Analyse exploratoire des cooccurrents de premier ordre dans un corpus technique

Ann Bertels[1], Dirk Speelman[2]

[1] KU Leuven – ann.bertels@ilt.kuleuven.be
[2] KU Leuven – dirk.speelman@arts.kuleuven.be

Abstract

This paper addresses the methodology and results of an exploratory co-occurrence data analysis in a technical corpus, within the framework of distributional semantics. Multidimensional scaling analysis is carried out in order to cluster first-order co-occurrences of a technical node, with respect to shared second order and third order co-occurrences. Our study aims at plotting first-order co-occurrences on a 2D-plot, as a way to find semantically related co-occurrences.

First-order co-occurrences are semantically similar if they appear in similar contexts or if they share second-order co-occurrences. By taking into account the association values between relevant first and second-order co-occurrences, semantic similarities and dissimilarities can be determined, as well as proximities and distances on a 2D-plot. This paper discusses the importance of various thresholds, the effect of a relevant selection of first-order co-occurrences and the added value of third-order co-occurrences. Several technical nodes, semantically homogeneous and heterogeneous, are taken into account, as a complement to previous quantitative semantic analysis.

Résumé

Cette contribution s'inscrit dans le cadre de la sémantique distributionnelle et présente la méthodologie et les résultats d'une analyse exploratoire des données de cooccurrence dans un corpus technique. A l'aide d'une analyse statistique de positionnement multidimensionnel (*Multidimensional Scaling* ou MDS), nous procédons au regroupement des cooccurrents de premier ordre d'un mot-pôle technique, en fonction des cooccurrents de deuxième ordre et de troisième ordre partagés. L'objectif de notre étude est de positionner les cooccurrents de premier ordre les uns par rapport aux autres en 2D, et ce faisant de cerner des groupes de cooccurrents sémantiquement liés.

Les cooccurrents de premier ordre sont sémantiquement similaires s'ils figurent dans des contextes similaires ou s'ils partagent des cooccurrents de deuxième ordre. La prise en compte des valeurs d'association entre les cooccurrents de premier et deuxième ordre pertinents permet de déterminer les similarités et dissimilarités sémantiques, ainsi que les proximités et distances en 2D. Nous discutons l'importance du seuillage, l'effet d'une sélection pertinente de cooccurrents de premier ordre et la valeur ajoutée des cooccurrents de troisième ordre. Dans cet article, nous considérons plusieurs mots-pôles techniques, sémantiquement homogènes et hétérogènes, dans le but de compléter des analyses sémantiques quantitatives antérieures.

Mots-clés : positionnement multidimensionnel, exploration visuelle, proximité sémantique, analyse des cooccurrences, cooccurrents de deuxième ordre et de troisième ordre

1. Introduction

Cette contribution décrit la méthodologie et les résultats d'une analyse exploratoire des cooccurrents dans un corpus spécialisé, qui relève du domaine technique des machines-outils pour l'usinage des métaux (1,7 million d'occurrences). L'analyse exploratoire consiste à regrouper les cooccurrents de premier ordre d'un mot-pôle en fonction des cooccurrents de deuxième ordre partagés et à les positionner les uns par rapport aux autres en 2D. Le but est de cerner des groupes de cooccurrents de premier ordre sémantiquement liés, qui permettent d'accéder à la sémantique du mot-pôle technique. Les analyses de regroupement (*clustering*)

et de visualisation (*plotting*) des cooccurrents de premier ordre font suite à une étude sémantique quantitative effectuée sur le corpus technique. Cette étude consistait notamment à développer une mesure de monosémie dans le but de calculer le degré de monosémie ou d'homogénéité sémantique des unités lexicales spécifiques en fonction des cooccurrents de deuxième ordre partagés (Bertels et al., 2010). La mesure de monosémie permet déjà de détecter l'homogénéité sémantique des unités lexicales monosémiques[1] ainsi que l'hétérogénéité sémantique des unités lexicales polysémiques[2], mais nous n'arrivons pas encore à dissocier les sens polysémiques, ni à faire la distinction entre la polysémie et le vague. En vue d'affiner l'analyse sémantique quantitative, nous procédons à des analyses exploratoires de regroupement et de visualisation des cooccurrents de premier ordre. En effet, la visualisation des proximités et distances sémantiques entre les cooccurrents de premier ordre d'un mot-pôle permettra de mieux interpréter son degré d'hétérogénéité sémantique.

Dans une étude préliminaire (Bertels et Speelman, 2013), nous avons analysé les cooccurrents de premier ordre du mot-pôle *tour*, qui se caractérise par un degré d'hétérogénéité sémantique élevé. *Tour* a plusieurs sens techniques dans le corpus technique, notamment « machine-outil pour l'usinage des pièces » et « rotation ». Les premières expérimentations de regroupement et de visualisation ont été effectuées sur un extrait du corpus technique de 320 000 mots. Elles visaient principalement à évaluer l'effet de plusieurs mesures d'association pour l'analyse des cooccurrences et l'effet de plusieurs métriques de distance pour le regroupement. La visualisation de la répartition des cooccurrents de premier ordre du mot-pôle *tour* a montré quelques groupes de cooccurrents et quelques cooccurrents isolés, qui reflétaient bien les différents sens de ce mot-pôle à la fois homonymique[3] et polysémique dans le corpus analysé.

Dans le présent article, nous discutons les résultats d'une analyse exploratoire des cooccurrents de premier ordre de plusieurs mots-pôles techniques, aussi bien sémantiquement homogènes que sémantiquement hétérogènes, dans l'ensemble du corpus technique de 1,7 million d'occurrences. Nous expliquons d'abord les principes méthodologiques de l'analyse statistique de positionnement multidimensionnel pour le regroupement et la visualisation des cooccurrents de premier ordre (section 2). Ensuite, nous discutons l'importance du seuillage, l'effet d'une sélection plus pertinente de cooccurrents de premier ordre et la valeur ajoutée de la prise en compte des cooccurrents de troisième ordre (section 3). Nous terminons par une conclusion et des pistes de recherches futures (section 4).

