Élaboration automatique d'un dictionnaire de cooccurrences grand public

Simon CHAREST, Éric BRUNELLE, Jean FONTAINE, Bertrand PELLETIER
Druide informatique inc.
1435, rue St-Alexandre, bureau 1040
Montréal (Québec) H3A 2G4, Canada
developpement@druide.com

Résumé. Antidote RX, un logiciel d'aide à la rédaction grand public, comporte un nouveau dictionnaire de 800 000 cooccurrences, élaboré essentiellement automatiquement. Nous l'avons créé par l'analyse syntaxique détaillée d'un vaste corpus et par la sélection automatique des cooccurrences les plus pertinentes à l'aide d'un test statistique, le rapport de vraisemblance. Chaque cooccurrence est illustrée par des exemples de phrases également tirés du corpus automatiquement. Les cooccurrences et les exemples extraits ont été révisés par des linguistes. Nous examinons les choix d'interface que nous avons faits pour présenter ces données complexes à un public non spécialisé. Enfin, nous montrons comment nous avons intégré les cooccurrences au correcteur d'Antidote pour améliorer ses performances.

Abstract. Antidote is a complete set of software reference tools for writing French that includes an advanced grammar checker. Antidote RX boasts a new dictionary of 800,000 co-occurrences created mostly automatically. The approach we chose is based on the syntactic parsing of a large corpus and the automatic selection of the most relevant co-occurrences using a statistical test, the log-likelihood ratio. Example sentences illustrating each co-occurrence in context are also automatically selected. The extracted co-occurrences and examples were revised by linguists. We examine the various choices that were made to present this complex data to a non-specialized public. We then show how we use the co-occurrence data to improve the performance of Antidote's grammar checker.

Mots-clés : antidote, cooccurrences, collocations, corpus, analyseur, correcteur.
Keywords: antidote, co-occurrences, collocations, corpus, parser, grammar checker.

1 Introduction

Antidote RX est la sixième édition d'Antidote, un logiciel d'aide à la rédaction développé et commercialisé par la société Druide informatique. Antidote RX comporte un correcteur grammatical avancé, dix dictionnaires de consultation et dix guides linguistiques. Parmi les dix dictionnaires de l'édition RX figure un nouveau dictionnaire de cooccurrences, constitué essentiellement automatiquement au moyen d'outils de traitement de la langue à couverture large. Le dictionnaire de cooccurrences d'Antidote a la particularité d'être destiné au grand public, pour qui il établira souvent le premier contact avec la notion de cooccurrence.

Par *cooccurrence*, nous entendons la présence simultanée et statistiquement significative, dans un corpus, de deux unités linguistiques en relation syntaxique. Notre définition correspond aux *cooccurrences relationnelles* d'Evert (2005), aussi appelées *cooccurrences syntaxiques*, par opposition aux *cooccurrences positionnelles*, dont les mots apparaissent simplement fréquemment dans une certaine proximité.

Notre concept de cooccurrence englobe des combinaisons lexicales dont le degré de figement est variable : nous y incluons à la fois des combinaisons libres (*entendre un cri*), des combinaisons semi-figées ou *collocations* au sens strict (*pousser un cri*) et des *locutions* figées courantes (*cri du cœur*) ou terminologiques (*cri primal*).

Comme Antidote comporte déjà un dictionnaire de locutions, nous avions d'abord pensé nous restreindre aux *collocations*, mais deux motifs nous ont incités à élargir notre cible :

1. La frontière entre collocation et locution figée, d'une part, et entre collocation et combinaison libre, d'autre part, n'est pas toujours nette. Par exemple, dans certains cas limites, il n'est pas clair si une combinaison de mots acquiert vraiment un sens différent de la composition des sens de chacune de ses parties.

2. Pour obtenir la description la plus complète possible de l'usage d'un mot, il nous est apparu plus intéressant de présenter toutes les combinaisons les plus fortes de ce mot, incluant les locutions figées et les combinaisons libres statistiquement significatives.

