CORPUS

Corpus

11 | 2012
La cooccurrence, du fait statistique au fait textuel

La contribution des cooccurrences de deuxième ordre à l'analyse sémantique

Ann Bertels et Dirk Speelman

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2184
DOI : 10.4000/corpus.2184
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 1 janvier 2012
ISSN : 1638-9808

Référence électronique
Ann Bertels et Dirk Speelman, « La contribution des cooccurrences de deuxième ordre à l'analyse sémantique », *Corpus* [En ligne], 11 | 2012, mis en ligne le 18 juin 2013, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/2184 ; DOI : https://doi.org/10.4000/corpus.2184

La contribution des cooccurrences de deuxième ordre à l'analyse sémantique

Ann BERTELS*, Dirk SPEELMAN**
* ILT, KU Leuven, Belgique
** QLVL, KU Leuven, Belgique

1. Introduction

Cet article s'inscrit dans le cadre d'une étude sémantique du vocabulaire spécifique d'un corpus technique. L'étude vise à vérifier si les unités lexicales les plus spécifiques du corpus technique sont effectivement les unités les plus monosémiques dans ce corpus. A cet effet, elle s'appuie sur une double approche quantitative, qui consiste d'une part à identifier les unités lexicales spécifiques et leur degré de spécificité et d'autre part à calculer leur degré de monosémie. Ces données quantitatives font l'objet d'une analyse statistique de régression simple, qui permet d'étudier la corrélation entre le rang de spécificité des unités lexicales spécifiques et leur rang de monosémie. Certaines unités lexicales spécifiques dans le corpus technique sont effectivement des mots à sens multiples. Citons par exemple le mot *découpe* qui signifie (1) « action de découper » et (2) « résultat de la découpe (pièce découpée) ».

Le corpus technique (1,7 million d'occurrences) relève du domaine spécialisé restreint des machines-outils pour l'usinage des métaux. Il a été lemmatisé et étiqueté par Cordial 7 Analyseur et consiste en 4 sous-corpus, datant de 1996 à 2002 : revues électroniques (800.000 occurrences), fiches techniques (300.000), normes ISO et directives (300.000) et 4 manuels numérisés (360.000). Les textes des quatre sous-corpus se situent à différents niveaux de normalisation (normes et manuels) et de vulgarisation (revues et manuels), ce qui garantit la représentativité du corpus technique. Le corpus de référence est

Corpus n°11, *La cooccurrence : du fait statistique au fait textuel* (2012), 147-165

constitué de textes journalistiques (*Le Monde* 1998). Il a aussi été lemmatisé et étiqueté et comprend 15,3 millions d'occurrences. Les fichiers étiquetés par Cordial se composent de trois colonnes, avec une occurrence par ligne : (1) forme fléchie ou forme graphique, (2) lemme ou forme canonique et (3) code Cordial, comparable à un POS-tag et indiquant la classe lexicale.

Le premier volet de la double approche quantitative est consacré à l'analyse des unités spécifiques du corpus technique. A cet effet, plusieurs approches méthodologiques sont envisageables, telles que le calcul des spécificités et l'analyse des mots-clés (Scott & Tribble 2006). Nous optons pour l'analyse des mots-clés, qui s'appuie sur la mesure statistique du log de vraisemblance (*Log-Likelihood Ratio* ou LLR) (Dunning, 1993) et qui est implémentée notamment dans le logiciel AV Frequency List Tool. A partir de deux listes de fréquence des lemmes des deux corpus, réalisées à l'aide de scripts en Python, le logiciel génère une liste de mots-clés ou de lemmes spécifiques du corpus technique. Il indique aussi leur degré de spécificité, la valeur du LLR (*keyness*), et une valeur p associée ($p < 0,05$). Plusieurs études antérieures ont validé la mesure du LLR et démontré la pertinence de l'analyse des mots-clés pour relever le vocabulaire spécifique d'un domaine particulier (Paquot *et al*. 2009 ; Kwary 2011). Pour l'instant, les unités spécifiques du corpus technique sont identifiées au niveau des unités simples (*fraisage, commande, machine*). Des recherches futures devront certainement porter sur les unités polylexicales, telles que *machine à fraiser* ou *commande numérique*, puisque la plupart des unités terminologiques d'un domaine spécialisé se situent au niveau des unités complexes. Le degré de spécificité permet de classer les unités spécifiques par ordre de spécificité décroissante. Après suppression des hapax, des mots grammaticaux et des noms propres, nous recensons 4 717 unités lexicales spécifiques. Dans le deuxième volet de l'étude, elles se voient attribuer un degré de monosémie, ce qui permet d'étudier la corrélation entre spécificité et monosémie.

Dans cet article, nous nous concentrons sur le deuxième volet et plus particulièrement sur la façon dont les cooccurrences de deuxième ordre peuvent être exploitées dans le cadre d'une analyse sémantique quantitative. Nous expliquons d'abord la

méthodologie de l'analyse des cooccurrences et les principes de la mesure de monosémie ou de recoupement (section 2). Dans la section 3, nous discutons quelques expérimentations qui font varier différents paramètres, tels que la fenêtre d'observation et le seuil de significativité. Finalement, nous abordons la question de l'intégration des étiquettes morphosyntaxiques dans l'analyse des cooccurrences (section 4) et nous terminons par les conclusions et perspectives (section 5).

