SÉMANTIQUE DISTRIBUTIONNELLE EN LINGUISTIQUE DE CORPUS

Kris Heylen, Ann Bertels

Armand Colin | « Langages »

2016/1 N° 201 | pages 51 à 64

ISSN 0458-726X
ISBN 9782200930394

Article disponible en ligne à l'adresse :

http://www.cairn.info/revue-langages-2016-1-page-51.htm

Pour citer cet article :

Kris Heylen, Ann Bertels, « Sémantique distributionnelle en linguistique de corpus », *Langages* 2016/1 (N° 201), p. 51-64.
DOI 10.3917/lang.201.0051

Kris Heylen
KU Leuven & Quantitative Lexicology and Variational Linguistics (QLVL)

Ann Bertels
KU Leuven & Leuven Language Institute (ILT) – Quantitative Lexicology and Variational Linguistics (QLVL)

Sémantique distributionnelle en linguistique de corpus

1. INTRODUCTION

Les analyses à base de corpus ont une longue tradition en sémantique lexicale. Depuis l'émergence de projets dictionnairiques à grande échelle au XIX^{e} siècle, les chercheurs en sémantique lexicale se sont appuyés sur des indices textuels attestés dans le langage réel pour dégager et organiser les différents sens et usages d'un mot. Alors que dans les années 1950, la recherche syntaxique se détournait des données authentiques, les idées de J. R. Firth (1957), Z. Harris (1954) et W. Weaver (1955) ont mené à des approches qui considéraient les données authentiques comme la base empirique naturelle pour les descriptions sémantiques [1]. Initialement, la collecte et l'analyse des données de corpus se faisait manuellement. Grâce à l'informatique et à la disponibilité de corpus électroniques de taille plus importante, les lexicologues et lexicographes disposent à présent de grandes quantités de données authentiques, qui constituent une énorme base de données empirique pour leurs travaux descriptifs. Pour analyser cette abondance de données, les chercheurs en sémantique lexicale recourent à des outils d'analyse statistique qui facilitent deux étapes dans l'analyse de corpus. D'une part, les méthodes statistiques permettent l'identification, dans les données de corpus, d'indices contextuels pour étudier la signification d'un lexème, tels que des mots cooccurrents (collocations) et des patrons syntaxiques (collocations grammaticales ou colligations). D'autre part, ces méthodes permettent la classification des occurrences d'un lexème en différents sens et usages en fonction des indices contextuels.

1. Pour un historique plus approfondi des approches récentes en linguistique de corpus pour la sémantique lexicale, voir Geeraerts (2010 : 165-178).

La première approche, *i.e.* l'identification d'indices contextuels à partir de méthodes statistiques, a été associée à la tradition britannique en linguistique de corpus. Créée par J. Sinclair (1991), cette approche consistait à décrire la signification lexicale d'un mot en fonction des mots typiques (collocations) et des patrons syntaxiques (colligations) avec lesquels il apparaît. K. Church et P. Hanks (1989) ont introduit des mesures statistiques, notamment le *t-score*, pour identifier les collocations et colligations pertinentes et informatives à partir des distributions de fréquence dans un texte. Par la suite, ces mesures ont été mises au point et optimisées [2]. Aujourd'hui, elles sont largement utilisées dans plusieurs sous-disciplines linguistiques.

La deuxième approche, à savoir le regroupement statistique d'usages, est très présente dans les récents développements en Sémantique Cognitive. Plus particulièrement, l'approche du *Behavioural Profile* a récemment introduit des techniques statistiques multivariées pour classer automatiquement les occurrences d'un mot en plusieurs sens et usages distinctifs à partir des données de corpus. D. Glynn (2010) recourt à l'analyse factorielle des correspondances (AFC) pour visualiser comment les occurrences du verbe *to bother* ('déranger') sont regroupées en usages distincts en fonction de leur comportement syntaxique ou de leurs caractéristiques sémantiques, comme par exemple « affect ».