2. Analyse statistique de positionnement multidimensionnel

2.1. Objet d'analyse : les cooccurrents de premier ordre

La plupart des analyses en sémantique distributionnelle (Sahlgren, 2006 et 2008 ; Turney et Pantel, 2010) étudient la proximité sémantique entre mots. Deux mots sont sémantiquement similaires s'ils figurent dans des contextes similaires, c'est-à-dire s'ils partagent soit des contextes syntaxiques (Morlane-Hondère, 2013 ; Morardo et Villemonte de La Clergerie, 2013) soit des cooccurrents de premier ordre (Sahlgren, 2008 ; Peirsman et Geeraerts, 2009 ;

[1] La monosémie caractérise les unités linguistiques qui n'ont qu'un seul sens.

[2] La polysémie caractérise les unités linguistiques à plusieurs sens, généralement apparentés ou reliés entre eux par métaphore, par métonymie, par restriction de sens ou par extension de sens.

[3] L'homonymie explique le phénomène par lequel deux mots (étymologiquement) différents coïncident formellement. Un signifiant (une forme graphique ou sonore) correspond à deux signifiés.

Ferret, 2010 ; Heylen et al., 2012 ; Wielfaert et al., 2013). Ces dernières analyses déterminent la proximité sémantique entre mots à partir des contextes qu'ils partagent dans un corpus donné, sous forme d'espaces de mots ou d'espaces vectoriels sémantiques (*Semantic Vector Spaces*). Elles s'appuient sur des mesures d'association pour déterminer les cooccurrents statistiquement pertinents et sur des métriques de distance pour positionner les mots les uns par rapport aux autres en fonction des cooccurrents de premier ordre qu'ils partagent. Les mots qui apparaissent souvent avec les mêmes cooccurrents se retrouvent regroupés dans un espace de mots, dont la représentation graphique permet de visualiser des groupes de synonymes (Ferret, 2010) ou de mots sémantiquement liés (Peirsman et Geeraerts, 2009). Si les données au niveau des cooccurrents de premier ordre sont rares (*data sparseness*), il est fait appel aux cooccurrents de deuxième ordre (Schütze, 1998 ; Lemaire et Denhière, 2006).

Dans nos analyses, nous cherchons à mieux comprendre le degré d'hétérogénéité sémantique d'un mot-pôle technique. Nous nous intéressons dès lors aux rapports sémantiques entre ses cooccurrents de premier ordre. Par conséquent, l'objet d'analyse se situe à un ordre supérieur par rapport à l'objet d'analyse des études en sémantique distributionnelle évoquées ci-dessus. L'objectif de notre étude est de positionner les cooccurrents de premier ordre d'un mot-pôle les uns par rapport aux autres en fonction des cooccurrents de deuxième ordre partagés, et ce faisant de discerner des groupes de cooccurrents de premier ordre sémantiquement liés. Les cooccurrents de deuxième ordre sont donc définis comme les cooccurrents des cooccurrents de premier ordre du mot-pôle. Si la visualisation des cooccurrents de premier ordre montre par exemple deux groupes nettement délimités, le mot-pôle se caractérise par deux sens différents. Pour remédier à un problème de rareté de données au niveau des cooccurrents de deuxième ordre, nous faisons appel aux cooccurrents d'un ordre supérieur (Grefenstette, 1994), c'est-à-dire aux cooccurrents de troisième ordre.

2.2. Identification des cooccurrents pertinents

Pour identifier les cooccurrents statistiquement pertinents, il faut s'appuyer sur des mesures d'association, telles que le log du rapport de vraisemblance (*Log-Likelihood Ratio* ou LLR) (Dunning, 1993), l'information mutuelle (*Pointwise Mutual Information* ou PMI) (Church et Hanks, 1990), le Z-score, le chi-carré, etc. (pour un aperçu détaillé, voir Evert, 2007). La mesure du LLR est couramment utilisée, mais elle souffre d'un biais de fréquence, parce qu'elle gonfle la valeur d'association de deux mots (très) fréquents. La mesure de la PMI est plus fiable pour des cooccurrences qui apparaissent souvent ensemble, mais moins fiable pour des cooccurrences dont la co-fréquence est faible, parce que la mesure de la PMI a tendance à surestimer la valeur d'association des mots rares (Evert, 2007).

Pour les analyses discutées dans cet article, nous aurons recours à la mesure de la PMI et nous introduirons un seuil de co-fréquence minimale (cf. section 3.1). Les cooccurrents de premier ordre sont repérés dans une fenêtre d'observation (*span*) de 5 mots à gauche et à droite du mot-pôle, les cooccurrents de deuxième ordre dans une fenêtre de 5 mots à gauche et à droite des cooccurrents de premier ordre. Des expérimentations préalables ont permis de constater que cette fenêtre recense des cooccurrents sémantiquement intéressants sans introduire trop de bruit. Les cooccurrents ne sont pas considérés au niveau des lemmes, mais au niveau des formes graphiques, qui véhiculent des informations sémantiques plus riches, comme la distinction entre *pièce usinée* (« résultat ») et *pièce à usiner* (« avant le processus d'usinage »), par exemple. Les mots grammaticaux sont conservés parmi les cooccurrents de premier et deuxième ordre, parce qu'ils sont susceptibles d'apporter des informations sémantiques utiles, par exemple *pendant* indique qu'il s'agit d'un processus.