2. Les cooccurrences de deuxième ordre et leur recoupement

2.1 L'analyse des cooccurrences

Les études et analyses cooccurrentielles sont nombreuses et elles se situent dans de nombreux domaines, allant de la lexicologie et lexicographie (Blumenthal & Hausmann 2006) à la terminologie et extraction de termes, en passant par le TAL (traitement automatique des langues) et l'ADT (analyse des données textuelles) (cf. Mayaffre 2008). Par contre, les études faisant intervenir les cooccurrences des cooccurrences, c'est-à-dire les cooccurrences de deuxième ordre ou les cooccurrences indirectes, sont beaucoup moins nombreuses.

Dans notre étude, nous recourons à la réitération de l'analyse des cooccurrences dans le but de quantifier l'analyse sémantique et de déterminer le degré de monosémie des unités lexicales spécifiques du corpus technique. A cet effet, nous proposons d'implémenter la monosémie en termes d'homogénéité sémantique (Ferret 2004 ; Condamines 2005 ; Habert *et al*. 2005). Une unité lexicale monosémique apparaît dans des contextes plutôt homogènes sémantiquement, parce qu'elle se caractérise par des cooccurrents qui appartiennent à des champs sémantiques plutôt similaires. En revanche, une unité lexicale polysémique se caractérise par des cooccurrents plus hétérogènes sémantiquement, qui appartiennent à des champs sémantiques différents (Véronis 2003 ; Habert *et al*. 2004). Habert *et al*. (2004) avancent l'hypothèse qu'un mot à sens multiple « aurait des voisins moins proches entre eux qu'un mot univoque » (Habert *et al.* 2004 : 570). Les mots homonymiques, polysémiques et vagues auraient donc des cooccurrents sémantiquement plus hétérogènes.

Toutefois, en sémantique distributionnelle et contextuelle, deux problèmes majeurs se posent. D'une part, la distribution des différents sens d'un mot dans le corpus est souvent irrégulière et, d'autre part, « la répartition des traits permettant de classer les mots est souvent très éparpillée » (Habert *et al*. 2004 : 573). Pour remédier à ces problèmes, on peut recourir aux cooccurrents de deuxième ordre (Grefenstette 1994). De plus, les cooccurrences de premier ordre sont généralement des cooccurrences syntagmatiques du mot-pôle et parfois des co-occurrences paradigmatiques. Par contre, les cooccurrences de deuxième ordre se caractérisent principalement par des relations paradigmatiques avec le mot-pôle (hyponymes, hyperonymes, synonymes, antonymes).

Jusqu'à présent, l'analyse des cooccurrences de premier et deuxième ordre a été adoptée principalement dans des études de désambiguïsation sémantique (Schütze 1998) et de recherche de similarités sémantiques ou synonymes (Martinez 2000 ; Pezik 2005). Dans notre étude, nous recourons aux cooccurrences de deuxième ordre dans le but de quantifier l'analyse sémantique et de mesurer le degré de monosémie. L'idée de degré de monosémie ou de polysémie est proposée également par Nerlich *et al*. (2003), où elle est exprimée en termes de « polysémie graduée ». Nerlich *et al*. (2003) relèvent des patrons sémantiques flexibles et avancent l'hypothèse que chaque mot est plus ou moins polysémique, avec des sens liés à un prototype par un ensemble de principes relationnels sémantiques, plus ou moins flexibles.

2.2 L'analyse des cooccurrences de deuxième ordre

Comme nous l'avons évoqué ci-dessus, l'accès à la sémantique des cooccurrents de premier ordre (ou *c*) d'un mot-pôle pourra se faire à partir de leurs propres cooccurrents, c'est-à-dire à partir des cooccurrents de deuxième ordre (ou *cc*) du pôle. Si les cooccurrents de premier ordre partagent beaucoup de co-occurrents de deuxième ordre, ces derniers se recoupent formellement, ce qui est une indication de l'homogénéité sémantique des cooccurrents de premier ordre et, dès lors, du mot-pôle. Le degré de similarité lexicale des cooccurrents d'un mot-pôle reflète le degré d'homogénéité sémantique ou de monosémie de ce mot-pôle. La similarité distributionnelle reflète clairement la

similarité sémantique. Par conséquent, un recoupement important des cooccurrents de deuxième ordre révèle un degré plus important de monosémie du mot-pôle. Par contre, si les cooccurrents de deuxième ordre sont formellement (très) différents, ils se recoupent (très) peu. Il s'ensuit que les cooccurrents de premier ordre sont sémantiquement plus diversifiés et par conséquent, le mot-pôle sera plus hétérogène sémantiquement.

