Analyse sémantique distributionnelle dans un corpus technique : les enjeux sémantiques dans un jeu de statistiques

Ann Bertels[1,2], Dirk Speelman[2]

[1]ILT, KU Leuven, Leuven – Belgique

[2]QLVL, KU Leuven, Leuven – Belgique

Abstract

This paper addresses the methodology of a distributional analysis in a relatively small corpus in a very specific domain. The aim is to show the semantic issues at stake in a series of statistical analyses. By fine-tuning these statistical analyses and by using enriched distributional data, we try to come to more interesting and more relevant semantic interpretations. In a small technical corpus, first-order co-occurrences of a French polysemous and homonymous node (*tour*) are clustered with respect to shared second-order and third-order co-occurrences. A Multidimensional Scaling analysis is carried out to plot first-order co-occurrences on a 2D-plot, in order to show semantically related co-occurrences. In this paper, we discuss several experiments in various configurations with varying parameters, taking into account POS-tags of co-occurrences and lemmas of first-order co-occurrences. Furthermore, the co-occurrence data matrix is enriched by weighting information based on the semantically most relevant co-occurrences. Finally, we show the outcome of an alternative cluster analysis and 3D-plots. In this technical corpus, with its specific semantic characteristics, we look for parameter combinations which are both efficient from a statistical point of view and relevant from a semantic point of view.

Résumé

Cette communication présente la méthodologie d'une analyse sémantique distributionnelle dans un corpus de taille modeste relevant d'un domaine très spécialisé. Le but est de montrer les enjeux sémantiques dans un ensemble d'analyses statistiques. En affinant les analyses statistiques et en enrichissant les données distributionnelles, nous essayons d'aboutir à des interprétations sémantiques plus intéressantes et plus pertinentes. Dans un corpus technique de petite taille, nous procédons au regroupement des cooccurrents de premier ordre d'un mot-pôle polysémique et homonymique (*tour*), en fonction des cooccurrents de deuxième et de troisième ordre partagés. L'analyse statistique de positionnement multidimensionnel permet de positionner les cooccurrents de premier ordre les uns par rapport aux autres, pour ainsi visualiser en 2D des groupes de cooccurrents sémantiquement liés. Nous présentons plusieurs expérimentations pour la mise au point du paramétrage des configurations, prenant en compte notamment la catégorie grammaticale des cooccurrents et les lemmes des cooccurrents de premier ordre. Dans un souci d'enrichir la matrice des cooccurrences, nous recourons également à un facteur de pondération en fonction des cooccurrents sémantiquement plus pertinents. Finalement, nous discutons quelques analyses statistiques de regroupement et de visualisation alternatives. Dans ce corpus technique, avec ses particularités sémantiques, nous sommes à la recherche de combinaisons de paramètres performantes du point de vue statistique et pertinentes du point de vue sémantique.

Key words : Distributional semantic analysis, cluster analysis, second-order and third-order co-occurrences.

1. Introduction

Cette communication s'inscrit dans le cadre de la sémantique distributionnelle et vise à montrer les enjeux sémantiques dans un ensemble d'analyses statistiques. A cet effet, nous procédons à des analyses sémantiques distributionnelles dans un corpus technique spécialisé de taille modeste (360 000 occurrences), relevant du domaine des machines-outils pour l'usinage des

métaux. Ces analyses consistent à regrouper les cooccurrents de premier ordre d'un mot-pôle polysémique et homonymique (*tour*), en fonction des cooccurrents de deuxième et de troisième ordre partagés. Ensuite, elles permettent de positionner les cooccurrents de premier ordre les uns par rapport aux autres, pour visualiser en 2D ou en 3D des groupes de cooccurrents sémantiquement liés. L'objectif linguistique n'est pas de classer les sens du mot-pôle, mais de proposer des regroupements sémantiques pertinents, susceptibles de refléter les différents sens de *tour*, et donc de mieux comprendre son caractère sémantiquement hétérogène. Nous procédons à plusieurs expérimentations pour enrichir les données distributionnelles et nous évaluons l'impact des enrichissements respectifs en fonction de critères quantitatifs (statistiques) et qualitatifs (linguistiques et sémantiques).

