Pour un Traitement Automatique Optimisant la Consultation de Corpus Électroniques en Lexicographie

Nathalie Gasiglia
SILEX (UMR 8528 du CNRS), Université Lille 3
B.P.149 ; 59653 VILLENEUVE D'ASCQ Cedex ; FRANCE
gasiglia@univ-lille3.fr

Abstract

In my presentation, I would like to concentrate on three points: First of all, to evaluate how the availability of electronic corpora (corpora representative of one language or sample of specialised languages corpora) can contribute to the richness of dictionaries. Secondly, to explain how we can optimise the output of KWIC lists consultations (concordances) by adding automatic analyses and sorting procedures according to these similarities (I will pay particular attention to syntactic and semantic automatic analyses of each extraction (each sentence in each line) of a KWIC list. I call "syntactic analysis" a combination of synonymous information taken out from an electronic synonym dictionary and information extracted from semantic theories.). At last, I evaluate, in the case of bilingual aligned corpora, how the comparison of syntactic and semantic analyses of each extraction in each two languages (the source and the target) can ameliorate KWIC lists which are more quickly and easily examined.

0 Introduction

Après avoir dressé un rapide état des lieux concernant le faible recours aux corpus dans les pratiques des lexicographes français (lexicographes appartenant à des équipes de recherche sur fonds publics ou à des maisons d'édition privées), cette étude propose, d'évaluer comment la consultation de corpus électroniques peut aujourd'hui contribuer à la richesse des articles lexicographiques produits. J'évaluerai, au sein des ressources disponibles, lesquelles sont effectivement nécessaires pour les lexicographes. Je proposerai ensuite, et c'est là le point principal de ma contribution, une méthode de dépouillement automatique de ces corpus qui ne se limite pas à la production de concordances ("KWIC lists") ou de listes de collocations, mais effectue un post-traitement de ces occurrences en contexte consistant en une analyse syntaxique des phrases extraites lors de l'établissement des "KWIC lists" combinée à une recherche de synonymes comme à la prise en compte de propriétés mises en évidence par des sémanticiens.

Le point de vue adopté ici n'est pas celui d'une lexicographe professionnelle mais d'une spécialiste du TAL visant à mettre son expérience au service des lexicographes afin de leur permettre de ne plus consulter de simples "KWIC lists" mais des extractions automatiquement triées, organisées, pour chaque item recherché, en fonction des cadres de sous-catégorisation intégrés ou d'appartenance à des classes sémantiques construites à partir de relations de synonymie et d'informations syntaxiques et sémantiques. Les contraintes de place ne permettant pas ici d'exemplifier mes propos au fil du texte, l'exposé restera de nature programmatique.

1 Peu de Place pour les Corpus dans la Pratique des Lexicographes Français

Je résumerai la situation en disant qu'au moins cinq facteurs expliquent la faible intégration des consultations de corpus électroniques dans la pratique des lexicographes français:
(i) des **difficultés linguistiques** inhérentes à l'objet étudié, qui est malaisément descriptible dans des "scripts de fouille": les objets cherchés par les lexicographes sont plus diversifiés et polymorphes (il peut s'agir de locutions ou d'unités lexicales autonomes mono- ou polylexicales[1] aux structures syntaxiques très diverses, de constituants d'unités polylexicales ou de morphèmes constitutifs de mots construits), et moins strictement identifiés et structurés qu'en terminographie (où les candidats-termes ont des structures syntaxiques repérables : *N Prep N* ou *N Adj* par exemple);
(ii) des **carences documentaires**: les recherches des lexicographes se font en corpus de langue "commune"[2], or aucun corpus de "référence"[3] n'existe pour le français;
(iii) une notable **insuffisance qualitative des outils informatiques** utilisés pour faire des extractions automatiques d'informations: les lexicographes français consultent des concordances ("KWIC lists") ou des listes de collocations [Fléchon 1998][4];
(iv) l'existence d'une **tradition lexicographique ancienne**: d'importants ouvrages de référence existent déjà dans les grandes maisons d'édition françaises sans qu'on voie une urgence à les réviser fondamentalement;
(v) et le fait que, vers 1985, les deux principales maisons d'édition dictionnairique, Larousse et Robert, rachetées par un grand groupe absorbé aujourd'hui dans Vivendi Universal, soient entrées dans une logique de remise à flot financière par la rentabilisation des catalogues et donc l'intensification de la dérivation de dictionnaires au détriment de créations innovantes [Corbin 1991, 1998 et à paraître]

Les corpus occupent peu de place dans la pratique actuelle des lexicographes français, mais l'outil informatique peut apporter beaucoup et le meilleur argument pour inciter les recours aux corpus électroniques est certainement d'offrir des ressources et des interfaces adaptées aux besoins effectifs, intégrant des outils de consultation efficaces qui permettraient aux lexicographes de tenir les cadences qui leur sont imposées tout en effectuant des recherches systématiques.

