Analyse de positionnement multidimensionnel sur le corpus spécialisé TALN

Ann Bertels[1, 2] Dirk Speelman[2]
(1) ILT, KU Leuven, Dekenstraat 6, B-3000 Leuven (Belgique)
(2) QLVL, KU Leuven, Blijde-Inkomststraat 21, B-3000 Leuven (Belgique)
ann.bertels@ilt.kuleuven.be, dirk.speelman@arts.kuleuven.be

Résumé. Cet article présente la méthodologie et les résultats d'une analyse sémantique distributionnelle, développée sur un corpus technique pour l'exploration visuelle de la proximité sémantique entre les cooccurrents d'un mot-pôle. Ici, nous utilisons cette approche sur un corpus relevant d'un autre domaine spécialisé, pour la mettre à l'épreuve et pour comparer les résultats à d'autres approches. A l'aide d'une analyse statistique de positionnement multidimensionnel (*Multidimensional Scaling* ou MDS), nous procédons au regroupement des cooccurrents de premier ordre de huit mots-pôles sélectionnés, en fonction des cooccurrents de deuxième et troisième ordre partagés. La visualisation par mot-pôle permet de cerner des groupes de cooccurrents sémantiquement similaires. Ces analyses exploratoires sur le corpus TALN visent non seulement à vérifier ce que nous apprend notre approche sur les nouvelles données, mais également à découvrir ce que ces données nous apprennent sur notre approche, dans le souci de la mettre au point.

Abstract. This paper addresses the methodology and results of a distributional semantic analysis, developed on a technical corpus for the visual exploration of the semantic proximity between the collocates of a node. We now use this approach on a corpus from another specialised domain, in order to put it to the test and compare the results to other approaches. Multidimensional scaling analysis (MDS) is carried out in order to cluster first-order co-occurrences of eight selected nodes, with respect to shared second and third-order co-occurrences. Visualisation for each node shows interesting groupings of semantically related collocates. The aim of this exploratory analysis on the TALN-corpus is not only to find out what our approach says about the new data, but also to discover what these data teach us about our approach and how we can improve and refine it.

Mots-clés : Analyse de cooccurrences, cooccurrents de deuxième et troisième ordre, positionnement multidimensionnel, regroupement, exploration sémantique visuelle.
Keywords : Co-occurrence analysis, second and third-order collocates, Multidimensional Scaling (MDS), clustering, visual semantic exploration.

1 Introduction

Cet article présente la méthodologie et les résultats d'une analyse de positionnement multidimensionnel effectuée sur le corpus TALN. Ce corpus spécialisé de petite taille (environ 2 millions d'occurrences) relève du domaine du traitement automatique des langues naturelles et contient une sélection d'articles issus des conférences TALN et RECITAL au cours de la période 2007-2013. Il a été mis à la disposition des participants à l'atelier SemDis 2014 du colloque TALN 2014. L'archive numérique a été constituée dans le but de regrouper des articles scientifiques publiés dans le domaine du TAL et d'offrir un portail facilitant l'accès à ces publications (Boudin, 2013). Les analyses décrites dans le présent article sont conduites sur la version lemmatisée et annotée du corpus TALN (Urieli, Tanguy, 2013).

La méthodologie adoptée s'inscrit dans le contexte de la sémantique distributionnelle. Elle a été développée dans le cadre d'une analyse exploratoire des données de cooccurrence dans un corpus technique relevant du domaine des machines-outils pour l'usinage des métaux, de taille comparable (environ 1,7 million d'occurrences). L'analyse exploratoire (cf. Bertels, Speelman, 2013) porte sur les cooccurrents d'un mot-pôle technique, c'est-à-dire les mots qui figurent dans une fenêtre d'observation (*span*) de 5 mots à gauche et à droite du mot-pôle. Ces cooccurrents de premier ordre du mot-pôle sont regroupés et visualisés en 2D en fonction des cooccurrents de deuxième et troisième ordre qu'ils partagent. Les cooccurrents de deuxième ordre sont définis comme les cooccurrents des cooccurrents de premier ordre du mot-pôle et les cooccurrents de troisième ordre comme les cooccurrents de ces cooccurrents de deuxième ordre. Le but de l'analyse exploratoire est de cerner des groupes de cooccurrents de premier ordre sémantiquement liés pour 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 précédemment sur le corpus technique (Bertels et al., 2010). Dans cette étude précédente, nous avons développé une mesure de monosémie qui consiste à implémenter la monosémie en termes d'homogénéité sémantique (Habert et al., 2005). Une unité lexicale monosémique apparaît dans des contextes sémantiquement plutôt homogènes, parce qu'elle se caractérise par des cooccurrents qui appartiennent à des champs sémantiques similaires. La similarité distributionnelle reflète la similarité sémantique. Par contre, une unité lexicale polysémique se caractérise par des cooccurrents sémantiquement plus hétérogènes. L'accès à la sémantique des cooccurrents d'un mot-pôle se fait à partir de leurs cooccurrents, c'est-à-dire à partir des cooccurrents de deuxième ordre (Grefenstette, 1994). Ceux-ci permettent aussi de remédier au problème de la distribution irrégulière des sens du mot-pôle dans un corpus (Habert et al., 2004).

Dans nos premières analyses de positionnement multidimensionnel sur le corpus technique, nous avons constaté que la visualisation des proximités et distances sémantiques entre les cooccurrents de premier ordre d'un mot-pôle sémantiquement hétérogène permet de mieux comprendre et interpréter son degré d'hétérogénéité sémantique (pour plus de détails, voir Bertels, Speelman, 2013). Pour le mot-pôle technique *tour*, par exemple, la visualisation de la répartition des cooccurrents de premier ordre montre quelques groupes de cooccurrents sémantiquement liés et quelques cooccurrents plutôt isolés, qui reflètent bien les différents sens de ce mot-pôle à la fois homonymique et polysémique dans le corpus technique analysé (Bertels, Speelman, 2013).

Dans le présent article, nous expliquons d'abord la problématique et les questions de recherche pour le corpus TALN (section 2), ainsi que l'approche méthodologique (section 3). Ensuite, nous discutons les résultats de l'analyse distributionnelle pour la sélection de huit mots-pôles, en prenant en considération non seulement les interprétations sémantiques, mais également les répercussions méthodologiques (section 4). Nous terminons par quelques conclusions pour l'analyse sémantique distributionnelle sur le corpus TALN et pour notre approche méthodologique (section 5).

2 Problématique

Dans le cadre de la tâche exploratoire sur le corpus TALN, nous nous proposons de mettre à l'épreuve une approche de sémantique distributionnelle. A cet effet, nous considérons les huit mots sélectionnés (*calculer*, *complexe*, *précis*, *fréquence*, *graphe*, *méthode*, *sémantique*, *trait*) comme mots-pôles et nous essayons de faire ressortir leurs propriétés sémantiques. Dans un premier temps, nous aimerions savoir ce que nous apprend notre approche sur les données du corpus TALN, c'est-à-dire ce qu'elle permet d'en retirer. Dans un deuxième temps, nous aimerions découvrir ce que nous apprennent ces données sur notre approche et comment elles pourraient contribuer à sa mise au point.

