TALN 2010, Montréal, 19-23 juillet 2010

Au-delà de la paire de mots : extraction de cooccurrences syntaxiques multilexémiques

Simon CHAREST, Éric BRUNELLE, Jean FONTAINE

Druide informatique inc.
1435, rue Saint-Alexandre, bureau 1040
Montréal (Québec) H3A 2G4, Canada
developpement@druide.com

Résumé

Cet article décrit l'élaboration de la deuxième édition du dictionnaire de cooccurrences du logiciel d'aide à la rédaction Antidote. Cette nouvelle mouture est le résultat d'une refonte complète du processus d'extraction, ayant principalement pour but l'extraction de cooccurrences de plus de deux unités lexicales. La principale contribution de cet article est la description d'une technique originale pour l'extraction de cooccurrences de plus de deux mots conservant une structure syntaxique complète.

Abstract

This article describes the elaboration of the second edition of the co-occurrence dictionary included in Antidote HD, a commercial software tool for writing in French. This second edition is the result of a complete overhaul of the extraction process, with the objective of extracting co-occurrences of more than two lexical units. The main contribution of this article is the description of an original method for extracting co-occurrences of more than two words retaining their full syntactic structure.

Mots-clés : Antidote, cooccurrences, collocations, expressions multimots.
Keywords: Antidote, co-occurrences, collocations, multi-word expressions (MWE).

1 Introduction

En 2006 paraissait avec Antidote RX[1] le premier dictionnaire électronique grand public de cooccurrences[2] du français (Charest et coll., 2007). Ce dictionnaire a été élaboré en utilisant l'analyseur syntaxique d'Antidote pour extraire, d'un corpus de 500 millions de mots, 17 millions de paires de mots différentes liées par diverses relations syntaxiques. Un score correspondant au rapport de vraisemblance (*log-likelihood ratio*) avait été calculé sur chacune de ces cooccurrences candidates afin d'en dégager les plus significatives. Au final, après validation par une équipe de linguistes, 800 000 cooccurrences avaient été retenues, réparties en 36 000 entrées, classées par sens et relations syntaxiques et illustrées par 880 000 exemples tirés du corpus.

Parmi les limitations de ce dictionnaire, nous avions noté la présence de cooccurrences incomplètes : l'extracteur ne considérant que les associations de deux mots, il lui arrivait de générer des combinaisons incomplètes, auxquelles il manque un élément essentiel, comme dans *prendre le taureau* ___, *nager en* ___ *délire* ou ___ *le feu aux poudres*. Ces cooccurrences avaient été soit rejetées par le linguiste, soit augmentées manuellement par l'ajout du mot manquant.

Pour la deuxième édition de notre dictionnaire de cooccurrences, parue en 2009 avec Antidote HD, nous avons revu notre processus d'extraction afin d'extraire automatiquement des cooccurrences de plus de deux mots. Le présent article décrit la démarche entreprise pour arriver à ce résultat.

2 Travaux antérieurs

La plupart des travaux sur les collocations (en anglais *multi-word expression [MWE]*) ont porté sur les combinaisons de deux mots. Tutin (2008) analyse les collocations de plus de deux mots et conclut qu'elles peuvent le plus souvent être considérées comme des combinaisons binaires respectant une structure prédicat-argument. Selon elle, la majorité des collocations *n*-aires peuvent être analysées comme des superpositions de collocations ou comme des collocations récursives, tandis que quelques cas sont de vraies collocations *n*-aires, où la séquence ne peut pas être décomposée.

Dans la tâche d'extraction de collocations à partir de corpus, les mesures statistiques d'association employées pour quantifier le degré de cohésion – et donc de pertinence – des combinaisons candidates sont pour la plupart formulées pour évaluer les associations de deux éléments (ou bigrammes). De nombreuses études ont comparé les performances de diverses mesures d'association. Dans l'une des plus ambitieuses, Pecina et Schlesinger (2006) ont évalué 82 mesures dans une tâche d'extraction de collocations en tchèque, et ont combiné les plus performantes à l'aide de techniques d'apprentissage.