2 Quels Corpus pour les Lexicographes?

2.1 Un Corpus de Référence Souhaitable pour Travailler en Lexicographie

Les lexicographes qui doivent décrire la langue ont un effectif besoin de ces corpus équilibrés et de taille conséquente appelés "corpus de référence" ([Habert et al. 1997] ; [Habert et al. 1998]), et ce afin de pallier les carrences de l'introspection qu'il s'agisse de sur- ou sous-évaluation des usages effectifs: le recours au corpus doit, dans cette perspective, permettre de fournir des données satisfaisantes sur l'usage linguistique réel. Un corpus électronique destiné à des études lexicographiques est donc une compilation de données linguistiques de taille importante (plusieurs dizaines voire centaines de millions de mots) dont les "documents primaires"[5] intégrés sont diversifiés et échantillonnés afin d'obtenir un équilibrage quantitatif et qualitatif[6] et d'avoir une couverture acceptable des phénomènes à observer en langue "commune".

Les lexicographes ont besoin de corpus de référence, mais les efforts de constitution déjà accomplis sont variables d'une communauté langagière à l'autre, ce qui fait que nous ne disposons pas de la même richesse en français[7] qu'en anglais par exemple.

2.2 Corpus de Langue "Commune" *vs* Corpus de Langue de Spécialité

Dans leur pratique, les lexicographes vont, lors de la constitution d'un dictionnaire spécialisé (médical ou de botanique par exemple) ou lors du traitement d'items plus scientifiques ou techniques intégrés dans un ouvrage de référence, ne plus considérer la langue "commune" dans sa globalité mais une partie d'une ou plusieurs langue(s) de spécialité, et ainsi adopter une optique plus terminographique. Les corpus électroniques pouvant être explorés pourront alors être de taille plus restreinte, constitués de textes de vulgarisation ou de documents relativement peu techniques et ne présenteront plus les mêmes exigences d'équilibrages quantitatifs et qualitatifs: tout sera fonction des pratiques observées, du domaine de spécialité étudié et des objectifs visés par l'ouvrage en chantier.

L'absence de corpus de référence pour le français ne favorise pas l'installation d'un automatique "recours aux corpus" chez les lexicographes exerçant en France. Cependant, ils peuvent, dès à présent, exploiter des corpus de différentes langues de spécialités, de différents types textuels[8], *etc.*, afin de disposer de larges échantillons d'observables dans des domaines de spécialité qui leur sont éventuellement peu connus.

3 Quelle Interface de Consultation pour les Lexicographes Français ?

Les besoins exprimés par les lexicographes français sont d'ordres divers. Certains me semblent déconcertants dans la mesure où ils traduisent les piètres performances des outils utilisés quand ils ne révèlent pas leur absence cruelle. Ils relèvent des paramétrages du concordancier au moyen duquel les extractions sont accomplies et consultées. Ce sont des questions techniques sur lesquelles je ne m'attarderai pas ici: pouvoir spécifier le nombre d'occurrences en contexte (= de lignes de concordance) souhaitées; pouvoir, à partir d'un contexte réduit à quelques caractères à droite et à gauche de la chaîne cherchée, accéder à un contexte large (plusieurs lignes voire un paragraphe entier ou le texte source dans son intégralité); disposer de manière directe des références des textes source[9]; ou encore pouvoir sélectionner des sous-corpus. D'autres sont corrélés aux moyens mis en œuvre pour mener à bien les consultations ou à la présence ou non de pré-traitement des corpus consultés. L'interface de consultation du corpus devrait autoriser les recherches avec caractères génériques c'est-à-dire permettre que toutes les adresses et sous-adresses puissent potentiellement faire l'objet d'une exploration, qu'il s'agisse de mots ou de morphèmes à isoler – préfixes, suffixes ou morphèmes "cachés" à l'intérieur d'une chaîne de caractères longue[10] – en traitant les caractères voisins avec un "joker"[11]. L'interface devrait également proposer une évaluation des fréquences d'emploi en corpus, c'est-à-dire extraire un index, une liste de "mots"[12] et de "n-gram"[13] (de deux, trois, ... *n* "mots") avec mention de leur fréquence relative, afin d'aider les lexicographes à évaluer l'importance de chacun dans la communication et donc à déterminer la juste place[14] à accorder aux unités lexicales correspondantes dans un dictionnaire en chantier. Enfin, sans pré-traitement, les extractions faites sont le résultat de simples recherches de chaînes de caractères et il sera par exemple nécessaire de chercher successivement les différentes formes fléchies qui correspondent à une unité lexicale étudiée