Tout d'abord, nous nous demandons si notre approche fonctionne sur un autre corpus spécialisé. Permet-elle de générer des résultats interprétables d'un point de vue sémantique, quand elle est appliquée à des données relevant d'un autre domaine que celui du corpus technique, avec ses particularités thématiques, stylistiques et sémantiques ? Si oui, quels sont ensuite les résultats de l'analyse distributionnelle sur le corpus TALN pour les huit mots-pôles proposés ? Plus particulièrement, quelles sont les conclusions sémantiques qu'on pourra en tirer ? Et, finalement, quels sont les enseignements méthodologiques ? Quelles sont les mises au point nécessaires et quels sont les paramétrages requis pour préciser les résultats et pour peaufiner la méthode ?

3 Approche méthodologique

3.1 Sémantique distributionnelle

La plupart des analyses en sémantique distributionnelle (Sahlgren, 2006 et 2008 ; Turney, 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, Villemonte de La Clergerie, 2013) soit des cooccurrents de premier ordre (Sahlgren, 2008 ; Peirsman, Geeraerts, 2009 ; Ferret, 2010 ; Heylen et al., 2012 ; Wielfaert et al., 2013). Ces dernières analyses 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 des mots sémantiquement liés (Peirsman, 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, Denhière, 2006). Dans nos analyses sur le corpus technique, nous cherchions à mieux comprendre le degré d'hétérogénéité sémantique d'un mot-pôle technique. Nous étions donc

intéressés par les rapports sémantiques entre ses cooccurrents de premier ordre. Par conséquent, dans notre approche méthodologique, 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. Le but 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 et/ou de troisième ordre partagés, pour ainsi cerner des groupes de cooccurrents de premier ordre sémantiquement similaires.

3.2 Analyse de positionnement multidimensionnel des cooccurrents de premier ordre

Pour chacun des huit mots-pôles sélectionnés sur le corpus spécialisé TALN, nous procédons d'abord à une analyse de cooccurrences, à trois reprises, pour déterminer les cooccurrents de premier ordre pertinents, ainsi que les cooccurrents de deuxième et troisième ordre pertinents. Pour identifier les cooccurrents pertinents, nous nous appuyons sur la mesure d'association de l'information mutuelle (*Pointwise Mutual Information* ou PMI) (Church, Hanks, 1990). Or, la mesure de la PMI a tendance à surestimer la valeur d'association des mots rares et de ce fait elle est moins fiable pour des cooccurrents à faible co-fréquence. Pour remédier à ce problème de fiabilité, il est conseillé de respecter un seuil de co-fréquence minimale supérieur ou égal à 5 (Evert, 2007), ce qui signifie que le mot-pôle et le cooccurrent doivent apparaître ensemble au moins cinq fois.

Pour le regroupement et la visualisation des cooccurrents de premier ordre d'un mot-pôle, en fonction des cooccurrents de deuxième et troisième ordre partagés, nous recourons à l'analyse statistique de positionnement multidimensionnel (*MultiDimensional Scaling* ou MDS) (Kruskal, Wish, 1978 ; Cox, Cox, 2001 ; Venables, Ripley, 2002). La technique de MDS[1] est implémentée dans le logiciel d'analyse statistique R[2]. 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 valeurs dans la matrice sont les valeurs d'association PMI respectives entre les *c* et 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* par mot-pôle, nous générons une matrice de distance dans le logiciel R, en calculant les distances par paire d'observations avec la métrique de l'angle du cosinus[3] (*cosine angle*). Cette métrique de distance 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. Les rangées de la matrice de base *c* × *cc*, à savoir les cooccurrents de premier ordre, sont conçues comme des vecteurs avec une valeur par colonne. 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. La matrice de distance est ensuite soumise à une analyse de positionnement multidimensionnel (MDS). Celle-ci consiste à regrouper les cooccurrents de premier ordre (*c*) d'un mot-pôle en fonction des valeurs d'association PMI similaires avec des *cc* similaires et à visualiser ces proximités et distances sémantiques en 2D. Ainsi, elle permet d'accéder à la sémantique du mot-pôle.

3.3 Configurations de paramètres pour la matrice de cooccurrences

Les analyses MDS discutées ci-dessous prennent comme point de départ une matrice de cooccurrences par mot-pôle. Celle-ci est réalisée à l'aide de scripts en Python à partir de la version lemmatisée et annotée du corpus TALN. Dans le

[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.

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

fichier *.txt mis à disposition, les scripts reprennent les colonnes deux et trois avec respectivement les formes graphiques et les lemmes, ainsi que les colonnes quatre et cinq avec les indications de classe lexicale, respectivement des formes graphiques et des lemmes. Les indications de classe lexicale permettent d'enrichir les informations sémantiques et/ou de cibler les analyses en fonction de la classe lexicale du mot-pôle ou des cooccurrents. Nous considérons plusieurs configurations de paramètres pour la matrice de cooccurrences des mots-pôles (cf. table 1), dans le but 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. S'il s'avère que les caractéristiques du mot-pôle (sa fréquence, sa classe lexicale, ses particularités sémantiques, etc.) affectent les résultats, il sera intéressant d'évaluer l'impact des caractéristiques linguistiques sur l'approche méthodologique et d'en ajuster le paramétrage. A cet effet, 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 et l'interprétation sémantique des représentations visuelles.

Paramètres	Configurations
Seuil de co-fréquence minimale	5 ou 10 ou 20 ou 50 en fonction de la fréquence du mot-pôle
Forme graphique (W) ou lemme (L) des *c* et *cc* et *ccc*	LWW versus LLL
Taille de la fenêtre d'observation (*span*)	5L5R versus 3L3R

TABLE 1: Configurations de paramètres pour la matrice de cooccurrences

Tout d'abord, nous tenons à expliquer l'importance du seuillage pour l'analyse de cooccurrences pendant la constitution de la matrice de cooccurrences. Nous introduisons un seuil inférieur de co-fréquence minimale pour tous les mots-pôles, aussi bien peu fréquents que plus fréquents, voire même très fréquents. Pour les mots-pôles peu fréquents (p.ex. *précis*, avec une fréquence de 378), nous appliquons le seuil minimal de co-fréquence minimale de 5 (cf. section 3.2). En-dessous de ce seuil, les résultats ne sont plus statistiquement fiables (Evert, 2007). Pour les mots-pôles plus fréquents (p.ex. *calculer* et *trait*, avec une fréquence d'environ 1000 occurrences), le seuil de co-fréquence est plus élevé, par exemple 10 ou 20. Pour le mot-pôle le plus fréquent (*méthode*, avec une fréquence de plus de 3800 occurrences), le seuil est fixé à une co-fréquence supérieure ou égale à 50. En général, un mot-pôle plus fréquent, voire très fréquent, se caractérise par un nombre plus élevé de cooccurrents pertinents. Dès lors, un seuil plutôt faible (par exemple ≥ 5 ou 10) recenserait énormément de *c* pertinents. Or, un nombre trop important de *c* rendrait la visualisation trop dense et dès lors illisible. Un seuil de co-fréquence plus élevé (par exemple ≥ 20 ou 50) permet de relever généralement 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 parmi les *c* poserait aussi un problème d'interprétation. Par conséquent, les mots grammaticaux sont supprimés parmi les *c,* dans les rangées de la matrice de cooccurrences, avant l'analyse MDS. Dans les colonnes de la matrice de cooccurrences, les mots grammaticaux sont conservés, parce qu'ils sont susceptibles d'apporter des informations sémantiques utiles, par exemple *pendant* indique un processus. Ce n'est pas la probabilité de la contribution sémantique des mots grammaticaux qui est différente entre les rangées et les colonnes de la matrice de cooccurrences, mais plutôt l'impact des mots grammaticaux sur la complexité de l'analyse et de la visualisation. Dans les colonnes, la présence des mots grammaticaux parmi les *cc* est parfaitement gérable. Par contre, dans les rangées, leur présence parmi les *c* rendrait la visualisation trop dense et dès lors illisible.