2.3. Positionnement multidimensionnel

Pour le regroupement et la visualisation des cooccurrents de premier ordre d'un mot-pôle, en fonction des cooccurrents de deuxième ordre partagés, nous recourons à l'analyse statistique de positionnement multidimensionnel (*MultiDimensional Scaling* ou MDS) (Kruskal et Wish, 1978 ; Cox et Cox, 2001 ; Venables et Ripley, 2002). La technique de MDS[4] est implémentée dans le logiciel d'analyse statistique R[5]. Dans nos analyses, nous utilisons le positionnement non métrique `isoMDS`, disponible dans le paquet `MASS`. Cette technique permet d'analyser une matrice pour un ensemble de données disposées en rangées (ici : les cooccurrents de premier ordre ou les *c*) à partir de leurs valeurs pour plusieurs variables disposées en colonnes (ici : les cooccurrents de deuxième ordre ou les *cc*). Les données de la matrice *c* × *cc* sont réarrangées de façon à obtenir la configuration visuelle qui représente le mieux possible les distances observées entre les *c*. La meilleure représentation visuelle est celle qui maximise la qualité de l'ajustement (*goodness-of-fit*) et qui minimise la distorsion, lors de la réduction de l'ensemble des dimensions aux deux dimensions visualisées (*plot*). La qualité de la représentation visuelle est évaluée à l'aide du *stress*. Le pourcentage de stress est un indicateur de la qualité de l'ajustement (Desbois, 2005). Il doit être minimal pour garantir la fiabilité de la représentation visuelle par rapport aux données disposées dans la matrice d'origine. En règle générale, un pourcentage de stress inférieur à 10% est excellent et un pourcentage supérieur à 15% est inacceptable (Clarke, 1993 ; Borg et Groenen, 2005).

À partir de la matrice *c* × *cc*, nous générons une matrice de dissimilarité, en calculant les distances par paire d'observations. La métrique de distance par défaut dans l'analyse `isoMDS` est la distance euclidienne. Elle mesure la distance spatiale ou la distance en ligne droite entre les deux observations dans un espace 2D. Comme nous l'avons constaté dans nos expérimentations précédentes, la distance euclidienne ne convient pas bien lorsque la matrice contient des valeurs d'association faibles et élevées, parce que les valeurs faibles perdent tout leur poids face aux valeurs plus élevées (Bertels et Speelman, 2013). Une métrique de distance alternative est l'angle du cosinus[6] (*cosine angle*). Cette métrique est couramment utilisée en sémantique distributionnelle (Padó et Lapata, 2007 ; Sahlgren, 2008). Elle s'applique à des observations représentées par des vecteurs et elle détermine la similarité entre les observations par le calcul de l'angle entre leurs vecteurs. Des observations avec des vecteurs similaires se situent plus près les unes des autres dans l'espace multidimensionnel et se caractérisent donc par un angle plus petit. Même si l'ordre de grandeur des valeurs dans les vecteurs change, l'angle entre les vecteurs n'est pas affecté (van der Laan et Pollard, 2003). En analyse sémantique distributionnelle, il est souvent fait appel aux modèles vectoriels et à l'angle du cosinus pour établir la similarité entre deux mots (Peirsman et Geeraerts, 2009). Dans notre étude précédente, l'angle du cosinus permet d'aboutir à des résultats visuels satisfaisants (Bertels et Speelman, 2013). Les lignes de la matrice de base *c* × *cc*, à savoir les cooccurrents de premier ordre, sont conçues comme des vecteurs avec une valeur par colonne.

[4] Le MDS est une méthode d'analyse multivariée descriptive, comme 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).

[5] R : www.r-project.org.

[6] Dans R, l'angle du cosinus est implémenté dans la fonction `distancematrix` du paquet `hopach`.

Pour ces vecteurs, la similarité est calculée en fonction des valeurs d'association PMI dans les différentes colonnes, c'est-à-dire avec les différents cooccurrents de deuxième ordre.

3. Mises au point méthodologiques

Les analyses MDS discutées ci-dessous prennent comme point de départ une matrice *c* × *cc* par mot-pôle, qui est réalisée à l'aide de scripts en Python. Nous considérons des mots-pôles sémantiquement homogènes et sémantiquement hétérogènes dans l'ensemble du corpus technique de 1,7 million d'occurrences, à savoir *tour*, *machine*, *usinage*, *usiner*, *avance* et *Iso*. À partir de la matrice *c* × *cc*, la métrique de l'angle du cosinus permet de générer une matrice de dissimilarité, qui est ensuite soumise à une analyse MDS. Le but est de regrouper les cooccurrents de premier ordre (*c*) de chaque mot-pôle en fonction des valeurs d'association PMI similaires avec des *cc* similaires et de visualiser ces proximités et distances sémantiques en 2D, pour ainsi accéder à la sémantique du mot-pôle.

Nous procédons à plusieurs mises au point méthodologiques afin de trouver la configuration de paramètres la plus efficace d'un point de vue statistique et la plus intéressante d'un point de vue sémantique. Lors de la discussion des résultats, nous prenons en considération des critères quantitatifs, comme le nombre de cooccurrents visualisés et le pourcentage de stress, ainsi que des critères qualitatifs, tels que la lisibilité des représentations visuelles.