Considérons quelques phrases-exemples du mot-pôle *broche* (cf. phrases 1 et 2 ci-dessous). Le mot *broche* signifie d'une part « partie tournante d'une machine-outil qui porte un outil ou une pièce à usiner » et d'autre part « outil comportant des dents étagées pour usiner des trous dans des pièces métalliques ». Les cooccurrents *vitesse*, *tr/mn* et *outil* (cf. phrase 1) indiquent clairement le sens « axe » ou « partie tournante d'une machine-outil qui porte un outil ou une pièce à usiner » (en anglais « a spindle »). En revanche, les cooccurrents *copeaux* et *à travers* (cf. phrase 2) indiquent le deuxième sens (en anglais « a broach »). Un être humain sait bien interpréter le sens de ces cooccurrents de premier ordre, mais une machine requiert des indices plus objectifs et mesurables, d'où la réitération de l'analyse des cooccurrences. On fait donc appel aux cooccurrents de deuxième ordre, dans la même phrase (par exemple *trous* et *soufflés* dans la phrase 2 comme cooccurrents de *copeaux*) et dans tous les autres contextes (par exemple *enlèvement* dans la phrase 3 comme cooccurrent de *copeaux*). Si les cooccurrents de deuxième ordre (*trous, soufflés, enlèvement,* etc.) sont plus nombreux à se recouper ou à être partagés par plus de cooccurrents de premier ordre, le mot-pôle (*broche*) sera plus homogène sémantiquement.

(1) *La **vitesse** de **broche** (n, **tr/mn**) est le nombre de tours que l'**outil** monté sur la **broche** effectue par minute.*
= « axe », « partie tournante d'une machine-outil qui porte un outil ou une pièce à usiner »

(2) *Dans le cas de trous débouchants, les **copeaux** sont soufflés **à travers** la **broche** en direction d'un bac collecteur.*
= « outil comportant des dents étagées pour usiner des trous dans des pièces métalliques »

(3) *Les gains de productivité dans le domaine de l'usinage des métaux par enlèvement de **copeaux** proviennent ... des progrès réalisés au niveau des machines-outils.*

2.3 La mesure de recoupement

Pour déterminer le degré de recoupement des cooccurrents de deuxième ordre, nous avons développé une mesure de recoupement (cf. figure 1). Elle s'appuie sur le recoupement formel des cooccurrents de deuxième ordre (ou cc) d'un mot-pôle et fait intervenir, en sommant pour tous les cc, les paramètres suivants :

- dans le numérateur, la fréquence d'un cc dans la liste des cc (= le nombre de cooccurrents (ou c) qui apparaissent avec ce cc)
- dans le dénominateur, le nombre total de c et le nombre total de cc.

$$\sum_{cc} \frac{\text{fq } cc}{\text{nbr total } c \cdot \text{nbr total } cc}$$

Figure 1. Formule de la mesure de recoupement

Expliquons, par souci de clarté, les paramètres de la mesure. Un mot-pôle, tel que *broche*, se caractérise par exemple par 5 cooccurrents de premier ordre (c), notamment *vitesse* et *copeaux*. A leur tour, ils ont chacun 5 cooccurrents, qui sont des cooccurrents de deuxième ordre (cc) du mot-pôle *broche*. Au total, le mot *broche* a donc 5 c différents et 25 cc. Certains cc figurent plusieurs fois dans la liste de tous les cc, parce qu'ils sont partagés par plusieurs c. Le recoupement de chaque cc sera d'autant plus important qu'il figure plusieurs fois dans la liste des cc. Ainsi, un cc qui est partagé par 2 c des 5 c aura un degré de recoupement de 2/5. Un cc qui figure 1 fois dans la liste et qui n'est pas partagé, ne sera pas important pour le recoupement moyen global. Par contre, un cc qui figure 3 fois dans la liste des cc et qui est donc partagé par 3 c des 5 c, pèse plus lourd sur le recoupement moyen global. En exprimant pour chaque cc le recoupement par la fraction *nombre de c avec le cc* (ou *fq cc*) divisé par *nombre total de c*, le résultat se situe toujours entre 0

(pas ou peu de recoupement) et 1 (recoupement important ou parfait) et par conséquent, le résultat est facilement interprétable. Comme on somme pour tous les cc, le dénominateur comprend aussi le nombre total de cc, car on considère tous les cc, y compris les doublons responsables du recoupement formel.

L'analyse des cooccurrences est effectuée, de façon récurrente, dans une fenêtre d'observation (ou *span*) de 5 mots à gauche et à droite, sans informations de position ni d'orientation. Cette fenêtre apporte suffisamment d'informations sémantiques pertinentes, sans introduire trop de bruit, et elle permet un traitement informatique efficace. Les mots grammaticaux sont conservés, parce qu'ils sont susceptibles d'apporter des informations sémantiques (par exemple *pendant* pour indiquer qu'il s'agit d'un processus). Les cooccurrents de premier et de deuxième ordre sont considérés au niveau des formes graphiques et non pas au niveau des lemmes. De cette façon, la mesure de recoupement permet de faire la distinction entre, par exemple, *pièce usinée* (« résultat ») et *pièce à usiner* (« avant le processus d'usinage ») et dès lors de tenir compte de la différence sémantique. Le mot-pôle sur lequel portent les analyses est le lemme, pour assurer l'appariement ultérieur des informations sémantiques aux informations de spécificité. La mesure d'association utilisée pour déterminer les cooccurrents statistiquement pertinents est la mesure statistique du LLR. Le seuil de significativité très sévère (0,9999 ou une valeur $p < 0{,}0001$) permet de relever uniquement les cooccurrents sémantiquement pertinents. La mesure de recoupement est concrétisée à l'aide de scripts en Python, afin de faire varier différents paramètres, tels que la fenêtre d'observation et le seuil de significativité, au niveau du repérage des cooccurrents de premier ordre et de deuxième ordre (cf. section 3). Pour les 4 717 unités lexicales spécifiques du corpus technique, nous calculons ainsi le degré de recoupement et donc le degré de monosémie, qui nous permet de les situer sur un continuum de monosémie, en fonction de leur rang de monosémie.