Ces deux approches ont été utilisées indépendamment l'une de l'autre dans ces traditions différentes. Les analyses de collocations ont automatisé le repérage d'indices contextuels au moyen d'analyses statistiques, mais elles confient la classification des occurrences et des contextes typiques à l'analyse manuelle. Par contre, les analyses de profils comportementaux *(behavioural profiles)* ont procédé à l'automatisation statistique de la classification des occurrences et contextes typiques d'un lexème en différents sens, mais elles s'appuient principalement sur des données où les indices contextuels ont été encodés manuellement. Le Tableau 1 *infra* propose un aperçu des différentes approches en sémantique lexicale et indique si le repérage des indices contextuels et des sens se fait manuellement ou par le biais d'analyses statistiques. Il est clair que les études en philologie classique réalisaient les deux étapes manuellement. Les études de collocations, d'une part, et les analyses de profils comportementaux, d'autre part, ont automatisé une seule des deux étapes. Dans cet article, nous présentons les méthodes d'analyse distributionnelle, que nous considérons comme une extension logique de l'état de l'art statistique sur lequel s'appuient les analyses sémantiques lexicales. Les méthodes d'analyse distributionnelle combinent de façon systématique des mesures d'association et des méthodes statistiques multivariées dans le but d'explorer des structures sémantiques lexicales dans des corpus textuels de taille importante. De ce fait, elles permettent de compléter le motif qui se dégage dans la présentation des outils statistiques utilisés en sémantique lexicale (cf. Tableau 1).

2. Voir Evert (2004) et Wiechmann (2008) pour un aperçu.

Tableau 1 : Aperçu des outils statistiques utilisés en sémantique lexicale

	Identification d'indices contextuels	Classification d'occurrences
Philologie classique	manuelle	manuelle
Analyse de collocations	statistique	manuelle
Analyse de profils comportementaux	manuelle	statistique
Méthodes d'analyse distributionnelle	statistique	statistique

L'introduction de méthodes statistiques dans les deux étapes du processus d'analyse de données constitue une extension logique, voire indispensable, et ceci, pour deux raisons. Premièrement, les lexicologues et lexicographes pourront bénéficier du soutien supplémentaire des techniques pour le repérage de patrons statistiques, parce qu'il leur permet de faire face à l'abondance de données auxquelles ils sont confrontés dans leurs travaux descriptifs. Il est tout simplement impossible d'encoder manuellement, de classer et de décrire plusieurs milliers de concordances d'un lexème. Les analyses de données statistiques pourront aider à prélever un échantillon représentatif de différents usages, qui pourra par la suite faire l'objet d'analyses plus approfondies. Deuxièmement, la disponibilité de *Big Data* ou de gros volumes de données suggère une extension de la focalisation traditionnelle du travail lexicologique et lexicographique. Un environnement *Big Data* permet aux chercheurs d'examiner des tendances et motifs qu'ils ne dégageraient pas en analysant de petits corpus, comme par exemple la répartition de nouveaux mots ou de nouveaux usages de mots existants à travers les réseaux sociaux. Pour détecter ce type de tendances, les techniques de fouilles de données sont indispensables et requièrent un recours aux techniques quantitatives appropriées.

Dans cet article, nous procédons d'abord à une présentation non technique des méthodes d'analyse distributionnelle pour la modélisation sémantique distributionnelle (§ 2). Pour différentes approches et sous-types, nous présentons les caractéristiques principales et les références les plus importantes, et nous expliquons les principes de l'analyse distributionnelle à partir d'un modèle distributionnel particulier. Ensuite, nous discutons une application lexicologique pour l'analyse de la polysémie (§ 3). Finalement, nous présentons une méthode de visualisation qui permet aux experts humains d'interpréter les structures sémantiques cernées par les modèles sémantiques distributionnels (§ 4).

2. MÉTHODES D'ANALYSE DISTRIBUTIONNELLE

2.1. Historique

Les méthodes d'analyse distributionnelle, aussi appelées modèles sémantiques distributionnels *(Word Space Models)* étaient initialement utilisées en Psychologie

Cognitive pour modéliser la mémoire lexicale (Landauer & Dumais 1997 ; Lund & Burgess 1996). Ensuite, elles ont été développées en Linguistique Computationnelle, où elles sont maintenant largement répandues pour modéliser l'analyse sémantique dans le domaine du Traitement Automatique des Langues (TAL) *(Statistical Natural Language Processing)* (voir Turney & Pantel 2010, pour un aperçu). Pour le français, les travaux précurseurs à la fin des années 90 étaient principalement consacrés à l'application des méthodes distributionnelles au traitement de corpus spécialisés (Bouaud *et al.* 1997 ; Habert & Zweigenbaum 2002). Évoquons également l'organisation d'une journée ATALA en 1999, intitulée *Approche distributionnelle de l'analyse sémantique* (voir Fabre *et al.* 2014a). Actuellement, l'analyse distributionnelle est devenue un mode de représentation et d'exploitation incontournable dans les travaux sur le lexique et sur la sémantique lexicale (Fabre *et al.* 2014b), comme en témoignent les deux éditions de l'atelier « Sémantique Distributionnelle », SemDis 2013 (Les Sables-d'Olonne) et SemDis 2014 (Marseille). L'édition 2013 visait à rassembler les études et travaux sur le français dans le domaine de l'analyse distributionnelle. L'édition 2014 permettait la confrontation de plusieurs méthodes distributionnelles en proposant une tâche compétitive de substitution lexicale sur un corpus de taille importante et une tâche exploratoire sur un petit corpus spécialisé.