Des expérimentations de regroupement et de visualisation effectuées sur un extrait du corpus technique de 320 000 mots (Bertels, Speelman, 2013) et des expérimentations plus récentes sur le corpus technique entier (Bertels, Speelman, 2014) ont démontré la valeur ajoutée de la prise en considération des cooccurrents de troisième ordre (ou *ccc*). Les résultats d'une matrice de cooccurrences *c* × *ccc* sont nettement meilleurs que ceux d'une matrice *c* × *cc*, avec un pourcentage de stress inférieur et une interprétation sémantique plus intéressante. En effet, la matrice *c* × *cc* souffre souvent d'un 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). Dans une telle 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 PMI, 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* pour chaque *c* permet de

calculer la somme par colonne pour ainsi générer un vecteur à *n* dimensions, qui permet de remplir les *n* cases de la rangée *c* de la matrice *c* × *ccc* (cf. table 2). La matrice *c* × *ccc* est moins creuse et donc plus intéressante pour visualiser les *c* en fonction des informations sémantiques véhiculées par tous les *ccc* de tous les *cc* de ces *c*.

Dans une prochaine étape, nous envisageons de pondérer les différentes lignes de la matrice *cc* × *ccc* en fonction de la valeur d'association PMI entre les *c* et les *cc*. La prise en compte d'une somme pondérée permettrait d'accorder plus d'importance aux *ccc* des *cc* les plus fortement associés aux *c*. Une telle pondération constitue une mise au point très intéressante, mais une simple somme est justifiée, provisoirement, comme procédure simplifiée.

	ccc_1	ccc_2	ccc_3	ccc_4	ccc_5	ccc_6
c_1	somme de colonne pour la colonne ccc_1 dans la matrice *cc* × *ccc* pour c_1	somme de colonne pour la colonne ccc_2 dans la matrice *cc* × *ccc* pour c_1				
c_2	somme de colonne pour la colonne ccc_1 dans la matrice *cc* × *ccc* pour c_2					
c_3						
c_4						

TABLE 2: Exemple simplifié d'une matrice de cooccurrences *c* × *ccc*

Les mots-pôles de la sélection sont tous considérés au niveau des lemmes. Pour les *c* et et pour les *cc* et *ccc*, nous envisageons les deux possibilités. Lorsque les cooccurrents sont considérés au niveau des formes graphiques, ils sont susceptibles de véhiculer des informations sémantiques plus riches, comme par exemple la distinction entre *pièce usinée* (« résultat ») et *pièce à usiner* (« avant le processus d'usinage »). L'extraction des cooccurrents au niveau des formes graphiques donne lieu à la configuration LWW (*lemma – word form – word form*) (cf. table 1). Par contre, lorsque les cooccurrents sont considérés au niveau des lemmes, dans la configuration LLL (*lemma – lemma – lemma*), la visualisation des *c* gagne en lisibilité, parce que toutes les formes fléchies et conjuguées sont ramenées sous le lemme correspondant. Les *c* sont repérés dans une fenêtre d'observation (*span*) de 5 mots à gauche et à droite (5L5R) du mot-pôle, ensuite les *cc* dans une fenêtre de 5 mots à gauche et à droite des *c* et, finalement, les *ccc* dans une fenêtre de 5 mots à gauche et à droite des *cc*. Des expérimentations préalables ont démontré que cette fenêtre recense des cooccurrents sémantiquement intéressants sans introduire trop de bruit (Bertels, Speelman, 2013). Or, pour certains mots-pôles, une fenêtre plus petite de 3 mots à gauche et à droite (3L3R) s'avère plus intéressante (cf. section 4). Dans nos analyses sur le corpus TALN, les deux fenêtres seront prises en considération (cf. table 1) afin d'en évaluer l'effet.

Il est à noter qu'il aurait été intéressant d'inclure dans les configurations de paramètres aussi des informations de dépendance syntaxique, ce qui aurait permis d'évaluer la différence entre les cooccurrents de surface, tels que nous les considérons à présent, et les cooccurrents syntaxiques. L'avantage des cooccurrents syntaxiques, c'est qu'ils sont à la fois moins sensibles au bruit et au silence (Evert, 2007). La prise en compte des dépendances syntaxiques constitue certainement une piste de recherche future et permettra de continuer la mise au point de notre approche méthodologique. Les analyses de positionnement multidimensionnel décrites dans le présent article bénéficient déjà d'une mise au point par rapport à l'analyse exploratoire mentionnée ci-dessus (Bertels, Speelman, 2013), parce qu'elles prennent en considération la différence entre la forme graphique et le lemme des cooccurrents et qu'elles exploitent les indications de classe lexicale. Ces indications constituent une première étape dans la prise en compte des informations syntaxiques. Par ailleurs, notre approche méthodologique repose sur le principe de « similarité sémantique lexicale » (Ferret, 2010). Elle vise à déterminer des similarités sémantiques à partir de similarités distributionnelles et s'appuie sur des mesures d'association pour identifier les cooccurrents statistiquement pertinents, à l'instar d'autres études en sémantique distributionnelle (cf. Peirsman, Geeraerts, 2009 ; Heylen et al., 2012 ; Wielfaert et al., 2013).

4 Discussion des résultats

Les matrices de cooccurrences générées sur le corpus TALN pour les huit mots-pôles sont ensuite soumises à une analyse de positionnement multidimensionnel (*Multidimensional Scaling* ou MDS). Rappelons que celle-ci vise à regrouper les cooccurrents de premier ordre (*c*) de chaque mot-pôle en fonction des cooccurrents de deuxième ordre (*cc*) et de troisième ordre (*ccc*) partagés et à générer une représentation visuelle des *c* par mot-pôle. Les *c* qui ont des valeurs d'association PMI similaires avec des *cc* et *ccc* similaires et qui se caractérisent dès lors par des similarités distributionnelles, se retrouveront ensemble ou à de faibles distances sur la visualisation en 2D, ce qui reflète des similarités sémantiques. Ces proximités et distances en 2D permettent d'accéder à la sémantique du mot-pôle.

Comme la méthode a été développée dans un souci de mieux comprendre le phénomène d'hétérogénéité sémantique dans un corpus technique par le biais de la visualisation des distances distributionnelles entre les cooccurrents, nous présentons d'abord les résultats pour le mot-pôle polysémique *trait*, qui a de multiples usages dans le corpus TALN. Il est non seulement employé dans des sens linguistiques très spécifiques (*trait distinctif, trait sémantique*), mais également dans un sens plus général de signe de ponctuation (*trait d'union*) et dans des locutions adverbiales telles que *d'un seul trait* (« sans interruption »).

