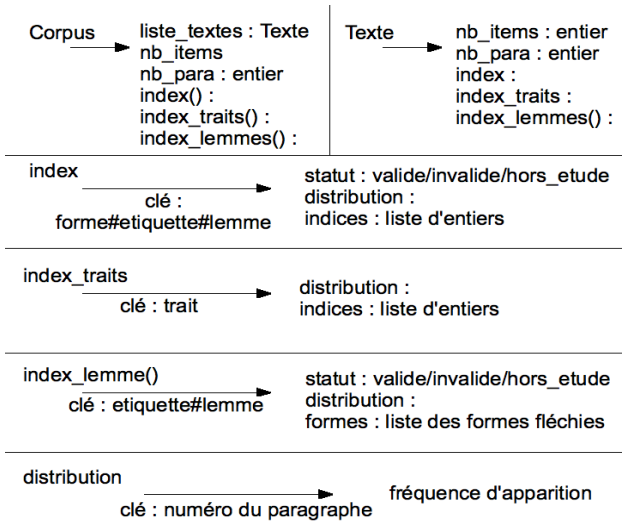
Figure 1 : La représentation du lexique

[1] En 2006, Etienne Petitjean a réalisé une extraction et une normalisation à partir des entrées du TLFi, ce qui a permis de produire une première version du lexique de sèmes utilisé, SEMEME. L'extraction consiste en la sélection des mots pleins des définitions associées aux mots vedettes du dictionnaire.

1.2 Représentation distributionnelle des textes

Pour chaque texte, deux index sont créés, l'un recensant tous les éléments textuels de surface[2], et l'autre étant l'index des traits sémantiques à savoir le résultat de l'annotation sémantique des items restants après le filtrage morphologique.

Au sein des index, chaque élément textuel ou trait sémantique est associé à sa représentation distributionnelle dans le texte et ses paragraphes[3]. Cela simplifie grandement, d'une part, l'analyse statistique qu'il est possible de réaliser sur le texte et ses items, et d'autre part, le développement de la structure logicielle. En effet, avec ces deux index (les seuls à être enregistrés), il est ensuite possible d'en générer d'autres comme par exemple l'index des lemmes d'un texte, mais aussi l'index d'un corpus, ou encore l'index des lemmes d'un corpus.



Cette représentation des corpus et des textes est synthétisée dans la figure 2.

Figure 2 : La représentation des textes, des corpus et des index.

1.3 Fonctionnement

La première version permet de finaliser plusieurs opérations essentielles à l'outil en développement. Ainsi, à ce jour il permet principalement trois opérations (explicitées par la figure 3).

Il permet d'importer un texte depuis un fichier (ou un corpus depuis plusieurs fichiers) , en réalisant plusieurs sous-tâches séquentielles : le nettoyer, l'étiqueter[4] et le filtrer[5] morphologiquement, annoter sémantiquement tous les items textuels cibles (avec par défaut les expressions figées qui sont répertoriées et repérées dans les textes) et le représenter informatiquement et de façon distributionnelle.

---

[2] Tous les éléments de la forme brute du texte sont répertoriés à l'aide d'une clé unique : les mots, certaines expressions figées, la ponctuation, les changements de paragraphes sans perte d'information afin de pouvoir recréer la forme originale du texte à tout moment.

[3] Par exemple si le mot 'chat' apparaît cinq fois dans un texte de 3 paragraphes, une fois dans le n° 1 et quatre fois dans le n° 2, alors sa représentation distributionnelle est {0:1,1:4,2:0}.

[4] L'étiquetage morphologique s'effectue avec l'étiqueteur TreeTagger : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[5] Par défaut, les noms communs, les verbes, les adjectifs, les adverbes et les syntagmes sont étudiés, au contraire de la ponctuation ou des mots fonctionnels.

Une fois cette première étape réalisée, il est possible d'exporter le texte (ou les textes) sous plusieurs formats, ainsi que leurs index et distributions. On peut également calculer et exporter des mesures statistiques sur les distributions.

Enfin on peut comparer « distributionnellement » deux textes entre-eux, et exporter les résultats.

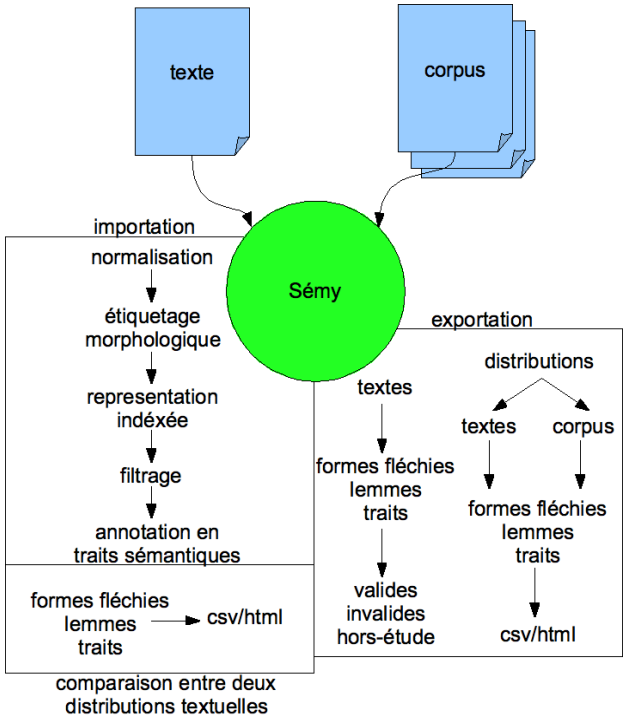


Figure 3 : schéma de fonctionnement

2. Applications sur corpus

La plateforme a déjà été utilisée pour annoter plusieurs corpus. Lors d'une première expérience, nous avons étudié les traits sémantiques d'une sélection d'articles du Monde Diplomatique [Auteur, 2007]. D'autre part, nous relatons une seconde expérience. La plateforme d'annotation a été utilisée sur un second corpus constitué de 300 titres de la presse française à propos de Ségolène Royal et Nicolas Sarkozy, soit un corpus assez homogène pour chacun des personnages, 9117 mots pour Sarkozy et 9477 pour Royal [Auteur2, 2007]. L'annotation sémantique en traits sémantiques assurée par la plateforme fait apparaître plusieurs éléments. Comme la plateforme permet d'étudier les fréquences tant des formes que des traits, il a été possible d'extraire facilement les verbes les plus fréquents autour de chaque personnage. Il apparaît alors une forte dissymétrie entre eux. Les verbes fréquents au côté de Royal décrivent des activités très générales (*révéler, montrer, dire, signifier, déclarer, signaler, énoncer*) alors que les verbes fréquents autour de Sarkozy décrivent des activités précises comme *embaucher, signer*. Du point de vue des champs sémantiques impliqués, on note une forte présence des verbes relationnels (*annoncer, susciter*) et de campagne (*élire, voter*) chez Royal, à comparer avec les champs sémantiques impliqués avec Sarkozy, la réalisation intellectuelle ou concrète (*réaliser, décider*) ou la présence d'un danger (*défendre, inquiéter*).

D'autre part, la même plateforme a permis d'étudier les fréquences de traits autour des deux personnages. Une comparaison des traits fréquents associés aux verbes fréquemment cooccurrents de chaque candidat entre en résonance avec les sondages d'opinion qui ont été

largement diffusés à cette période : la compétence et la stature d'un homme d'état pour Sarkozy, l'écoute pour Royal. En analysant les fréquences de traits sémantiques calculées par la plateforme, Sarkozy apparaît comme un entrepreneur (/administrer/, /remplir/, /charger/, /exécuter/, etc.) et Royal comme une militante, autrement dit comme caractérisée par une qualité associée habituellement à la condition féminine.

Perspectives et objectifs

Les développements futurs de la plateforme vont se concentrer, principalement autour du module de statistiques (en apprentissage, classification, reconnaissance, recherche et extraction d'informations), en restant cohérent avec les contraintes linguistiques issues de la sémantique interprétative et celles de la linguistique de corpus. Étudier l'aspect de la mise-à-jour (ou l'enrichissement ou encore le raffinage) du lexique mais aussi des textes et des corpus annotés, soit en d'autres termes de l'aspect évolutif de la plateforme elle-même, apparaît également comme intéressant du point de vue théorique. L'idée serait de se rapprocher itérativement du couple (lexique, corpus de référence) le plus cohérent, le plus efficace sur les plans de la sémantique, la linguistique, l'informatique et des statistiques.

A l'heure actuelle, la plateforme sait extraire du TLFi un premier lexique et annoter des textes avec celui-ci. Un stage a eu pour objet la réduction de l'hétérogénéité du lexique[6] sur les plans lexicologiques et lexicographiques. Nous envisageons des procédures pour classifier le lexique en s'appuyant sur [Valette et al, 2006], mais aussi les corpus (et les textes). A plus long terme, nous souhaitons détecter les isotopies textuelles[7] et lier de façon plus fine les corpus et le lexique.

Bibliographie

[Auteur1, 2007] Auteur1 (2007) :

[Auteur2, 2007] Auteur2. (2007) : « Féminisation des noms de métiers, discours journalistique : Une grande victoire ou une petite concession ? » *SILF 2007 : XXXIe colloque international de linguistique fonctionnelle*, Université Saint-Jacques de Compostelle, Espagne, Septembre 2007.

[M. Caillet, J-F. Pessiot, M-R. Amini, P. Gallinari, 2004] Unsupervised Learning with term clustering for Thematic segmentation of texts, In RIAO 2004, 26-28 Avril 2004, Avignon, France.

[P. Enjalbert, 2005] Sémantique et traitement automatique des langues Hermès Sciences.

[P. Enjalbert & B. Victorri, 2005] Les paliers de la sémantique. Chapitre 2 du document [Enjalbert, 2005].

[M. Rossignol & P. Sébillot, 2006] Acquisition sur corpus non spécialisés de classes sémantiques thématisées, In Jean-Marie Viprey, editor, 8èmes Journées internationales d'Analyse Statistiques des Données Textuelles (JADT 2006), Besançon, France.