2.2. Plusieurs approches et sous-types

Les méthodes d'analyse distributionnelle s'appuient sur l'hypothèse distributionnelle (Firth 1957) selon laquelle des mots qui se trouvent dans des contextes d'apparition similaires tendent à avoir des sens similaires. Elles regroupent plusieurs approches et sous-types. Trois dimensions sont importantes à cet effet : le type de contexte, le niveau d'analyse et le mode de représentation computationnelle.

2.2.1. Type de contexte

Premièrement, les méthodes et modèles diffèrent en fonction du type de contexte pris en considération pour représenter le sens des mots. On fait la distinction entre les modèles dits « à base de documents » *(document-based)* et les modèles dits « à base de mots » *(word-based)*. Dans les modèles à base de documents, l'analyse sémantique d'un mot *x* s'appuie sur l'observation de l'occurrence de *x* dans le contexte du document entier. Les modèles anglosaxons originels sont des modèles à base de documents, par exemple l'analyse sémantique latente (*Latent Semantic Analysis* ou LSA) (Landauer & Dumais 1997). Les modèles à base de mots, par contre, prennent comme point de départ l'occurrence d'un mot *x* dans le contexte d'un autre mot. Certains modèles à base de mots représentent chaque occurrence de *x* dans le corpus par ses « cooccurrents graphiques » dans une fenêtre de contexte donnée (Heylen, Speelman & Geeraerts 2012 ; Ferret 2010, 2014 ; Bernier-Colborne 2014 ; Bertels & Speelman 2013, 2014 ; Périnet & Hamon 2014). D'autres modèles à

base de mots prennent en considération les relations de dépendance syntaxique entre le mot *x* et ses « cooccurrents syntaxiques » (Morlane-Hondère 2013 ; Morardo & Villemonte de La Clergerie 2013 ; Fabre *et al.* 2014b). Les modèles à base de documents sont plus appropriés pour modéliser des relations syntagmatiques et associatives, par exemple entre *médecin* et *hôpital* ou entre *voiture* et *rouler*. Les modèles à base de mots, et plus particulièrement ceux qui s'appuient sur des informations de dépendance syntaxique, sont plus précis et parviennent à mieux saisir des relations paradigmatiques, comme les quasi-synonymes *hôpital* et *clinique* (Sahlgren 2006). Certains travaux vont au-delà des contextes d'apparition dans les textes (mots ou documents) pour intégrer également dans l'analyse distributionnelle des « mots visuels » ou des images extraites à partir de techniques visuelles computationnelles, ce qui aboutit à la « sémantique distributionnelle multimodale » (Bruni, Tran & Baroni 2014).

2.2.2. Niveau d'analyse

Deuxièmement, les modèles distributionnels permettent d'étudier les mots soit au niveau des types *(type-level)* (Navigli 2012), soit au niveau des occurrences individuelles *(token-level)* (Dinu, Thater & Laue 2012). Les modèles au niveau des types *(type-level)* tentent non seulement de distinguer un mot d'un autre dans le but de détecter la synonymie et les relations lexicales liées (Curran & Moens 2002) mais aussi de donner un aperçu des différents sens d'un mot polysémique (Bertels & Speelman 2013). Or, pour déterminer le sens d'une seule occurrence particulière, les modèles au niveau des occurrences *(token-level)* essaient de distinguer un usage d'un mot donné d'un autre usage de ce même mot, pour ainsi étudier sa polysémie (Heylen *et al.* 2014).