Les choses se compliquent lorsque vient le temps de mesurer la force d'association de combinaisons de plus de deux mots. Deux grandes approches sont possibles : 1) employer des mesures statistiques plus

[1] Logiciel d'aide à la rédaction grand pulic (www.antidote.info), sixième édition.

[2] Par cooccurrence, nous entendons la présence simultanée et statistiquement significative, dans un corpus, de deux unités linguistiques ou plus en relation syntaxique. 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 coeur*) ou terminologiques (*cri primal*).

complexes formulées pour évaluer les associations d'un nombre arbitraire d'éléments; ou 2) employer des mesures d'association de bigrammes sur des groupes de mots plutôt que sur des mots individuels.

La première approche a été suivie par Dias et coll. (2000), qui ont proposé une normalisation de quatre mesures d'association courantes (le coefficient de Dice, l'information mutuelle spécifique, Φ^2 et le rapport de vraisemblance) afin de pouvoir calculer la cohésion de *n*-grammes pour $n >= 2$. La normalisation est effectuée en considérant toutes les façons possibles de diviser un *n*-gramme en 2 sous-groupes (créant ainsi des « pseudobigrammes »), et en prenant la moyenne arithmétique des fréquences de toutes les combinaisons possibles. Ils ont en outre introduit une nouvelle mesure, l'expectative mutuelle, calculée selon le même principe et donnant, selon leurs tests, de meilleurs résultats que les mesures classiques.

Dans la même veine, Blaheta et Johnson (2001) ont employé un modèle log-linéaire pour l'extraction de locutions verbales en anglais. Les modèles log-linéaires permettent de mettre en évidence des interactions complexes entre plusieurs variables. Villada Moiron (2005) a comparé des modèles bigrammes et trigrammes pour l'extraction de locutions verbales et prépositionnelles en néerlandais. Parmi les modèles trigrammes, elle a testé deux mesures (l'information mutuelle spécifique et le test du Chi-carré de Pearson) qu'elle a étendues aux trigrammes, de même que le modèle log-linéaire de Blaheta et Johnson (2001). D'après ses tests, le modèle bigramme utilisant le rapport de vraisemblance a donné les meilleurs résultats.

À propos du rapport de vraisemblance, il est intéressant de noter qu'il est en fait mathématiquement très similaire à l'information mutuelle[3]. Dans la formulation de Dunning (1993), reprise dans Manning et Schütze (1999), le rapport de vraisemblance est exprimé comme le rapport entre les probabilités de deux hypothèses, modélisant respectivement l'indépendance et la dépendance de deux évènements. Comme telle, la formule s'applique mal à plus de deux évènements. Mais, comme le mentionnent Moore (2004) et Evert (2005), le rapport de vraisemblance (G^2) peut être reformulé, par une série de dérivations, sous une forme presque équivalente à celle de l'information mutuelle de deux variables aléatoires (*M.I.*).

Équation 1 : information mutuelle de deux variables aléatoires

$$M.I. = \frac{1}{N}\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\widehat{f}_{ij}}$$

*Équation 2 : rapport de vraisemblance (*log-likelihood ratio*) pour deux évènements*

$$G^2 = 2\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\widehat{f}_{ij}}$$

L'avantage de cette reformulation est qu'elle peut s'étendre aux *n*-grammes de degré supérieur. Cependant, avec $n >= 3$, il existe plus d'une façon de modéliser l'hypothèse nulle (Banerjee et Pedersen, 2003). Par exemple, pour $n = 3$, on peut considérer le modèle où les trois mots sont mutuellement indépendants, mais aussi trois autres modèles où une paire de mots est dépendante, mais indépendante du troisième.

Le nombre de modèles à considérer, de même que le nombre de fréquences marginales à calculer, augmente rapidement avec *n*, ce qui rend l'approche *n*-gramme plus complexe. C'est pourquoi plusieurs auteurs ont préféré le modèle bigramme.

---

[3] À ne pas confondre avec l'« information mutuelle spécifique » (*PMI*, *pointwise mutual information*), qui est une mesure de dépendance entre deux événements X = x et Y = y. En théorie de l'information, l'information mutuelle mesure la dépendance de deux variables aléatoires X et Y, et correspond à la valeur espérée des PMI sur l'ensemble de toutes les éventualités de X et Y.