2.4 Validation de la mesure de recoupement

Comme il n'existe pas de mesure de référence pour évaluer les résultats quantitatifs de notre mesure de recoupement, nous avons

procédé à une validation interne à partir de l'analyse manuelle des cooccurrents les plus pertinents, ainsi qu'à une validation externe au moyen de plusieurs dictionnaires spécialisés.

Pour la validation manuelle, nous avons relevé dans le corpus technique tous les cooccurrents statistiquement les plus pertinents (au seuil de significativité le plus sévère de 1) d'une vingtaine d'unités lexicales spécifiques, intuitivement polysémiques ou homonymiques, par exemple *machine, outil, tour, avance, arête* et intuitivement monosémiques, par exemple *m/min, Iso*. Les mots intuitivement polysémiques (*avance*) ou homonymiques (*tour*) se caractérisent effectivement par une hétérogénéité plus importante de leurs cooccurrents statistiquement très significatifs, ce qui confirme le degré de monosémie (très) faible de ces mots-pôles. Ainsi, on retrouve pour le mot-pôle *tour*, d'une part les cooccurrents *minute, mille* (sens : « rotation ») et d'autre part les cooccurrents *centre, horizontal, bi-broche* (sens : « machine-outil pour l'usinage de pièces »). Par contre, les mots intuitivement monosémiques (*m/min, Iso*) ont moins de cooccurrents statistiquement très pertinents ou ils ont des cooccurrents sémantiquement plus homogènes, ce qui permet de confirmer leur degré de monosémie plus important.

La validation par confrontation avec les dictionnaires techniques a également permis de confirmer les résultats de notre mesure, pour un échantillon de 50 unités lexicales représentatives. Il convient toutefois de signaler que les mots les plus fréquents, comme *machine* et *outil*, entrent très souvent dans la composition d'unités polylexicales, telles que *machine à fraiser, machine à usiner,* etc., mentionnées également dans les dictionnaires spécialisés. Cette caractéristique pourrait en partie expliquer l'hétérogénéité sémantique de ces mots les plus fréquents, étant donné que le deuxième composant (ou le cooccurrent) entraîne une certaine désambiguïsation du mot-pôle. Comme nous l'avons évoqué ci-dessus, les unités polylexicales constituent une piste à explorer dans nos futures recherches.

3. Mises au point méthodologiques

Dans le but de déterminer la configuration de paramètres la plus stable, qui fournit les informations sémantiques les plus stables

et les plus fiables, nous avons effectué quelques mises au point méthodologiques. Ces expérimentations portent essentiellement sur trois paramètres, à savoir la forme graphique ou le lemme des cooccurrents de premier et de deuxième ordre, la taille de la fenêtre d'observation et le seuil de significativité (cf. Martinez 2005 et 2011 ; Mayaffre 2008). Les expérimentations font aussi varier simultanément plusieurs paramètres, pour évaluer l'impact combiné.

3.1 Forme graphique versus lemme

Le premier paramètre oppose la forme graphique (ou forme fléchie) au lemme, pour les cooccurrents de premier et/ou de deuxième ordre. L'impact de ce paramètre est analysé pour un échantillon du corpus technique d'environ 320.000 occurrences (tec02). Les expérimentations portent sur les 25 unités lexicales les plus spécifiques du corpus technique entier. La taille de la fenêtre d'observation (5 mots à gauche et à droite) et le seuil de significativité ($p < 0{,}0001$) restent inchangés. Dans la configuration préconisée, les mots-pôles se situent au niveau des lemmes (*lemmas* ou L) et les cooccurrents de premier ordre (c) et de deuxième ordre (cc) au niveau des formes graphiques (*word forms* ou W). Nous examinons si la prise en considération du lemme des c ou du lemme des c et des cc a une influence sur le degré de recoupement ou de monosémie et dès lors sur le rang de monosémie des unités lexicales spécifiques.

A cet effet, nous comparons, dans un premier temps, le degré de monosémie et le rang de monosémie des 25 unités lexicales les plus spécifiques à travers trois configurations : (1) lemme – forme graphique – forme graphique (ou LWW), (2) lemme – lemme – forme graphique (LLW) et (3) lemme – lemme – lemme (LLL). Signalons d'emblée qu'au niveau du lemme, on recense moins de c différents et moins de cc différents, étant donné que les c et les cc sont regroupés sous leur lemme correspondant. Par conséquent, les cc (lemmes) manifestent un degré de recoupement ou de monosémie plus élevé, ce qui se traduit par un degré d'homogénéité sémantique plus élevé du mot-pôle. Cependant, les différences de degré de recoupement ou de monosémie ne se traduisent pas toujours par des différences de rang de monosémie. Le rang de monosémie

d'un mot change, non seulement en fonction de son propre degré de monosémie, qui est plus ou moins élevé dans les différentes configurations, mais également en fonction du rapport entre son propre degré de monosémie et le degré de monosémie des autres mots de la sélection.