2.2.3. Représentation computationnelle

Troisièmement, les modèles distributionnels diffèrent en fonction du mode de représentation computationnelle des données distributionnelles, à savoir une représentation sous forme de graphe *(graph-based)* (Morardo & Villemonte de La Clergerie 2013 ; Desalle *et al.* 2014 ; Claveau, Kijak & Ferret 2014) ou une représentation vectorielle *(vector-based)*. Les modèles vectoriels reposent sur la représentation matricielle des mots dans un espace vectoriel. Il s'agit par exemple d'une matrice mots × mots, qui vise à caractériser les mots à partir de leurs cooccurrents graphiques, ou d'une matrice mots × relations de dépendance, qui permet de caractériser les mots en fonction de leurs cooccurrents syntaxiques. Il est important de réduire le nombre élevé de caractéristiques contextuelles *(context features)* à un nombre limité de « dimensions sémantiques », soit pour un traitement computationnel efficace (p. ex. Ferret 2010, 2014), soit pour une visualisation qui permet aux chercheurs-linguistes d'accéder plus facilement aux relations sémantiques qui se dégagent (p. ex. Heylen, Speelman & Geeraerts 2012). Dans la description de l'application lexicologique (cf. § 3), nous discutons, en guise d'exemple, la technique du positionnement multidimensionnel.

2.3. Explication détaillée des principes de base

Pour expliquer les principes de base des modèles sémantiques distributionnels, nous recourons à un modèle distributionnel particulier, à savoir un modèle à base de mots *(word-based)* qui considère les cooccurrents graphiques au niveau des types *(type-level)* sous forme de représentation vectorielle *(vector-based)*. L'application lexicologique discutée dans la section suivante repose sur ce même modèle.

Nous faisons appel à un exemple pour étudier trois mots-cibles, à savoir *chien*, *chat* et *café*, dans un petit corpus « jouet » de six phrases-exemples.

(1) Le *chien* aboie fort contre les passants.
(2) Le vétérinaire prend le *chien* par le cou.

(3) Le *chat* ronronne fort.
(4) Le vétérinaire s'est fait griffer par le *chat* en le prenant.

(5) On boit plus de *café* que de thé.
(6) Le vétérinaire prend sa tasse et boit son *café*.

Ce petit corpus « jouet » peut être représenté sous forme de table de fréquence, avec les mots-cibles en ligne et les caractéristiques contextuelles ou les « mots-contextes » en colonne. Nous recourons au modèle le plus simple possible, *i.e.* celui qui ignore les relations de dépendance syntaxique et qui recense simplement les fréquences absolues des noms, verbes et adverbes. Les fréquences sont indiquées dans la matrice de cooccurrence *infra* (cf. Tableau 2). Le fait d'ignorer les relations syntaxiques est qualifié d'approche « sac de mots » *(Bag-of-Words approach)* en recherche d'informations. Dans un exemple issu du monde réel et dans un grand corpus, il y a beaucoup plus de mots-contextes (dans les colonnes), généralement plusieurs milliers, et bien sûr les fréquences sont largement supérieures à 2.

Tableau 2 : Matrice de cooccurrence pour les mots-cibles *chien*, *chat* et *café*

	aboyer	passant	vétérinaire	prendre	cou	ronronner	fort	griffer	boire	plus	thé	tasse
chien	1	1	1	1	1	0	1	0	0	0	0	0
chat	0	0	1	1	0	1	1	1	0	0	0	0
café	0	0	1	1	0	0	0	0	2	1	1	1

Les fréquences de cooccurrence peuvent être interprétées comme les coordonnées qui permettent de situer *chien*, *chat* et *café* dans un espace sémantique multidimensionnel, où chaque cooccurrent représente une dimension particulière. Si l'on poursuit les explications dans cette métaphore géométrique, on pourra également calculer la distance entre deux mots dans cet espace sémantique multidimensionnel pour ainsi mesurer la distance entre leur sens. En pratique, on fait appel à l'algèbre des vecteurs pour calculer la similarité entre les trois mots-cibles. En effet, chaque mot-cible peut être représenté par le vecteur de

ses cooccurrents ou mots-contextes. Les données distributionnelles prennent donc la forme d'un vecteur de cooccurrents. Les mots-cibles qui partagent des cooccurrents auront une représentation vectorielle semblable. Une mesure de similarité permet ensuite de comparer les vecteurs et de calculer leur distance dans l'espace vectoriel. La similarité entre les vecteurs des trois mots-cibles est calculée à partir du cosinus de l'angle entre eux. Le cosinus est une mesure de similarité standard en sémantique distributionnelle (Bullinaria & Levy 2007 ; Ferret 2010). Intuitivement, on s'attend à ce que l'angle entre deux concepts similaires (p. ex. *chien* et *chat*) soit plus petit que celui entre deux mots différents (p. ex. *chien* et *café*). En d'autres termes, quand les distributions de mots sont peu similaires, la valeur de similarité cosinus est inférieure :

cos(chien, chat) = 0,55
cos(chien, café) = 0,27
cos(chat, café) = 0,30

Le résultat correspond à l'intuition : 'chien' et 'chat' se ressemblent le plus, tandis que 'chien' et 'café' partagent le moins de cooccurrences. Il est à noter que 'chien' ressemble un peu moins (0,27) à 'café' que 'chat' (0,30), parce que 'chien' apparaît aussi avec 'aboyer' et 'passant', contrairement à 'chat' et 'café'.