Mettant à contribution l'analyseur syntaxique profond *Fips*, Seretan et coll. (2003) décrivent une méthode itérative pour extraire des cooccurrences multimots. Tous les bigrammes candidats sont d'abord extraits, puis, dans un processus itératif, des (n + 1)-grammes sont construits à partir de n-grammes partageant n - 1 mots. Les n-grammes ainsi combinés sont exclus des résultats, la procédure ne conservant que les combinaisons les plus longues. Le rapport de vraisemblance est employé comme mesure de pertinence des combinaisons extraites. Pour un n-gramme de plus de 2 mots, ce score est calculé à partir des fréquences des (n - 1)-grammes le composant. Par rapport à des techniques d'extraction basées sur des patrons syntaxiques, leur méthode a l'avantage d'être souple, car aucune restriction n'est imposée sur la structure syntaxique des combinaisons extraites. De plus, des relations syntaxiques éloignées sont prises en compte.

L'approche que nous avons employée se rapproche de celle de Seretan et coll. (2003). Comme eux, nous employons un analyseur syntaxique. Nous avons aussi choisi de nous en tenir au modèle bigramme, pour sa simplicité. La prochaine section décrit la méthodologie que nous avons mise au point pour extraire les cooccurrences multimots.

3 Méthodologie

3.1 Analyse syntaxique

De 500 millions de mots pour la première édition, nous avons augmenté le corpus à 1,8 milliard de mots ou 92 millions de phrases. Ces phrases ont été analysées par le moteur syntaxique d'Antidote, qui effectue une analyse en dépendance et génère des arbres syntaxiques complets. Lorsque plusieurs analyses sont trouvées pour une même phrase, l'arbre le plus probable, selon la pondération de l'analyseur, est choisi. Un arbre syntaxique complet correspond en fait à un ensemble de liens de dépendance entre les divers mots de la phrase. Chaque lien unit un gouverneur (que nous appelons mère) à un dépendant (que nous appelons fille), au moyen d'une relation syntaxique. Les liens correspondant aux relations syntaxiques les plus pertinentes (sujet, COD, COI, épithète, complément du nom, etc.) sont alors extraits de ces arbres et entreposés dans une base de données. Au total, 440 millions de liens sont ainsi extraits, correspondant à 60 millions de combinaisons <mère – relation – fille> différentes, chaque combinaison apparaissant en moyenne 7,3 fois dans le corpus. Par exemple, prenons la phrase « Le département joue un rôle de premier plan. ». L'arbre produit par l'analyseur syntaxique, ainsi que les relations pertinentes qui en sont extraites, sont illustrés à la Figure 1.

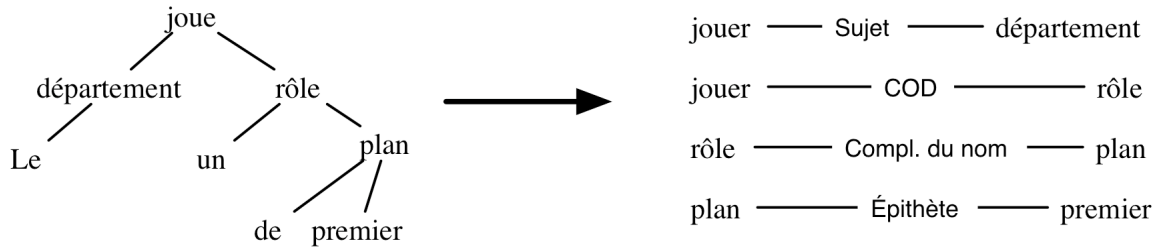

Figure 1 : extraction des relations pertinentes

Au-delà des relations syntaxiques directes, le système s'efforce d'extraire des relations plus profondes. Ainsi, on substitue à un pronom relatif son antécédent afin d'extraire la relation entre l'antécédent et le verbe de la relative (*la femme qu'il a séduite → séduire une femme*). Lorsqu'un nom agit comme agent

d'un verbe à l'infinitif, ce lien est transformé en relation « sujet » (*il incombe au juge de trancher → le juge tranche*). Lorsqu'une relation met en jeu un nom collectif ayant un complément, une relation directe avec le complément est aussi générée (*un groupe d'étudiants revendique → les étudiants revendiquent*).