ou présente dans une séquence donnée. Cette succession de recherches peut être évitée par une opération préalable: la lemmatisation, qui consiste en l'association de lemmes aux différentes formes fléchies du corpus. Lancer ensuite une recherche d'un lemme donné revient donc à extraire, en une passe, toutes les formes fléchies correspondantes. Mais le lexicographe à la recherche de séquences où une unité donnée est immédiatement suivie d'une préposition ou précédée d'un verbe à l'indicatif présent, par exemple, a besoin de procéder à des discriminations en fonction de la catégorie syntaxique ou de la flexion des items et donc que son corpus de travail soit morphosyntaxiquement étiqueté, c'est-à-dire qu'un code de partie du discours soit associé à chaque item ou paire forme fléchie - lemme.

4 Aller Encore Plus Loin dans l'Analyse Automatique des Extractions Produites par les Concordanciers à partir de Corpus Monolingues

4.1 Analyses Syntaxiques des Phrases des Concordances

Pour l'heure, les lexicographes qui dépouillent manuellement les concordances font des repérages "à la volée", en tentant d'identifier au fil de leur lecture les éléments pertinents. Procéder à des analyses syntaxiques systématiques et minutieuses des phrases présentes dans les concordances permet de déterminer quelles constructions syntaxiques sont intégrées par une chaîne de caractères, une séquence graphique pouvant potentiellement constituer plusieurs unités lexicales homographes, et *in fine* de définir quelle en est la valeur sémantique en fonction des différents contextes d'emploi. Un traitement automatique peut prendre en charge une part importante de ce travail d'analyse en sélectionnant, dans la concordance produite pour une unité à partir d'un corpus étiqueté, toutes les phrases qui la contiennent. S'il s'agit d'un nom ($N_{étudié}$), par exemple, il repérera les verbes dont cette unité est sujet ou complément puis identifiera et marquera le lien V - $N_{étudié}$, puis procédera de même pour les noms dont cette unité est complément et des adjectifs ou compléments de ce nom, *etc.*

Mais l'analyse automatique peut aller plus loin, dépasser cette étape d'analyse linéaire et proposer des regroupements d'occurrences en fonctions de propriétés communes, par exemple le fait d'être argument d'un même verbe ou qualifiées par un même adjectif, comme des dégroupements qui aideront le lexicographe à discriminer les unités lexicales homographes en homonymes, polysèmes ou variantes selon des préférences contextuelles. La présentation des regroupements peut être ordonnée par fréquences d'apparition en corpus ou selon des critères sémantiques.

4.2 Adjonction d'Informations Issues de Dictionnaires de Synonymes ou de Travaux de Sémanticiens

Ces analyses syntaxiques peuvent améliorer sensiblement les performances de consultation des lexicographes. Il est cependant souhaitable de les enrichir encore en adjoignant aux regroupements précédemment évoqués différentes informations issues de dictionnaires de synonymes ou de travaux de sémanticiens: observations effectuées au LLI[15] (afin de déterminer l'appartenance des unités à des classes sémantiques, les "classes d'objets" [Gross 1994], [Le Pesant 1994]) comme dans d'autres cadres (sémantique du prototype [Kleiber 1990] et sémantique indexicale [Cadiot & Nemo 1997a et 1997b]). Les enrichissements obtenus à ce jour, en exploitant un dictionnaire électronique de synonymes[16] sont

encourageants, dans la mesure où les "cliques" de synonymes (cf. [Ploux & Victorri 1998]) présentées aident à partitionner plus finement les analyses syntaxiques.

4.3 Concordances Bilingues Parallèles

Si la pertinence d'user de corpus bilingues pour identifier plus facilement la nature des problèmes de traduction ou encore pour mieux repérer toutes les séquences non strictement compositionnelles[17] n'est plus véritablement à discuter (cf. [Grundy 1996]) et que seul le manque de ressources explique la faible représentation de cette pratique, il me semble important d'évaluer dans quelle mesure les analyses syntaxiques et sémantiques réalisables pour les extractions de chaque langue peuvent être utilement combinées afin d'optimiser la lecture de ces concordances bilingues parallèles en comparant les regroupements d'extractions faits pour chacune des langues et ainsi pister les divergences d'usages entre les langues étudiées.