[M. Valette & al, 2006] M. Valette, A. Estacio-Moreno, E. Petitjean, E. Jacquey, « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens », Verbum ex machina, Actes

[6] Dans le cadre d'un stage de Master 2 Recherche, été 2007, Egle Ramdani a jeté les premiers jalons de ce travail.

[7] Les traits sémantiques récurrents qui participent à la cohésion sémantique d'un texte.

de la 13ème conférence sur le traitement automatique des langues naturelles (TALN 06). Piet Mertens, Cédrick Fairon, Anne Dister, Patrick Watrin (éds). Cahiers du CENTAL, 2.1, UCL Presses Universitaires de Louvain. Volume 1. Pages 357-366).

Noms d'objets imprimés : ambiguïté lexicale sémantique et proxémie

Évelyne Jacquey (1)
evelyne.jacquey@atilf.fr
Christiane Jadelot (1)
christiane.jadelot@atilf.fr

(1) ATILF Nancy Université & CNRS

Mots-clés : sémantique lexicale, ambiguïté lexicale sémantique, métonymie, proxémie, dictionnaire

Keywords: lexical semantics, lexical semantic ambiguity, metonymy, proxemy, dictionary

Résumé : Cet article étudie un genre particulier d'ambiguïté lexicale sémantique, la « polysémie logique » dans [Pustejovsky, 1995], sous-genre des métonymies, par exemple l'ambiguïté « contenant / contenu ». L'étude de ce type d'ambiguïté se concentre sur la classe des noms d'objets imprimés, par exemple le substantif *livre*. Dans cette perspective, nous utilisons le logiciel PROX qui représente les lexèmes du TLFi par un graphe « petit monde » dans lequel les lexèmes entretiennent des relations de proxémie, autrement dit de proximité supposée sémantique. Au travers de cette étude, nous avons donc aussi réalisé une évaluation qualitative des relations de proxémie obtenues, tant du point de vue lexicographique que de celui du comportement de ces noms en corpus.

Abstract : This article is concerned by a special kind of lexical semantic ambiguity, « logical polysemy » as introduced in [Pustejovsky, 1995], special kind of metonymy, for instance the « container / content » ambiguity. Our study is centered on the semantic class of printed object nouns, for instance the noun *book*. We use the PROX NLP system which gives a graphic representation of the whole set of lexemes in the TLFi dictionary. In the resulting graph, lexemes are bound by proxemy relations, that some kind of semantic proximity relations. In this study, we also evaluate in a qualitative way the proxemy relations between lexemes of the TLFi, from a lexicographic point of view as well as from the behavior in corpus.

Introduction

Cet article relate une recherche menée en collaboration entre l'étude des ambiguïtés lexicales sémantiques en français et une introspection sur la pratique des dictionnaires. En effet,

comme le souligne [Pustejovsky, 1995], l'ambiguïté lexicale sémantique reste un écueil pour la constitution de dictionnaires et de lexiques, que ceux-ci soient destinés à être exploités dans le domaine du Traitement Automatique des Langues (TAL) ou par des utilisateurs humains (locuteurs natifs, apprenants du français, chercheurs, etc). En effet, rendre compte de l'ambiguïté lexicale sémantique suppose de rendre compatibles deux directions a priori contradictoires que sont le principe d'économie et le principe d'exhaustivité. Cette difficulté provient du fait que l'on cherche à associer au lexème tous les emplois attestés dans lesquels il entre. Dans des travaux antérieurs [Jacquey, 2005], nous avons pu étayer l'hypothèse qu'un type particulier d'ambiguïté lexicale sémantique, la « polysémie logique »[1] ou « ambiguïté lexicale sémantique systématique » (ALSS) a de bonnes chances de permettre de faire collaborer le principe d'économie et le principe d'exhaustivité. A juste titre, la question s'est posée pour nous d'utiliser ces travaux antérieurs comme guide pour la constitution ou la restructuration de ressources lexicales. Cependant, si la pertinence de l'hypothèse de la structuration de l'ambiguïté lexicale sémantique autour des ALLS a été prouvée théoriquement, il reste nécessaire d'appréhender le phénomène plus spécifiquement en français dans la mesure où la plupart des travaux réalisés sur le sujet l'ont été sur l'anglais [Copestake & Briscoe, 1995], [Cruse, 1995]. Dans cette perspective, notre participation à l'ACI DiLan – Prox[2] a permis de constituer une liste de lexèmes centrée sur une classe particulière dans le type d'ambiguïté lexicale sémantique que nous cherchions à caractériser en français : les substantifs désignant des objets imprimés dont le représentant prototypique est *livre* en français et *book* en anglais[3].
A partir de la classe ainsi constituée, nous avons procédé d'une part à son analyse lexicographique, d'autre part à son analyse en corpus. L'enjeu de l'analyse lexicographique est de qualifier linguistiquement la proxémie. L'enjeu de l'analyse en corpus est de qualifier le type d'ambiguïté lexicale sémantique dont relèveraient les noms d'objets imprimés en français.

1. Les ambiguïtés lexicales sémantiques systématiques (ALSS) et la proxémie

1.1 Les noms d'objets imprimés comme classe prototypique des ALSS

Les ALSS sont connues sous d'autres noms dans la littérature : la « polysémie logique » selon [Pustejovsky, 1995, 1996], [Asher & Pustejovsky, 2000], [Pinkal & Kohlhase, 2000], la classe des « mots à facettes sémantiques » selon [Cruse, 1986, 1995], [Copestake & Briscoe,

[1]Le terme de polysémie logique est dû à [Pustejovsky, 1995] et désigne une sorte particulière de polysémie. Les lexèmes concernés comportent au moins deux sens identifiables, en cela, ils relèvent de la polysémie, mais ces sens ne sont pas forcément mutuellement exclusifs, et même lorsqu'ils le sont au sein d'un même énoncé, tous les sens de ce type de lexèmes restent accessibles. C'est pour cette raison, dans le principe, que Pustejvosky les a ainsi dénomés.

[2] Dans le cadre de cette ACI, l'ATILF a participé à la création et à la mise au point du logiciel PROX, un outil de calcul de proxémie. Cette ACI s'est poursuivie pendant deux ans entre 2004 et 2006.

[3] Le choix du nom *livre* est motivé par plusieurs raisons. Premièrement, le comportement de ce nom est très bien connu dans la mesure où il a largement été étudié et cela depuis de nombreuses années. Deuxièmement, dans la mesure où cet article a pour l'un de ses objectifs d'évaluer en quoi les relations calculées par PROX sont pertinentes du point de vue linguistique et lexicographique, il paraissait judicieux de s'appuyer sur un ensemble de lexèmes dont les propriétés étaient bien établies. Enfin, dans [Jacquey, 2005], nous avons pu établir, uniquement en nous appuyant sur le métalangage lexicographique du TLFi, que les lexèmes relevant de la polysémie logique, ALSS, concernaient environ 12% de l'ensemble des définitions des 54000 substantifs décrits dans ce dictionnaire.

1995] et [Kleiber, 1999] et de la classe des noms multitypés selon [Godard & Jayez, 1996]. Quelle que soit l'étiquette qu'on leur associe, les auteurs précités s'accordent sur le fait que les lexèmes étudiés, représentés par des noms comme *livre*, *ville*, *reproduction* ou *description*, partagent trois propriétés caractérisantes : la distinction entre les sens possibles de ces lexèmes, leur non exclusion mutuelle et l'influence de ces sens sur l'interprétation de la quantification lorsque celle-ci concerne de tels lexèmes. Dans la suite, nous nous concentrons sur le nom *livre* qui tiendra lieu de représentant prototypique à la fois pour les noms d'objets imprimés et pour les lexèmes relevant des ALSS.

Propriété de distinction : le nom *livre* est considéré comme ambigu entre deux sens au moins. Ce nom désigne un objet physique (1.a) ou un objet informationnel (1.b).

(1) a. Pierre a volé mon livre$_{[phys]}$.
b. Pierre a compris mon livre$_{[info]}$.

Propriété de non exclusion mutuelle : cette propriété concerne la coopération possible des différents sens de ces noms par le biais d'un phénomène de « coprédication ». Ce phénomène peut être vu comme l'association, dans une même phrase ou sur un même syntagme au travers de plusieurs phrases, de contextes différents reposant sur des prédicats qui sélectionnent chacun un sens particulier dans l'ensemble des sens possibles des lexèmes ambigus. Les exemples suivants montrent le caractère scalaire de cette propriété et les différentes configurations dans lesquelles elle apparaît. L'acceptabilité varie depuis un niveau quasiment nul avec le nom *plateau* (2), qui relève donc de la polysémie en général, jusqu'à une acceptabilité quasi systématique avec *livre* (3).

(2) a. ? ? Ce plateau$_{[ustensile\text{-}ménager]}$ est lourd. Il$_{[paysage]}$ est couvert de forêts.
b. ? ? Ce plateau, qui est lourd, est très peu peuplé.
c. ? ? Ce plateau, qui est très peu peuplé, est lourd.

(3) a. Ce livre est très lourd mais passionnant.
b. Ce livre est passionnant mais très lourd.

Interprétation variable de la quantification dans des phrases comportant ces noms particuliers sous le champ d'une quantification : (1) la quantification ne sélectionne pas toujours le même sens, et (2), le nombre d'entités désignées peut différer selon le sens qui est sous le champ de la quantification.

(4) a. Marie a déjà emballé tous les livres de cette étagère.
b. Marie a déjà traduit tous les livres de cette étagère.
c. Marie a déjà lu tous les livres de cette étagère.

Dans cet exemple, le groupe nominal *tous les livres de cette étagère* devrait normalement référer à l'ensemble des livres figurant sur l'étagère, chaque livre étant pris comme un objet matériel. Or, comme le montrent les phrases en (4), selon le type d'entités attendues par le verbe principal, cette interprétation varie. La phrase (4a) sera considérée comme vraie si plus aucun objet ressemblant physiquement à un livre ne se trouve sur l'étagère. A l'inverse, le nombre d'entités ressemblant physiquement à un livre n'est pas déterminant dans l'interprétation de la phrase (4b). Ce qui conditionne la valeur de vérité de cette phrase, c'est le fait que tous les livres, pris sous l'angle informationnel, aient été traduits, c'est-à-dire que toutes les oeuvres dont un exemplaire se trouvait sur l'étagère aient été traduites au moment de l'énonciation de cette phrase. Enfin, l'interprétation de la phrase (4c) est conditionnée par les deux aspects simultanément.