3.2 Inférence des cooccurrences intéressantes

L'idée principale de notre approche est la suivante : plutôt que d'utiliser, comme base de nos calculs statistiques, des combinaisons de deux mots simples, nous travaillons avec des combinaisons de deux arbres. Comme pour les combinaisons de mots, nous avons une mère et une fille, unies par une relation syntaxique. Or, si chaque occurrence d'une relation correspond à une seule combinaison de mots, elle peut correspondre à plusieurs combinaisons d'arbres. Par exemple, à la Figure 2, pour la relation entre B et D, nous obtenons 8 combinaisons différentes, selon les branches que l'on inclut sous B et D.

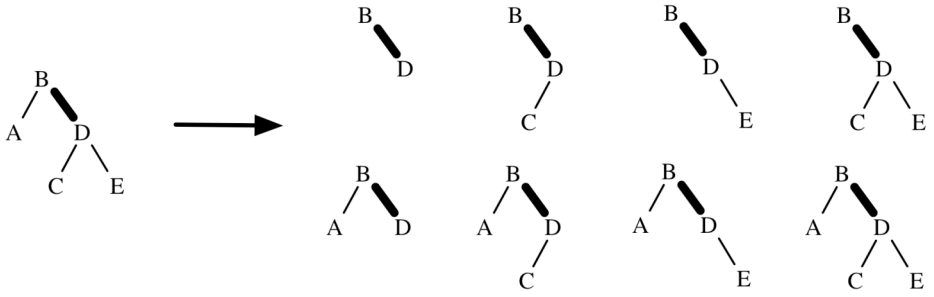

Figure 2 : toutes les combinaisons d'arbres pour une relation donnée

Notre approche considère toutes les combinaisons possibles de tous les sous-arbres ayant pour tête chacun des deux mots liés par une relation de dépendance. En théorie, la taille des arbres générés n'est limitée que par la longueur et la structure des phrases du corpus. En pratique, nous avons décidé de limiter les arbres considérés à 5 mots maximum, pour raccourcir le temps de calcul et parce que les combinaisons trop longues ont tendance à être moins intéressantes linguistiquement. Les 60 millions de combinaisons de mots différentes trouvées dans le corpus correspondent à 1,1 milliard de combinaisons d'arbres de longueur 5 ou moins, chaque combinaison d'arbres apparaissant en moyenne 1,5 fois.

3.3 Calcul de la force en tenant compte de la dispersion

Comme pour les combinaisons de mots, nous calculons la force des combinaisons d'arbres à l'aide de la version bigramme du rapport de vraisemblance. Nous avons fixé empiriquement un seuil, de façon à ne conserver que les combinaisons les plus fortes. Nous excluons aussi les combinaisons ayant une fréquence inférieure à 4, car les mesures d'association sont moins fiables lorsque les fréquences sont très faibles. En effet, Evert (2005) mentionne que les estimations de probabilités sont biaisées pour les évènements de faible fréquence, et suggère fortement de toujours exclure les données de fréquence 1 ou 2. Le biais diminue progressivement à mesure que monte la fréquence. À partir d'une fréquence de 5, le biais devient minime et les mesures d'association se comportent bien.

Nous avions remarqué, pour la première édition de notre dictionnaire, que certaines cooccurrences ayant une fréquence étonnamment élevée provenaient en fait d'un seul site Web (ou d'un nombre restreint de sites) traitant d'un sujet particulier. Ces cooccurrences sont souvent inintéressantes linguistiquement, car

elles ne sont pas le reflet de la langue en général, mais d'une terminologie associée à un contexte spécifique. Comme nous l'avions fait alors, nous avons adapté le calcul de la force pour tenir compte de la dispersion d'une combinaison à travers plusieurs sources, et ainsi réduire l'effet des cooccurrences trop dépendantes du choix du corpus. À l'époque, nous avions implémenté un calcul de fréquence pondérée qui correspondait à la somme des racines carrées des fréquences pour chaque source (Charest et coll., 2007). Or, cette technique, en plus de ne pas être fondée mathématiquement, avait le désavantage de compliquer le calcul des fréquences, en particulier pour la phase de l'élagage (voir section 3.4).