L'analyse sémantique distributionnelle repose sur l'hypothèse distributionnelle (Harris, 1954) selon laquelle des mots qui se trouvent dans des contextes d'apparition similaires tendent à avoir des sens similaires. Les méthodes sémantiques distributionnelles calculent la proximité sémantique entre mots sur la base des contextes qu'ils partagent dans un corpus donné. Les dernières années, elles sont devenues indispensables pour la modélisation de la sémantique lexicale (Turney et Pantel, 2010). Les modèles d'analyse distributionnelle se déclinent en plusieurs sous-types, selon le type de contexte, le niveau d'analyse et le mode de représentation computationnelle (pour un aperçu voir Heylen et Bertels, 2016). La plupart des analyses distributionnelles étudient la proximité sémantique entre mots. Certaines d'entre elles représentent les occurrences du mot *x* dans le corpus par ses « cooccurrents graphiques » dans une fenêtre de contexte donnée et s'appuient sur des mesures d'association pour déterminer les cooccurrents statistiquement pertinents (Heylen et al., 2012 ; Ferret, 2010 et 2014 ; Bernier-Colborne, 2014 ; Bertels et Speelman 2013, 2014a et 2014b ; Périnet et Hamon, 2014). D'autres études prennent en considération les relations de dépendance syntaxique entre le mot *x* et ses « cooccurrents syntaxiques » (Morlane-Hondère, 2013 ; Morardo et Villemonte de La Clergerie, 2013 ; Fabre et al., 2014). Les modèles distributionnels diffèrent aussi en fonction du mode de représentation computationnelle des données distributionnelles, soit une représentation sous forme de graphe (Morardo et Villemonte de La Clergerie, 2013, Desalle et al., 2014, Claveau et al., 2014), soit une représentation vectorielle. Les modèles vectoriels reposent sur la représentation matricielle des mots dans un espace vectoriel sémantique, par exemple une matrice mots × mots, qui caractérise les mots (en rangées) à partir de leurs cooccurrents graphiques (en colonnes). Le nombre élevé de caractéristiques contextuelles est réduit à un nombre limité de « dimensions sémantiques », notamment par la technique de positionnement multidimensionnel, pour une visualisation qui permet d'accéder plus facilement aux relations sémantiques qui se dégagent (Wielfaert et al., 2013).

Dans nos analyses distributionnelles, l'objet d'analyse se situe au niveau des cooccurrents de premier ordre d'un mot-pôle polysémique et homonymique dans le corpus technique, à savoir *tour*. Les cooccurrents de premier ordre sont caractérisés à partir de leurs cooccurrents graphiques, soit les cooccurrents de deuxième ordre, sous forme d'une matrice de données distributionnelles. Ensuite, les cooccurrents de premier ordre sont regroupés et positionnés les uns par rapport aux autres, en fonction des cooccurrents de deuxième ordre partagés. Cela permet de visualiser des groupes de cooccurrents sémantiquement liés et d'accéder à la sémantique du mot-pôle *tour*. Plus concrètement, l'analyse distributionnelle consiste en trois étapes, à savoir (a) la création d'une matrice de données distributionnelles, (b) l'analyse statistique de regroupement des cooccurrents de premier ordre et (c) la visualisation des résultats. Les expérimentations décrites dans cet article visent à enrichir et à affiner chacune

des étapes afin d'aboutir à des interprétations sémantiques plus intéressantes. Toutefois, mise au point statistique ne rime pas toujours avec mise au point sémantique.

La suite de l'article est organisée comme suit. Dans la section 2, nous expliquons la mise au point de la première étape de l'analyse sémantique distributionnelle, à savoir l'enrichissement des données distributionnelles ou l'enrichissement de la matrice des cooccurrences. Les expérimentations pour la mise au point du paramétrage des configurations prennent en considération non seulement la catégorie grammaticale des cooccurrents et les lemmes des cooccurrents de premier ordre, mais aussi un facteur de pondération en fonction des cooccurrents sémantiquement plus pertinents. Ensuite, nous discutons les résultats des analyses statistiques de positionnement multidimensionnel pour le regroupement (*clustering*) et la visualisation des cooccurrents de premier ordre (section 3). Nous présentons aussi des analyses statistiques alternatives. Finalement, la section 4 présente une conclusion et quelques pistes de recherches futures.