L'implémentation de cette technique aux exemples et au corpus du monde réel ne s'appuie pas sur des fréquences de cooccurrence absolues entre le mot-cible et ses cooccurrents. Les cooccurrents les plus fréquents ne donnent pas nécessairement le plus d'informations sur la signification du mot-cible. À l'instar des travaux sur les collocations, les méthodes d'analyse distributionnelle s'appuient sur des mesures d'association statistiques, ce qui permet de donner plus de poids aux cooccurrents qui apparaissent significativement plus souvent avec le mot-cible qu'attendu par le hasard. Ces cooccurrents avec un poids plus élevé fournissent plus d'informations sur le sens du mot-cible que les autres cooccurrents, quelle que soit leur fréquence absolue. Dans l'exemple avec le mot-cible 'chien', le cooccurrent 'vétérinaire' est sémantiquement plus proche du concept « animal » que le mot 'prendre', bien qu'il soit moins fréquent. Les mesures d'association qui sont utilisées dans les modèles sémantiques distributionnels comme fonctions de pondération sont empruntées aux analyses de collocations. Les schémas de pondération appliqués dans cette étude de cas reposent sur la mesure de l'information mutuelle ponctuelle (*Pointwise Mutual Information* ou *PMI*). Les détails techniques des mesures pour la pondération dépassent le cadre du présent article, mais on pourra se référer à A. Thanapoulos, N. Fakotakis et G. Kokkinakis (2002) pour une comparaison de plusieurs mesures statistiques. Pour calculer la valeur d'association, on s'appuie sur les fréquences de cooccurrence recueillies dans un grand corpus. Supposons que *vétérinaire* a un poids collocationnel de 4,3 s'il apparaît avec *chien*, de 3,5 avec *chat* et de 0,8 avec *café*. Ensuite, le calcul de la similarité cosinus à partir des poids collocationnels indique, d'une part, que *chien* et *chat* sont plus similaires que dans le calcul non pondéré et, d'autre part, qu'ils sont moins similaires à *café* :

cos(chien, chat) = 0,87
cos(chien, café) = 0,31
cos(chat, café) = 0,32

Le calcul pondéré de la similarité cosinus entre toutes les paires de mots-cibles permet de générer une matrice de similarité avec les mots-cibles, tant dans les lignes que dans les colonnes, et avec la valeur de similarité cosinus par paire de mots-cibles dans les cellules. Les éléments de la diagonale sont tous des '1', chaque mot-cible étant complètement similaire à lui-même. La matrice est symétrique, avec des valeurs identiques de part et d'autre de la diagonale, parce que la similarité cosinus entre les mots A et B est identique à celle entre les mots B et A. Le Tableau 3 *infra* visualise les similarités cosinus pour nos trois mots-cibles exemples. Dans un grand corpus du monde réel, on pourra trouver ainsi, pour chaque mot-cible, le mot le plus similaire dans le reste du vocabulaire. Étant donné que des mots très similaires sont souvent des (quasi-)synonymes, ce type de matrice de similarité mot-par-mot est souvent utilisé en linguistique computationnelle pour des tâches d'extraction automatique de synonymes.

Tableau 3 : Matrice des valeurs de similarité cosinus

	chat	café	chien
chat	1	0,32	0,87
café	0,32	1	0,31
chien	0,87	0,31	1

3. APPLICATION LEXICOLOGIQUE

Les méthodes d'analyse distributionnelle se prêtent à diverses applications lexicologiques et lexicographiques, notamment la sélection automatique de phrases-exemples pour les dictionnaires (Cook *et al.* 2013), l'étude diachronique de la sémantique (Sagi, Kaufmann & Clark 2009 ; Perek 2014) et l'étude de la variation socio-linguistique (Peirsman, Geeraerts & Speelman 2010). Nous présentons *infra* une étude de la polysémie dans la langue spécialisée. Le modèle distributionnel expliqué dans la section précédente permet non seulement de recenser des (dis)similarités sémantiques entre différents mots, tels que 'chien', 'chat' et 'café', mais également d'explorer les différents sens d'un mot particulier à partir des (dis)similarités sémantiques entre ses cooccurrents.