La configuration la plus intéressante pour le lemme ***trait*** (fréquence de 1804 dans le corpus TALN) permet de relever les cooccurrents au niveau des formes graphiques (LWW) dans une fenêtre de 5 mots à gauche et à droite (5L5R), avec les indications de classe lexicale (_wc) pour un seuil de co-fréquence minimale supérieur ou égal à 20. L'analyse MDS affiche un pourcentage de stress parfaitement acceptable de 12,12%. La visualisation ci-dessous montre le positionnement des 32 *c* lexicaux (cf. figure 1). Les 3 formes verbales (*permet, utilisant, utilisés*) sont indiqués en rouge, les 8 adjectifs en bleu et les 21 noms en noir. Ce qui saute aux yeux, ce sont les 3 *c* plutôt isolés en haut de la visualisation (*union*, *assignation* et *humain*). Le nom *union* réfère évidemment au terme *trait d'union*, le signe de ponctuation. *Assignation* se combine avec *trait* lorsque ce dernier est employé comme « signe distinctif » et indique le fait d'attribuer un trait. A droite, on retrouve d'ailleurs un cluster linguistique regroupant les *c* qui se combinent avec *trait* dans ce contexte linguistique spécifique : plusieurs noms (*nœuds*, *valeurs*, *catégories*, *structures*), mais surtout des adjectifs (*sémantiques*, *lexicaux*, *morphologiques*, *distinctifs*). Un seul adjectif (*humain*) se situe à l'écart à gauche en haut de la visualisation. Sa position particulière s'explique par le fait qu'il se combine avec des *cc* très particuliers, souvent des codes, tels que + et -. En bas, on retrouve des *c* plus généraux (*type*, *exemple*, *modèle*, *ensemble*, *nombre*). Les premières expérimentations sur le corpus spécialisé TALN semblent donc indiquer que notre approche fonctionne sur ce corpus et qu'elle permet de générer des résultats sémantiquement interprétables.

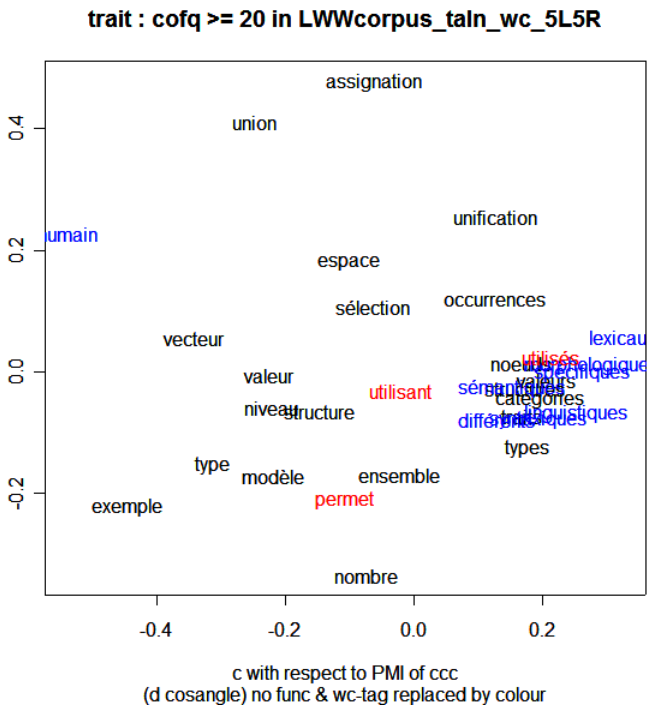

FIGURE 1: MDS des *c* de *trait* (mots lexicaux au seuil de co-fq ≥ 20)

Pour le lemme ***fréquence***, qui a une fréquence moins élevée (947), le seuil de co-fréquence a été fixé à 10, pour pouvoir relever suffisamment de *c* lexicaux. Une fenêtre plus restreinte de 3 mots à gauche et à droite (3L3R) permet de relever des *c* plus pertinents, qui sont pour la plupart des adjectifs. Il est donc plus judicieux de faire l'analyse MDS à partir des lemmes des cooccurrents (LLL). Pour cette configuration de paramètres avec 26 *c*, le pourcentage de stress est de 14,85%. La visualisation (cf. figure 2) montre les verbes en rouge, les adjectifs en bleu et les noms en noir. Un *c* est très éloigné des autres, à savoir *apparition*, qui s'emploie dans la combinaison privilégiée *fréquence d'apparition*. A gauche, on retrouve des *c* plutôt généraux qui expriment ce dont on étudie la fréquence, par exemple *la fréquence du nom / mot / terme dans le corpus*. En haut, vers la droite, on voit des verbes et noms déverbaux qui expriment ce qu'on fait avec les fréquences (*calculer*, *calcul*, *compte*). Finalement, le coin inférieur droit montre un cluster sémantique avec des adjectifs qui expriment le résultat du calcul ou qui indiquent l'importance de la fréquence, souvent par paire (*inférieur*, *supérieur*, *bas*, *haut*, *faible*, *élevé*). On observe un seul nom (*seuil*) qui marque le *cut off* de fréquence. Les adjectifs qui se situent plus au centre de la visualisation expriment des notions linguistiques plus spécifiques (*relatif*, *moyen*, *simple*).

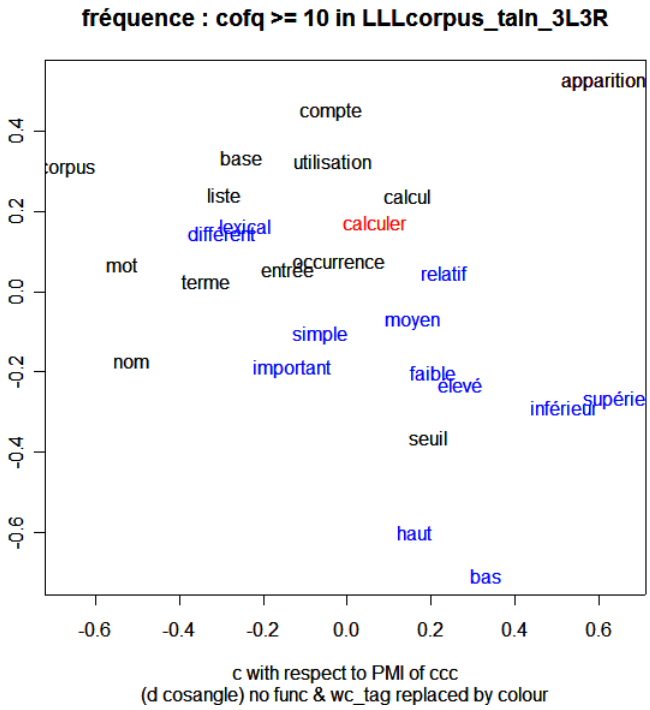

FIGURE 2: MDS des *c* de *fréquence* (mots lexicaux au seuil de co-fq ≥ 10)

Le mot-pôle le plus fréquent, *méthode*, se caractérise par une fréquence de 3808 dans le corpus TALN. Il permet donc d'évaluer l'impact de la fréquence du mot-pôle sur le paramétrage, plus particulièrement sur le seuil de co-fréquence minimale appliqué pendant la constitution de la matrice de cooccurrences. Au seuil de co-fréquence minimale supérieur ou égal à 20, nous relevons 149 *c*. Au seuil 10, nous en relevons 257. La visualisation de ces nombreux *c* afficherait une grande tache noire, même après suppression des *c* grammaticaux (environ un tiers). Il s'ensuit que plus le mot-pôle est fréquent dans le corpus, plus le seuil de co-fréquence minimale devra être élevé afin de garantir la lisibilité de la représentation visuelle.