3.1. L'importance du seuillage

Comme nous l'avons mentionné ci-dessus, la mesure d'association de la PMI est moins fiable pour des cooccurrents à faible co-fréquence (*cf.* section 2.2). Pour y remédier, il est conseillé de respecter un seuil de co-fréquence minimale supérieur ou égal à 5 (Evert, 2007). Dans nos analyses MDS sur l'ensemble du corpus technique, nous proposons d'introduire des seuils inférieurs de co-fréquence minimale suffisamment élevés, à savoir 10, 20, 50 et 100. Nous évaluons l'effet de plusieurs seuils inférieurs dans le but de trouver le seuil le plus efficace. Un seuil de co-fréquence plutôt faible (par exemple ≥ 10) permet de relever plus de *c* pertinents (*cf.* tableau 1 ci-dessous pour quelques exemples). Toutefois, un nombre trop important de *c* rendrait la visualisation de l'analyse MDS trop dense et dès lors illisible. Un seuil de co-fréquence plus élevé (par exemple ≥ 50 ou 100) relève moins de *c* pertinents, mais des *c* plus fréquents, qui sont souvent des mots grammaticaux. Un nombre trop faible de *c* ne donnerait pas assez d'informations pour l'interprétation sémantique de la visualisation. Une prédominance de mots grammaticaux poserait également un problème d'interprétation. Bien que certains *c* grammaticaux soient utiles pour l'interprétation sémantique (par exemple *pendant*), la plupart d'entre eux ne sont pas sémantiquement distinctifs. Comme le montre le tableau 1 ci-dessous, la configuration pour le mot-pôle *tour* au seuil 100 dépasse légèrement le seuil de stress acceptable de 15%. Or, cette configuration ne recense que 13 *c* : 2 *c* lexicaux (*CNC* et *poupée*) et 11 *c* grammaticaux sémantiquement vides (*pour*, *sur*, *des*, *les…*).

3.2. Vers une sélection plus pertinente de cooccurrents de premier ordre

Il est clair que le seuil inférieur de co-fréquence minimale ne suffit pas pour trouver une configuration avec un pourcentage de stress acceptable et des *c* sémantiquement intéressants. La prise en compte d'un seuil supérieur permettrait de supprimer les mots grammaticaux parmi les *c* visualisés. Parmi les *cc* dans la matrice *c* × *cc*, les mots grammaticaux sont préservés, parce qu'ils contribuent à la différenciation sémantique pendant le regroupement effectué dans l'analyse MDS. Plusieurs critères sont envisageables pour déterminer le seuil

supérieur. La co-fréquence des *c* avec le mot-pôle permettrait de supprimer beaucoup de *c* grammaticaux, mais on perdrait aussi les *c* lexicaux les plus importants. Il en va de même pour la fréquence absolue des *c*. Par conséquent, il est plus judicieux d'enlever manuellement tous les *c* grammaticaux dans la matrice *c* × *cc*.

La suppression manuelle des *c* grammaticaux permet non seulement de réduire le nombre de *c* sur la visualisation, mais elle permet aussi de montrer uniquement les *c* lexicaux qui sont sémantiquement plus distinctifs et donc plus pertinents. En outre, dans les configurations sans mots grammaticaux, il y a généralement moins de distorsion au moment de la représentation visuelle, ce qui se traduit par un pourcentage de stress légèrement moins élevé (cf. tableau 1). Or, ces pourcentages de stress sont toujours trop élevés, ce qui s'explique probablement par les propriétés de la matrice considérée (*c* × *cc*). Elle souffre d'un véritable problème de rareté de données, parce que de nombreux *cc* sont partagés par très peu de *c*. Par conséquent, la représentation visuelle est basée sur des informations très dispersées et de ce fait moins intéressantes. Pour enrichir la matrice, nous recourons aux cooccurrents de troisième ordre (ou *ccc*). Les informations sémantiques apportées par les cooccurrences d'un ordre supérieur sont généralement plus riches et plus robustes (Schütze, 1998).

Mot-pôle	Seuil inférieur de co-fréquence minimale	Nombre de cooccurrents (*c*)	Stress de la matrice *c* × *cc*	Stress de la matrice *c* × *ccc*
tour (fq 1476)	10	127	30%	16%
	20	62	29%	16%
	50	24	37%	13%
	100	13	**17%**	9%
	10 (sans *c* gramm.)	75	29%	**18%**
	20 (sans *c* gramm.)	33	26%	18%
usinage (fq 1577)	10	431	34%	21%
	20	203	33%	19%
	50	81	31%	17%
	100	38	23%	12%
	50 (sans *c* gramm.)	53	30%	23%
	100 (sans *c* gramm.)	18	19%	**17%**

Tableau 1. Tour et usinage : seuils inférieurs et stress des matrices c × cc et c × ccc

3.3. La valeur ajoutée des cooccurrents de troisième ordre

Dans la matrice *c* × *ccc* par mot-pôle, tous les *c* pertinents sont disposés en rangées et tous les *ccc* pertinents (pour tous les *c* pertinents et tous les *cc* pertinents) en colonnes. La valeur d'une case n'est pas simplement la valeur d'association entre un *c* et un *cc*, mais la somme de colonne d'une nouvelle matrice générée pour chaque *c* du mot-pôle, avec les *cc* en rangées et les *ccc* en colonnes. S'il y a *n ccc* au total pour tous les *cc* d'un *c*, la nouvelle matrice *cc* × *ccc* permet de calculer la somme par colonne pour générer un vecteur à *n* dimensions, qui permet

de remplir les *n* cases de la rangée *c* de la matrice *c* × *ccc*. La matrice *c* × *ccc* est moins creuse et plus intéressante pour visualiser les *c* en fonction des informations sémantiques véhiculées par tous les *ccc* de tous les *cc* de ces *c*. Dans les configurations *c* × *ccc*, les pourcentages de stress sont nettement inférieurs, comme le montre la dernière colonne du tableau 1 ci-dessus.