Nous procédons à l'analyse exploratoire de la polysémie ou de l'hétérogénéité sémantique du mot 'tour' dans un corpus technique, relevant du domaine spécialisé des machines-outils pour l'usinage des métaux (1,7 millions d'occurrences). Ce mot a en effet plusieurs sens techniques dans ce corpus spécialisé, à savoir « machine-outil pour l'usinage des pièces » et « rotation », ainsi que plusieurs sens généraux, notamment dans les expressions *tour d'horizon* et *à tour de rôle*. Comme nous l'avons indiqué *supra* (cf. § 2), la plupart des analyses en

sémantique distributionnelle étudient la proximité sémantique entre mots en fonction des mots-contextes partagés. Dans notre étude, l'objet d'analyse se situe à un niveau supérieur. Les mots-cibles (cf. Tableau 2), dont nous essayons d'étudier les (dis)similarités sémantiques, sont les cooccurrents de 'tour', parce que ceux-ci reflètent les différents sens et usages de 'tour' dans le corpus technique. Notre analyse consiste à regrouper les mots-cibles en fonction de leurs mots-contextes et à les positionner les uns par rapport aux autres en 2D. Ces analyses de regroupement *(clustering)* et de visualisation *(plotting)* des cooccurrents permettent de cerner des groupes de cooccurrents sémantiquement liés pour ainsi accéder à la sémantique du mot 'tour'.

Pour identifier les mots-cibles ou les cooccurrents pertinent de 'tour', ainsi que leurs mots-contextes pertinents, nous nous appuyons sur la mesure d'association de l'information mutuelle spécifique ou *Pointwise Mutual Information* (PMI). Pour le regroupement et la visualisation des mots-cibles, nous recourons à l'analyse de positionnement multidimensionnel (*MultiDimensional Scaling* ou MDS) (Kruskal & Wish 1978 ; Cox & Cox 2001 ; Venables & Ripley 2002 ; Borg & Groenen 2005). La technique de MDS [3] est implémentée dans le logiciel d'analyse statistique R [4]. Dans nos analyses, nous utilisons le positionnement non métrique isoMDS. Cette technique permet d'analyser une matrice pour un ensemble de données disposées en ligne (ici : les mots-cibles) à partir de leurs valeurs pour plusieurs variables disposées en colonne (ici : les mots-contextes) (cf. Tableau 4). Dans un premier temps, ces valeurs sont les valeurs d'association PMI entre un mot-cible et un mot-contexte avec lequel il apparaît. Cela permet de générer un vecteur par mot-cible avec toutes les valeurs d'association avec tous ses mots-contextes. Au cas où cette matrice serait trop creuse, nous pourrions enrichir les données en faisant appel aux cooccurrents des mots-contextes. Pour les détails sur l'enrichissement de la matrice, on pourra se référer à A. Bertels et D. Speelman (2013). Dans un deuxième temps, le calcul pondéré de la similarité cosinus entre toutes les paires de mots-cibles permet de générer une matrice de similarité, à l'instar du Tableau 3.

3. Le MDS est une méthode d'analyse multivariée descriptive, comme l'analyse factorielle des correspondances (AFC) ou l'analyse en composantes principales (ACP). À 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].

4. R [www.r-project.org]

Tableau 4 : Matrice de cooccurrence pondérée : mots-cibles par mots-contextes

	minute	***par***	***commande***	***broche***
mille	valeur d'association (PMI) entre *mille* et *minute*	PMI	/	/
numérique	/	/	PMI	/
inversé	/	/	/	PMI
horizon	/	/	/	/

La matrice de similarité est ensuite soumise à une analyse de positionnement multidimensionnel, qui consiste à regrouper les mots-cibles ou les cooccurrents de 'tour' en fonction des valeurs d'association similaires avec des mots-contextes similaires et à visualiser ces proximités et distances sémantiques en 2D. Cette visualisation permettra aux chercheurs-linguistes, non-informaticiens, d'accéder à la sémantique du mot 'tour'.