Par conséquent, pour ***méthode*** le seuil de co-fréquence minimale sera fixé à 50. Pour les cooccurrents relevés au niveau des formes graphiques dans une fenêtre de 5 mots à gauche et à droite de *méthode*, le pourcentage de stress s'élève à 11,17% pour 24 *c* lexicaux. Nous optons ici pour la configuration des formes graphiques, d'une part, parce que la configuration des lemmes se caractérise par un pourcentage de stress trop élevé (18,41%) et, d'autre part, parce que nous observons un phénomène particulier. Dans les visualisations précédentes, les *c* se regroupaient en clusters sémantiques (cf. figures 1 et 2). Dans la visualisation des cooccurrents de *méthode*, un mot plus général et plus fréquent, nous observons ce qu'on pourrait qualifier de « cluster syntagmatique », c'est-à-dire une combinaison syntagmatique ou une suite de mots effective (cf. figure 3). En effet, à droite de la visualisation, au centre, nous retrouvons le début d'un paragraphe : *dans cet article nous proposons / présentons une méthode qui permet* ou *dans cet article nous proposons une méthode d'évaluation / d'analyse*. Surtout les formes conjuguées (1ère personne pluriel) et les types de verbes sont

très représentatifs pour les articles scientifiques qui constituent le corpus. On observe en outre que les *c* plus spécifiques se regroupent à gauche dans la partie inférieure (*extraction*, *apprentissage*, *alignement*, *classification*, *segmentation*). Les rares adjectifs (en bleu) se situent également à gauche, avec l'adjectif *automatique* tout près du nom *extraction* ; les verbes (en rouge) se retrouvent majoritairement en bas de la visualisation.

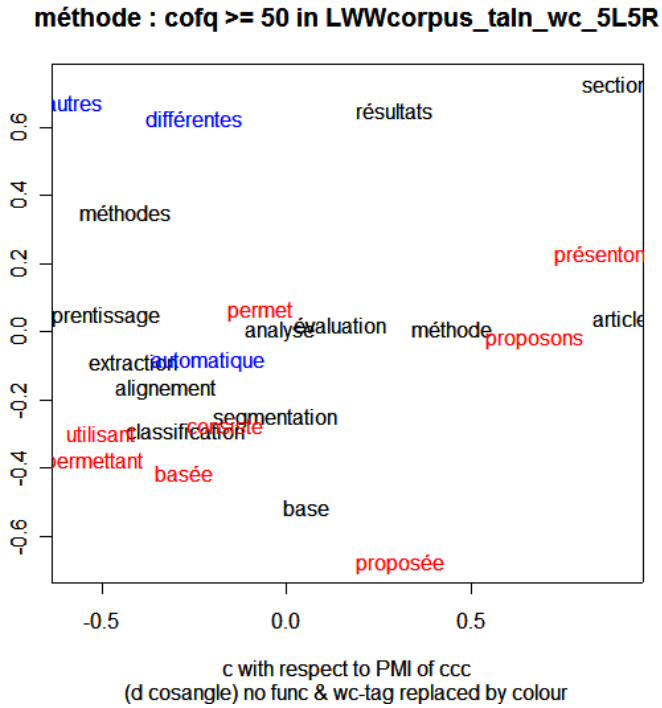

FIGURE 3: MDS des *c* de *méthode* (mots lexicaux au seuil de co-fq ≥ 50)

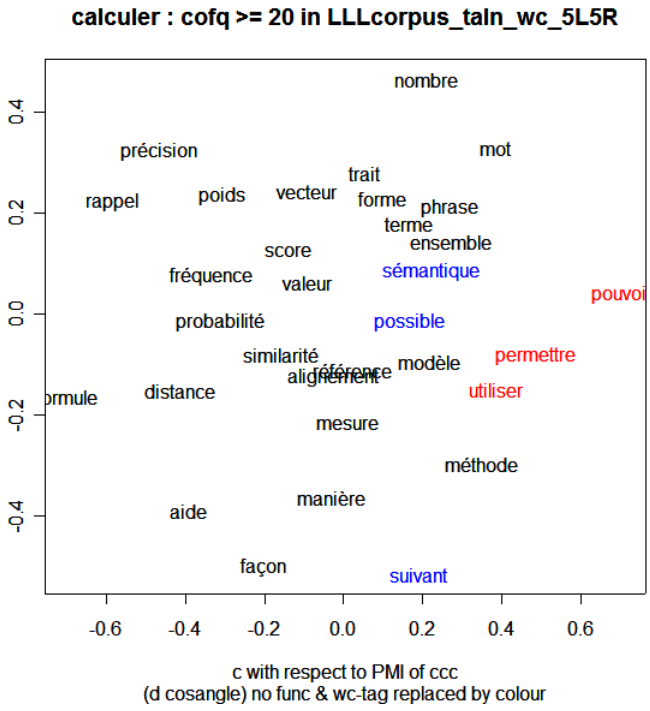

FIGURE 4: MDS des *c* de *calculer* (mots lexicaux au seuil de co-fq ≥ 20)

Le mot-pôle ***calculer*** (fréquence de 1236) est le seul verbe de la sélection de mots. Dans une fenêtre de 5 mots à gauche et à droite et au niveau des lemmes, nous recensons 32 *c* lexicaux au seuil 20 et 67 *c* lexicaux au seuil 10. Lorsque les 67 *c* sont visualisés en 2D, la visualisation est plutôt dense et de ce fait difficile à interpréter. Nous optons dès lors pour le seuil 20 et les 32 *c* lexicaux. Pour cette configuration, le pourcentage de stress (18,28%) dépasse le seuil de 15%. Les configurations avec un pourcentage de stress inférieur recensent trop peu de *c* lexicaux (p.ex. 12 ou 13 *c*) pour une interprétation sémantique intéressante. Pour le verbe *calculer* (cf. figure 4), nous observons clairement un cluster sémantique très spécifique à gauche en haut de la visualisation avec *précision* et *rappel*, et avec des *c* linguistiques tout proches, tels que *poids*, *fréquence*, *probabilité* et *similarité*. Signalons que les verbes se situent à droite : *pouvoir* et *permettre* expriment une possibilité et se retrouvent près de l'adjectif *possible*. Les *c* plus généraux se trouvent en bas de la visualisation (*aide*, *façon*, *manière*, *méthode*, *suivant*).

Le mot-pôle ***graphe*** (fréquence de 1116) se caractérise par une fréquence comparable à celle de *calculer* (1236) et de *fréquence* (947). A titre d'expérimentation, nous appliquons le seuil de co-fréquence minimale de 20 adopté pour *calculer*, ainsi que le seuil 10 adopté pour *fréquence*. Au seuil 20, le nombre de *c* lexicaux pertinents est nettement inférieur (25 *c*) à celui au seuil 10 (64 *c*). Dans la configuration au niveau des lemmes dans une fenêtre de 5 mots à gauche et à droite, l'analyse MDS au seuil 20 n'est pas satisfaisante (pourcentage de stress de 22,60%). Dans la configuration similaire au seuil 10, elle affiche un pourcentage de stress très bas de 3,33%. A première vue, ce pourcentage semble indiquer que la représentation visuelle en 2D est très fiable et qu'il y a très peu de distorsion pour représenter toutes les distances en 2D. Toutefois, lorsqu'on regarde la visualisation de près, on constate qu'il y a un seul *c* très périphérique, à savoir *acyclique*, et que tous les autres *c* se regroupent en un grand cluster de 63 *c* superposés, qu'il est impossible d'interpréter. Il est clair qu'il faudra éliminer ce *c* périphérique pour pouvoir appliquer notre approche et interpréter les résultats. Par ailleurs, le *c acyclique* constitue avec le mot-pôle *graphe* un terme classique en théorie des graphes et il est largement prédominant dans les combinaisons de mots avec *graphe*. Dans la configuration au seuil 10 avec indication de classe lexicale, l'analyse MDS pour les 64 *c* mène à un pourcentage de 12,19%. Le c *acyclique* se situe également à une position isolée, bien que moins extrêmement isolée. Les autres *c* sont fortement regroupés et la plupart d'entre eux sont difficiles à identifier.