Comme cela a été souligné dans l'introduction, les ALSS, et en particulier le nom *livre*, ont largement été étudiés pour l'anglais. Dans la mesure où cette classe d'ambiguïté pourrait permettre une représentation exhaustive et économique de l'ambiguïté lexicale sémantique, il semble intéressant d'étudier le phénomène sur le français. Dans des travaux récents, il nous a été possible d'étudier ce phénomène pour les noms d'action ambigus entre processus et artefact directement à partir du TLFi. En effet, les définitions de ce dictionnaire comportent

suffisamment d'indices réguliers pour qu'il soit possible d'établir une nomenclature puis d'étudier les lexèmes retenus. Pour la classe des noms d'objets imprimés en revanche, de tels indices n'ont pu être identifiés. Face à cette difficulté, l'un des enjeux de l'étude présentée dans cet article est d'évaluer dans quelle mesure les relations de proxémie calculées par le logiciel PROX (voir ci-dessous) permettent de constituer une nommenclature satisfaisante pour pouvoir étudier les ALSS en français sur le cas particulier, et prototypique d'après Pustejovsky, des noms d'objets imprimés.
Avant de poursuivre, précisons que le logiciel PROX a été privilégié pour plusieurs raisons par rapport à l'exploitation d'atlas linguistiques ou de bases de synonymes. Premièrement, les relations de proxémie que PROX permet d'obtenir sont plus diversifiées que les relations de synonymie. Deuxièmement, PROX utilise une information lexicographique en entrée plus structurée et plus riche que celle qui a été utilisée par exemple pour l'établissement de la base DICOSYN (transformation en cliques des relations synonymie fusionnées à partir de cinq dictionnaires et visibles dans leur état initial dans les ressources informatisées de l'ATILF). Enfin, les deux atlas linguistiques consultés au laboratoire ne donnent pas des informations exploitables dans le cadre de nos recherches (*ALG : Atlas linguistique de la Gascogne*, CNRS, 1954 ; *ALMC : Atlas linguistique et éthnographique du Massif Central*, CNRS, 1961) : tous deux donnent les variantes de la forme au pluriel « les livres », le dernier donne des variantes telles que *libre*.

1.2 Proxémie et sémantique lexicale

La notion de proxémie a été introduite par Bruno Gaume, chercheur au C.L.L.E.-E.R.S.S. [Gaume, 2004]. Ce terme s'appuie sur le nom de la procédure, PROX, qui est un modèle mathématico-computationnel particulièrement bien adapté selon l'auteur à l'exploitation de « graphes petits mondes » d'après son concepteur – cette sorte de graphes a des propriétés particulières différentes des grands graphes (aléatoires ou réguliers), selon [Gaume, 2006], ils ont une connectivité forte (il existe beaucoup de chemins courts entre deux sommets) et une tendance forte à l'agrégation (un taux de clustering fort : « p.2 *plus le graphe a tendance à posséder des agrégats denses en arêtes, plus le taux de clustering du graphe est proche de 1* ») . Dans le domaine de la lexicographie et de la sémantique lexicale, les relations établies par le logiciel PROX peuvent être considérées comme proches de l'analogie. De plus, [Gaume, 2004] a montré que ce type de graphes était particulièrement bien adapté pour avoir une représentation générale des liens qui pouvaient exister entre les lexèmes d'un dictionnaire par le truchement de leurs définitions. La méthode choisie pour extraire le graphe du dictionnaire consiste à prendre pour sommets du graphe les entrées du dictionnaire et d'admettre l'existence d'un arc d'un sommet A vers un sommet B si et seulement si l'entrée B apparaît dans la définition de l'entrée A. Dans le cadre de l'ACI DiLan – Prox, le logiciel PROX a été appliqué au Trésor de la Langue Française informatisé (TLFi) dans sa version XML. Cette version du dictionnaire étant par ailleurs étiquetée en morpho-syntaxe par l'étiqueteur Maucourt-Papin[4], les étiquettes catégorielles sont accessibles. Ainsi trois graphes ont pu être réalisés : un graphe ne contenant que des sommets correspondant à des substantifs, un graphe ne contenant que des sommes correspondant à des verbes et un graphe croisé substantif – verbe. Actuellement, les résultats des calculs de proxémie à partir du TLFi sont accessibles sur le site du CNRTL (www.cnrtl.fr, onglet portail lexical puis proxémie), centre de ressources adossé au laboratoire ATILF.
A partir des résultats fournis par PROX et accessibles sur le site du CNRTL, nous avons sélectionné les graphes de substantifs uniquement (puisque nous étudions une classe de

[4]La version XML et la version XML catégorisée du TLFi ont été réalisés par les soins de Jacques Dendien, IR CNRS à l'ATILF.

noms). Nous avons demandé une liste de proxèmes du nom *livre* puisque ce nom est considéré comme représentant prototypique de la classe des noms d'objets imprimés et des ALSS. Dans la liste fournie, les 50 premiers proxèmes ont été sélectionnés pour l'analyse lexicographique et les 10 premiers pour l'analyse en corpus.

2. Données et analyses

2.1 Analyse lexicographique des proxèmes de *livre*

Le logiciel Prox fournit la liste de 50 premiers mots suivants lorsque les proxèmes nominaux de *livre* sont recherchés par l'utilisateur : *agenda, atlas, barbouillage, bouquin, brochure, brouillard, cahier, calepin, carnet , catalogue, collection, elzévir, fascicule, grimoire, herbier, keepsake, libellé, liste, livraison, livre, livret, matricule, mémento, mémoire, oeuvre, ~~once~~, opuscule, ouvrage, partie, plaquette, portulan, ~~pound~~, publication, recueil, registre, revue, répertoire, rôle, tome, travail, volume, écrit, édition*[5]. Du point de vue linguistique, les relations de proxémie obtenues par le logiciel Prox sont supposées représenter des proximités sémantiques. Autrement dit, l'ensemble des proxèmes d'un mot donné devrait pouvoir être interprété comme une bonne approximation de la ou les classes sémantiques de ce mot. L'analyse lexicographique présentée dans cette section vise à apporter des éléments de réponse à cette hypothèse.

Dans l'ensemble des proxèmes de *livre*, en nous appuyant sur la préface du tome 1 du *TLF*, nous avons supposé que *livre* avait un statut de « classificateur ». Pour le vérifier, nous avons recherché dans le *TLFi* les définitions contenant le nom *livre* en position initiale, comme par exemple celle du nom *carnet* reproduite ici : ***A.—*** *Petit* ***livre*** *ou registre de poche où l'on inscrit des comptes ou des notes.* Le bilan de cette première recherche est assez décevant puisque très peu d'entrées de la liste (environ 7/50) possèdent des définitions qui reprennent le mot *livre*. Nous nous sommes ensuite demandés quels liens sémantiques reliaient les proxèmes proposés par PROX. Afin de répondre à cette question, nous avons consulté les articles concernant les 50 proxèmes et sélectionné les définitions qui, selon nous, ont un lien sémantique (possiblement indirect) avec *livre*. La deuxième liste de définitions ainsi obtenue fait apparaître des liens directs ou indirects avec *livre* lorsque *livre* apparaît sous forme lexicale. Ainsi, on trouve par exemple *agenda est un petit carnet* et *carnet est un petit livre*. De plus, *livre* peut être présent dans les définitions des proxèmes par le biais de traits sémantiques repérables dans les définitions du substantif *livre*. En effet, la définition qui domine l'ensemble de l'article de *livre* dans le TLFi peut être segmentée en deux parties : *Assemblage de feuilles en nombre plus ou moins élevé* et *portant des signes destinés à être lus*. Chacune de ses parties peuvent être associées respectivement au trait [+physique] pour la première, qui correspond à la description de l'objet, et au trait [+information] pour la seconde, qui s'apparente au contenu informatif du livre et à sa finalité typique, et enfin [+physique_informationnel] pour l'ensemble. A partir de cette analyse en traits, outre les liens directs et indirects constatés entre *livre* et ses proxèmes, nous avons pu qualifier la valeur de ceux-ci en partant de leurs définitions. A titre d'illustration, citons deux exemples : *agenda* dont le début de définition *petit carnet* incite à la qualifier par le trait [+physique_informationnel], ou encore *collection* dont le début de définition *recueil de textes*

[5]Les noms *once* et *pound* sont barrés car ils concernent la monnaie et non l'objet imprimé. Ils sont retournés par PROX car, pour le moment, ce logiciel opère l'union de toutes les définitions d'une même chaîne de caractères, que le mot correspondant ait fait l'objet d'une ou plusieurs entrées dans le dictionnaire. Ce point est en cours d'amélioration.

incite à choisir plutôt le trait [+information][6]. Sur les 50 proxèmes, tous sont qualifiables par l'un au moins des trois traits sémantiques issus de la définition principal de *livre*. Le calcul de proxémie fournit donc une bonne approximation de la classe sémantique des noms d'objets imprimés : *livre* pouvant être caractérisé par le trait [+physique_informationnel] et ce trait pouvant se subdiviser en [+phys], [+info] ou [+phys_info] d'après [Pustejovsky, 1995]. Afin de procéder à une observation en corpus les dix premiers proxèmes de *livre* ont été extraits de la liste initiale analysée du point de vue lexicographique. Chacun d'eux, *agenda, album, brochure, cahier, carnet, fascicule, livret, publication, recueil, registre*, a été étiqueté selon la distinction [+physique], [+informationnel] ou [+physique_informationnel], en fonction de sa définition la plus proche de *livre*.