2. Matrice des cooccurrences

Le point de départ de nos analyses sémantiques distributionnelles est la création d'une matrice des cooccurrences ou des données distributionnelles à partir du corpus technique, à l'aide de scripts en Python. Les rangées de cette matrice contiennent les cooccurrents de premier ordre (ou *c*) du mot-pôle *tour*. Les colonnes comprennent leurs cooccurrents, c'est-à-dire les cooccurrents de deuxième ordre (ou *cc*) du mot-pôle. Pour identifier les cooccurrents statistiquement pertinents, nous nous appuyons sur la mesure d'association de l'information mutuelle spécifique ou *Pointwise Mutual Information* (PMI) (Church et Hanks, 1990), qui est couramment adoptée en sémantique distributionnelle (Wielfaert et al., 2013 ; Bernier-Colborne, 2014). Nous respectons le seuil de co-fréquence minimale de 5 (Evert, 2007), dans une fenêtre d'observation de 5 mots à gauche et à droite. Parmi les *c*, les mots grammaticaux sont supprimés. Par contre, parmi les cooccurrents d'un ordre supérieur, ils sont conservés, étant susceptibles d'apporter des informations sémantiques utiles ; par exemple *pendant* indique qu'il s'agit d'un processus. La valeur d'une case dans la matrice des cooccurrences est la valeur d'association entre le *c* et le *cc*, soit la valeur de PMI.

Il résulte des expérimentations précédentes sur le corpus technique (Bertels et Speelman, 2014a) que la matrice *c* × *cc* souffre d'un problème de rareté des données. De nombreux *cc* sont partagés par très peu de *c*. L'analyse statistique et la représentation visuelle de ces données distributionnelles seraient basées sur des informations trop dispersées et de ce fait moins pertinentes. Pour y remédier, nous faisons appel aux cooccurrents d'un ordre supérieur (Grefenstette, 1994), c'est-à-dire aux cooccurrents de troisième ordre (ou *ccc*) du mot-pôle *tour*. Dans la matrice *c* × *ccc*, 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 correspond à 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 un souci d'aboutir à des interprétations sémantiques plus pointues et de ce fait plus pertinentes, nous décidons d'enrichir la matrice *c* × *ccc*. Nous prenons en considération plusieurs configurations de paramètres, en faisant varier un paramètre à la fois, pour évaluer l'impact des enrichissements respectifs.

2.1. Enrichissement linguistique

Premièrement, nous procédons à un enrichissement linguistique, qui consiste dans un premier temps à considérer les lemmes des cooccurrents de premier ordre, au lieu des formes fléchies. Dans un deuxième temps, il consiste à considérer la catégorie grammaticale (*POS*) de tous les cooccurrents pertinents (*c, cc* et *ccc*). Ces enrichissements linguistiques respectifs donnent lieu à quatre configurations de paramètres : la configuration de base dans laquelle les *c* sont des formes fléchies (1), la configuration avec les lemmes des *c* (2) et les deux configurations avec les indications de catégorie grammaticale, d'abord pour les formes fléchies des *c* (3) et ensuite pour les lemmes des *c* (4).

2.2. Enrichissement sémantique

Ensuite, nous proposons d'enrichir le contenu des matrices en intégrant un facteur de pondération en fonction des cooccurrents sémantiquement plus pertinents. Les valeurs distributionnelles dans une matrice *c* × *ccc* enrichie sémantiquement ne représenteront pas simplement la somme des valeurs d'association PMI entre les *ccc* et *cc*, mais elles seront pondérées par les valeurs d'association entre les *cc* et le *c*. La pondération permettra d'intégrer dans les données distributionnelles l'importance d'un *cc* fortement associé au *c* et donc sémantiquement plus pertinent, puisqu'il pèsera plus lourd. Cette information sera prise en considération dans les valeurs représentées dans la matrice *c* × *ccc*, pondérée en fonction de la PMI entre *cc* et *c*. Lorsque les *c* sont regroupés en fonction des *cc* et *ccc* partagés, l'analyse statistique et la visualisation seront basées sur des informations sémantiquement plus riches.

Nous procédons également à l'intégration d'un deuxième facteur de pondération, en fonction de la valeur d'association PMI entre le *c* et le mot-pôle. Un *c* fortement associé au mot-pôle et sémantiquement plus pertinent, sera caractérisé par des valeurs plus importantes dans la matrice *c* × *ccc*, qui sera pondérée à deux reprises. Lorsque les *c* sont regroupés et positionnés les uns par rapport aux autres, l'analyse tiendra compte des valeurs d'association entre *ccc* et *cc*, mais aussi entre *cc* et *c*, et même entre *c* et le mot-pôle. Ces enrichissements sémantiques permettent de générer trois configurations : pas de pondération (i), une pondération en fonction du premier facteur de pondération (PMI entre *cc* et *c*) (ii) et une pondération en fonction des deux facteurs de pondération (iii). La combinaison de tous les paramètres d'enrichissement linguistique et sémantique génère douze configurations de la matrice des cooccurrences.