Après suppression de cette observation très périphérique (*outlier*), l'analyse MDS pour les 63 *c* restants affiche un pourcentage de 13,28% dans la configuration au seuil 10 avec indication de classe lexicale et un pourcentage de 13,52% dans celle sans indication de classe lexicale, ce qui est parfaitement comparable. La représentation visuelle ci-dessous montre les résultats de cette première configuration (cf. figure 5). Les *c* les plus spécifiques du domaine (*nœud*, *arc*, *chemin*, *sommet*, *clique*) se regroupent en haut à droite : ils indiquent les éléments les plus importants d'un *graphe*, qu'on observe en haut à gauche.

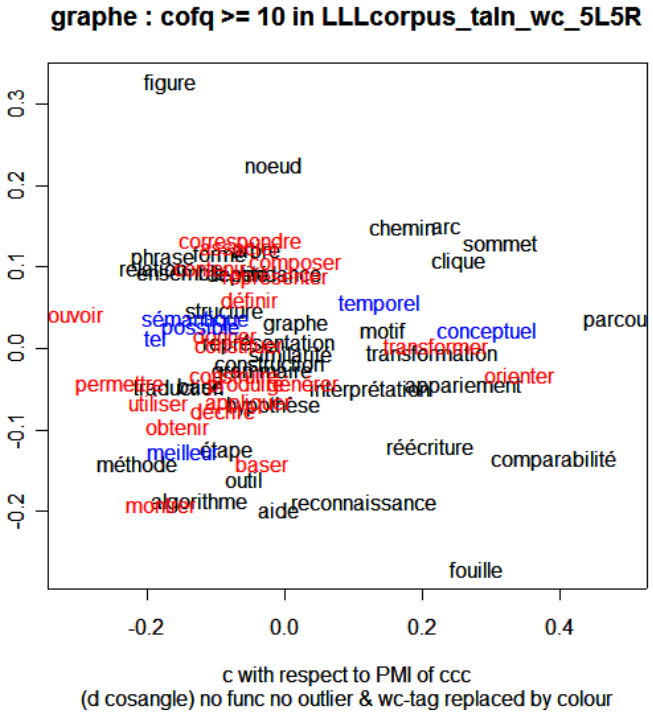

FIGURE 5: MDS des *c* de *graphe* (mots lexicaux au seuil de co-fq ≥ 10, après suppression de *acyclique*)

Pour le nom ***sémantique*** comme mot-pôle, il est indispensable d'intégrer les indications de classe lexicale pour le repérage des *c* pertinents du mot-pôle, afin de relever uniquement les occurrences (au niveau des lemmes) du nom *sémantique* (459) et pas celles de l'adjectif *sémantique* (3059). Comme le nom *sémantique* n'est pas très fréquent, nous adoptons le seuil de co-fréquence minimale de 5. Dans la configuration au niveau des lemmes dans une fenêtre de 5 mots à gauche et à droite du nom *sémantique*, l'analyse pour les 41 *c* lexicaux affiche un pourcentage de stress extrêmement bas de 0,63%. En effet, il y a une observation extrêmement périphérique, à savoir *Montague*, qui constitue une combinaison privilégiée (*la sémantique de Montague*). Après suppression de ce *c* périphérique, l'analyse MDS pour les 40 *c* lexicaux aboutit à un résultat satisfaisant de 13,78%. La visualisation ci-dessous (cf. figure 6) montre que les *c* du nom *sémantique* sont majoritairement des noms, indiqués en noir. Il n'y a que 5 adjectifs (*formel*, *temporel*, *lexical*, *syntaxique*, *sémantique*). A droite, on trouve des *c* sémantiquement liés (*syntaxe*, *sémantique*, *morphologie*). Le *c* *connecteur* se trouve à une position isolée à droite en bas de la visualisation. A gauche, on voit des *c* dont on pourra étudier la sémantique (*nom*, *mot*, *phrase*, *texte*). Les indications plus générales se regroupent en haut (*analyse*, *traitement*, *domaine*, *information*) et les indications plus spécifiques au centre, vers le bas (*arbre*, *sens*, *graphe*, *trait*, *discours*). Remarquons que *compte* et *calcul* se situent l'un près de l'autre au milieu de la visualisation !

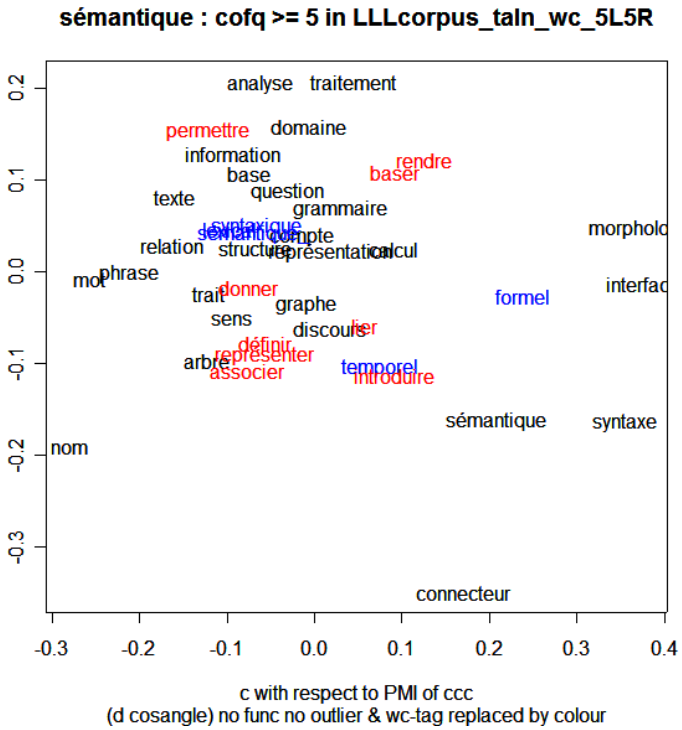

FIGURE 6: MDS des *c* de *sémantique* (mots lexicaux au seuil de co-fq ≥ 5, après suppression de *Montague*)

Les deux derniers mots-pôles discutés ci-dessous sont les deux adjectifs de la sélection de mots, à savoir *complexe* et *précis*. Pour l'interprétation de la visualisation des *c* de ces adjectifs, on s'intéressera surtout aux noms parmi les *c* visualisés, parce qu'un adjectif se combine de préférence avec un nom qu'il caractérise ou modifie. Il semble judicieux de séparer les visualisations en fonction des catégories, afin d'y voir plus clair, ce qui plaide également en faveur d'une prise en considération des relations syntaxiques dans nos futures recherches. Il est à noter que ces deux adjectifs ont un sens plutôt général, qu'on pourrait qualifier de vague. Pour cette raison, les résultats de notre approche pour ces deux adjectifs sont particulièrement intéressants. A l'origine, l'approche a été conçue pour observer comment se positionnent les cooccurrents d'un mot polysémique les uns par rapport aux autres, dans le but de mieux comprendre ce qui se cache derrière le phénomène d'hétérogénéité sémantique. Nous sommes donc curieux de voir si l'analyse MDS des cooccurrents de premier ordre fonctionne également pour ces deux adjectifs dans le corpus TALN.