Cette fois-ci, nous employons comme mesure de dispersion l'écart de proportions (*Deviation of Proportions [DP]*) de Gries (2008). Cette mesure, qui allie simplicité et flexibilité, se calcule en prenant la somme des différences entre les proportions espérée et observée d'un phénomène dans chacune des sources par rapport à l'ensemble du corpus.

Équation 3 : écart de proportions (DP)

$$DP = \frac{\sum_i \left| \frac{f_i}{\sum_i f_i} - \frac{t_i}{\sum_i t_i} \right|}{2} \qquad dispersion = 1 - DP$$

où f_i est la fréquence d'une combinaison dans la source i et t_i est la taille de la source i. Gries (2008) mentionne que la plupart des mesures de dispersion sont appliquées directement sur la fréquence pour obtenir des fréquences dites ajustées; cependant, il n'est pas clair pour lui ce que représentent réellement ces pseudofréquences. Plutôt que de modifier les fréquences, nous avons donc choisi de multiplier directement la valeur de la force (calculée par le rapport de vraisemblance avec les fréquences brutes) par la valeur de dispersion. Cette façon de faire, plus simple, nous apparait tout aussi valable.

Enfin, mentionnons que le score calculé correspond à la relation entre la tête d'un l'arbre syntaxique et l'une de ses filles. Lorsque la tête a plus d'une fille, chaque relation génère un score. Par exemple, dans *prendre ses jambes à son cou*, on a deux filles branchées sur la même mère. Ces deux branchements correspondent aux interprétations [prendre ses jambes] + [à son cou] et [prendre à son cou] + [ses jambes]. Pour le besoin du filtrage statistique des cooccurrences, un seul score suffit. Nous choisissons alors le plus élevé comme score de la cooccurrence.

3.4 Élagage

Certaines cooccurrences de plus de deux mots ne sont pas naturellement décomposables. Par exemple, même si les fréquences de *prendre le taureau* et *prendre par les cornes* sont nécessairement au moins aussi élevées que la fréquence de *prendre le taureau par les cornes*, seule cette dernière est intéressante.

Pour élaguer ces combinaisons moins intéressantes, Silva et coll. (1999) ont introduit un algorithme nommé LocalMaxs, qui sélectionne les combinaisons de n mots dont le score est à la fois supérieur ou égal à celui de toutes les combinaisons de n - 1 mots incluses dans celles-ci et supérieur à celui de toutes les combinaisons de n + 1 mots dérivées de celles-ci. Cet algorithme leur évite aussi de devoir fixer un seuil empiriquement : toutes les combinaisons satisfaisant à ces critères sont conservées, peu importe leur score.

De notre côté, nous avons plutôt choisi de réduire la force des cooccurrences incomplètes en modifiant leur fréquence brute. Lorsqu'une cooccurrence complexe a une fréquence relativement élevée par rapport à une cooccurrence plus simple incluse dans celle-ci, on recalculera le score de la cooccurrence plus simple en prenant une fréquence modifiée qui ne tient pas compte des occurrences de la cooccurrence complexe. Ainsi, pour *prendre le taureau*, on comptera seulement les occurrences où cette combinaison n'est pas accompagnée du complément *par les cornes*. Par contre, si aucune cooccurrence plus complexe n'a une fréquence assez élevée par rapport à une cooccurrence simple, sa fréquence n'est pas réduite. Par exemple, la fréquence de *jouer un rôle* ne sera pas affectée par *jouer un rôle important*, *jouer un grand rôle*, *jouer un rôle de premier plan*, etc.

3.5 Détermination des attributs de forme

Outre les mots formant la cooccurrence, nous notons certaines données morphosyntaxiques qui définissent la distribution de ses emplois. Nous retenons ainsi, pour chaque cooccurrence, les types des déterminants, le genre, le nombre et la casse de chaque mot, la position relative de ceux-ci, la présence de clitiques (*en*, *y*), de pronoms réfléchis (*se, s'*) ou de particules négatives autour des verbes, etc. Ces données déterminent la formulation la plus fréquente de la cooccurrence, qui sera utilisée pour générer la forme finale de la cooccurrence au moment de son affichage.