Dans la Figure 1 *infra*, la répartition des mots-cibles représente bien le caractère sémantiquement hétérogène de 'tour'. Les proximités et les distances sémantiques montrent quelques mots-cibles isolés et quelques groupes de mots-cibles sémantiquement liés. Les mots-cibles périphériques 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 mots-cibles qui attestent le sens technique particulier « tour inversé », avec *inversés*, *inversé* et *bibroches*. Le nuage de mots-cibles dans la partie inférieure à gauche visualise le sens technique « machine-outil pour l'usinage des pièces ». La présence de plusieurs petits regroupements de mots-cibles 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 mots-cibles moins spécialisés. Les mots-cibles plus généraux comme *manutention*, *ligne*, *générale*, *production* et *conception* se retrouvent dans la partie supérieure de ce nuage et font preuve d'un contexte d'apparition plus général.

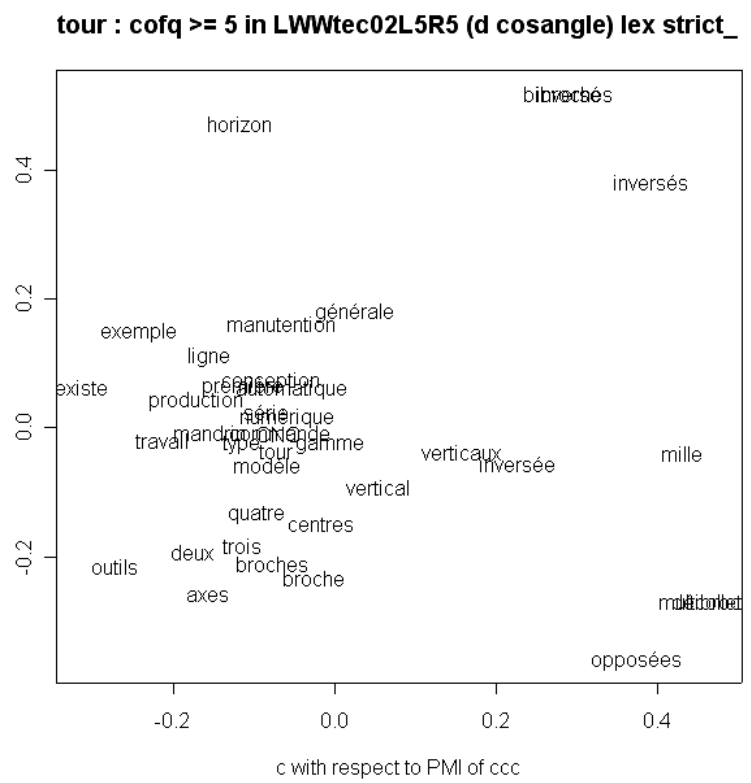

Figure 1 : Visualisation des mots-cibles ou cooccurrents de 'tour'

4. CONCLUSION

Dans cet article, nous avons présenté les possibilités de la modélisation sémantique distributionnelle pour les analyses linguistiques. Grâce à la disponibilité de vastes corpus électroniques, les lexicologues et lexicographes disposent de nos jours d'une énorme base de données empirique. Pour analyser cette abondance de données, ils pourront recourir à des outils d'analyse statistique et à des méthodes d'analyse distributionnelle qui leur permettront notamment de repérer des patrons et motifs.

Les méthodes d'analyse distributionnelle s'appuient sur l'hypothèse distributionnelle selon laquelle des mots qui se trouvent dans des contextes d'apparition similaires tendent à avoir des sens similaires. Elles regroupent plusieurs approches et sous-types en fonction du type de contexte, du niveau d'analyse et du mode de représentation computationnelle des données. Nous avons expliqué les principes de base de ces modèles à partir d'un modèle distributionnel particulier, relativement intuitif pour des linguistes, à savoir un modèle à base de

mots qui considère les cooccurrents graphiques au niveau des types sous forme de représentation vectorielle. Ce modèle distributionnel permet non seulement de recenser des (dis)similarités sémantiques entre différents mots, tels que les exemples-jouets 'chien', 'chat' et 'café', mais également d'explorer les différents sens d'un mot particulier à partir des (dis)similarités sémantiques entre ses cooccurrents. À titre d'exemple, nous avons discuté une application lexicologique qui consistait à regrouper les mots-cibles ou les cooccurrents du mot technique 'tour' en fonction des valeurs d'association similaires avec des mots-contextes similaires. Ces proximités et distances sémantiques sont visualisées en 2D pour permettre aux chercheurs-linguistes, non-informaticiens, d'accéder à la sémantique du mot 'tour'.

L'objectif de cet article était de montrer l'utilité des modèles sémantiques distributionnels pour les linguistes, lexicologues et lexicographes à l'aide d'une présentation non technique des principes de base et d'une application lexicologique concrète. Comme les différentes approches et les nombreux sous-types de modèles le suggèrent, des recherches futures devront s'atteler à la mise au point des nombreux paramètres pour aboutir au modèle qui soit le mieux adapté aux besoins spécifiques de la recherche lexicologique ou lexicographique nécessaire.