Pour le mot-pôle ***complexe*** (fréquence de 732), nous envisageons la configuration au seuil 10. Dans presque toutes les configurations, au niveau des lemmes et des formes graphiques, dans une fenêtre de 3 et de 5 mots à gauche et à droite du mot-pôle, le pourcentage de stress est supérieur à 15%. Nous décidons dès lors de supprimer les adjectifs et les verbes parmi les *c*, puisqu'ils sont moins susceptibles de pointer vers un des sens de l'adjectif *complexe*. Pour les 26

noms restants dans la configuration des formes graphiques dans une fenêtre de 5 mots à gauche et à droite, avec indication de classe lexicale, le pourcentage de stress s'élève à 14,44%. La configuration au niveau des lemmes s'avère inacceptable (20,15%), bien que les lemmes des *c* améliorent la lisibilité et l'interprétation des résultats. Comme le montre la visualisation pour la configuration acceptable (cf. figure 7), les formes du pluriel s'affichent à gauche et les formes du singulier à droite. En haut à droite, nous observons des *c* plutôt généraux (*problème*, *tâche*, *traitement*) et au milieu à droite des *c* plus spécifiques (*terme*, *structure*, *phrase*). Les formes du pluriel correspondantes se regroupent également en fonction de leurs caractéristiques sémantiques : *termes*, *phrases*, *expressions*, *formes*, *relations* et *règles* en bas à gauche et *requêtes*, *tâches*, *questions*, *phénomènes* et *modèles* en haut. Notre approche s'applique donc également à des adjectifs comme mots-pôles, mais elle requiert alors un ajustement des paramètres. Quand on se focalise sur les noms dans les rangées (des *c*) de la matrice de cooccurrences, les résultats sont nettement meilleurs, tant en termes de pourcentage de stress (critère quantitatif), qu'en termes d'interprétation sémantique (critère qualitatif).

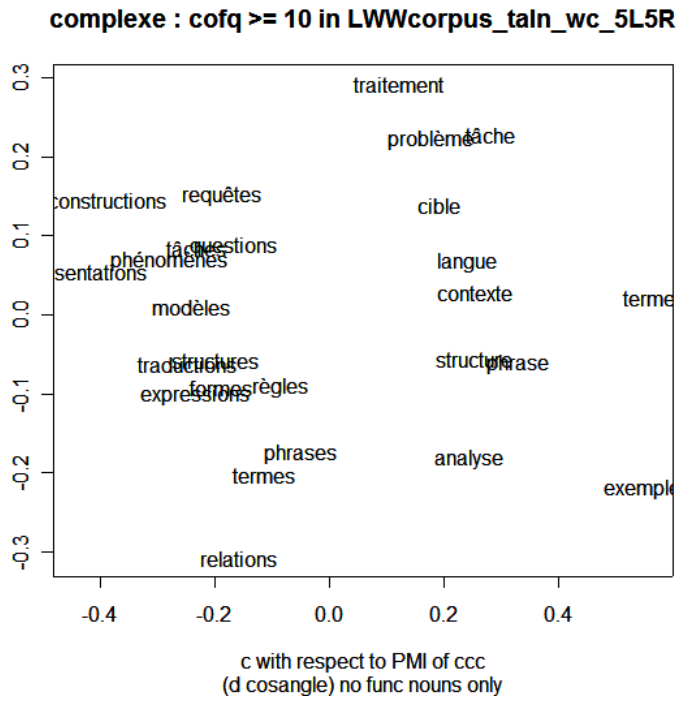

FIGURE 7: MDS des *c* de *complexe* (noms uniquement au seuil de co-fq ≥ 10)

Le dernier mot-pôle (***précis***) est un adjectif peu fréquent dans le corpus TALN (fréquence de 378). En fait, c'est le mot le moins fréquent de la sélection de mots. Nous appliquons dès lors le seuil minimal de 5 et nous considérons la configuration au niveau des lemmes, dans une fenêtre de 5 mots autour du mot-pôle pour 54 *c* lexicaux (stress de 16,64%). La configuration au niveau des formes graphiques donne lieu à des pourcentages de stress plus élevés, voire supérieurs à 20%. Une première visualisation de la configuration des lemmes montre un *c* très périphérique (*exhaustif*). Après suppression de cette observation aberrante, le pourcentage de stress ne s'améliore pas (18,96%). Après suppression des adjectifs et des verbes, donc en considérant uniquement les noms dans les rangées de la matrice de cooccurrences, comme nous l'avons fait pour *complexe* (cf. ci-dessus), l'analyse MDS pour les 38 noms génère un stress très légèrement supérieur au seuil (16,15%). La visualisation montre quelques groupes de *c* sémantiquement similaires : au milieu *manière* et *façon* qui se superposent, à gauche en bas *étude*, *recherche*, *évaluation*, *analyse*, en bas à droite *cas* et *exemple* (cf. figure 8 ci-dessus). Nous identifions également un groupe qui pourrait s'identifier comme un cluster syntagmatique, en haut à gauche, *identification et description d'un phénomène*.

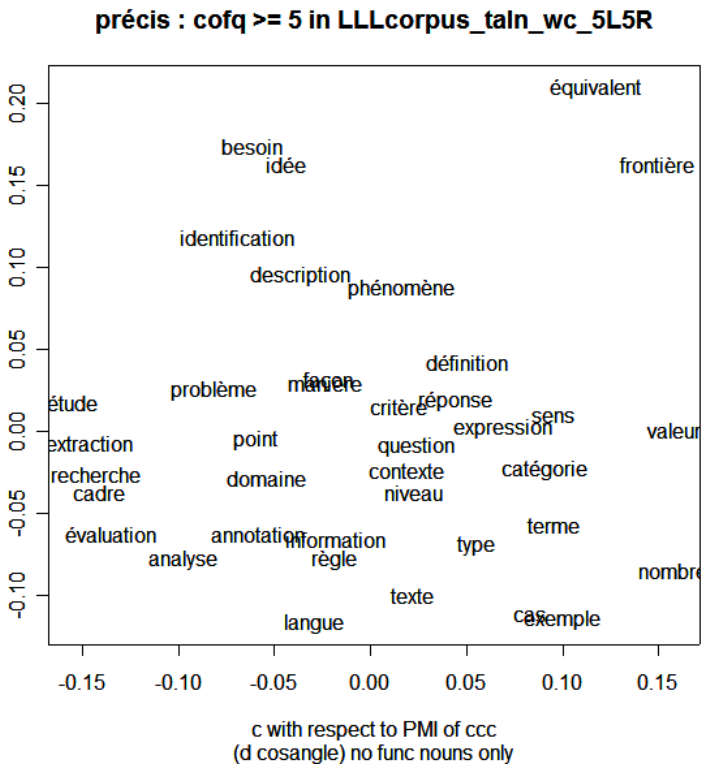

FIGURE 8: MDS des *c* de *précis* (noms uniquement au seuil de co-fq ≥ 5)

5 Conclusions

Dans le cadre de la tâche exploratoire sur le corpus TALN, nous avons appliqué une analyse de positionnement multidimensionnel des cooccurrents de premier ordre pour une sélection de huit mots-pôles. Cette approche a généré des résultats statistiquement acceptables et sémantiquement interprétables. En plus, les nouvelles données nous ont permis d'apprendre plus sur notre approche et de réajuster le paramétrage en fonction des caractéristiques du mot-pôle, afin de peaufiner les résultats.