Agenda	Album	Brochure	Cahier	Carnet	Fascicule	Livret	Publication	Recueil	Registre	Livre
[+phys/+phys_info]	[+phys_info]	[+phys_info]	[+phys]	[+phys/+phys_info]	[+phys]	[+phys/+phys_info]	[+processus/+phys_info]	[+phys_info]	[+phys_info/+phys]	[+phys_info]

Comme le montre ce tableau, la plupart des proxèmes possèdent le trait [+phys_info]. Deux proxèmes comportent uniquement le trait [+phys], il s'agit de *cahier* et *fascicule*. Trois autres comportent les traits [+phys_info] et [+phys], *agenda, carnet* et *livret*. Enfin, le proxème *publication* se distingue car il comporte le trait supplémentaire [+processus] (dans la définition de *publication*, on trouve *action de publier...*).

2.2 Dix proxèmes de *livre* en corpus

2.2.1 Le corpus

Munis de la liste des dix proxèmes de *livre* que nous avons choisi d'analyser, la base FRANTEXT catégorisée (www.atilf.fr/frantext.htm) nous a permis de constituer notre corpus de travail et d'en extraire les occurrences de ces dix noms d'objets imprimés. Le corpus de travail choisi ne comporte aucune restriction en genre. En revanche, il se limite aux ouvrages postérieurs à 1950. Une fois cette sélection faite, le corpus de travail utilisé compte 487 ouvrages et contient près de 3.400.000 mots.

Le nombre d'occurrences trouvées pour chacun des dix proxèmes en plus de *livre* est variable, de 70 pour *fascicule* à 7380 pour *livre*. Afin d'homogénéiser le nombre d'occurrences de chaque proxème, nous avons sélectionné au plus un tiers des occurrences trouvées avec un plancher de 100 occurrences minimum sauf lorsque cela était impossible (cas de *fascicule*). Les occurrences sélectionnées ont été extraites dans l'ensemble trouvé et trié par ordre chronologique décroissant. L'ensemble des occurrences ainsi extraites constitue un corpus d'environ 3500 extraits de FRANTEXT, soit un corpus d'environ 280.000 mots.

2.2.2 Analyse distributionnelle et recoupement avec l'information sémantique lexicale

Sur le corpus restreint de 280.000 mots, nous avons relevé les prédicats verbaux avec lesquels les proxèmes donnés pour *livre* était cooccurrents et où le proxème apparaissait dans l'une des positions grammaticales, sujet, objet ou objet indirect. Parmi ces relevés, quelques verbes seulement se démarquent par leur fréquence et leur répartition homogène pour l'ensemble des dix proxèmes étudiés en détail : *consulter, écrire, feuilleter, inscrire_sur/dans, lire, ouvrir, ranger, ranger_dans, refermer, prendre, sortir, trouver, paraître*. L'observation en corpus nous conduit à deux premières observations :

[6] Les définitions pertinentes de *agenda* et *collection* sont les suivantes respectivement : « *agenda :*

1. Bien que chaque verbe opère effectivement une sélection dans le contenu sémantique du proxème coocurrent, par exemple *consulter* sélectionne préférentiellement la facette [+info], les extraits montrent que, dans presque la moitié des cas, le contenu sémantique plus large du proxème [+phys_info] reste accessible. Sur les 191 extraits de FRANTEXT étiquetés, 89 sont globalement ambigus [+phys_info] (5), 71 sont globalement physique et seulement 26, globablement informationnel (6).

(5)*En se promenant dans le monastère, Simon feuilleta[+phys] son **agenda** où il avait jeté d'avance et à la hâte quelques idées[+phys_info] destinées aux trois articles qu'il avait l'intention de rapporter de Moscou*. FRANTEXT, ORMESSON.J D' /LE BONHEUR A SAN MINIATO/1987 Page 35 / I Le temps des épreuves

(6)[...] *tandis qu' en 1956, la fédération mondiale des travailleurs scientifiques, favorable aux thèses soviétiques, publie[+info] une **brochure** sur le danger des explosions expérimentales[+info]*. FRANTEXT, GOLDSCHMIDT.B /L'AVENTURE ATOMIQUE/1962 Pages 190-191 / 7 ARMES ET DESARMEMENT ATOMIQUES

2. Alors que la classe des noms d'objets imprimés telle qu'elle est constituée par le calcul de proxémie donne une impression d'homogénéité (presque tous les proxèmes comportent le trait [+phys_info]), la cooccurrence avec les verbes les plus fréquents pour dix d'entre eux semble moins homogène. Par exemple, le verbe *paraître* semble très sélectif sur le plan informationnel (on ne le trouve qu'avec les proxèmes *brochure, fascicule, livre*) et le verbe *ranger_dans* lui aussi sur le plan physique (on ne le trouve qu'avec *album, carnet, registre, livre*). Cette observation entre en résonance avec l'analyse fondée sur la notion de classes d'objets des objets, directs et indirects, du verbe *lire* [Le Pesant, 1994]. En effet, l'auteur établit un premier niveau de classification des compléments nominaux de ce verbe entre les objets relevant de la classe <è->, « les noms désignant de l'écriture » (*adresse, conte, décret, idéogramme, quatrain*) et ceux relevant de la classe <s-> « les noms désignant des supports d'écriture » (*ardoire, carnet, journal, pannonceau*). En tant que classes d'objets, ces classes sont définies par leurs verbes cooccurrents dans un ensemble défini de construction syntaxique. Ainsi, les noms désignant de l'écriture sont objet direct de verbes comme *lire, décrypter, déchiffrer, relire*, alors que les supports d'écriture sont introduits par les prépositions *dans, sur* avec les mêmes verbes. De plus, une partie des noms de supports d'écriture peuvent apparaître soit en position d'objet direct, soit introduits par *dans* ou *sur*. Les proxèmes de *livre* entre dans cette dernière catégorie. Les restrictions que nous avons observées dans notre corpus d'étude peuvent s'expliquer dans les termes de Le Pesant : les sujets nominaux de *paraître* pourraient être assimilés à des supports de publication (*bande dessinée, brochure, revue, magazine, plaquette, journal, incunable...*) et les objets de *ranger_dans* pourraient être assimilés à des supports d'écriture à plusieurs pages (*cahier, carnet, registre*,...).

Conclusion

En conclusion, cette étude a permis de montrer deux aspects importants pour l'observation et la modélisation des alternances lexicales sémantiques systématiques dont *livre* est un représentant prototypique en français comme en anglais. Premièrement, les relations de proxémie, même si elles n'ont pas encore été étudiées du point de vue précis des relations de sémantique lexicale (synonymie, hyponymie, etc), donnent néanmoins un point de départ intéressant et pertinent pour l'étude d'une classe sémantique du point de vue lexicographique et linguistique sur corpus. Au sens de [Le Pesant, 1994], la classe ainsi atteinte pourrait être

assimilée à celle des supports d'écriture. Dans le cas des noms d'objets imprimés, étudiés ici, la proxémie permet effectivement d'atteindre une classe sémantique cohérente du point de vue des traits sémantiques qui émergent des définitions des lexèmes ainsi réunis. Sur le plan lexicographique, la recherche présentée dans cet article a permis d'établir un parallèle entre les résultats de PROX et une cartographie supposée du/des rédacteur(s) liée à une connaissance implicite des liens sémantiques entre les mots. Concrètement, il s'agissait pour nous de voir comment ces liens avaient pu être traduits au niveau des définitions, si le lien se faisait directement ou s'il passait par une ou deux étapes, voire plus, et quelles acceptions de « livre » étaient mises en jeu dans la construction de ce réseau. En d'autres termes nous avons été amenés à une lecture plus sélective des articles du TLF, puisqu'elle visait à sélectionner les données lexicographiques dans le but d'établir des liens qui n'apparaissaient pas en surface et de construire un corpus à partir de ces données. Ainsi, PROX ouvre la perspective d'un outil permettant la construction d'une nomenclature de dictionnaire non plus linéaire mais organisée par champ sémantique, autrement dit à avoir une démarche à mi-chemin entre sémasiologie (puisqu'on part d'une liste de formes) et onomasiologie (puisqu'on part du sens). Sur le plan linguistique, cette étude, démarrée grâce à la proxémie, poursuivie grâce à la base FRANTEXT et à un repérage manuel dans les extraits, a permis d'observer *in situ* la propriété principale des ALSS, c'est-à-dire la coprédication. Faute de place, nous n'avons reproduit aucun autre extrait que celui en (5). Cela ouvre des perspectives intéressantes pour l'étude de ces ambiguïtés lexicales particulières. En effet, on peut faire l'hypothèse qu'à partir d'une liste exhaustive des lexèmes étudiés dans la littérature consacrée à ce type d'ambiguïté, il sera possible, en utilisant le logiciel PROX, d'étudier en corpus les classes sémantiques auxquelles ces lexèmes particuliers appartiennent. Ainsi, il serait possible d'établir une analyse du phénomène en français, en extension et en intension.

Bibliographie

[Asher & Pustejovsky, 2000] Asher, N. et Pustejovsky, J. (2000) « The metaphysics of words in context » à paraître dans *Journal of Logic, Language and Information*, sur le site http://www.dla.utexas.edu/depts/philosophy/faculty/asher/papers/MWC.ps

[Jacquey, 2005] Jacquey, E. (2005) : « Ambiguïté lexicale et modélisation unifiée de la polysémie logique » dans *Interpréter en contexte* sous la direction de Corblin, F. et Gardent, C. Hermès Sciences.

[Copestake & Briscoe, 1995] Copestake, A. et Briscoe, T. (1995) « Semi-productive polysemy and sense extension » dans *Journal of Semantics*, 12 : 15,67.

[Cruse, 1986] Cruse, D.A. (1986) *Lexical Semantics*. Cambridge : Cambridge University Press.

[Cruse, 1995] Cruse, D.A. (1995) "Polysemy and related phenomena from a cognitive linguistic point of view" dans *Computational Lexical Semantics* (eds. Saint-Dizier, P. et Viegas, E.).

[Gaume, 2004] Gaume, B. (2004) : « Balades aléatoires dans les Petits Mondes Lexicaux » dans *I3 Information, Interaction Intelligence, vol4 n°2*. CEPADUES édition (Computer sciences)

[Gaume, 2006] Gaume, B. (2006) : « Cartographier la forme du sens dans les petits mondes lexicaux », *JADT 2006*, p. 541 – 465

[Godard & Jayez, 1996] Godard, D. et Jayez, J. (1996) « Types nominaux et anaphores ; le cas des objets et des événements » *dans Anaphores temporelles et (in-)cohérences, Cahiers Chronos*.