3. Analyses statistiques de regroupement

3.1. Analyses de positionnement multidimensionnel

Pour le regroupement et la visualisation des *c*, nous recourons à l'analyse de positionnement multidimensionnel (*MultiDimensional Scaling* ou MDS) (Kruskal et Wish, 1978, Cox et Cox, 2001, Venables et Ripley, 2002, Borg et Groenen, 2005). La technique de MDS[1] est implémentée dans le paquet `MASS` du logiciel d'analyse statistique R[2]. Dans nos analyses, nous

[1] 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).

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

utilisons le positionnement non métrique `isoMDS`. Cette technique permet d'analyser une matrice pour un ensemble de données disposées en rangées (ici : les *c*) à partir de leurs valeurs pour plusieurs variables disposées en colonnes (ici : les *ccc*). Les données de la matrice sont réarrangées de façon à obtenir la configuration visuelle qui représente le mieux les distances observées entre les *c*. La qualité est évaluée à l'aide du pourcentage de stress ; celui-ci doit être minimal pour garantir la fiabilité de la représentation visuelle par rapport aux données de la matrice. En règle générale, un pourcentage de stress inférieur à 10% est excellent et un pourcentage supérieur à 15% est inacceptable (Borg et Groenen, 2005). À partir de la matrice *c* × *ccc*, nous générons une matrice de similarité en nous appuyant sur la métrique du cosinus, jugée performante en sémantique distributionnelle (Padó et Lapata, 2007 ; Sahlgren, 2008).

3.1.1. Analyse MDS en 2D

La matrice de similarité est soumise à une analyse MDS, qui consiste à regrouper les cooccurrents de *tour* en fonction des valeurs d'association similaires et à visualiser ces proximités et distances sémantiques dans un espace à 2 ou 3 dimensions, par défaut en 2D. Le MDS est donc une technique de réduction de dimensions. Le tableau 1 montre les pourcentages de stress de l'analyse MDS en 2D pour les douze configurations présentées dans la section 2.

	MDS_2D	(i) pas de pondération	(ii) pondération PMI(*cc-c*)	(iii) pondération PMI(*cc-c*) et PMI(*c-tour*)
1	*c* = formes fléchies (38*c*)	**14,81%**	19,13%	19,13%
2	*c* = lemmes (47*c*)	16,67%	23,30%	23,30%
3	*c* = formes fléchies (35*c*) avec POS des *c, cc, ccc*	17,61%	18,31%	18,31%
4	*c* = lemmes (48*c*) avec POS des *c, cc, ccc*	18,18%	23,16%	23,16%

Tableau 1 : pourcentages de stress de l'analyse MDS en 2D

La configuration (1i) est la configuration de base, sans enrichissement linguistique et sans pondération. Le tableau 1 ci-dessus montre que c'est la seule configuration avec un pourcentage de stress acceptable, même s'il frôle le seuil de 15%. La visualisation en 2D (cf. figure 1), représente bien le caractère sémantiquement hétérogène du mot-pôle *tour*. Les proximités et les distances sémantiques montrent quelques cooccurrents isolés et quelques groupes de cooccurrents sémantiquement liés. Les *c* 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 dans la partie inférieure à gauche visualise le sens technique « machine-outil pour l'usinage des pièces ». La présence de plusieurs petits regroupements *c* témoigne de la variation des contextes d'apparition. On observe un contexte spécialisé « tour (à) deux broches » à gauche en bas : *axes*, *broches*, *broche*, *outils*, *deux*, *trois*, *centres* et même au milieu *vertical* et *verticaux*. À gauche au milieu on retrouve *commande*, *numérique*, *CNC*, *gamme*, *série*, *type* et quelques *c* moins spécialisés. Les *c* plus généraux comme *manutention*, *ligne*, *générale*, *production* et *conception* se retrouvent dans la partie supérieure de ce nuage et témoignent d'un contexte d'apparition plus général.

Comme les résultats des configurations avec pondération dépassent largement le seuil de 15%, nous avons intégré les informations d'association entre le *c* et le mot-pôle *tour* dans le graphe de la configuration de base, en rajoutant des couleurs. Les *c* qui sont associés le plus fortement

au mot-pôle (PMI>6) sont visualisés en mauve : ils sont sémantiquement très pertinents et ils se situent à des positions plutôt périphériques. Les *c* de la tranche suivante (PMI>5) sont visualisés en cyan. Les trois tranches suivantes sont visualisées respectivement en bleu foncé, vert et rouge. Les *c* moins pertinents sont indiqués en noir. Il semble que l'analyse MDS en 2D n'arrive pas à gérer le surplus d'informations dans les configurations enrichies, tant au niveau linguistique (i.e. lemmes et POS) qu'au niveau sémantique (i.e. pondération(s)). Plus la matrice des cooccurrences s'enrichit, plus il est difficile, d'un point de vue technique et statistique, de passer d'un espace hautement dimensionnel, à savoir espace à *n* dimensions pour les *n* colonnes dans la matrice des cooccurrences, et de cerner et de représenter toutes ces informations en 2D. L'étape suivante dans notre recherche méthodologique à travers les différentes analyses statistiques consistera à effectuer une analyse MDS en 3D. Une dimension supplémentaire permettra de représenter plus d'informations lors de la réduction des dimensions.