3.4. Discussion des résultats

Pour le mot-pôle *tour*, les configurations aux seuils 50 et 100 contiennent trop peu de *c* lexicaux pour une interprétation sémantique, en dépit du faible pourcentage de stress. La configuration au seuil 10, sans mots grammaticaux, visualisée ci-dessous (cf. figure 1), permet de formuler des conclusions sémantiques intéressantes. Le nuage de *c* dans la partie inférieure à droite (*outils, usinage, deux, broches, axes, centres*) visualise le sens technique « machine-outil pour l'usinage des pièces ». Les *c* plus généraux se retrouvent dans la partie supérieure à droite (*équipé, premier, propose, banc*). Le *c* périphérique *horizon* est une indication d'un des sens généraux du mot-pôle *tour* (p.ex. *un tour d'horizon*). Le *c mille* occupe également une position plutôt périphérique et renvoie au sens « rotation » (p.ex. *dix mille tours par minute*). Or, la visualisation montre deux autres *c* périphériques, à savoir *monobroche* et *frontaux*, qui se caractérisent par des valeurs plus élevées pour des *ccc* particuliers. Comme ces valeurs représentent la somme des valeurs d'association de tous les *cc* et *ccc* par *c*, il est difficile de retracer l'origine de leur position périphérique.

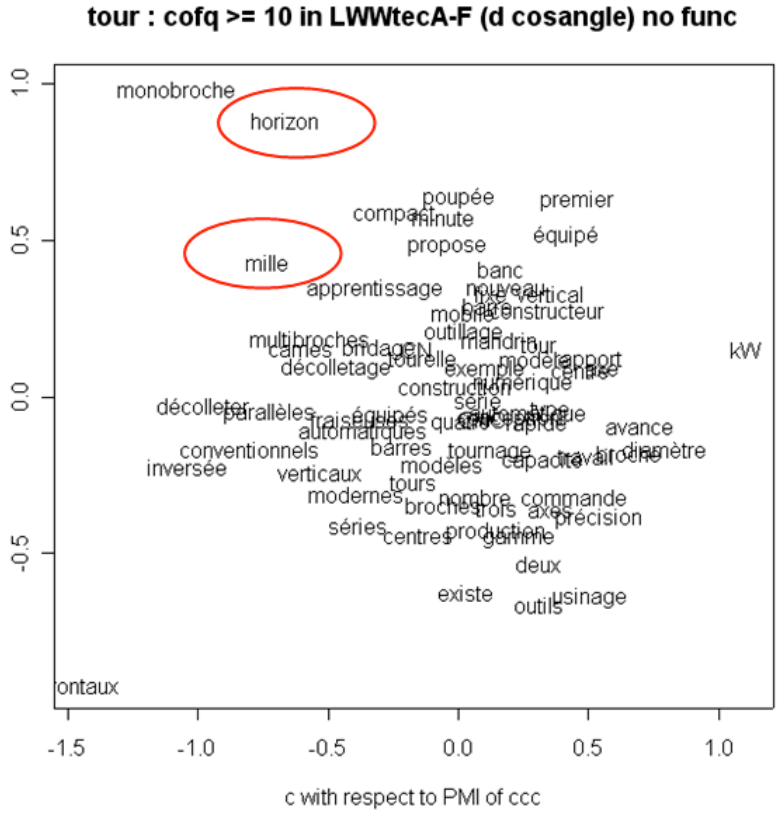

Figure 1. MDS des c *de* tour *par rapport aux* ccc *(mots lexicaux au seuil de co-fq ≥ 10)*

Apparemment, cette configuration sans *c* grammaticaux au seuil 10 (75 *c* et 18% de stress) présente plus de distorsion que celle avec des *c* grammaticaux (127 *c* et 16% de stress). Cela s'explique par le fait que les *c* grammaticaux se caractérisent par des valeurs de PMI comparables avec tous types de *cc* et que leur comportement est relativement stable à travers

le corpus. Les *c* lexicaux ont des préférences d'association plus diversifiées et donc des valeurs de PMI plus hétérogènes, ce qui rend la réduction à deux dimensions plus difficile.

Il est à signaler que la visualisation des *c* de *tour* par rapport aux *ccc* dans le sous-corpus des revues techniques se prête à une interprétation sémantique plus claire (*cf.* (Bertels et Speelman, 2013)). La figure 2 ci-dessous montre des *c* périphériques qui pointent vers des sens particuliers, à savoir *horizon* (sens général dans « tour d'horizon ») et *mille* (sens technique dans « mille tours par minute »). Dans la partie supérieure à droite, on retrouve des *c* qui attestent le sens technique particulier « tour inversé », avec *inversés, inversé* et *bibroches*. Le nuage de *c* dans la partie inférieure à gauche visualise le sens technique « machine-outil pour l'usinage des pièces ». La visualisation réalisée sur l'ensemble du corpus (cf. figure 1 ci-dessus) pourrait être considérée comme la combinaison ou la superposition de 4 visualisations pour chacun des 4 sous-corpus (revues techniques, fiches techniques, normes et directives, manuels). Cela nous incite à effectuer les analyses MDS par sous-corpus, parce que les figures 1 et 2 indiquent que les cooccurrents pertinents de *tour*, qui sont différents selon le sous-corpus analysé, ont un comportement cooccurrentiel différent selon le sous-corpus.