Tutin (2005) examine l'intérêt d'un dictionnaire de cooccurrences et constate les lacunes des dictionnaires généralistes en la matière. Comme outil de consultation, un dictionnaire de cooccurrences énumère les contextes d'usage d'un mot. En production de texte, il complète le dictionnaire de synonymes pour retrouver le mot juste ou le tour idiomatique. Il aide enfin à l'apprentissage du français langue seconde, car les locutions et collocations, qui reflètent des emplois figés ou semi-figés, sont difficilement prédictibles pour un locuteur étranger.

Le dictionnaire de cooccurrences d'Antidote RX a été extrait automatiquement à partir d'un corpus de 500 millions de mots. L'analyseur syntaxique de haut niveau d'Antidote a été mis à contribution pour recenser plus de 17 millions de paires de mots liées par diverses relations syntaxiques. Les cooccurrences les plus significatives ont été dégagées à l'aide d'un filtre statistique et d'une révision manuelle. Le résultat est un dictionnaire de 800 000 cooccurrences illustrées par plus de 2 millions de phrases exemples tirées du corpus. À notre connaissance, il s'agit du plus vaste dictionnaire de cooccurrences du français à ce jour. Le présent article décrit le processus d'élaboration de ce dictionnaire.

2 Travaux antérieurs

Les dictionnaires de cooccurrences en français sont peu nombreux. Un pionnier fut Ulysse Lacroix avec *Les mots et les idées : dictionnaire des termes cadrant avec les idées*, dont la première édition remonte à 1931. Plus récemment, le *Dictionnaire des cooccurrences* de Beauchesne (2001) recense environ 150 000 cooccurrences. Le site Web *Dictionnaire des collocations* de Rodriguez, lancé en 2004, affichait 31 400 combinaisons à la fin de 2006. Ces trois dictionnaires ont été élaborés à partir d'une collecte manuelle des cooccurrences. Le *Dictionnaire combinatoire du français* de Zinglé et Brobeck-Zinglé (2003) a été créé à partir d'un corpus dont ont été tirées 65 000 expressions (34 000 en version imprimée).

Parmi les dictionnaires universitaires, mentionnons le *Dictionnaire explicatif et combinatoire du français contemporain* (Mel'čuk et coll., 1984, 1988, 1992, 1999), dont une version électronique simplifiée développée par Mel'čuk et Polguère peut être consultée sur le site *DiCouèbe*. On peut aussi consulter, sur le site de l'Équipe de recherche en syntaxe et sémantique (ERSS), la base lexicale distributionnelle *Les voisins de Le Monde*, construite automatiquement à partir d'un corpus comprenant l'ensemble des articles du quotidien *Le Monde* sur une période de dix ans (1991-2000), soit environ 200 millions de mots.

L'extraction automatique de cooccurrences à partir d'un corpus a fait l'objet de nombreux articles. Plusieurs traitent de cooccurrences positionnelles, basées sur la seule proximité des mots à l'intérieur d'un intervalle donné (par ex. de 5 positions). Parmi les travaux qui mentionnent un traitement syntaxique, soulignons ceux de Lin (1998), qui utilise un parseur et filtre les cooccurrences en fonction de l'information mutuelle ; ceux de Kilgarriff et Tugwell (2001) pour *Word Sketch*, un module du *Sketch Engine* accessible en ligne, où les cooccurrences sont extraites du *British National Corpus* à l'aide d'un étiqueteur et d'une grammaire de type *pattern matching*, puis filtrées selon une mesure intégrant l'information mutuelle ; et enfin les travaux du Laboratoire d'Analyse et de Technologie du Langage (LATL) de l'Université de Genève (Seretan, Wehrli, 2006), qui utilisent le parseur multilingue *Fips* et filtrent les résultats en employant la mesure du rapport de vraisemblance.

3 Méthodologie

3.1 Constitution d'un corpus

Le matériau brut de notre dictionnaire étant un corpus, il est essentiel que celui-ci soit de grande taille, afin de refléter un large éventail d'usage des mots, des plus fréquents aux plus rares. Il faut aussi varier les styles d'écriture ainsi que les domaines des textes (voir Tableau 1). Nous récoltons en outre des écrits de diverses régions de la francophonie dans le but d'extraire des cooccurrences propres aux locuteurs de chacune de ces régions.