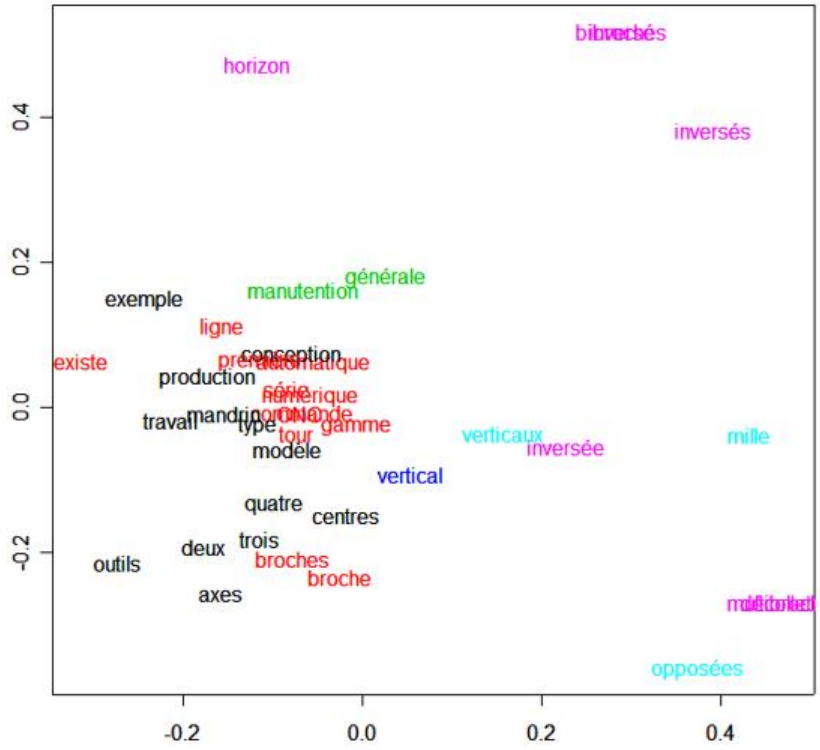

Figure 1 : MDS en 2D dans la configuration de base (1i) avec couleurs des valeurs PMI (c-tour)

3.1.2. Analyse MDS en 3D

Le tableau 2 ci-dessous montre les pourcentages de stress de l'analyse MDS en 3D pour les douze configurations envisagées. Il est clair que les pourcentages sont nettement inférieurs aux pourcentages de stress de l'analyse MDS en 2D (cf. tableau 1) et qu'ils sont presque partout acceptables (visualisé en vert). Le pourcentage le plus bas par configuration linguistique (par rangée du tableau 2) est indiqué en gras. Le tableau 2 montre clairement que l'analyse MDS en 3D sait mieux gérer les informations enrichies linguistiquement et sémantiquement. La configuration la plus performante du tableau est toujours la configuration de base (1i), mais on voit que les configurations pondérées pour les formes fléchies des *c* avec indication de classe lexicale (3ii et 3iii) affichent des pourcentages de stress très intéressants! Les configurations pour les lemmes des *c* avec indication de classe lexicale (4ii et 4iii) frôlent le seuil de 15%, mais restent acceptables statistiquement.

	MDS_3D	(i) pas de pondération	(ii) pondération PMI(*cc-c*)	(iii) pondération PMI(*cc-c*) et PMI(*c-tour*)
1	*c* = formes fléchies (38*c*)	**10,63%**	12,47%	12,47%
2	*c* = lemmes (47*c*)	11,68%	16,38%	16,38%
3	*c* = formes fléchies (35*c*) avec POS des *c, cc, ccc*	12,58%	**11,98%**	**11,98%**
4	*c* = lemmes (48*c*) avec POS des *c, cc, ccc*	**11,86%**	15,09%	15,09%