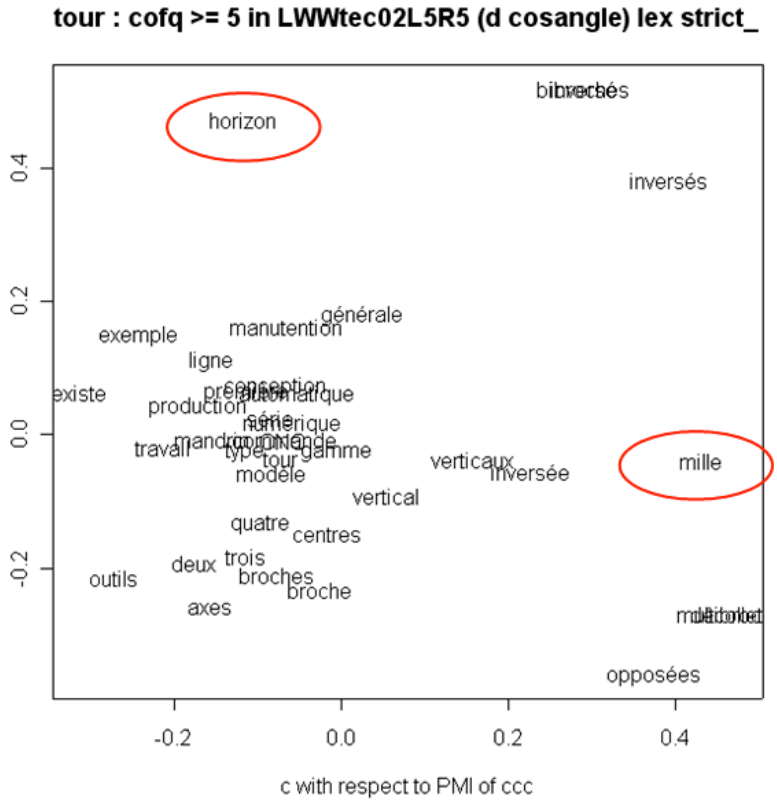

Figure 2. MDS des c *de* tour *par rapport aux* ccc *(mots lexicaux au seuil de co-fq ≥ 5) : revues*

Les configurations intéressantes pour le mot-pôle *usinage*, aux seuils 50 et 100 et sans *c* grammaticaux, se caractérisent par un stress plutôt élevé, de 23% et de 17% respectivement (cf. tableau 1). La figure 3 ci-dessous de la répartition des *c* pertinents du mot-pôle *usinage* au seuil 100 montre que les *c* se regroupent en fonction des unités polylexicales qu'ils constituent, par exemple *enlèvement de copeaux*, *haute précision*, *grande vitesse*.

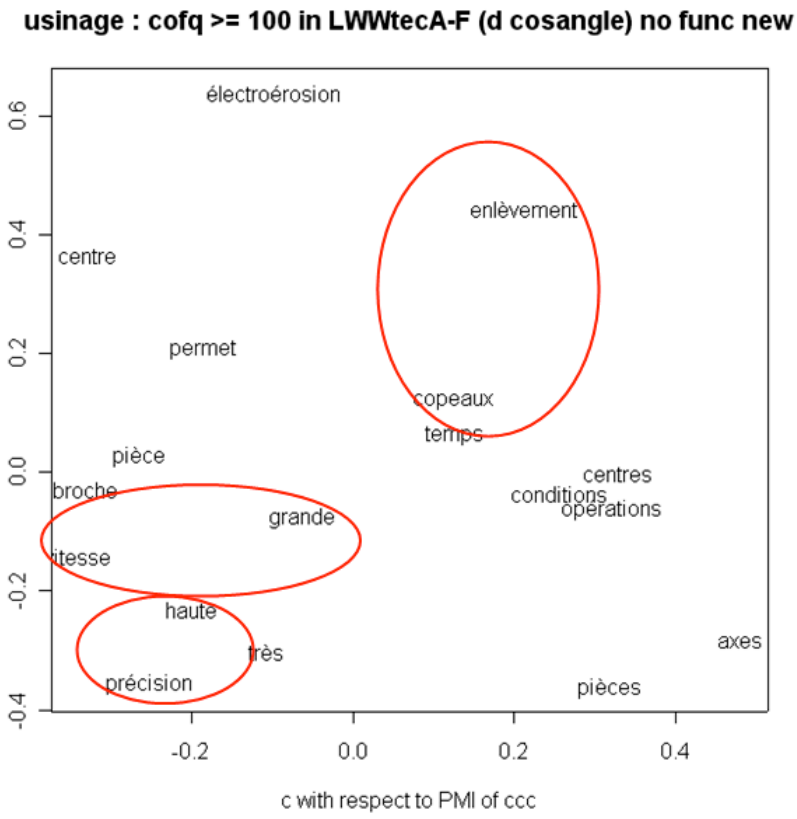

Figure 3. MDS des c *de* usinage *par rapport aux* ccc *(mots lexicaux au seuil de co-fq ≥ 10)*