[Kleiber, 1999] Kleiber, G. (1999) : *Problèmes de sémantique : la polysémie en questions*, Sens et structures, Villeneuve d'Ascq : Presses universitaires du Spetentrion.

[Le Pesant, 1994] Le Pesant, D. (1994) « Les compléments nominaux du verbe *lire*, une illustration de la notion de 'classe d'objets' », Revue *Langages, n°115*, p. 31-46.

[Pinkal & Kohlhase, 2000] Pinkal, M. et Kohlhase M. (2000) "Feature logic for dotted types : A formalism for complex words meanings" dans *ACL*.

[Pustejovsky, 1995] Pustejovsky, J. (1995) : *The Generative Lexicon*.

[Pustejovsky 1996] Pustejovsky, J. (1996) « The semantics of complex types » dans *Langue française*.

"Caught in the Web of Words" : la lexicographie et la Toile

Russon Wooldridge (1)
wulfric@chass.utoronto.ca

(1) ATILF Nancy Université & CNRS

Introduction

Je voudrais parler de la macrolexicographie et de la microlexicographie et de certains aspects de chacune. La macrolexicographie correspond à ce que M. Quemada a appelé la dictionnairique, c'est-à-dire la description organisée, et diffusée, de tout le lexique d'une langue. La microlexicographie est ce qu'on appelle souvent la lexicologie, c'est-à-dire l'étude et la description d'un secteur du lexique. Je vais parler d'abord de la macrolexicographie sous deux angles : statique et dynamique.

1. La macrolexicographie statique

Il y a un peu plus de cinquante ans, on a pu assister à Strasbourg au mariage de la lexicographie, de la philologie, de la linguistique et de l'informatique qui a donné lieu à la dernière belle floraison de la lexicographie traditionnelle à travers la continuation enrichie des dictionnaires Robert, la parution de plusieurs dictionnaires de genre nouveau chez Larousse et la pièce maîtresse qu'est le *Trésor de la langue française*.

De par ses dimensions, un volume du *TLF* rappelle les grandes bibles in-folio et ce n'est pas un hasard si le colloque de Wolfenbüttel sur la lexicographie française du XVIe au XVIIIe siècle s'est tenu dans la salle des bibles de la Herzog August Bibliothek. Cependant, dans le cas du *TLF*, il faudrait aligner seize pupitres pour pouvoir le consulter convenablement et on peut se demander qui, en dehors des bibliophiles et des philologues, consulte encore le *TLF* papier.

C'est que, heureusement, il y a le *TLFI*, la version du *TLF* intelligemment informatisé par l'équipe de Jacques Dendien, édition accessible gratuitement en ligne et c'est bien entendu cette version qui est consultée par les étudiants, professeurs et chercheurs au Canada et ailleurs dans le monde. Une autre belle initiative, qui contribue à alléger la sacoche de l'élève et de l'étudiant, est la mise en ligne du *Petit Robert sur CD-ROM*. Tout membre de l'Université de Toronto, par exemple, peut consulter gratuitement, soit à la maison, soit à la bibliothèque, le *PR sur CD-ROM* accessible par le biais du World Wide Web.

La différence essentielle entre le dictionnaire papier et le dictionnaire électronique, c'est que dans le premier on a accès aux informations à travers les mots-vedettes du lexicographe alors que l'accès aux informations dans le second se fait, soit en partant des mots-vedettes, soit à travers les mots-clés choisis par le consulteur. Dans les deux cas, il s'agit toujours de la lexicographie statique dans le sens que le contenu du dictionnaire ne bouge pas d'une édition à l'autre.

2. La macrolexicographie dynamique

La macrolexicographie dynamique n'est possible qu'à travers un réseau électronique tel qu'Internet et y est devenue réalité grâce à ce qu'on appelle le Web 2.0, c'est-à-dire le Web interactif, et ce grâce au wiki. Un wiki est un "Site web dynamique dont les pages peuvent être modifiées par tout visiteur. Un wiki permet non seulement de communiquer et diffuser rapidement des informations, mais aussi de les structurer pour permettre de naviguer aisément. Wikipédia est l'exemple le plus célèbre d'un wiki." (article du *Wiktionnaire*)

Wikipedia, toutes langues confondues, serait, d'après les statistiques du 3 octobre 2007 de la page "Top 500" d'Alexa, compagnie d'informations sur le Web, le neuvième site web le plus consulté dans le monde, derrière des sites comme Yahoo, Google, YouTube ou Facebook. Pour la France, Wikipédia, logiquement surtout dans sa version française, serait au rang du treizième site le plus consulté.

Les critiques adressées à l'endroit de Wikipédia ont été nombreuses, surtout au sujet de l'anonymat des articles. Cependant, une étude sur échantillonnage menée par la prestigieuse revue scientifique *Nature* en décembre 2005 conclut que la qualité des articles de Wikipédia est à peu près la même que celle des articles de l'*Encyclopaedia Britannica*.

Wikipédia, en principe une encyclopédie mais en réalité un dictionnaire encyclopédique, a été lancé, dans sa version française, en mars 2001. À la date du 3 octobre 2007, en fin d'après-midi, la version française contenait 565 248 articles (contre 2 034 274 pour la version anglaise). On peut suivre, du moins depuis un an et demi environ, sur le site de Wikipédia l'évolution de l'article consacré à un terme donné, en comparant les différentes versions et en lisant le détail des modifications apportées à tel ou tel état de l'article. Pour illustrer mon propos, je prendrai l'exemple du mot *pipolisation*/*peopolisation*.

En juin 2006, l'essentiel de l'article était comme suit :

> La *peopleisation* (souvent transcrite **pipolisation** par les médias) est un néologisme péjoratif. Il peut se mettre en relation avec l'expression « presse people ». Il désigne l'introduction progressive de la vie privée dans la vie publique. [...] Voir aussi *starisation*.

Le 5 septembre 2006, l'article commençait de la façon suivante :

> **Peoplisation** est un néologisme français dérivé de l'anglicisme « people » (désignant les « gens » célèbres). On peut le trouver écrit sous les formes « peopleisation » ou « pipolisation ».
>
> Le terme est apparu dans les années 2000 avec le développement de la presse people en France. Jusqu'alors, la notion de « people » ne s'appliquait qu'aux personnalités du show business et aux médias spécialisés traitant de l'actualité de celles-ci. Il s'inscrit dans la continuité du phénomène de « starisation » constaté dans les années quatre-vingt.

Le 3 octobre 2007 enfin, l'article commençait comme suit :

> **Peoplisation** est un néologisme français dérivé de « people » (désignant les « gens » célèbres). On peut le trouver écrit « peopleisation », « pipeulisation », « pipolisation »

> et sous sa forme la plus courante « peopolisation ».
>
> Le terme est apparu en France dans les années 2000 avec le développement de la presse people. Jusqu'alors, le terme de « people » ne s'appliquait qu'aux personnalités du show business et aux médias spécialisés traitant de l'actualité de celles-ci. Il est désormais utilisé pour décrire aux moins deux phénomènes, l'un concernant la politique, l'autre les médias.

Wikipédia donne aussi, entre autres, des statistiques, tirées de Google, sur l'emploi quantitatif des variantes graphiques du mot *pipolistion*/*peopolisation*. Je reviendrai là-dessus en parlant de la microlexicographie.

Mais avant de quitter le monde wiki, je voudrais dire un mot sur une application dérivée, le Wiktionnaire. Le Wiktionnaire, comme son nom le suggère, est un dictionnaire et non une encyclopédie, "un dictionnaire libre et gratuit que chacun peut améliorer" (page d'accueil du Wiktionaire). Tout comme Wikipédia, le Wiktionnaire est dynamique et permet de retrouver les différents états antérieurs d'un article.

La page "Wiktionnaire : À propos" décrit les objectifs et la méthode de ce dictionnaire :

> Le Wiktionnaire a pour but de devenir un dictionnaire universel libre basé sur des wikis.
>
> En tant que dictionnaire, il vise à donner des renseignements les plus complets et les plus neutres possibles sur tous les mots ou locutions, actuellement ou anciennement utilisés, oralement ou par écrit, dans toutes les langues, vivantes ou mortes. Son objectif est seulement descriptif : il ne s'agit ni de défendre le français ou une autre langue, ni d'être normatif. Il ne juge donc pas la valeur des mots et n'essaie pas de leur donner ou de leur refuser son aval.
>
> Voici ce que vous trouverez dans le Wiktionnaire :
>
> - le ou les types de mot ou de locution
> - ses variations (conjugaisons, accords)
> - la ou les définitions
> - la ou les prononciations
> - la ou les étymologies
> - des traductions (ou équivalents) en d'autres langues
> - les synonymes et les antonymes
> - les mots apparentés
> - les mots et locutions dérivés (dont les expressions)
> - les homophones
> - les paronymes
> - les hyperonymes et les hyponymes
> - les holonymes et les méronymes
>
> Le Wiktionnaire peut en outre donner :
>
> - des exemples et des citations
> - les registres de langue
> - les diverses terminologies spécialisées

- les différents usages selon les régions
- les différentes prononciations selon les régions
- des illustrations
- des enregistrements de prononciations
- des liens pertinents vers l'encyclopédie libre Wikipédia

Ainsi que tous les affixes et caractères utilisés pour écrire les mots.

Pour le mot *pipolisation*, l'article du Wiktionnaire du 3 octobre 2007, en plus des rubriques Étymologie (les mêmes informations que dans Wikipédia), Variantes orthographiques et "Voir aussi" (mots apparentés), donne une présentation dictionnairique d'informations concernant la prononciation, la partie du discours et le sens.

pipolisation /pi.p□li.za.sj□/ *féminin*

1. Médiatisation, voulue ou non, de la vie privée d'une personnalité extérieure au show business. On parle ainsi de la « peopolisation du politique » avec la multiplication dans la presse écrite et les médias d'information en général de sujets mettant en avant la vie privée (famille, amis, vacances...) des responsables politiques
2. Utilisation à des fins médiatiques de l'image de personnalités célèbres par des associations, des entreprises ou des hommes politiques
3. Tendance des médias généralistes à traiter de l'actualité des personnalités du show business et à aborder certains aspects de leur vie privée, au même titre que la presse people.