Tableau 2 : pourcentages de stress de l'analyse MDS en 3D

La visualisation de l'analyse MDS en 3D se présente sous forme de graphe en 3D[3], comme le montre la figure 2 ci-dessous pour la configuration 3iii. Les résultats de l'analyse MDS en 3D visualisés ci-dessous confirment la position périphérique de *mille* et *exemple* (cf. figure 1). Nous retrouvons le contexte spécialisé « tour (à) deux broches » en bas de ce graphe en 3D, avec les cooccurrents *axes*, *broches*, *broche*, *outils*, *deux*, *trois*, *centres*. Le cooccurrent *horizon* occupe également une position plutôt périphérique, mais ce n'est pas bien visible sur ce graphe en 3D. Premièrement, les indications de classe lexicale, par exemple |nom, alourdissent les informations visualisées sur le graphe. Deuxièmement, même si les trois dimensions sont intégrées dans les résultats, le graphe affiché ci-dessous (cf. figure 2) ne permet pas de montrer toute la richesse des informations distributionnelles représentées, puisqu'il ne permet pas d'accéder visuellement aux trois dimensions représentées. Il faudrait faire pivoter le graphe sur l'axe vertical pour mieux visualiser les distances sur l'axe de la profondeur, notamment la position plutôt périphérique du cooccurrent *horizon*.

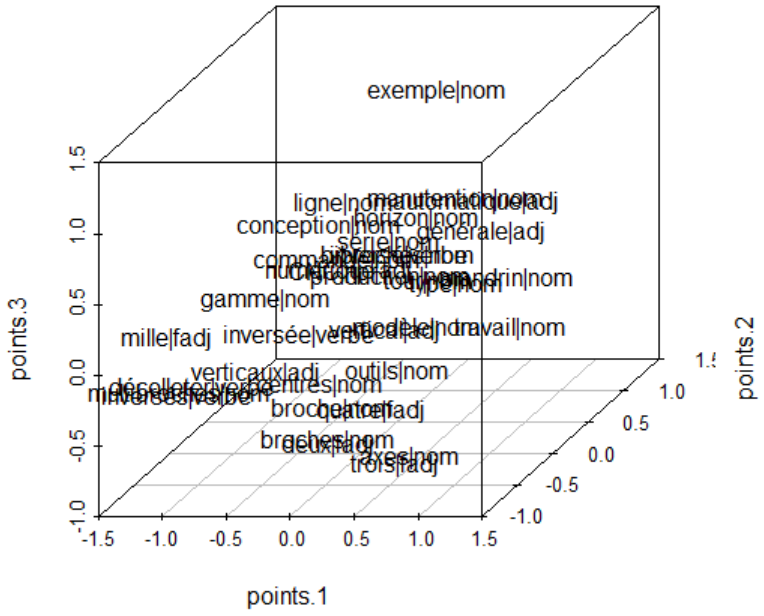

Figure 2 : MDS en 3D dans la configuration enrichie (3iii) (formes fléchies des c*, POS, pondérations)*

[3] Graphe réalisé à l'aide du paquet `scatterplot3d` dans R.

3.2. Analyses statistiques alternatives

3.2.1. Visualisation 3D rotative de l'analyse MDS en 3D

Pour mieux représenter visuellement les résultats de l'analyse MDS en 3D dans les configurations enrichies, nous recourons à une visualisation 3D dynamique et rotative[4]. Dans R, on peut la faire pivoter dans tous les sens pour regarder les cooccurrents positionnés en 3D sous tous les angles. Ces rotations facilitent considérablement la compréhension des données. Il n'est pas possible d'inclure la présentation animée de la visualisation 3D rotative dans cet article, mais elle sera montrée et commentée pendant la présentation orale.

3.2.2. Visualisation 3D rotative de l'analyse ACP

Dans le but de comparer les résultats de plusieurs analyses statistiques et de plusieurs métriques, nous avons aussi réalisé une visualisation 3D rotative à partir d'une analyse en composantes principales (ACP ou `PCA` dans R), effectuée sur la matrice des cooccurrences dans les douze configurations. Les résultats visuels sont très intéressants, surtout pour les configurations pondérées (ii et iii), où les *c* sémantiquement particuliers (*horizon*, *mille*) se situent à des positions plus périphériques. Toutefois, l'ACP repose sur la distance euclidienne et non sur le cosinus, recommandé pour l'analyse sémantique distributionnelle (Turney et Pantel, 2010).