3.6 Révision manuelle

Les cooccurrences qui ne faisaient pas partie de la première édition du dictionnaire ont toutes été révisées par une équipe de linguistes. Les cooccurrences incomplètes, inintéressantes, délicates (offensantes, vulgaires, etc.), fautives (impropriétés, calques, pléonasmes, etc.) ou résultant d'une mauvaise analyse syntaxique ont été retranchées. Comme on pouvait s'y attendre, les mauvaises analyses sont plus fréquentes pour les cooccurrences de plus de deux mots. Même si une forme peut sembler correcte en surface, lorsque l'on examine en détail la structure syntaxique, on découvre parfois un mauvais branchement ou la substitution d'un mot par son homographe. Dans ce dernier cas, un mécanisme permet au linguiste de corriger la situation en transférant la cooccurrence sur le bon mot. Le linguiste a aussi la possibilité de modifier les attributs des mots de la cooccurrence qui ont été choisis automatiquement (les déterminants, la flexion, la casse, etc.). Enfin, le linguiste a aussi pour tâche de classer les cooccurrences des mots polysémiques selon leurs divers sens.

4 Résultats

Du 1,1 milliard de combinaisons d'arbres identifiées par le système, 4,2 millions sont à la fois assez fréquentes et assez fortes pour être conservées dans notre base de données. Ces cooccurrences se répartissent comme suit : 49,0 % sont de 2 mots, 42,1 % de 3 mots, 7,5 % de 4 mots et 1,4 % de 5 mots. À ce jour, 952 000 cooccurrences ont été revues par un linguiste, en procédant par ordre décroissant de force. Parmi celles-ci, 55 000 (5,8 %) ont été rejetées, car inintéressantes; 25 000 (2,6 %) ont été identifiées comme étant de mauvaises analyses; 10 000 (1,1 %) ont été jugées incomplètes. Au total, 862 000 cooccurrences, dont 783 000 de deux mots (90,8 %)[4], ont été retenues et font partie de la présente édition.

[4] La proportion de cooccurrences de deux mots est plus élevée pour les cooccurrences retenues (90,8 %) que pour les cooccurrences brutes (49,0 %), car la majorité d'entre elles faisait partie de la première édition et a été conservée.

5 Présentation des cooccurrences

Nous avons maintenu les principes de base que nous avions déterminés pour la présentation des cooccurrences de la première édition et qui ont fait leurs preuves : affichage des cooccurrences complètes avec leurs attributs morphosyntaxiques réels, histogramme illustrant la force, regroupement par sens puis par contexte syntaxique, affichage initial des 5 premières cooccurrences de chaque groupe avec possibilité d'afficher la suite en un clic, exemples réels tirés du corpus présentés dans un panneau secondaire.