5 Conclusion

A partir de ce qui a été observé dans les pratiques des lexicographes français, deux grands objectifs se dessinent:
I) envisager la création d'un corpus de référence pour le français comme il en existe pour diverses autres langues, et, dès à présent, parallèlement à ce travail de création, développer l'usage de corpus plus modestes, relatifs à des domaines de spécialité, dont la consultation permettrait de pallier les limites de compétence des lexicographes (cf. § 2).
II) constatant le manque d'outils d'exploration de corpus électroniques sophistiqués et le déficit d'emploi des outils disponibles par les lexicographes français: (i) mettre minimalement à leur disposition une interface de consultation aussi riche que les moyens informatiques disponibles et combinables le permettent et (ii) chercher à optimiser encore les modalités d'exploration de corpus en proposant un outil nouveau, complet, prenant en compte maximalement les besoins explicités par les lexicographes (cf. § 3.) et accélérant la consultation des extractions faites en facilitant la lecture des résultats pertinents (cf. § 4.).

Acknowledgements

Je voudrais, ici, remercier Pierre Corbin, qui m'a initiée au monde de la lexicographie en m'invitant à travailler avec lui dans le cadre du DESS de "Lexicographie et Terminographie" (Univ. Lille 3): la richesse de ses interrogations a motivé le recentrage de mes compétences en TAL sur des questionnements lexicographiques. Nos échanges ont nourri le présent travail, mais il n'aurait pas pu se faire sans la disponibilité des professionnels intervenant dans le DESS, qui, en évoquant leur expérience, m'ont permis de prendre la mesure des pratiques effectives. Je pense en particulier à Valerie Grundy, Henri Béjoint, Dominique Le Fur, Martyn Back, Marta Krol, Laurent Catach, Jean Binon, Serge Verlinde. Qu'ils en soient vivement remerciés.

Endnotes

[1] Je considère ici que, du point de vue du TAL, les unités polylexicales sont constituées de plusieurs chaînes de caractères alors que les unités monolexicales n'en comptent qu'une. Du point de vue de la morphologie constructionnelle, les unités polylexicales sont toutes construites alors que les unités monolexicales peuvent ou non être construites selon qu'elles sont ou non constituées de plusieurs morphèmes combinés afin de créer une unité sémantique particulière.

[2] La langue considérée ici comme "commune" est la langue non spécialisée enrichie de la langue spécialisée observable dans des textes vulgarisateurs et strictement non scientifiques ou techniques.
[3] Cf. § 2.
[4] Si Oxford University Press dispose maintenant, pour traiter l'anglais, d'outils qui génèrent des analyses syntaxiques partielles, rien de semblable n'est utilisé, ni chez Oxford, ni en France, à ma connaissance, pour les corpus de français.
[5] Les "documents primaires" sont les documents (textes, transcriptions, etc.) constitutifs du corpus.
[6] Il faut particulièrement veiller, pour la lexicographie, à disposer d'une réelle diversité des modes d'expression et registres de langue (avec une bonne part de transcriptions d'oral), d'une richesse lexicale certaine, d'une actualisation constante des données, d'une masse conséquente de données récentes, *etc.*
[7] Pour le français: *Frantext* (corpus de l'ILF (ex-INaLF), http://zeus.inalf.cnrs.fr/noncateg.htm et http://zeus.inalf.cnrs.fr/categ.htm, pour la version étiquetée morpho-syntaxiquement.), des CD-Rom d'articles de presse et quelques corpus de taille modeste consultés localement, mais pas de corpus de référence.
[8] Il est tout à fait concevable que, sans construire un corpus de référence, une équipe de lexicographes cherche à pallier les déficits de qualification dans un domaine de spécialité donné et le manque de ressources électroniques en compilant les données de CD-Rom d'articles de presse, de versions électroniques d'œuvres littéraires, de rapports d'activités de programmes de recherches, de comptes rendus d'audiences, de contenus ("aspirés") de sites Web, *etc.*
[9] Quand une unité lexicale ou une construction présente une fréquence élevée, si toutes les occurrences sont issues de la même source, le nombre perd sa signification: ce n'est pas représentatif d'un phénomène en langue mais d'un particularisme propre à un rédacteur, un type d'écrit, *etc.*
[10] Le repérage automatique de ces unités infralexicales est évidemment borné par les interactions de la phonologie avec la morphologie (troncations, haplologies...) qui déforment les éléments concaténés.
[11] Il s'agit de caractères génériques: ? ou * pour remplacer un ou plusieurs caractères respectivement.
[12] Pour un traitement informatique, un "mot" est, minimalement, une chaîne de caractères délimitée par deux espaces ou un espace et un signe de ponctuation.
[13] Pour un traitement informatique, un "n-gram" est une séquences composée de n (n>1) "mots", c'est-à-dire de n chaînes de caractères délimitées par deux espaces (ou un espace et un signe de ponctuation pour l'élément extérieur droit), qu'il s'agisse d'une séquence constituant une unité lexicale, une collocation ou que ce ne soit qu'un enchaînement sans intérêt lexical particulier.
[14] En fonction des consultants et utilisateurs visés par le projet lexicographique, tous les mots à haute fréquence n'ont pas toujours besoin d'être décrits avec précision: les lexicographes se doivent donc d'équilibrer précision et concision.
[15] Laboratoire de Linguistique Informatique, Université Paris XIII.
[16] Ce dictionnaire est consultable en ligne: http://elsap.unicaen.fr/dicosyn.html. Il est le fruit de la collaboration de trois équipes : le CRISCO (Caen), l'ISC (Lyon) et le LaTTICe (ENS - Université Paris VII).
[17] Il s'agit de constructions syntaxiques, collocations et autres moyens fixes mais imprévisibles de combiner des lexèmes pour produire du sens dans une langue, surtout si ces séquences ne sont pas symétriques dans les deux langues, cf. [Grundy 1996].