3.2.3. Analyse statistique sans réduction de dimensions

Comme les techniques de réduction de dimensions, principalement l'analyse MDS en 2D, ont du mal à capter les informations distributionnelles enrichies dans la matrice des cooccurrences, nous faisons appel à une analyse statistique de regroupement sans réduction de dimensions. La technique de pvclust (paquet `pvclust` dans R) (Suzuki et Shimodaira , 2006) est une analyse de classification hiérarchique qui repose sur la méthode de Ward. Elle génère un diagramme arborescent, appelé dendrogramme, avec des rectangles rouges qui indiquent les bordures des clusters identifiés. La pertinence des clusters pourra être évaluée par le calcul des valeurs p à partir d'un ré-échantillonnage bootstrap multi-échelle. Des valeurs p élevées indiquent des clusters bien soutenues par les données. La technique pvclust regroupe les colonnes d'une matrice, pas les rangées. Il est donc important de transposer les données de la matrice des cooccurrences pour pouvoir regrouper les cooccurrents ou *c* de *tour*. Dans la configuration de base (1i), la seule configuration avec un pourcentage de stress acceptable dans l'analyse MDS en 2D, la technique de pvclust ne trouve que deux grands clusters de *c*, ce qui n'est pas intéressant d'un point de vue sémantique. Dans les configurations avec pondération (1ii et 1iii), avec des informations distributionnelles sémantiquement enrichies, on trouve 5 clusters. En règle générale, des configurations enrichies génèrent des clusters plus nombreux et plus intéressants sémantiquement. Cette technique de classification hiérarchique sans réduction de dimensions s'avère donc plus intéressante pour les configurations sémantiquement enrichies.

La figure 3 ci-dessous permet de combiner visuellement les résultats de deux analyses, à savoir l'analyse MDS en 3D et la technique de pvclust. Le graphe en 3D visualise l'analyse MDS en 3D dans la configuration enrichie 3iii, comme le montrait également la figure 2 ci-dessus, et les différentes couleurs représentent chacun des clusters repérés dans cette configuration avec la technique de pvclust. Par ailleurs, les couleurs de la figure 3 montrent que les *c* des clusters pvclust se situent les uns près des autres dans le graphe et qu'ils sont sémantiquement similaires, par exemple *commande, numérique, CNC* en cyan, au milieu, et *trois, axes, deux, quatre,*

[4] Visualisation animée réalisée à l'aide du paquet `rgl` dans R.

broches en vert clair et en rose en bas du graphe. Les *c* indiqués en noir ne font pas partie d'un cluster pvclust. Les résultats de la technique de pvclust semblent donc confirmer les résultats de l'analyse MDS en 3D.

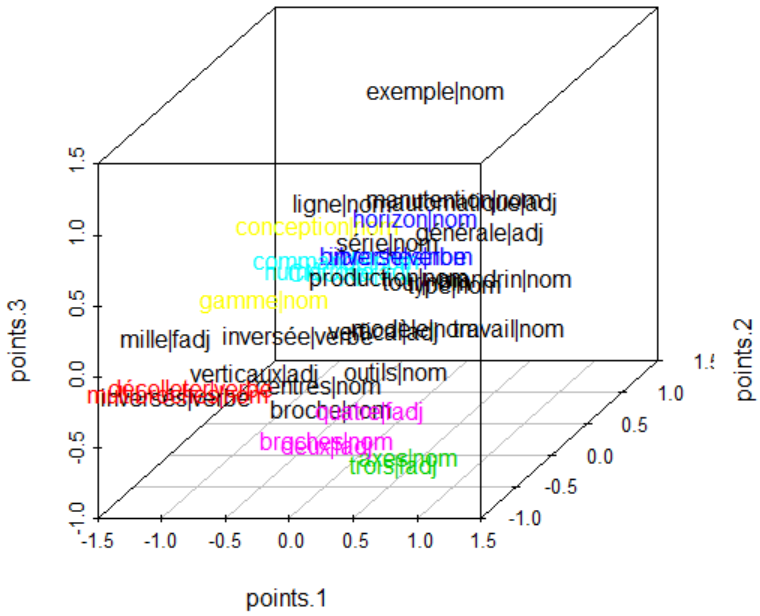

Figure 3 : MDS en 3D dans la configuration enrichie (3iii) : avec couleurs des clusters de pvclust

4. Conclusion et perspectives

Dans cet article, nous avons discuté les interprétations sémantiques dans un ensemble d'analyses statistiques. A cet effet, nous avons procédé à plusieurs expérimentations d'analyse distributionnelle dans un petit corpus technique (360 000 occurrences). En enrichissant les données distributionnelles et en affinant les analyses statistiques, nous avons réussi à trouver des interprétations sémantiques pertinentes et fiables.