L'article remonte au 6 septembre 2006, sinon au-delà. Dans l'article du 6 septembre 2006, la partie Définition (la prononciation n'était pas donnée) était la suivante :

La peopolisation peut désigner :

- la médiatisation, voulue ou non, de la vie privée d'une personnalité exterieure au show business. On parle ainsi de la « peopolisation du politique » avec la multiplication dans la presse écrite et les médias d'information en général de sujets mettant en avant la vie privée (famille, amis, vacances...) des responsables politiques ;
- l'utilisation à des fins médiatiques de l'image de personnalités célèbres par des associations, des entreprises ou des hommes politiques. En France, suite au soutien public apporté par les chanteurs Johnny Hallyday et Doc Gyneco, en août 2006, à Nicolas Sarkozy pour l'élection présidentielle de 2007, le quotidien *Libération* décrivait une « peopolisation de la campagne électorale » du président de l'UMP ;
- la tendance des médias généralistes à traiter de l'actualité des personnalités du show business et à aborder certains aspects de leur vie privée, au même titre que la presse people.

3. La macrolexicographie : conclusion

La dictionnairique offre des descriptions du lexique l'une langue. Le lexique et la langue sont dynamiques, pluriels et anonymes. À ce titre, la macrolexicographie dynamique, plurielle et

anonyme, est bien plus proche de la langue que la macrolexicographie traditionnelle, qui, elle, en plus d'être statique et donc obsolescente, est le fait de quelques auteurs-autorités nommés.

4. "Caught in the Web of Words"

En guise d'introduction de ma discussion de la microlexicographie et pour faire le pont entre la macrolexicographie et la microlexicographie, je voudrais commenter la première partie du titre de mon propos. "Caught in the Web of Words" veut dire « Pris dans la toile des mots » et on peut noter qu'au Canada français on dit la Toile pour parler du World Wide Web. C'est notamment le titre d'un livre écrit par Elizabeth Murray sur la vie lexicographique de son grand-père James Murray, éditeur en chef du grand dictionnaire d'Oxford, le *Oxford English Dictionary* ou *OED*.

Les rédacteurs ou scripteurs, puisqu'ils travaillaient dans un Scriptorium, exploraient les réseaux étymologiques, lexicaux et sémantiques des mots qu'ils traitaient. Poursuivant ces pistes, ils rangaient au fur et à mesure les fiches qu'ils rédigeaient à leur place alphabétique pour y revenir plus amplement lorsque ce serait le tour de traiter la lettre en question. Comme des mouches, ils étaient pris dans la toile des mots que leur avait tissée la langue anglaise.

5. La microlexicographie

L'image de la toile appliquée à la langue est une métaphore très forte. Ce n'est pas par hasard que Tim Berners-Lee nomme "World Wide Web" le système qu'il commence à mettre en place vers 1990. Le contenu du Web est virtuel ; il devient une réalité partielle visible, une toile de mots, à travers les mots-clés des requêtes adressées aux moteurs de recherche, aidés par les hyperliens des sites portails et des documents de tout genre. Tout comme la langue, le Web est dynamique, pluriel et anonyme et aux limites floues. Il est énorme, beaucoup plus vaste que tout corpus sciemment constitué à l'aide de financements importants, et il est gratuit. Le Web est représentatif de la langue et offre au curieux, amateur ou spécialiste, un terrain riche pour toutes sortes d'études microlexicographiques. Je vais illustrer par quelques exemples quelques-uns des types d'exploitations que l'on peut faire à partir du Web.

Les chiffres : *pipolisation/peopolisation/peoplisation*

	28 octobre 2006 *	**30 janvier 2007 ****	**4 octobre 2007 *****
peoplisation	804	870	14 400
peopolisation	12 900	55 900	38 500
pipolisation	20 400	14 400	34 200

* Observation de l'auteur de l'article *peoplisation* de Wikipédia.
** Observation faite par RW.
*** Observation faite par RW excluant les pages mentionnant "wikipédia" ou "wiktionnaire".

Les chiffres bruts n'ont de valeur que relative, mais on fait attention à éviter les homonymes fréquents dans ce type de relevé : si on étudie les noms d'animaux, par exemple, on prendra des précautions à l'endroit des *chats* et des *souris* du Web. Le tableau ci-dessus amène au

moins deux remarques : d'abord, il permet d'observer l'hésitation entre deux façons de représenter la prononciation /pip□lizasj□/, l'une diachronique indiquant l'étymologie du mot, l'autre synchronique se conformant aux règles de la langue française ; deuxièmement, on peut se demander pourquoi Wikipédia choisit une forme minoritaire pour la vedette de son article.

La clarification lexicale : *flexicurité* et *flexsécurité*

Dans les premiers mois de l'année 2006, le gouvernement Villepin menait une discussion du CPE, ou Contrat première embauche. Il s'agissait de marier selon le modèle dit danois la flexibilité dite anglo-saxonne et la sécurité française. Pour exprimer ce concept, au début on utilisait surtout le mot *flexicurité*. Cependant, on aurait vu dans ce mot plus de flexibilité que de sécurité. Au mois d'avril 2006, Google montre que la forme *flexsécurité*, plus claire, serait devenue la plus utilisée. (Voir tableau.)

23 janvier 2006		**10 avril 2006**	
flexicurité	636	*flexicurité*	20 200
flex-sécurité	307	*flex-sécurité*	997
flexsécurité	304	*flexsécurité*	27 800
flexisécurité	256	*flexisécurité*	613
flexi-sécurité	183	*flexi-sécurité*	14 500
flexicarité	12	*flexicarité*	11

Chiffres d'après les pages francophones de Google.

Les variantes régionales : *pourriel* et *spam*

Il y a le bon courriel et le mauvais courriel. Ce dernier se nomme *spam* en anglais. Le français a emprunté ce mot, mais les Canadiens français, soucieux d'éviter les anglicismes, ont eu tendance à le remplacer par le mot *pourriel*, le courriel pourri que l'on jette à la poubelle. Cette différence, plus prononcée aujourd'hui qu'il y quelques années, se montre clairement lorsqu'on regarde les sites du domaine français (`.fr`) et du domaine canadien (`.ca`).

	5 avril 2003		**4 octobre 2007**	
	pourriel(s)	spam	pourriel(s)	spam
Canada `.ca`	592	3 990	593 000	124 000
France `.fr`	192	18 500	44 200	2 520 000

Chiffres d'après les pages francophones de Google.

Le Web et le dictionnaire (1) : *douet*

Sillonnant les petites routes de l'Auge en Normandie en 2003, j'ai été intrigué par le panneau "Route des Douets" qui se rencontrait plusieurs fois au bord de la route. Ayant fui un instant la canicule en nous installant sur la terrasse ombragée de l'auberge de Saint-Hymer à côté du ruisseau qui nous rafraîchissait en même temps qu'il alimentait un joli lavoir, j'ai demandé à la

jeune serveuse ce que voulait dire ce mot de *douet*. Elle nous a répondu que c'était le nom du ruisseau.

De retour à la maison, j'ai eu confirmation de la nature régionale du mot en ne le trouvant ni dans le *Petit Robert*. ni dans le *TLF*. Je l'aurais sûrement trouvé à la bibliothèque dans un dictionnaire spécialisé ou un glossaire, mais le Web m'a informé tout de suite et de façon plus complète que n'importe quel dictionnaire imprimé. Je trouve une dizaine de pages pertinentes, plusieurs accompagnées d'images, dont une consacrée à Saint-Hymer qui me dit que "La route des « douets » (ruisseaux en patois Normand) passe par ici et alimente un très ancien lavoir" ; une autre sur Villers-sur-Mer m'informe que "La Route des Douets [...] vous fera connaître les eaux-vives du Pays d'Auge, ces « douets », petits ruisseaux allant se jeter dans la Touques, et quelques rivières telles que l'Yvie et la Calonne" ; une troisième enfin sur la famille Douet me dit que "Le nom de famille Douet est le plus souvent un nom de village, de lieu-dit, surtout dans l'Ouest. Il est issu du latin *ductus* et signifie : un courant d'eau, une source, un lavoir."

Wikipédia, depuis août 2007, contient l'articls suivant : "Un **douet** désigne un ruisseau en augeron, dialecte normand du Pays d'Auge."

Le Web et le dictionnaire (2) : *doudou*

Dans sa chanson *Caroline*, le chanteur québécois Richard Desjardins dit : "Prends ta robe et ton bijou, / dis bye bye à ta doudou." Selon le *Petit Robert* et le *TLF*, une doudou est une femme antillaise chérie. Caroline aurait-elle une femme de chambre? C'est peu probable. Le Web m'apprend dans des dizaines de milliers de pages ce qu'est la doudou canadienne et le doudou français, en en montrant plusieurs exemples différents. Le Wiktionnaire me dit, entre autres, qu'un "**doudou**, *masculin (féminin au Canada)*, [est un] Objet procurant un réconfort psychologique à un petit enfant (généralement une couverture ou une peluche)."

L'homophonie : *sous les meilleurs auspices/hospices*

Le 6 octobre 2002 sur le site Web de la chaîne de télévision française France 2, on pouvait lire la phrase "Ce résultat laisse donc envisager l'avenir sous les meilleurs hospices.", vite corrigée par la rédaction – suite peut-être à des coups de téléphone ou de courriel de la part de quelques lecteurs – en "Ce résultat laisse donc envisager l'avenir sous les meilleurs auspices."

Une interrogation du Web avec Google faite le 20 octobre 2002 a rapidement confirmé l'hypothèse de la fréquence de la faute :

A. "sous les meilleurs auspices" = 2540 documents = 90,39% de A+B
B. "sous les meilleurs hospices" = 270 documents = 9,61% de A+B

On peut faire le même genre de comparaison quantitative pour des paires comme "exaucer/exhausser un voeu", "nu comme un ver/verre/vers/vert" ou, dans le domaine des paronymes, "dans la conjoncture/conjecture actuelle" ou "se perdre en conjectures/conjonctures", par exemple.