Références

[Atkins 1990] Atkins, B.T.S., 1990. Corpus Lexicography: The Bilingual Dimension, in: *Linguistica Computazionale, VI.*
[Blank 1995] Blank, I., 1995. Sentence alignment: methods and implementation, in: *TAL, 36:1/2.*

[Cadiot & Nemo 1997a] Cadiot, P. & F. Nemo, 1997. Propriétés extrinsèques en sémantique lexicale, in: *Journal of French Linguistic Studies.*

[Cadiot & Nemo 1997b] Cadiot, P. & F. Nemo, 1997. Pour une sémiogénèse du nom, in: *Langue Française, 113.*

[Corbin 1991] Corbin, P., 1991. Le maquis lexicographique. Aperçus sur l'activité lexicographique monolingue dans le domaine français à la fin du XX^{e} siècle, in: *Le français aujourd'hui, 94.*

[Corbin 1998] Corbin, P., 1998. La lexicographie française est-elle en panne?, in: *Cicle de Conferències 96-97. Lèxic, corpus i diccionaris.* Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

[Corbin à paraître.] Corbin, P., à paraître. Lexicographie et linguistique : une articulation difficile. L'exemple du domaine français, in: Actes de la journée d'étude «De la lexicologie à la lexicographie» (Université d'Utrecht, Pays-Bas, 18 juin 1999).

[Fléchon 1998]. Fléchon, G., 1998. Expérience de rédaction. La mise au point de quelques rubriques synchroniques dans le *Trésor de la Langue Française*, in: *International Journal of Lexicography, 11.2 et 11.3.*

[Gross 1994] Gross, G., 1994. Classes d'objets et descriptions des verbes, in: *Langages, 115.*

[Grundy 1996] Grundy, V., 1996. L'utilisation d'un corpus dans la rédaction du dictionnaire bilingue, in: H. Béjoint & Ph. Thoiron *Les dictionnaires bilingues.* Duculot, Louvain-la-Neuve.

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

[Habert et al. 1998] Habert, B., C. Fabre & F. Issac, 1998. *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques.* Interedition, Paris.

[Ide & Véronis 1994] Ide, N. & J. Véronis, 1994. MULTEXT (Multilingual Tools and Corpora). *Proceedings of the 14th International Conference on Computational Linguistics (COLING'94).*

[Kleiber 1990] Kleiber, G., 1990. *La sémantique du prototype. Catégories et sens lexical.* Presses Universitaires de France, Paris.

[Le Pesant 1994] Le Pesant, D., 1994. Les compléments nominaux du verbe *lire*, une illustration de la notion de "classe d'objets". *Langages, 115.*

[Melamed 2000] Melamed, I.D., 2000. Bitext maps and alignments via pattern recognition, in: J. Véronis (ed.) *Parallel Text Processing.* Kluwer: Dordrecht.

[Ploux & Victorri 1998] Ploux, S. & B. Victorri, 1998. Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes, in: *TAL, 39:1.*

[Véronis & Khouri 1995] Véronis, J. & L. Khouri, 1995. Étiquetage grammatical multilingue : le projet MULTEXT, in: *TAL, 36:1/2.*