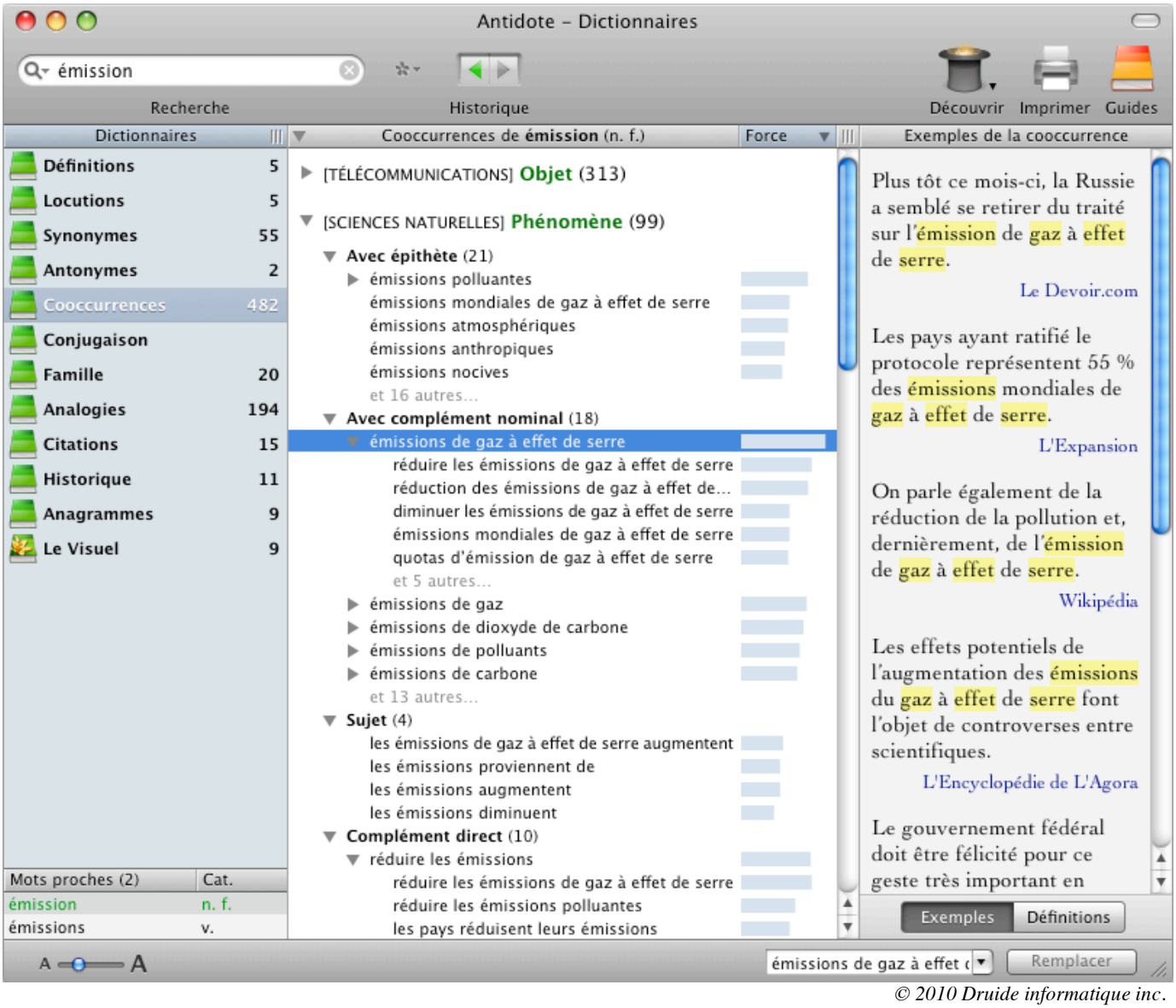

Figure 3 : interface du dictionnaire de cooccurrences d'Antidote HD

La seule adaptation que nous avons effectuée concerne l'affichage des cooccurrences de plus de deux mots. Nous avons choisi de regrouper celles-ci sous une cooccurrence plus simple s'il en existe une dont la force n'est pas trop faible par rapport à la force des cooccurrences plus complexes. Par exemple, à la Figure 3, nous avons, sous le deuxième sens de l'entrée *émission*, à la relation « avec complément nominal », la cooccurrence de 4 mots *émissions de gaz à effet de serre*, dont la force est très élevée. Sous elle, diverses cooccurrences de 5 mots comme *réduire les émissions de gaz à effet de serre*. Cette dernière

cooccurrence se retrouve aussi à la relation « complément direct », sous la cooccurrence de 3 mots *réduire les émissions*. En effet, le mot-vedette *émission* est à la fois complément direct de *réduire* et mère du complément nominal *de gaz à effet de serre*. Notons que la cooccurrence de 4 mots *émissions de gaz à effet de serre* n'est pas classée sous la cooccurrence plus simple *émission de gaz*, car sa force est significativement plus élevée que celle de cette dernière.