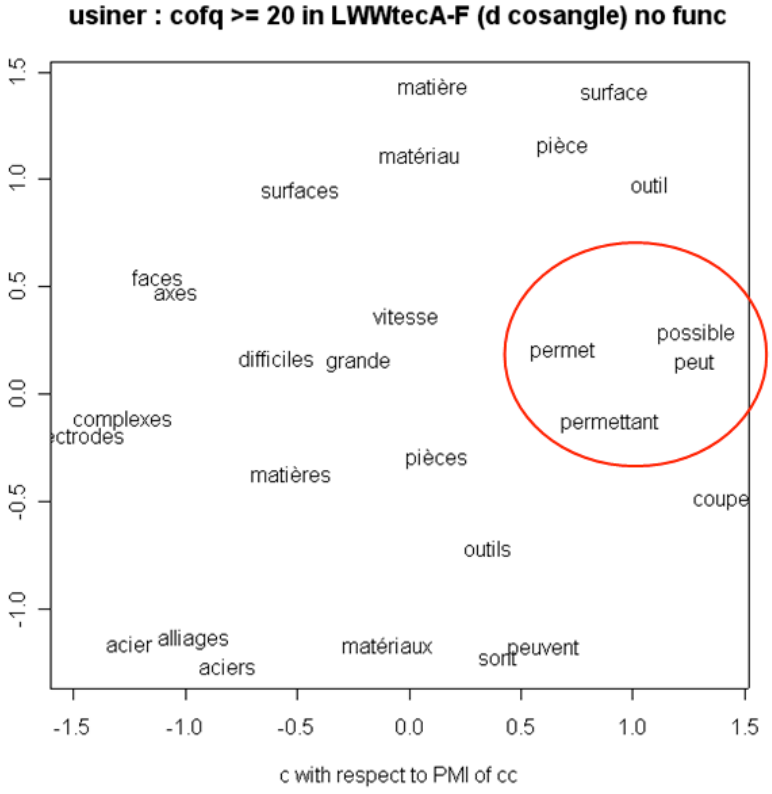

Figure 4. MDS des c *de* usiner *par rapport aux* cc *(mots lexicaux au seuil de co-fq ≥ 20)*

Le mot-pôle *usiner* est sémantiquement vague : il a un sens plutôt général qui est précisé selon le complément. La configuration *c* × *cc* au seuil 20 (avec 27 *c*) dépasse le seuil de stress (24%), mais elle montre tout de même un cluster intéressant de « possibilité » (cf. figure 4). De plus, elle suggère qu'il serait plus judicieux de considérer les *c* au niveau des lemmes, puisque les formes du singulier se retrouvent majoritairement en haut de la figure 4 (*matière*, *matériau*, *outil*, *pièce*) et les formes du pluriel en bas (*matières*, *matériaux*, *outils*, *pièces*). Les autres figures ci-dessus confirment l'idée du repérage des *c* au niveau des lemmes. Cela permettrait une interprétation sémantique plus univoque de l'analyse MDS. Au niveau des *cc* et *ccc*, les formes graphiques seraient maintenues, parce qu'elles véhiculent des informations utiles pour la désambiguïsation, sans compliquer l'interprétation de la visualisation.

Finalement, nous discutons la visualisation des *c* pertinents du mot-pôle *Iso*, sémantiquement plutôt homogène. La configuration *c* × *ccc*, au seuil 10 et sans mots grammaticaux, relève 32 *c* (stress acceptable de 12%). La figure 5 ci-dessous montre que les indications chiffrées et les renvois aux normes se regroupent à gauche de la visualisation. Les indications *9001*, *9002* et *9000* se regroupent en bas de la visualisation, tout comme *certifiée* et *certification* : ces *c* réfèrent à la certification de qualité, un usage particulier du mot-pôle *Iso*.

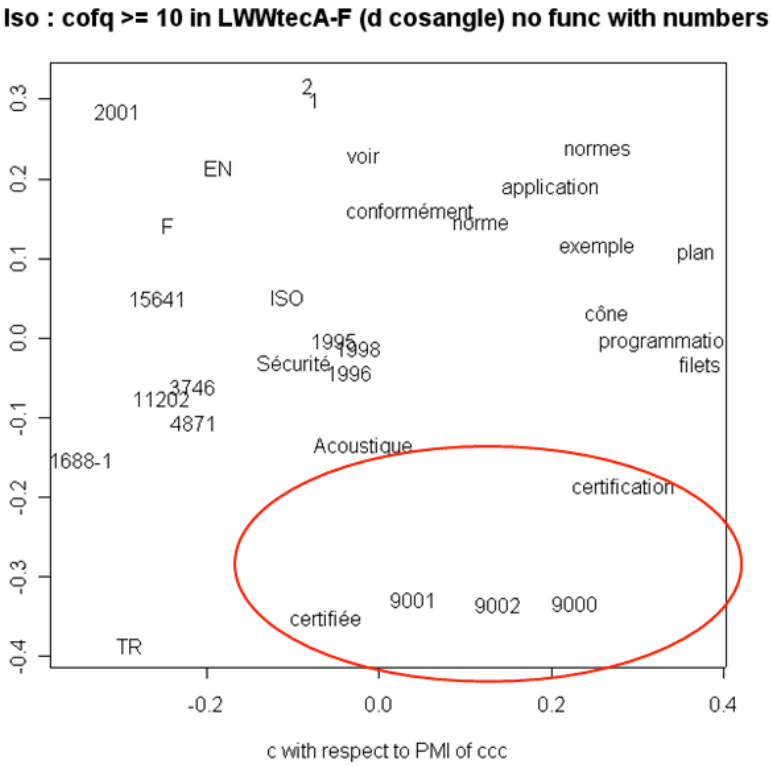

Figure 5. MDS des c *de* Iso *par rapport aux* ccc *(mots lexicaux au seuil de co-fq ≥ 10)*

Les configurations pour le mot-pôle *avance* (*c* lexicaux par rapport aux *ccc*) dépassent le seuil de stress acceptable. Pour le mot-pôle *machine*, qui est très fréquent dans le corpus technique (fq 12671), nous recensons plusieurs centaines de *c* pertinents. Par conséquent, les analyses MDS de regroupement des *c* devraient se faire en fonction de la classe lexicale des *c*, c'est-à-dire une analyse réalisée pour les adjectifs, pour les substantifs, pour les infinitifs (par exemple *machine* + *à* + *usiner*, *fraiser*, *meuler*, *mesurer*, …). Ces analyses MDS par classe lexicale des cooccurrents de premier ordre permettraient de constituer la transition vers l'analyse des unités polylexicales.