Il s'avère que les mots les plus hétérogènes sémantiquement (*machine, outil*) ont un degré de recoupement ou de monosémie plus faible dans les trois configurations. Les mots les plus homogènes sémantiquement (*Fig, mm*) ont un degré de monosémie plus élevé, généralement dans les trois configurations. Cette stabilité à travers les trois configurations se confirme pour le rang de monosémie de ces mots les plus hétérogènes et les plus homogènes sémantiquement. Pour les trois configurations, les différences de rang de monosémie les plus importantes s'observent pour *copeau* et *acier*. Le mot-pôle *acier*, qui était plutôt hétérogène sémantiquement dans la configuration préconisée (LWW), devient plus homogène sémantiquement dans la configuration des lemmes (LLL). Il s'ensuit que les lemmes des cc du mot-pôle *acier* se recoupent beaucoup plus que les formes graphiques de ses cc. Si certains mots acquièrent un rang plus monosémique en passant des formes graphiques aux lemmes, d'autres se voient accorder, de ce fait, un rang plus polysémique, par exemple *usinage*.

Les différences en termes de degré de monosémie et de rang de monosémie sont certes indicatives pour certaines unités lexicales spécifiques, mais elles ne permettent pas de visualiser les tendances globales pour les 25 unités lexicales spécifiques à travers les trois configurations. Nous recourons dès lors à la technique de *Multidimensional Scaling* (MDS)[1] ou de positionnement multidimensionnel, implémentée dans le logiciel R[2].

1 Le MDS est une méthode d'analyse multivariée descriptive, telle que l'analyse factorielle des correspondances (AFC) ou l'analyse en composantes principales (ACP). A la différence de ces techniques, le MDS permet d'analyser tout type de matrice de (dis)similarité, si les (dis)similarités sont évidentes. Le MDS n'impose pas de restrictions, telles que des relations linéaires entre les données sous-jacentes, leur distribution normale multivariée ou la matrice de corrélation (http://www.statsoft.com/textbook/stmulsca.html).

2 R : http://www.r-project.org

Elle permet de visualiser en 2D la distance entre les mots, représentés par leur rang de monosémie, en faisant varier les configurations.

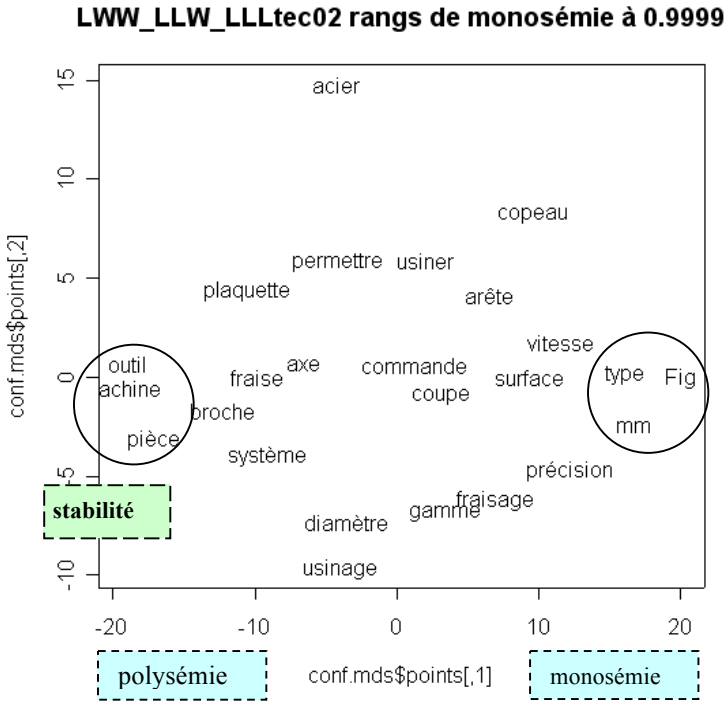

Figure 2. Résultat du MDS des 25 unités lexicales spécifiques (3 configurations)

La visualisation ci-dessus du MDS pour les 25 unités spécifiques dans les trois configurations (Cf. figure 2) montre que les mots les plus polysémiques se regroupent à gauche de la visualisation (*machine, outil*). Les mots les plus monosémiques se regroupent à droite (*Fig, mm, type*). Dans cette représentation visuelle du MDS, l'axe horizontal peut s'interpréter comme l'axe sémantique, allant des mots plus polysémiques à gauche aux mots plus monosémiques à droite. L'axe vertical peut s'interpréter comme l'axe de la stabilité dans les trois configurations. Les mots avec des différences de rang importantes dans les trois configurations se trouvent à une distance plus

importante des autres mots. Ainsi, le mot *acier* se situe en haut de la visualisation, puisqu'il est plutôt polysémique dans la configuration préconisée et monosémique quand on considère les c et cc au niveau du lemme. Le mot *usinage* en revanche se situe en bas de la visualisation, parce qu'il se trouve parmi les mots les plus monosémiques dans la configuration LWW et qu'il devient plutôt hétérogène sémantiquement. La plupart des mots se trouvent bien au milieu de la visualisation et se caractérisent par une relative stabilité de leur rang de monosémie à travers les trois configurations.