6 Conclusion et perspectives

Nous venons de présenter une technique originale pour l'extraction de cooccurrences de plus de deux mots conservant une structure syntaxique complète. Notre approche se situe dans la même veine que celle de Seretan et coll. (2003), en ce sens qu'elle allie la sortie d'une analyse syntaxique profonde à un test statistique appliqué sur deux groupes de mots, pour dégager des cooccurrences saillantes de plusieurs mots pour lesquelles aucune restriction de structure syntaxique n'est imposée. Comme la leur, notre approche se démarque donc des méthodes d'extraction multimot basées sur du texte brut ou simplement annoté grammaticalement, qui par nature sont beaucoup plus restreintes syntaxiquement (Seretan, 2008).

Notre approche se distingue par contre à plusieurs points de vue. D'abord, la méthode que nous proposons n'est pas récursive ni itérative : les cooccurrences plus complexes ne sont pas construites à partir de cooccurrences plus simples. Nous avons plutôt choisi l'approche combinatoire : pour une relation syntaxique entre deux mots dans l'analyse d'une phrase, toutes les combinaisons de tous les sous-arbres ayant pour tête chacun des deux mots sont évaluées. Le calcul des fréquences pour la mesure d'association est aussi différent. L'approche récursive de Seretan et coll. (2003) considère une cooccurrence complexe comme étant formée de deux cooccurrences plus simples en intersection, et ce sont les fréquences de ces deux cooccurrences qui sont utilisées dans la table de contingence. De notre côté, les fréquences utilisées sont celles des deux arbres formant la cooccurrence, arbres qui sont nécessairement disjoints.

Parmi les améliorations envisagées, nous aimerions étendre le concept de cooccurrences à des généralisations des arguments des fonctions syntaxiques. Par exemple, la cooccurrence *avoir besoin pour vivre*, telle qu'extraite par notre système, semble incomplète. Elle a dû être augmentée manuellement par un linguiste pour devenir *avoir besoin <u>de qqch.</u> pour vivre*. En généralisant les compléments que prennent les verbes à l'aide de pronoms génériques comme *qqch.* et *qqn*, nous espérons pouvoir inférer ces ajouts automatiquement.

Remerciements

Nous tenons à remercier Mala Bergevin, Jean Saint-Germain, Maud Pironneau et toute l'équipe des druides sans qui Antidote HD n'aurait pu voir le jour.

Références

BANERJEE S., PEDERSEN T. (2003). The Design, Implementation and Use of the Ngram Statistics Package. *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 370–381.

BLAHETA D., JOHNSON M. (2001). Unsupervised learning of multi-word verbs. *Proceedings of the ACL Workshop on Collocations*, 54–60.

CHAREST S., BRUNELLE É., FONTAINE J., PELLETIER, B. (2007). Élaboration automatique d'un dictionnaire de cooccurrences grand public. Actes de la *14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, 283–292.

DIAS G., GUILLORE S., PEREIRA LOPES, J. G. (2000). Normalization of Association Measures for Multiword Lexical Unit Extraction. *International Conference on Artificial and Computational Intelligence for Decision Control and Automation in Engineering and Industrial Applications (ACIDCA 2000)*, 207-216.

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*. Thèse de doctorat, Université de Stuttgart.

GRIES, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics,* vol. 13, no 4, 403–437.

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

MOORE R. C. (2004). On log-likelihood-ratios and the significance of rare events. *Proceedings of the 2004 Conference on EMNLP*, 333-340.

PECINA P., SCHLESINGER P. (2006). Combining association measures for collocation extraction. *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, 651-658.

SERETAN V. (2008). *Collocation extraction based on syntactic parsing*. Thèse de doctorat, Université de Genève.

SERETAN V., NERIMA L., WEHRLI E. (2003). Extraction of Multi-Word Collocations Using Syntactic Bigram Composition. *Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003)*, 424–431.

SILVA J., DIAS G., GUILLORE S., PEREIRA LOPES, J. G. (1999). Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. *Proceedings of 9th Portuguese Conference in Artificial Intelligence*, 21-24.

TUTIN A. (2008). For an extended definition of lexical collocations. *Proceedings of the XIII Euralex International Congress (Barcelona, 15-19 July 2008)*, 1453-1460.

VILLADA MOIRON B. (2005). *Data-driven Identification of fixed expressions and their modifiability*. Thèse de doctorat, Université de Groningue.