4. Conclusion et recherches futures

Dans cet article, nous avons décrit la méthodologie pour le regroupement et la visualisation des cooccurrents de premier ordre, en fonction des cooccurrents de deuxième ordre et de troisième ordre partagés, pour plusieurs mots-pôles techniques sémantiquement homogènes et hétérogènes dans le corpus technique comportant 1,7 million d'occurrences. Nous avons discuté l'importance du seuil inférieur pour la pertinence des cooccurrents et l'importance de la suppression des mots grammaticaux pour l'interprétation sémantique. Par ailleurs, nous avons démontré que l'intégration des cooccurrents de troisième ordre partagés permet d'enrichir les informations sémantiques véhiculées, de réduire le pourcentage de stress de l'analyse MDS et de faciliter l'interprétation sémantique. Les visualisations en 2D pour plusieurs mots-pôles nous ont permis de discerner des groupes de cooccurrents sémantiquement liés, et ainsi mieux de comprendre l'hétérogénéité sémantique de ces mots-pôles analysés.

Les expérimentations et analyses exploratoires nous incitent à envisager plusieurs pistes de recherches futures. Premièrement, les cooccurrents de premier ordre pourront être identifiés au niveau des lemmes et pas au niveau des formes fléchies, ce qui permettra une interprétation sémantique plus cohérente des visualisations. Ensuite, les analyses MDS seront effectuées par sous-corpus, afin de comparer la répartition visuelle des cooccurrents de premier ordre en fonction des cooccurrents de deuxième et troisième ordre partagés dans les différents sous-corpus, qui ont chacun leurs particularités stylistiques et thématiques. De plus, les analyses MDS devront se faire en fonction de la classe lexicale des cooccurrents de premier ordre et aussi en fonction de la classe lexicale des cooccurrents de deuxième et troisième ordre partagés. Cette piste de recherche permettra de combiner des informations de cooccurrence statistique avec des informations de cooccurrence syntaxique, en termes de patrons syntaxiques privilégiés. Elle permettra ainsi de faire un premier pas vers l'identification et l'analyse des unités polylexicales, importantes dans un corpus de langue spécialisée.

Références

Bertels A. et Speelman D. (2013). Exploration sémantique visuelle à partir des cooccurrences de deuxième et troisième ordre. In *Actes de TALN 2013 (Traitement Automatique des Langues Naturelles) Atelier SemDis (Sémantique Distributionnelle)*, pp. 126-139.

Bertels A., Speelman D. et Geeraerts D. (2010). La corrélation entre la spécificité et la sémantique dans un corpus spécialisé. *Revue de Sémantique et de Pragmatique*, vol.(27): 79-102.

Church K.W. et Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, vol.(16-1): 22-29.

Cox T.F. et Cox M.A.A. (2001). *Multidimensional Scaling*. Boca Raton, FL. Chapman & Hall.

Desbois D. (2005). Une introduction au positionnement multidimensionnel. *Modulad*, vol.(32): 1-28.

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

Evert S. (2007). *Corpora and Collocations*. Extended Manuscript of Chapter 58 of Lüdeling A. & M. Kytö, 2008, Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin. http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf.

Fabre C., Habert B. et Labbé D. (1997). La polysémie dans la langue générale et les discours spécialisés. *Sémiotiques*, vol.(13): 15-30.

Ferret O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. In *Actes de TALN 2010 (Traitement Automatique des Langues Naturelles)*, http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_77.pdf.

Grefenstette G. (1994). Corpus-derived first, second and third-order word affinities. In Martin W., Meijs W. e.a., editors, *Proceedings of Euralex 1994. International Congress on Lexicography,* pp. 279-290.

Heylen K., Speelman D. et Geeraerts D. (2012). Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pp. 16-24.

Kruskal J.B. et Wish M. (1978). *Multidimensional Scaling. Sage University Paper series on Quantitative Applications in the Social Sciences*, number 07-011. Newbury Park, CA. Sage Publications.

Lemaire B. et Denhière G. (2006). Effects of High-Order Co-occurrences on Word Semantic Similarity. *Current Psychology Letters*, vol.(18-1). http://cpl.revues.org/index471.html.

Morardo M. et Villemonte de La Clergerie E. (20013). Vers un environnement de production et de validation de ressources lexicales sémantiques. In *Actes de TALN 2013 (Traitement Automatique des Langues Naturelles) Atelier SemDis (Sémantique Distributionnelle)*, pp. 167-180.

Morlane-Hondère F. (2013). Utiliser une base distributionnelle pour filtrer un dictionnaire de synonymes. In *Actes de TALN 2013 (Traitement Automatique des Langues Naturelles) Atelier SemDis (Sémantique Distributionnelle)*, pp. 112-125.

Padó S. et Lapata M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, vol.(33-2): 161-199.

Peirsman Y. et Geeraerts D. (2009). Predicting Strong Associations on the Basis of Corpus Data. In *Proceedings of EACL-2009 (European Chapter of the Association for Computational Linguistics)*, pp. 648-656.

Roy T. (2007). *Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires.* Thèse de doctorat de l'Université de Caen Basse-Normandie, France.

Sahlgren M. (2006). *The Word-Space Model*. Ph.D. thesis, Stockholm University, Sweden.

Sahlgren M. (2008). The Distributional Hypothesis. *Rivista di Linguistica*, vol.(20-1): 33-53.

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

Turney P.D. et Pantel P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, vol.(37): 141-188.

Van der laan M.J. et Pollard K.S. (2003). A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference,* vol.(117): 275-303.

Venables W.N. et Ripley B.D. (2002). *Modern Applied Statistics with S* (Fourth edition). New York. Springer-Verlag.

Wielfaert T., Heylen K. et Speelman D. (2013). Interactive visualizations of Semantic Vector Spaces for lexicological analysis. In *Actes de TALN 2013 (Traitement Automatique des Langues Naturelles) Atelier SemDis (Sémantique Distributionnelle)*, pp. 154-166.