3.2 La taille de la fenêtre d'observation

Afin de déterminer la taille idéale de la fenêtre d'observation (*span*), nous procédons à une comparaison de plusieurs tailles. Les expérimentations sont conduites dans la configuration préconisée (LWW ou lemme – forme graphique – forme graphique), au seuil de significativité sévère ($p < 0{,}0001$). Les 11 fenêtres d'observation comparées sont de taille 1, 2, 3, 4, 5, 6, 8, 10, 12, 15 et 3-15 (à partir du 3^{e} mot à gauche et à droite jusqu'au 15^{e} mot inclus). Cette dernière fenêtre d'observation est intéressante, parce qu'elle permet d'exclure les cooccurrents syntaxiques et de se concentrer surtout sur les cooccurrents lexicaux. Les expérimentations visent à vérifier si la taille de la fenêtre d'observation préconisée de 5 mots à gauche et à droite du mot-pôle n'est pas périphérique par rapport aux autres tailles. Une fenêtre plus petite entraîne certes moins de bruit, mais aussi moins de c (et moins de cc) sémantiquement pertinents (notamment des collocations) et plus de c (et cc) syntaxiquement dépendants. Une fenêtre plus large apporte plus de c (et cc) sémantiquement pertinents, mais risque d'inclure plus ou peut-être même trop de bruit. Une analyse MDS des distances entre les différentes tailles (*span*) pour les rangs de monosémie des 25 unités spécifiques confirme les résultats précédents. En effet, les mots les plus polysémiques et les plus monosémiques se caractérisent par des rangs de monosémie stables à travers les différentes tailles de fenêtre d'observation. Le MDS indique que la taille préconisée de 5 mots à gauche et à droite se situe bien au centre des différentes configurations de taille et qu'elle n'est pas périphérique. Les résultats en termes de rang de

monosémie dans cette fenêtre ne présentent pas d'écart important par rapport aux autres fenêtres d'observation. Il est à remarquer que la fenêtre plus particulière de 3-15 s'avère très périphérique par rapport aux autres.

3.3 Le seuil de significativité

A l'instar des expérimentations précédentes et des analyses MDS, nous procédons aussi à la comparaison de plusieurs seuils de significativité, pour vérifier si le seuil préconisé (0,9999 ou $p < 0,0001$) n'est pas trop périphérique. Les expérimentations sont conduites pour les 25 unités lexicales les plus spécifiques, dans la configuration préconisée LWW et pour une fenêtre d'observation de 5 mots à gauche et à droite. Elles font intervenir quatre seuils de significativité : 0,95 ($p < 0,05$), 0,99 ($p < 0,01$), 0,999 ($p < 0,001$) et 0,9999 ($p < 0,0001$). Notons que moins on est sévère (seuil de 0,95), plus de c et de cc seront significatifs, et, dès lors, pris en considération pour le calcul du recoupement. Ces cc supplémentaires (moins significatifs et moins pertinents sémantiquement), pourront soit augmenter le degré de recoupement moyen, s'ils sont identiques à d'autres cc plus significatifs, soit diminuer le degré de recoupement, s'ils sont formellement différents des autres cc plus significatifs. Un seuil plus sévère (seuil de 0,9999) permet de relever des c et des cc sémantiquement plus pertinents. Le MDS indique que les seuils de significativité moins sévères (0,99 et 0,95) génèrent des résultats similaires en termes de rang de monosémie pour les 25 mots analysés. Ces deux seuils incluent plus de c et de cc et ils se situent loin des deux autres seuils qui incluent des c et des cc plus saillants et plus pertinents. Force est de constater que les deux seuils plus « sévères » (0,999 et 0,9999) se situent à une distance considérable l'un de l'autre, ce qui indique tout de même des dissimilarités importantes quant au rang de monosémie.

3.4 Plusieurs paramètres

Nous décidons dès lors de faire varier simultanément plusieurs paramètres. Etant donné que nous préférons considérer les cooccurrents de premier et de deuxième ordre au niveau des formes graphiques, sémantiquement plus riches, nous proposons d'inclure dans l'analyse MDS les deux autres paramètres,

à savoir la taille de la fenêtre d'observation et le seuil de significativité. Le MDS prendra donc en considération les 11 tailles (cf. section 3.2) et les 2 seuils de significativité les plus sévères, à savoir 0,9999 et 0,999 (cf. section 3.3). Dans le but de ne pas trop compliquer la visualisation du MDS, les tailles au seuil de significativité 0,9999 seront dénommées span+ et celles au seuil de significativité 0,999 span- (cf. figure 3). Il est clair que la taille préconisée (span5) n'est pas trop périphérique par rapport aux autres tailles, tant au seuil 0,999 (span5-) qu'au seuil plus sévère 0,9999 (span5+). Les choix méthodologiques se voient donc confirmés par les résultats du MDS.