Domaine	Proportion	Exemples de sources
Littéraire	30 %	Gallica, Projet Gutenberg, Les Éditions Québec Amérique
Journalistique	40 %	Le Devoir, Voir, L'Express, Libération, La Tribune de Genève
Autres	30 %	Wikipédia, CyberSciences, LégiFrance, Université de Montréal

Tableau 1: principales sources du corpus

Le Web, mine quasi inépuisable de textes de toute sorte, est la principale source de notre corpus. Mais le Web a aussi ses défauts. Par exemple, il est fréquent de trouver des phrases qui apparaissent de manière récurrente dans plusieurs pages d'un même site. On trouve aussi des phrases, provenant de citations ou de dépêches journalistiques, qui se retrouvent sur plusieurs sites différents, parfois telles quelles, parfois légèrement reformulées. Ces phrases récurrentes faussent les statistiques de fréquence. Il a donc fallu identifier et éliminer automatiquement les phrases identiques ou trop similaires.

Au final, nous avons constitué un corpus de 500 millions de mots, répartis sur 25 millions de phrases. C'est de ce matériau brut que nous avons extrait nos cooccurrences.

3.2 Extraction des cooccurrences

L'analyseur d'Antidote, fruit de plus de 10 années de développement intensif, reconnait un large éventail de structures syntaxiques du français. Nous en avons créé une version adaptée, optimisée pour analyser en traitement distribué la masse énorme du corpus et en extraire les cooccurrences. Sur une grappe de 15 ordinateurs utilisant la technologie de distribution XGrid d'Apple, il lui a fallu 1100 heures pour analyser les 500 millions de mots du corpus.

L'analyseur effectue une analyse en dépendance et génère des arbres syntaxiques complets, desquels les cooccurrences sont extraites directement. Lorsque plusieurs analyses sont trouvées pour une même phrase, l'arbre le plus probable, selon la pondération de l'analyseur, est choisi. Nous avons sélectionné les relations syntaxiques les plus pertinentes pour un dictionnaire de cooccurrences (voir Tableau 2). Pour cette première version, nous n'avons considéré que les cooccurrences à deux membres.

Categ 1	Categ 2	Relation	Proportion	Exemple
Nom	Adjectif	Épithète	18 %	*jeune fille*
Nom	Nom	Apposition	1 %	*site internet*
Nom	Nom/Verbe	Complément du nom	25 %	*coup d'œil*
Verbe	Adverbe	Modificateur	4 %	*aller loin*
Verbe	Nom	Sujet	12 %	*le vent souffle*
Verbe	Nom	Complément direct	10 %	*jouer un rôle*
Verbe	Nom	Autres compléments[1]	26 %	*perdre de vue*
Verbe	Nom/Adj	Attribut	1 %	*retenir prisonnier*
Verbe	Verbe	Complément verbal	1 %	*entendre parler*
Adjectif	Adverbe	Modificateur	1 %	*gravement malade*
Adjectif	Nom	Complément de l'adjectif	1 %	*âgé de x ans*

Tableau 2 : principales relations extraites

Au-delà des relations syntaxiques directes, le système s'efforce d'extraire des cooccurrences en franchissant des relations plus profondes. Par exemple :

- Coordinations : « De temps en temps cette clameur et ce bruit redoublaient. » → *la clameur redouble* ; *le bruit redouble*
- Relatives : « Or, sa réouverture est cruciale pour acheminer l'aide humanitaire dont a désespérément besoin la population. » → *avoir besoin d'aide*
- Agents : « Nous laissons à nos lecteurs le soin de deviner quel genre de surprise cette chute apporterait aux habitants. » → *le lecteur devine*

[1] Les « autres compléments » incluent les compléments indirects (COI) et les compléments adverbiaux ou circonstanciels.

- Collectifs : « Il a ouvert une infinité de routes, toutes raboteuses, qu'il a fallu ensuite aplanir. » → *ouvrir la route*

Nous avons considéré certaines locutions verbales figées (p. ex. *faire face*, *laisser place*, *avoir besoin*) comme des verbes à part entière, ce qui nous a permis d'extraire des cooccurrences comme *faire face aux défis*, *laisser place au doute* et *avoir besoin d'aide*.