La question de recherche principale, à l'origine de notre approche de sémantique distributionnelle, était celle de savoir si le degré plus ou moins élevé d'hétérogénéité sémantique d'un mot-pôle se reflète dans la visualisation des distances entre ses cooccurrents de premier ordre. Pour le mot-pôle *trait* dans le corpus TALN, nous avons effectivement constaté que la dispersion des cooccurrents sur la visualisation en 2D correspond aux divers sens de *trait*. Aussi pour les autres mots de la sélection, nous avons pu distinguer clairement des clusters de *c* sémantiquement similaires. Pour les mots plutôt généraux, comme *méthode* et *précis*, nous avons observé des clusters syntagmatiques de *c*. Comme nous avons également trouvé des clusters syntagmatiques dans la visualisation des résultats MDS sur le corpus technique, il s'agit d'une piste méthodologique intéressante à creuser dans nos recherches futures. Affiner le modèle permettrait peut-être de voir plus clair dans les phénomènes de proximité sémantique et de proximité syntagmatique, soit par des analyses et des visualisations par classe lexicale, soit par la prise en compte des relations syntaxiques.

Les analyses MDS sur le corpus TALN nous ont également incités à procéder à des mises au point de notre approche et à des réajustements du paramétrage pour améliorer les résultats. Nous avons effectivement pu tirer des enseignements méthodologiques intéressants sur notre approche. Tout d'abord, il s'est avéré que la fréquence du mot-pôle a une influence considérable sur les paramètres, plus particulièrement sur le seuil de co-fréquence appliqué pour la matrice de cooccurrences. Plus le mot-pôle est fréquent (cf. *méthode*), plus le seuil doit être élevé. Ensuite, pour certains mots-pôles (*graphe* et *sémantique*), la suppression d'un seul cooccurrent extrêmement périphérique, qui constitue un terme ou une combinaison privilégiée avec le mot-pôle, a contribué à un meilleur résultat pour tous les autres cooccurrents et à une meilleure lisibilité de la visualisation. Finalement, il a été plus intéressant de se focaliser sur les cooccurrents d'une

seule classe lexicale et ce en fonction de la classe lexicale du mot-pôle. Pour les adjectifs de la sélection de mots (*complexe* et *précis*), nous avons uniquement considéré les noms parmi les cooccurrents, ce qui a nettement amélioré les résultats. Notre approche méthodologique et les informations sémantiques extraites pourraient s'avérer utiles dans plusieurs domaines d'application, tels que la désambiguïsation ou la description lexicographique.

Dans la plupart des analyses, nous avons appliqué une fenêtre d'observation de 5 mots à gauche et à droite du mot-pôle. Elle contient suffisamment de cooccurrents sémantiquement intéressants sans introduire trop de bruit. Généralement, il est plus intéressant de considérer les cooccurrents au niveau des lemmes, puisque le niveau des formes graphiques alourdit la visualisation, souvent inutilement, avec des formes fléchies au singulier et au pluriel ou avec plusieurs formes conjuguées des verbes. Enfin, il s'est avéré que la prise en considération des indications de classe lexicale, tant pour les mots-pôles (cf. le nom *sémantique*) que pour les cooccurrents de premier, deuxième et troisième ordre, permet d'enrichir les analyses et de préciser les résultats.

Références

BERTELS A., SPEELMAN D., GEERAERTS D. (2010). La corrélation entre la spécificité et la sémantique dans un corpus spécialisé. *Revue de Sémantique et de Pragmatique* n°27, 79-102.

BERTELS A., SPEELMAN D. (2013). Exploration sémantique visuelle à partir des cooccurrences de deuxième et troisième ordre. Actes de *TALN 2013 (volume 3 Atelier SemDis 2013)*, 126-139.

BERTELS A., SPEELMAN D. (2014). Analyse exploratoire des cooccurrents de premier ordre dans un corpus technique. Actes de *JADT 2014*, (sous presse).

BORG I., GROENEN P. (2005). *Modern Multidimensional Scaling: theory and applications* (Second edition). New York : Springer-Verlag.

BOUDIN F. (2013). TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue. Actes de *TALN 2013 (volume 1 TALN)*, 507-514.

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

CLARKE K.R. (1993). Non-parametric multivariate analyses of change in community structure. *Australian Journal of Ecology* n°18, 117-143.

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

DESBOIS D. (2005). Une introduction au positionnement multidimensionnel. *Modulad* n°32, 1-28.

EVERT S. (2007). *Corpora and collocations.* Extended Manuscript of Chapter 58 of Lüdeling A. et Kytö M., 2008, *Corpus Linguistics. An International Handbook.* Berlin : Mouton de Gruyter.

FERRET O. (2010) Similarité sémantique et extraction de synonymes à partir de corpus. Actes de *TALN 2010*. http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_77.pdf. [consulté le 08/04/2014].

GREFENSTETTE G. (1994), Corpus-derived first, second and third-order word affinities. Proceedings of *Euralex '94.* Amsterdam, 279-290.

HABERT B., ILLOUZ G. FOLCH, H. (2004), Dégrouper les sens : pourquoi ? comment ? Actes de *JADT 2004,* Louvain-la-Neuve, 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 : Hermès-Science, 277–318.

HEYLEN K., SPEELMAN D., GEERAERTS D. (2012). Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. Proceedings of *the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 16-24.

KRUSKAL J.B., 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., DENHIÈRE G. (2006). Effects of High-Order Co-occurrences on Word Semantic Similarity. *Current Psychology Letters* n°18(1). `http://cpl.revues.org/index471.html`. [consulté le 08/04/2014].

MORARDO M., VILLEMONTE DE LA CLERGERIE E. (2013). Vers un environnement de production et de validation de ressources lexicales sémantiques. Actes de *TALN 2013 (volume 3 Atelier SemDis 2013)*, 167-180.

MORLANE-HONDÈRE F. (2013). Utiliser une base distributionnelle pour filtrer un dictionnaire de synonymes. Actes de *TALN 2013 (volume 3 Atelier SemDis 2013)*, 112-125.

PEIRSMAN Y., GEERAERTS D. (2009). Predicting Strong Associations on the Basis of Corpus Data. Proceedings of *EACL-2009*, 648-656.

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

SAHLGREN M. (2008). The Distributional Hypothesis. *Rivista di Linguistica* n°20(1), 33-53.

SCHÜTZE H. (1998), Automatic Word Sense Discrimination. *Computational Linguistics* n°24(1), 97-123.

TURNEY P.D., PANTEL P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* n°37, 141-188.

URIELI A., TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur *Talismane*. Actes de *TALN 2013 (volume 1 TALN)*, 188-201.

VENABLES W.N., RIPLEY B.D. (2002). *Modern Applied Statistics with S*, (Fourth edition). New York : Springer-Verlag.

WIELFAERT T., HEYLEN K., SPEELMAN D. (2013). Interactive visualizations of Semantic Vector Spaces for lexicological analysis. Actes de *TALN 2013 (volume 3 Atelier SemDis 2013)*, 154-166.