Les expérimentations d'analyse distributionnelle consistent à regrouper les cooccurrents de premier ordre du mot-pôle *tour* et à les positionner les uns par rapport aux autres, en fonction des cooccurrents de deuxième et de troisième ordre partagés, dans le but de trouver des groupes de cooccurrents sémantiquement liés et d'accéder à la sémantique du mot-pôle. Nous avons procédé à plusieurs enrichissements de la matrice des cooccurrences, tant sur le plan linguistique (lemmes et indications de POS) que sur le plan sémantique (pondérations). Il s'est avéré que l'analyse MDS en 2D n'arrive pas à capter toute la richesse des informations distributionnelles dans les configurations sémantiquement enrichies. Le recours à une dimension supplémentaire permet de mieux gérer la richesse des informations dans la matrice des cooccurrences. L'analyse MDS en 3D permet effectivement d'aboutir à des résultats statistiques nettement meilleurs et la visualisation rotative en 3D constitue une aide précieuse à l'interprétation sémantique des résultats de l'analyse MDS en 3D. Par ailleurs, une analyse statistique alternative, à savoir la technique de pvclust qui procède par classification hiérarchique, confirme les résultats de l'analyse MDS en 3D.

Dans nos recherches futures, nous aimerions approfondir le problème de la rareté des données dans la matrice des cooccurrences pour vérifier la position des *c* qui se combinent très fréquemment avec très peu de *cc* très spécifiques. Nous envisageons également de conduire ces

analyses MDS et pvclust sur l'ensemble du corpus technique (1,7 million de mots) et sur des corpus relevant d'autres domaines. Finalement, nous aimerions transposer la méthodologie, adoptée dans cet article pour le mot simple *tour* dans le corpus technique, au niveau des unités polylexicales telles que, par exemple, *tour d'horizon* et *tour inversé*.

Références

Bernier-Colborne G. (2014). Analyse distributionnelle de corpus spécialisés pour l'identification de relations lexico-sémantiques. In *Actes de TALN 2014 (Traitement Automatique des Langues Naturelles) Atelier SemDis*, pages 238-251.

Bertels A. and 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*, pages 126-139.

Bertels A. and Speelman D. (2014a). Analyse exploratoire des cooccurrents de premier ordre dans un corpus technique. In *Actes de JADT 2014 (Journées internationales d'Analyse statistique des Données Textuelles)*, pages 67-78.

Bertels A. and Speelman D. (2014b). Analyse de positionnement multidimensionnel sur le corpus spécialisé TALN. In *Actes de TALN 2014 (Traitement Automatique des Langues Naturelles) Atelier SemDis*, pages 252-265.

Borg I. and Groenen P. (2005). *Modern Multidimensional Scaling: theory and applications*. (2e édition) New York: Springer-Verlag.

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

Claveau V., Kijak E. and Ferret O. (2014). Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. In *Actes de TALN 2014 (Traitement Automatique des Langues Naturelles) Atelier SemDis*, pages 220-231.

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

Desalle Y., Navarro E., Chudy Y., Magistry P. and Gaume B. (2014). BACANAL : Balades Aléatoires Courtes pour ANAlyses Lexicales Application à la substitution lexicale. In *Actes de TALN 2014 (Traitement Automatique des Langues Naturelles) Atelier SemDis*, pages 206-217.

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., Hathout N., Sajous F. and Tanguy L. (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *Actes de TALN 2014 (Traitement Automatique des Langues Naturelles) Atelier SemDis*, pages 266-279.

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.

Ferret O. (2014). Utiliser un modèle neuronal générique pour la substitution lexicale. In *Actes de TALN 2014 (Traitement Automatique des Langues Naturelles) Atelier SemDis*, pages 218-227.

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

Harris Z. (1954). Distributional structure. *Word*, vol.(10 n°23): 146-162.

Heylen K. and Bertels A. (2016). Sémantique distributionnelle en linguistique de corpus. *Langages*, vol.(201): (à paraître).

Heylen K., Speelman D. and 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*, pages 16-24.

Heylen K., Wielfaert T., Geeraerts D. and Speelman D. (2014). Monitoring Polysemy: Word Space Models as a Tool for Large-Scale Lexical Semantic Analysis. *Lingua: International Review of General Linguistics,* vol.(157):153-172.

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

Morardo M. and Villemonte de La Clergerie E. (2013). Vers un environnement de production et de validation de ressources lexicales sémantiques. In *Actes de TALN 2013 (Traitement Automatique des Langues Naturelles) Atelier SemDis*, pages 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*, pages 112-125.

Périnet A. and Hamon T. (2014). Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité. In *Actes de JADT 2014 (Journées internationales d'Analyse statistique des Données Textuelles)*, pages 507-518.

Suzuki R. and Shimodaira H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, vol.(22 n°12): 1540-1542.

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

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

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