Outre les deux mots formant la cooccurrence, nous notons certaines données morpho-syntaxiques qui définissent la distribution de ses emplois. Nous retenons ainsi, pour chaque cooccurrence, les types des déterminants, le genre et le nombre de chaque mot, et la position relative de ceux-ci. Ces données déterminent la formulation la plus fréquente de la cooccurrence, qui sera utilisée notamment pour l'affichage.

Au total, plus de 17 millions de cooccurrences distinctes ont été extraites, chacune apparaissant en moyenne 4 fois dans le corpus.

3.3 Sélection des cooccurrences

3.3.1 Sélection automatique

Un test statistique permet de faire un premier tri des cooccurrences. Le test le plus simple serait d'utiliser directement les fréquences, en ne retenant que les cooccurrences d'une fréquence minimale donnée. Mais un tel test peut difficilement s'appliquer à l'ensemble des mots d'une langue, car les fréquences varient énormément d'un mot à l'autre. Plusieurs autres *mesures d'association* ont toutefois été proposées pour quantifier la force d'une combinaison de mots, dont le rapport de vraisemblance (*log-likelihood ratio*), l'information mutuelle, le test *t* et le test du khi-carré. Pour une description approfondie de diverses mesures d'association, voir (Evert, 2005), Manning & Schütze (1999) ou Dunning (1993).

Selon Dunning (1993), l'information mutuelle et le test *t* ont tendance à surestimer la force des combinaisons de faible fréquence ou dont un des composants est rare. En revanche, le rapport de vraisemblance s'appuie sur des fondements statistiques solides pour comparer directement l'importance d'évènements rares et d'évènements fréquents. De plus, selon Orliac (2004), il s'agit de la mesure la plus apte à isoler les collocations d'un ensemble de combinaisons. Pour ces raisons, nous avons choisi le rapport de vraisemblance comme mesure de la force de nos cooccurrences.

Nous l'avons mentionné à la section 3.1, la diversité du corpus a un impact direct sur les cooccurrences extraites. Lors de nos premiers essais, certaines cooccurrences à priori peu intéressantes mais présentant une valeur de force anormalement élevée provenaient souvent d'un seul et même texte ou site Web. Nous avons dû élaborer une euristique pour tenir compte de la dispersion d'une combinaison à travers plusieurs sources. Plutôt que la fréquence brute, nous employons la somme des racines carrées des fréquences pour chaque source. Ainsi, une combinaison apparaissant 9 fois dans une même source aura le même poids que si elle apparaissait une fois dans 3 sources distinctes.

Après avoir calculé la force des combinaisons candidates, nous avons filtré les moins intéressantes en fixant empiriquement un seuil pour chaque type de relation syntaxique. Cet ensemble de seuils a ainsi retranché 93,5 % des 17 millions de cooccurrences extraites. De

plus, les cooccurrences de fréquence 1 (apax) et celles formées de mots banals, comme le verbe « être » et certains adverbes (« très », « trop »…), ont été filtrées automatiquement, retirant un autre 1,5 %. Au total, la sélection automatique retient donc un peu plus de 850 000 cooccurrences « brutes ».

3.3.2 Révision manuelle

Les cooccurrences brutes ont ensuite fait l'objet d'une révision « manuelle » par des linguistes, afin de vérifier la qualité des résultats et de rejeter les cooccurrences jugées indésirables. Ces rejets concernent :

- Des cooccurrences mal analysées : des phrases complexes peuvent donner du fil à retordre à l'analyseur et lui faire générer des analyses incorrectes. Par exemple, *président du trésor*, mauvaise analyse de *président du Conseil du trésor*.

- Des cooccurrences incomplètes : l'extracteur ne considérant que deux mots principaux par cooccurrence, il lui arrive de générer une cooccurrence incomplète comme *emploi à temps*, à laquelle il manque un adjectif essentiel (*plein* ou *partiel*). Parfois, le linguiste peut y remédier en insérant manuellement certains éléments récurrents (*nager en délire > nager en plein délire*) ou paires binaires de tels éléments (*blague de gout > blague de bon/mauvais gout*).