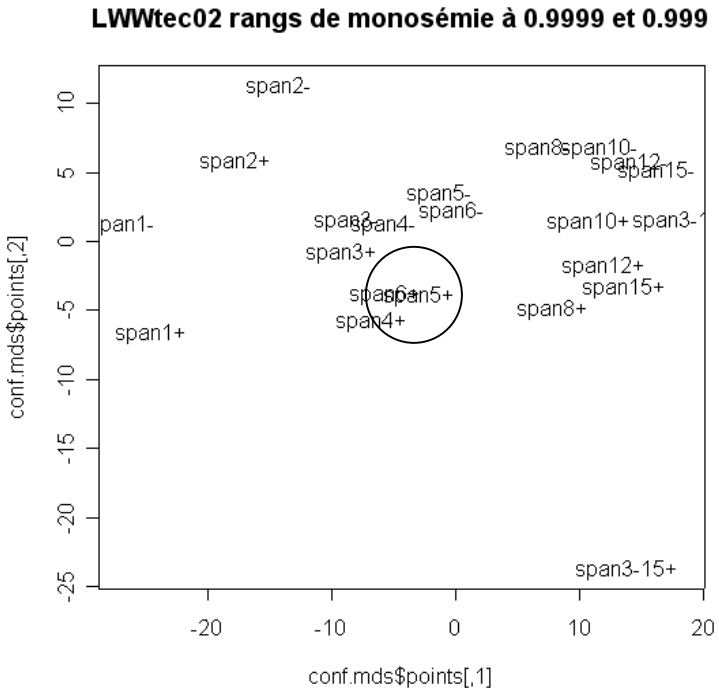

Figure 3. Résultat du MDS des 2 seuils de significativité et des 11 tailles

4. Intégration des étiquettes morphosyntaxiques

Pour examiner si les unités lexicales les plus spécifiques du corpus technique sont effectivement les plus monosémiques, nous procédons à une analyse statistique de régression simple, qui évalue la corrélation entre leur rang de spécificité et leur rang de monosémie. Les résultats montrent une corrélation négative (coefficient de corrélation Pearson de -0,72) et un pourcentage de variation expliquée R^2 de 51,57% ($p < 2e^{-16}$). Il s'avère que les unités lexicales les plus spécifiques du corpus technique ne sont pas les plus monosémiques, mais, au contraire, les plus hétérogènes sémantiquement (par exemple *machine, pièce, tour*). En plus, les unités lexicales les moins spécifiques du corpus technique sont les plus homogènes sémantiquement (par exemple *rationnellement, télédiagnostic*).

Dans le but d'affiner ces résultats, nous proposons d'enrichir la mesure de recoupement en y intégrant des informations de classe lexicale par le biais des étiquettes morphosyntaxiques, tant pour l'identification des unités lexicales spécifiques que pour le repérage des cooccurrents de premier et de deuxième ordre, c'est-à-dire pour le calcul du degré de monosémie. Le but est de dissocier les occurrences des lemmes homonymiques (par exemple *abrasif|nom* et *abrasif|adj*), afin d'améliorer la précision et la fiabilité de l'analyse sémantique quantitative ainsi que de vérifier les effets sur les résultats quantitatifs et statistiques.

L'analyse statistique de régression simple pour les 5 207 unités lexicales spécifiques enrichies (avec classe lexicale) confirme les résultats de l'analyse de régression précédente et montre une corrélation négative similaire (R^2 de 51,63 %). L'effet global de la prise en compte de la classe lexicale est donc très limité, voire négligeable. Apparemment, l'analyse des cooccurrences, qui se situe en amont du calcul du degré de monosémie, filtre déjà la classe lexicale. Nous observons également un effet local en termes de degrés de recoupement ou de monosémie. Premièrement, il s'avère que les lemmes homonymiques dissociés (par exemple *abrasif|adj* et *abrasif|nom*) deviennent plus homogènes sémantiquement. Ensuite, les mots intuitivement monosémiques (tels que *Iso, t/m*) restent plutôt monosémiques, lorsqu'on intègre la classe lexicale : leur degré

de monosémie change à peine. Finalement, les mots intuitivement polysémiques (*avance, tour*) ou les mots vagues (*usinage*), qui sont hétérogènes sémantiquement dans la liste des 4 717 mots spécifiques, se caractérisent également par de faibles différences entre les deux degrés de monosémie. L'hétérogénéité sémantique de ces mots intuitivement polysémiques est très importante, mais elle reste cachée.

5. Conclusions et perspectives

Dans cet article, nous avons discuté l'importance des cooccurrences de deuxième ordre pour automatiser et quantifier l'analyse sémantique. En implémentant la monosémie en termes d'homogénéité sémantique, nous avons calculé le degré de monosémie d'un mot-pôle à partir du degré de recoupement de ses cooccurrences de deuxième ordre. Il convient toutefois de signaler que des recherches supplémentaires s'imposent pour vérifier si et à quel point la monosémie traditionnelle correspond au degré de monosémie « monosémique » calculé par notre mesure de recoupement. Nous recourons à cette mesure dans le but opérationnel de développer un critère mesurable permettant de quantifier l'analyse sémantique. Sans recherches supplémentaires, il serait impossible d'affirmer que les degrés de monosémie calculés dans notre étude correspondent parfaitement à ce que les terminologues traditionnels considèrent comme monosémie. Dans un souci de précision et d'efficacité, les monosémistes de la terminologie traditionnelle prescriptive préconisent la monoréférentialité (chaque terme a un seul référent) et la monosémie (chaque terme a un seul sens). Ce sens unique est généralement prescrit par des ouvrages normatifs et expliqué (et/ou délimité) à l'aide d'une définition dans des normes ou dans un dictionnaire spécialisé. Néanmoins, il n'est pas toujours clair si ce sens prescrit s'applique effectivement à tous les contextes d'usage de l'unité terminologique.