Le lexique du passé : *élève* "élevage"

En novembre 2003, rencontrant les expressions "l'élève du poisson" et "l'élève des poissons", dans une *Analyse du rapport fait à l'académie des sciences par M. Coste, le 7 Février 1853,*

sur l'élève et la multiplication du poisson mise en version numérique dans la Bibliothèque électronique de Lisieux, j'ai consulté d'abord le *Trésor de la langue française informatisé*, qui dit que c'est un substantif féminin synonyme vieilli d'*élevage* attesté dans le *Dictionnaire de l'Académie française* de 1932 et dans des textes de 1853 et de 1880 dans les syntagmes "élève des chevaux", "élève des bestiaux" et "élève du melon". Le 10 novembre 2003, Google m'a donné une vingtaine de contextes pour les cinq syntagmes ; la plupart des occurrences se trouvent dans des textes datés, dont le premier en date remonte à 1835 et le dernier à 1868. Voici quelques exemples :

> "l'élève des chevaux de troupe" dans une note dans M. Cailleux, *Des Causes de la diminution du commerce des chevaux en Normandie ; Des moyens de le rétablir et Instruction sur les chevaux nouvellement castrés* (Caen, 1835), autre document publié dans la Bibliothèque électronique de Lisieux
>
> "l'élève des chevaux et des bêtes à cornes" dans des "Extraits de la FRANCE PITTORESQUE, Département des Côtes-du-Nord (Ci-devant Basse-Bretagne), Par Abel HUGO - 1835"
>
> "un mémoire sur les remontes actuelles de la cavalerie, relativement à l'élève des chevaux [...] par m. Flavien d'Aldéguier. [...] Toulouse, J.B. Paya; Paris, Maison Anselin, 1843" sur une fiche de la J.T.I. Preston Library, Virginia Military Institute, Lexington, Virginia
>
> "l'élève des bestiaux" dans ROYER-COLLARD, "Organoplastie hygiénique, ou essai d'hygiène comparée, sur les moyens de modifier artificiellement les formes vivantes par le régime", Mémoire lu à l'Académie de médecine le 6 décembre 1843, HPML, XXIX, (1843), p. 213.
>
> "l'élève des bestiaux" dans "Les Colonies Françaises" (Extraits) Par J. Rambosson 1868

Le Web d'aujourd'hui peut donc être utile pour la compréhension et la description du lexique du passé.

Conclusion

Pour résumer l'essentiel de ce que j'ai dit, avec l'avènement et l'expansion du World Wide Web la lexicographie s'est donné les moyens de s'approcher de la langue et du sujet parlant. Le corpus du Web, le plus grand qui ait jamais existé et accessible à tout le monde, est, comme la langue, toujours changeant et toujours à jour. La lexicologie s'en trouve transformée, étant à la portée de tout curieux de la langue et du lexique. Ce corpus, travaillé désormais par une rédaction nombreuse mise en réseau, a permis la création de la dictionnairique dynamique, qui est à même de suivre le mouvement de la langue et de prétendre à l'exhaustivité.

Le Supplément du Trésor de la langue française

Pascale Bernard (1)
pascale.bernard@atilf.fr
Geneviève Fléchon (1)
genevieve.flechon@atilf.fr
Marie-Josèphe Mathieu (1)
marie-josephe.mathieu@wanadoo.fr

(1) ATILF Nancy Université & CNRS (Axe Lexiques)

Mots-clés : lexicographie, lexicologie, linguistique de corpus, lexique, lexicographie informatisée

Keywords: lexicography, lexicology, corpora linguistic, lexicon, computerized lexicography

Résumé : Une fois achevé, le *Trésor de la langue française* (*TLF*) requérait un *Supplément* qui devait être le tome 17 du *TLF*. Il devait contenir des informations annexes et parfois cachées dans les divers volumes et mettre à jour l'information bibliographique. Le *Supplément* devait enrichir la nomenclature du *TLF* grâce à quelques articles complétant les articles du *TLF*, et surtout grâce à des articles nouveaux. Ce dix-septième volume ne paraîtra jamais en version papier mais le laboratoire ATILF a pris la décision de l'informatiser. En examinant de près les données nous avons constaté que certains mots posaient quelques problèmes tant du point de vue synchronique que diachronique ; nous avons donc entrepris une vague de corrections et amendements afin que le *Supplément* soit le point de départ d'une nouvelle base lexicographique.

Abstract : Once finished, the *Trésor de la langue française* (*TLF*) required a *Supplement* which should be the volume 17 of the *TLF*. It had to contain information secondary and sometimes hidden in the diverse volumes and also it had to update bibliographical information. The *Supplement* had to enrich the word list of the *TLF* thanks to some articles which complement the articles of the *TLF*, and specially thanks to new articles.This seventeenth volume will never be published and the laboratory ATILF decided to computerize it. By having a close look at its data, we have noticed that some words posed problems in synchrony and diachrony as well; therefore we started a wave of corrections so that the *Supplement* is the starting point of a new lexicological database.

Introduction

Tout d'abord appelé *Complément*, ce volume finit par prendre le nom de *Supplément*. Ce volume non publié commençait par cet « Avis au lecteur » écrit en septembre 1998 par Gérard Gorcy :

« L'importance du *Trésor de la langue française* (*TLF*) (16 volumes représentant 24 000 pages d'articles), la durée de son élaboration (une trentaine d'années), le rythme de sa publication échelonnée sur plus de vingt ans (1971 à 1994) ont impliqué à la fois une évolution des méthodes de description et des modifications dans l'usage observé et l'image de langue de culture. A l'instar de toute œuvre de longue haleine, et singulièrement des grands dictionnaires et encyclopédies, le *TLF* parvenu à son terme requérait un *Supplément*.

Cet ouvrage doit remplir une double fonction : d'une part, faciliter la consultation du *TLF* en regroupant des informations annexes et parfois « cachées » dans les divers volumes ; d'autre part et surtout, compléter le *TLF*, en mettant à jour l'information générale tant pour les usages littéraires que techniques de mots bien installés en langue depuis 1965 (auxquels seuls les quatre derniers volumes du *TLF* ont commencé à faire une place satisfaisante), en mettant à jour aussi l'information bibliographique. Ainsi, pour ne prendre qu'un exemple, le tome 6 du *TLF*, publié en 1978, n'enregistre pas *décideur*, pourtant déjà attesté avec son sens bien connu dans le domaine de l'administration et de la politique depuis 1973 et pouvant être daté de 1969. Pour le *Supplément* on s'est efforcé de combler ces déficits ou ces lacunes. On notera que les articles complétant les articles du *TLF*, ont un effectif réduit. Le *Supplément* enrichit surtout la nomenclature du *TLF*, comme on le précisera plus loin et compte donc principalement des articles nouveaux ».

1. Un volume non paru

Ce volume de *Supplément*, qui devait être le tome 17 du *TLF*, promis aux souscripteurs de la version papier, était donc fortement attendu comme une mise à jour puisqu'il devait compléter des articles existants, combler les déficits et les lacunes du *TLF*. Mais il n'est jamais paru car la décision de ne pas le publier en version papier fut prise par le dernier directeur de l'INaLF, Bernard Cerquiglini, pour des raisons financières d'une part, Gallimard demandait une somme exorbitante pour l'impression de ce tome, et pour des raisons scientifiques d'autre part (manque de rigueur dans la nomenclature, mots trop anciens ou encyclopédiques).

2. Ses écueils en synchronie

Gérard Gorcy note que les articles qui complètent les articles du *TLF* sont peu nombreux, mais il est difficile de comprendre pourquoi des articles déjà traités dans le *TLF* ressurgissent dans le *Supplément*. Parfois ce sont des articles traités sous un élément formant du *TLF* qui bénéficient d'une vedette à part entière sans qu'aucune raison apparente n'explique ce nouveau traitement. Ainsi *éperonnier* est traité en dérivé sous *éperon* dans le *TLF*, *éphiggère* sous *-gère* et *épiaison* en dérivé sous *épier1*, *diguette* sous *–ette*, etc… . *Artaban* se voit attribué une vedette autonome dans le *Supplément* alors qu'il n'apparaît que sous la locution «fier comme Artaban» et qu'elle est bien traitée sous le mot *fier* dans le *TLF*. *Introspectif* est mieux rédigé dans le *TLF* que dans le *Supplément*, et *hypersustentateur* est déjà traité dans le

TLF avec le même exemple. On comprend mal pourquoi ce type de mots encombre la nomenclature du *Supplément*, alors que d'autres manquent cruellement (nous avons dû ajouter *canopée* par exemple.)

Lorsque nous avons étudié de près les articles de ce volume, nous avons ressenti la nécessité d'ajouter des exemples à des sens existants, car trop souvent la définition est empruntée à un dictionnaire. En outre, on constate également que de nombreux exemples sont issus de la presse, non pas que celle-ci ne puisse constituer une source d'exemples en soi, mais pourquoi la citer systématiquement lorsque nous avons de bons exemples dans nos bases littéraires ?

Un complément d'informations pour un sens est souvent nécessaire. En effet, un mot est traité mais le lecteur reste en attente d'informations parce qu'une forme ou une acception plus moderne ou plus commune sont manquantes, ainsi *dressing*, bien plus usité que *dressing-room*, manque sous cette entrée. Une *déneigeuse* n'est pas que la machine thermique à faire fondre la neige, c'est aussi un chasse-neige, nous avons donc dû la définir et l'illustrer.

Un complément d'informations pour une forme est également attendue pour des mots comme *DJ* (avec ses variantes graphiques) qui est plus employé que *disc-jockey* et qui manque sous cette entrée. Nous apportons ces informations.

Le *Supplément* complète des familles de mots traités dans le *TLF* mais il reproduit le même schéma, il entre une nouvelle famille de mots sans entrer tous les mots de la famille, ainsi on y trouve *hélitreuiller* mais pas *hélitreuillage* qui est beaucoup plus fréquemment employé. Nous le rattrapons en dérivé.

Nous montrerons comment des conditions d'emplois doivent être ajoutées, et comment des indications de domaines doivent être supprimées.