- Des cooccurrences peu intéressantes : certaines cooccurrences offrent peu d'intérêt pour un dictionnaire malgré leur relative fréquence. Par exemple, des combinaisons libres avec un gentilé (*ambassadeur américain*, *français*, etc.), ou un verbe de sens très générique (*avoir un chien*, *un livre*, etc.).

- Des cooccurrences délicates : cooccurrences de caractère offensant, expressions de registre vulgaire, anglicismes condamnés par Antidote, etc.

Un autre objet de la révision humaine est de vérifier la formulation choisie de manière automatique et de rectifier, le cas échéant, certains attributs morphosyntaxiques. Par exemple, le linguiste peut remplacer un article par un possessif (*rencontrer sur le passage > rencontrer sur son passage*), insérer une négation (*la pluie discontinue > la pluie ne discontinue pas*), modifier le genre (*abandonné par son mari > abandonnée par son mari*), etc.

Après huit mois de révision, 5 % seulement des cooccurrences ont ainsi été rejetées manuellement. Nous obtenons donc 800 000 cooccurrences qui constituent le dictionnaire final.

3.4 Choix des exemples

Nous avons choisi d'illustrer chaque cooccurrence par des exemples réels tirés du corpus. Un algorithme vorace sélectionne un ensemble minimal de phrases qui serviront d'exemples. L'algorithme s'efforce de minimiser la longueur totale des phrases tout en maximisant leur « qualité ». La qualité d'une phrase tient compte de plusieurs paramètres, dont :

- sa longueur (ni trop courte ni trop longue) ;

- la qualité de la source d'où elle provient ;

- le nombre de fautes identifiées par l'analyseur ;
- le nombre de noms propres (l'idée étant de minimiser les noms propres).

L'algorithme tente aussi de maximiser la diversité des sources, pour éviter de choisir, pour une même cooccurrence, plusieurs phrases du même ouvrage, du même auteur, ou du même site Web. En revanche, l'algorithme essaie de réutiliser le plus possible une même phrase pour illustrer plusieurs cooccurrences distinctes, afin de minimiser la taille totale des données. Enfin, les phrases trop similaires pour une même cooccurrence sont identifiées et coupées.

Parmi 11 millions de phrases candidates, 870 000 phrases sont ainsi sélectionnées, représentant plus de 2 millions d'exemples.

De ce nombre, 300 000 phrases ont été identifiées automatiquement pour être révisées par des linguistes, dans le but de rejeter les phrases jugées indésirables selon certains critères, parmi lesquels :

- Présence d'erreurs non détectées par l'analyseur ; phrases de mauvaise qualité, trop « orales », parsemées d'anglais, en ancien français, etc.
- Présence de mots offensants, inconvenants, vulgaires ; propos non neutres sur des sujets délicats.
- Mention d'une personne non publique ; présence de coordonnées, de numéros de téléphone, etc. ; phrases à caractère publicitaire.

Environ le quart des phrases révisées ont été ainsi rejetées, puis remplacées et révisées à nouveau, en un processus itératif de sélection automatique et de révision manuelle.

4 Présentation des cooccurrences

Présenter efficacement une masse de plusieurs centaines de cooccurrences à un utilisateur non spécialisé présente plusieurs défis. La Figure 1, ouverte sur le début des cooccurrences du mot *grève*, donne un aperçu des choix que nous avons faits et que nous discutons ci-dessous.

Nous estimons que des cooccurrences complètes ont plus d'impact et sont plus faciles à lire que celles où le mot-vedette est omis ou remplacé par un symbole. De plus, nous jugeons qu'il est utile de présenter les cooccurrences avec leurs attributs morphosyntaxiques réels. Nous affichons donc *déclencher une grève*, *mettre fin à la grève*, *voter en faveur d'une grève*, ce qui correspond aux formulations les plus fréquentes. D'autres formulations ont pu être rencontrées dans le corpus (*déclencher la grève*, *déclencher cette grève*, *les grèves ont été déclenchées*…) ; les exemples d'occurrences peuvent donner un aperçu de cette variabilité. Lorsqu'une cooccurrence est sélectionnée, des exemples de phrases tirées du corpus s'affichent en effet dans le panneau de droite, avec surlignement des mots de la cooccurrence.