Notre étude a montré que l'impact de l'intégration des informations de classe lexicale est plutôt limité. Il s'est avéré que l'analyse des cooccurrences de deuxième ordre, qui se situe en amont du calcul du degré de monosémie, filtre déjà la classe lexicale. Nous avons ainsi réussi à détecter l'hétérogénéité sé-

mantique des homonymes et l'homogénéité sémantique de leurs lemmes dissociés. Nous avons également réussi à détecter l'hétérogénéité sémantique des mots polysémiques. Cependant, notre mesure ne permet pas encore de dissocier les différents sens polysémiques, ni de faire la distinction entre la polysémie et le vague. Dans nos recherches futures, nous envisageons de remédier à ce problème en complétant notre mesure de monosémie par des analyses statistiques multivariées de regroupement (*cluster analysis*). Celles-ci permettraient de regrouper les cooccurrents de premier ordre d'un mot-pôle à partir des cooccurrents de deuxième ordre qu'ils partagent. Ainsi, les analyses de regroupement pourraient conduire à mieux comprendre le phénomène de l'hétérogénéité sémantique et à opérer des distinctions sémantiques plus fines. Elles nous permettraient peut-être aussi de vérifier si une grande proximité entre les cooccurrents de premier ordre implique toujours un degré de monosémie plus important du mot-pôle. En effet, il n'est pas exclu qu'un mot-pôle soit polysémique tout en ayant des cooccurrences de premier ordre sémantiquement plutôt homogènes dans le corpus à l'étude.

Dans cet article, nous avons démontré que l'analyse sémantique quantitative comporte de nombreux avantages. D'abord, elle permet de procéder à l'analyse sémantique simultanée de plusieurs milliers d'unités lexicales. Ensuite, les données quantitatives se prêtent facilement à des analyses statistiques, qui conduisent à des résultats objectifs. Enfin, les approches méthodologiques élaborées dans notre étude pourront facilement être appliquées à d'autres corpus spécialisés, relevant d'autres domaines techniques ou scientifiques ou relevant de domaines politiques ou juridiques. On pourrait aussi envisager une analyse sémantique quantitative, par le biais d'une analyse du recoupement des cooccurrences de deuxième ordre, pour un corpus de langue générale. Finalement, il est clair que nos recherches futures passent inévitablement par les unités polylexicales, omniprésentes dans les corpus spécialisés.

Références

Blumenthal P. & Hausmann F.J. (dir.) (2006). Collocations, corpus, dictionnaires. *Langue française* 150.

Condamines A (éd.) (2005). *Sémantique et corpus*. Paris : Hermes-Science.

Dunning T. (1993). « Accurate methods for the statistics of surprise and coincidence », *Computational Linguistics* 19 (1) : 61-74.

Ferret O. (2004). « Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales », *Actes de TALN 2004* : 183-192.

Grefenstette G. (1994). « Corpus-derived first, second and third-order word affinities », *in* Martin W., Meijs W. *et al.* (eds), *Proceedings of Euralex '94. International Congress on Lexicography, Amsterdam* : 279-290.

Habert B., Illouz G. & Folch H. (2004). « Dégrouper les sens : pourquoi ? comment ? », *Actes des JADT 2004* : 565-576.

Habert B., Illouz G. & Folch H. (2005). « Des décalages de distribution aux divergences d'acception », *in* Condamines A. (éd.). *Sémantique et corpus*. Paris : Hermes-Science, 277-318.

Kwary D.A. (2011). « A hybrid method for determining technical vocabulary », *SYSTEM* 39, 2 : 175-185.

Martinez W. (2000). « Mise en évidence de rapports synonymiques par la méthode des cooccurrences », *Actes des JADT 2000 :* 78-84.

Martinez W. (2005). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de Doctorat en Sciences du Langage, Université de la Sorbonne nouvelle – Paris 3, sous la direction d'André Salem, Paris.

Martinez W. (2011). « Vers une cartographie géo-lexicale », in *Situ* 15, mis en ligne le 29 juin 2011. (http://insitu.revues.org/590)

Mayaffre D. (2008). « Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence », *Actes des JADT 2008* : 811-822.

Nerlich B., Todd Z., Herman V. & Clarke D. (2003). *Polysemy. Flexible patterns of meaning in mind and language.* Berlin / New York : Mouton de Gruyter.

Paquot M. & Bestgen Y. (2009). « Distinctive words in academic writing : a comparison of three statistical tests for keyword extraction », *in* Jucker A., Schreier D. & Hundt M. (eds), *Corpora : Pragmatics and Discourse*. Amsterdam : Rodopi, 247-269.

Pezik P. (2005). « You shall know a word by the company it keeps. A comparative study of co-occurrence statistics », Paper presented at *PALC 2005, Practical applications in language and computers*. Lodz : Poland.

Schütze H. (1998). « Automatic Word Sense Discrimination », *Computational Linguistics* 24, 1 : 97-123.

Scott M. & Tribble C. (2006). *Textual Patterns. Key words and corpus analysis in language education. Studies in Corpus Linguistics* 22. Amsterdam : Benjamins.

Veronis J. (2003). « Cartographie lexicale pour la recherche d'informations », *Actes de TALN 2003* : 265-274.