On y trouve aussi des mots trop techniques (*catathermomètre*) ou très encyclopédiques (*pothamothérium*) directement empruntés à des index de dictionnaires techniques, des mots trop anciens pour un volume de *Supplément* et souvent déjà qualifiés de « vieux » dans les dictionnaires du 20^{e} siècle, ou des mots de dictionnaires (*pointis* uniquement chez Larousse) qui n'ont pas leur place dans un volume complémentaire du *TLF*.

La féminisation des noms de sportifs ou de profession a été complètement occultée, comme dans le *TLF*, mais ne trouver en ski que des descendeurs et ne faire de *designer* qu'un métier d'homme, est dépassé. Il nous a fallu rattraper tout ceci.

En outre, certaines définitions, comme celle de *dance music*, sont bien obscures et ont demandé à être revues. Entrer un tel mot a nécessité de refaire intégralement la définition de la *dance* qui était absente.

Il pose également des problèmes de fond à traiter globalement, ainsi en ce qui concerne les noms latins du vocabulaire botanique. Une décision globale est à prendre pour un traitement uniforme et pour lequel nous avons déjà contacté un spécialiste en botanique.

3. Ses écueils en diachronie

Un fait récurrent dans les étymologies et histoires des mots est le manque de précision et de rigueur dans les articles. Le *FEW* a été mal interprété, les ouvrages disponibles au laboratoire n'ont pas été consultés, voire mal lus, et une mauvaise date a été attribuée à certaines sources.

4. Ses atouts

Après ces critiques qui paraissent sévères, il faut reconnaître au *Supplément* ses atouts, sinon, quel intérêt aurions-nous à le valoriser pour l'informatiser et le diffuser.
C'est un volume qui compte 10 000 mots qui n'entrent pas tous dans les catégories énoncées plus haut.
Certains mots sont des rattrapages obligés soit parce qu'ils ont été oubliés dans le *TLF*, (*dérivé*), d'autres complètent des familles (*disjoncter*, *évaporimétrie*, *vanillerie*…) et un cas très fréquent est le rattrapage de mots présents dans des définitions ou des exemples alors qu'ils sont absents de la nomenclature du *TLF*, le *Supplément* en récupère ainsi un certain nombre. Nous donnerons quelques exemples de ce genre, *dessuinter*, est absent du *TLF* mais *dégraissage* est défini dans le *TLF* comme l' «action de dessuinter les laines», *calcer* est cité et glosé dans un exemple etc.
Sa grande qualité réside dans les mots qui y ont vraiment leur place parce que ce sont des mots qui complètent réellement le *TLF* et qui sont des mots attendus et incontournables tant par leur pertinence que par la qualité de leur rédaction.
Après cet état des lieux, il faut savoir en tirer le meilleur parti et le considérer, une fois corrigé de ses imperfections, comme un volume supplémentaire qui augmentera la nomenclature du TLF de 10 000 mots environ.
On ne peut plus faire l'impasse sur son informatisation et sa diffusion car sa nomenclature est incluse dans Morphalou qui est un lexique ouvert des formes fléchies du français et dont les données initiales proviennent du TLFnome, la nomenclature du Trésor de la Langue Française.
En outre, il est cité dans la refonte de la lettre A du FEW (notamment t.25, 1182 b (Auvergne), 1186 b (auxesis).

5. Vers une nouvelle base lexicographique

Une fois informatisé, il constituera une nouvelle base de données que nous voulons être le point de départ d'une grande base lexicographique permettant d'intégrer des ressources complémentaires, notamment celles développées au laboratoire.

Conclusion et objectifs

De nombreux utilisateurs du *TLF* demandent une mise à jour des données et c'est notre devoir de recherche de nous atteler à cette tâche afin de donner à ce volume de *Supplément* la valeur scientifique auquel il peut prétendre.

Trop de mots récents sont absents car la rédaction du *Supplément* s'est arrêtée il y aura bientôt plus de dix ans et quel que soit le moyen choisi pour augmenter la nomenclature de cette base lexicographique (cf les travaux de l'équipe Lexiques, demande des utilisateurs du *TLF*, travail de veille des lexicographes ou du centre documentaire, intégration de données existantes au laboratoire prêtes à être informatisées), cette base lexicographique qui inclura le *Supplément* ne doit pas s'arrêter aujourd'hui.

Bibliographie

Introduction du *Supplément*, Gérard Gorcy, 1998, non publiée.

Le dictionnaire comme genre ou comme ressource

Philippe Gréa (1)
grea@u-paris10.fr
Sylvain Loiseau (2)
Loiseau@limsi.fr

(1) *Université Paris 10*
(2) LIMSI-CNRS

Résumé

De nombreuses expériences dans le domaine de la linguistique informatique recourent aux dictionnaires comme à des « corpus » à partir desquels il devient possible d'inférer des relations et des structures sémantiques à l'échelle de la langue. Ce recours aux dictionnaires – et plus particulièrement aux dictionnaires de synonymes – dans le traitement automatique des langues s'est accru récemment avec la diffusion de méthodologies issues de la théorie des graphes (Gaume *et al*., 2002*;* Ploux & Victorri, 1998, etc.).
Dans ces expériences, les dictionnaires ne sont plus utilisés comme l'aboutissement et la stabilisation d'une description (la description lexicologique), mais comme une « matière première » à informer, sur laquelle se fondent de nouvelles descriptions impliquant d'autres objectifs (notamment psycholinguistiques et cognitifs). Un aboutissement descriptif devient donc le point de départ d'une nouvelle construction ; ce qui est, somme toute, une situation banale de la description en sciences humaines, où l'on n'accède jamais à une matière qui n'ait pas déjà été informée d'interprétations.
Or, autant les règles de lecture et d'interprétation d'un dictionnaire en tant que résultat d'une description sont bien connues, et s'appuient sur une longue tradition d'élaboration de conventions, autant les règles d'interprétation et d'utilisation d'un dictionnaire comme source d'un dispositif descriptif quantitatif sont largement méconnues. Le dictionnaire n'est pas une abstraction mais un texte relevant d'un genre, et à ce titre, il est difficile d'en dériver des taxinomies spontanées sans tenir compte des propriétés liées à ce genre.

Cette proposition de communication se donne donc pour objectif d'examiner l'incidence des propriétés génériques du dictionnaire sur la construction de dispositifs descriptifs quantitatifs recourant à des dictionnaires. On montrera que l'examen de ces dispositifs dans cette perspective permet, en retour, de proposer de nouveaux indicateurs pour caractériser et décrire les dictionnaires en tant que textes.
Les expériences auxquelles nous nous limiterons recourent à la théorie des graphes comme outil de modélisation du dictionnaire. Nous montrerons tout d'abord qu'un grand nombre de graphes peuvent être dérivés d'un dictionnaire (il peut représenter les relations « mots vedettes » - mots de la glose » ou les relations entre mots d'une même glose ; il peut être orienté, bipartite ; se fonder sur différentes parties de chaque article, etc.). De plus, les

différents types de dictionnaires (trésors, généralistes, pour enfants) produisent également des graphes ayant des topologies différentes.
Notre analyse s'appuiera sur des indicateurs couramment utilisés dans la manipulation de graphes, comme les distributions des degrés des noeuds, le taux de clustering ou l'intermédiarité (Newman, 2003) ou la distribution du nombre de définitions (Cooper, 2005). Certains de ces indicateurs reprennent d'ailleurs des observations déjà connues. Les relations entre fréquence d'un mot et nombre de définitions ont par exemple été observé pour la première fois par G. Zipf (1949, cf. Manning & Schütze, 1999, p. 27).

Sur ces bases comparatives, nous montrerons que la recherche d'une correspondance entre catégories sémantiques et propriétés topologiques des graphes de dictionnaire est difficile à faire aboutir. Alors que de nombreuses interprétations sur la structure profonde du lexique ont été produites sur la base de graphes de dictionnaires (hypothèse continuiste de S. Ploux (1998), hypothèse sur l'acquisition de K. Duvignau (2002)), aucune mise en correspondance fine de la topologie d'un graphe de dictionnaire avec une typologie sémantique n'a encore été proposée. La recherche d'une telle articulation entre le modèle mathématique et les outils descriptifs, qui permettrait de raisonner et de contrôler l'interprétation de ces données, montre qu'un graphe de dictionnaire est un « mélange » de plusieurs graphes, une superposition de plusieurs typologies (structuration en domaines lexicaux, en classes sémantiques locales, en niveaux diaphasiques, etc.), qui reflète ses propriétés en tant que genre.

Bibliographie

Batagelj V., Mrvar A., Zaveršnik M. (2002) "Network analysis of dictionaries", *Jezikovne tehnologije [Language Technologies]*, pp. 135-142, Ljubljana.
Cooper M. C. (2005) "A Mathematical Model of Historical Semantics and the Grouping of Word Meanings into Concept", *Computational Linguistics*, 31(2), pp. 227-248.
Fox Keller E. (2005) « Revisiting "scale-free" networks », *BioEssays*, 27(10), pp. 1060-1068.
Gaume B., Duvignau K., Gasquet O. & Gineste M.-D. (2002) « Forms of meaning, meaning of forms », *Journal of Experimental & Theoretical Artificial Intelligence*, 14(1), pp. 61-74.
Gaume B., Hathout N., Muller P. (2004) "Word Sense Disambiguation using a dictionary for sense similarity measure", *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1194-1200.
Manning C. & Schütze H (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge.
Newman M. E. J. (2003) "Power laws, Pareto distributions and Zipf's law", *Contemporary Physics*, 46(5), pp. 323-351.
Newman M. E. J. (2005) "The structure and function of complex networks", *SIAM Review*, 45, pp. 167-256.
Ploux S. & Victorri B. (1998) « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *TAL*, 39 (1), pp. 161--182.
Zipf G. (1949) *Human Behaviour and the Principle of least effort: An Introduction to Human Ecology*, Hafner, New York.

ATILF / CNRS Nancy-Université
44, avenue de la Libération BP 30687 54 063 Nancy Cedex
contact@atilf.fr - Téléphone : 03 83 96 21 76 - Télécopie : 03 83 97 24 56
www.atilf.fr www.cnrtl.fr

atilf ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Nancy-Université