Dans le cas d'un mot polysémique, comme *grève*, les cooccurrences sont regroupées sous chacun des sens, affichés en vert, afin de faciliter la consultation[2]. Un triangle de dévoilement permet de réduire les listes sous chaque sens afin d'accéder rapidement au sens désiré.

Les cooccurrences sont ensuite classées par relation syntaxique. Un histogramme discret illustre la force relative de chaque cooccurrence. Au départ, seules les cinq cooccurrences les plus fortes de chaque relation sont affichées, pour donner une meilleure vue d'ensemble. Le lien *et x autres,* à la fin de chaque liste, permet de dévoiler la suite. Un triangle de dévoilement permet de réduire chacune des listes individuellement.

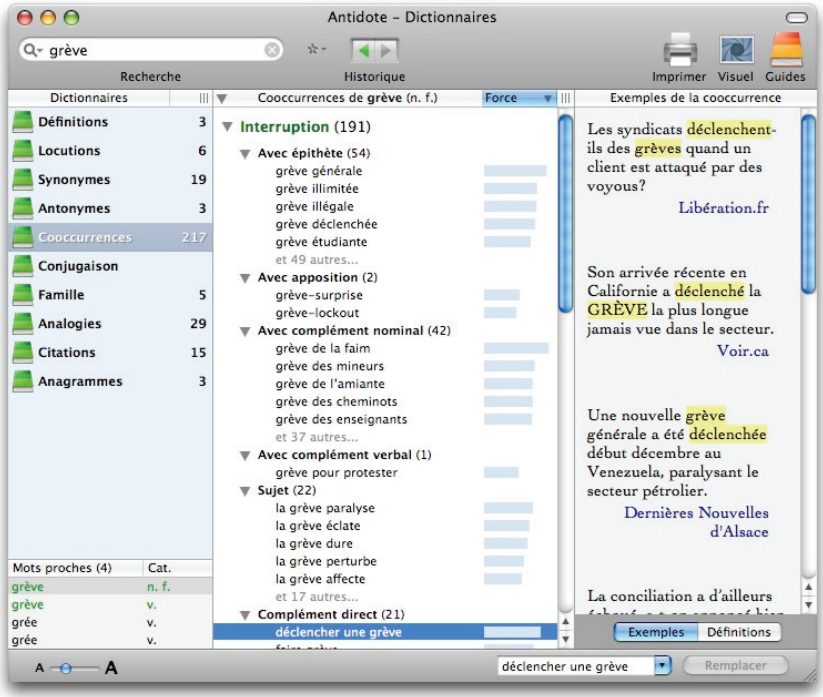

Figure 1: interface du dictionnaire de cooccurrences d'Antidote RX

[2] Pour l'affichage, les cooccurrences dont les bases sont polysémiques ont été manuellement classées sous les divers sens de la base (environ 4000 sens sous 900 bases). Par exemple, les cooccurrences de grève au sens d'« interruption de travail » ont été séparées de celles de grève au sens de « plage ».

5 Utilisation dans le correcteur

Le dictionnaire de cooccurrences d'Antidote RX n'est pas seulement un outil de consultation, il est aussi utilisé par le correcteur pour raffiner l'analyse syntaxique, détecter et corriger des fautes sémantiques et éliminer certaines alertes inutiles.

Les cooccurrences guident l'analyseur en lui permettant d'éliminer des branchements syntaxiques moins probables si un branchement plus fort est détecté. Ce faisant, l'analyseur purement symbolique devient un analyseur hybride intégrant aussi des notions statistiques. Cette avancée a permis de réduire d'environ 10 % le nombre d'arbres syntaxiques produits, augmentant du même coup la vitesse d'analyse.

Les erreurs de nature sémantique, comme **tache ingrate* ou **perpétrer la tradition*, ne peuvent être repérées par la seule analyse syntaxique. Mais comme ces erreurs proviennent souvent de la confusion entre homonymes ou paronymes, il devient possible, en consultant les cooccurrences, de déterminer si un homonyme ou un paronyme est statistiquement plus probable et de proposer la correction le cas échéant. Antidote peut corriger 25 000 confusions de ce type.

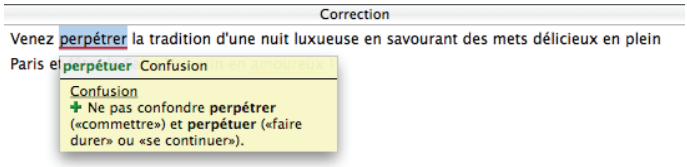

Figure 2 : correction d'une erreur de nature sémantique

Inversement, si le correcteur trouve deux mots en relation de cooccurrence forte, il peut taire certaines alertes de nature sémantique qu'il aurait autrement générées. Par exemple, sur un mot qui peut être un faux ami (ou anglicisme sémantique), comme *digital* au sens de *numérique*, le correcteur n'alertera pas l'utilisateur si le mot est employé dans un contexte de cooccurrence forte, comme *empreinte digitale*.

6 Conclusion

Nous avons vu comment a été élaboré un dictionnaire de cooccurrences commercial, au moyen d'outils de traitement automatique de la langue (TAL) à couverture large appliqués à un vaste corpus du français. Nous avons vu également comment l'analyseur syntaxique utilisé pour extraire les cooccurrences en a lui-même tiré profit pour améliorer ses performances d'analyse et de correction.

Le dictionnaire de cooccurrences d'Antidote RX est aujourd'hui entre les mains de milliers d'utilisateurs. Plusieurs ont manifesté leur satisfaction, et même noté que, de tous les outils d'Antidote RX, le dictionnaire de cooccurrences est déjà devenu celui qu'ils consultent le plus souvent. D'autres ont été agréablement surpris par une correction « sémantique » pertinente. Ces résultats montrent que le grand public peut profiter pleinement des avancées du TAL, et qu'il est prêt à accueillir de nouveaux outils linguistiques avancés et inédits.

Remerciements

Nous tenons à remercier Mala Bergevin, Jean Saint-Germain, Jasmin Lapalme, Marie-Hélène Gaudreault, Sophie Campbell, Sara-Anne Leblanc, Ophélie Tremblay, Guy Lapalme, Alain Polguère et toute l'équipe des druides sans qui Antidote RX n'aurait pu voir le jour.

Références

DUNNING, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, 61-74.

EVERT S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

KILGARRIFF A., TUGWELL D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. *Proc. Collocations workshop, ACL 2001, Toulouse*, 32-38.

LIN, D. (1998). Extracting Collocations from Text Corpora. *First Workshop on Computational Terminology, COLING-ACL '98, Montréal*, 57-63.

MANNING C., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge : The MIT Press.

ORLIAC B. (2004). *Automatisation du repérage et de l'encodage des collocations en langue de spécialité.* Thèse de doctorat présentée à l'Université de Montréal.

SERETAN V., WEHRLI E. (2006). Accurate collocation extraction using a multilingual parser. *Proceedings of COLING-ACL 2006, Sydney, Australia*, 953-960.

TUTIN A. (2005). Le dictionnaire de collocations est-il indispensable ? *Revue française de linguistique appliquée,* X-2, 31-48.

Dictionnaires

BEAUCHESNE, J. (2001). *Dictionnaire des cooccurrences.* Montréal : Guérin.

LACROIX, U. (1931). *Les mots et les idées. Dictionnaire des termes cadrant avec les idées.* Paris/Bruxelles.

MEL'ČUK I. *et coll.* (1984, 1988, 1992, 2000). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I, II, III, IV*. Montréal : Les Presses de l'Université de Montréal.

MEL'ČUK I., POLGUÈRE A. (Sous presse). *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Louvain-la-Neuve : Duculot.

ZINGLÉ H., BROBECK-ZINGLÉ M.-L. (2003). *Dictionnaire combinatoire du français. Expression, locutions et constructions.* Paris : La Maison du Dictionnaire.

DiCouèbe, dictionnaire en ligne de combinatoire du français : http://olst.ling.umontreal.ca/dicouebe

Dictionnaire des collocations : http://www.tonitraduction.net

*Les voisins de Le Mond*e : http://w3.univ-tlse2.fr/erss/voisinsdelemonde