Références

Bernier-Colborne G. (2014), « Analyse distributionnelle de corpus spécialisés pour l'identification de relations lexico-sémantiques », *Actes du 21e Traitement Automatique des Langues Naturelles (TALN2014), Atelier SemDis 2014*, Marseille, France, 238-251.

Bertels A. & Speelman D. (2013), « Exploration sémantique visuelle à partir des cooccurrences de deuxième et troisième ordre », *Actes du 20e Traitement Automatique des Langues Naturelles (TALN2013), Atelier SemDis 2013*, Les Sables-d'Olonne, France, 126-139.

Bertels A. & Speelman D. (2014), « Analyse de positionnement multidimensionnel sur le corpus spécialisé TALN », *Actes du 21e Traitement Automatique des Langues Naturelles (TALN2014), Atelier SemDis 2014*, Marseille, France, 252-265.

Borg I. & Groenen P. (2005[2]), *Modern Multidimensional Scaling: theory and applications*, New York: Springer-Verlag.

Bouaud J. et al. (1997), « Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles », *Actes des 1res journées Ingénierie des Connaissances*, Roskoff, France, 207-223.

Bruni E., Tran N. K. & Baroni M. (2014), "Multimodal Distributional Semantics", *Journal of Artificial Intelligence Research* 49 (1), 1-47.

Bullinaria J. A. & Levy J. P. (2007), "Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study", *Behavior Research Methods* 39, 510-526.

Church K. W. & Hanks P. (1989), "Word association norms, mutual information and lexicography", *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*, Vancouver: British Columbia, 76-83.

Claveau V., Kijak E. & Ferret O. (2014), « Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels », *Actes du 21e Traitement Automatique des Langues Naturelles (TALN2014)*, Marseille, France, 220-231.

COOK P. *et al.* (2013), "A lexicographic appraisal of an automatic approach for detecting new word-senses", *Proceedings of the eLex 2013 Conference*, Tallinn, Estonia, 49-65.

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

CURRAN J. R. & MOENS M. (2002), "Improvements in automatic thesaurus extraction", *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, Vol. 9, Stroudsburg (PA), USA, 59-66.

DESALLE Y. *et al.* (2014), « BACANAL : Balades Aléatoires Courtes pour ANAlyses Lexicales Application à la substitution lexicale », *Actes du 21e Traitement Automatique des Langues Naturelles (TALN2014), Atelier SemDis 2014*, Marseille, France, 206-217.

DINU G., THATER S. & LAUE S. (2012), "A comparison of models of word meaning in context", *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta (GA), USA, 611-615.

EVERT S. (2004), *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

FABRE C. *et al.* (2014a), « Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés », *Actes du 21e Traitement Automatique des Langues Naturelles (TALN2014), Atelier SemDis 2014*, Marseille, France, 196-205.

FABRE C. *et al.* (2014b), « Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille », *Actes du 21e Traitement Automatique des Langues Naturelles (TALN2014), Atelier SemDis 2014*, Marseille, France, 266-279.

FERRET O. (2010), « Similarité sémantique et extraction de synonymes à partir de corpus », *Actes du 17e Traitement Automatique des Langues Naturelles (TALN2010)*, Montréal, Canada. [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 », *Actes du 21e Traitement Automatique des Langues Naturelles (TALN2014), Atelier SemDis 2014*, Marseille, France, 218-227.

FIRTH J. R. (1957), *Papers in Linguistics, 1934-1951*, Oxford: Oxford University Press.

GEERAERTS D. (2010), *Theories of Lexical Semantics*, Oxford/New York: Oxford University Press.

GLYNN D. (2010), "Testing the Hypothesis. Objectivity and Verification in Usage-Based Cognitive Semantics", *in* D. Glynn & K. Fischer (eds), *Quantitative Methods in Cognitive Semantics. Corpus-driven approaches*, Berlin/New York: Mouton de Gruyter, 239-270.

HABERT B. & ZWEIGENBAUM P. (2002), "Contextual acquisition of information categories: what has been done and what can be done automatically?", *in* B. Nevin & S. Johnson (eds), *The Legacy of Zellig Harris. Language and information into the 21st century*, Amsterdam: John Benjamins, 203-231.

HARRIS Z. (1954), "Distributional structure", *Word* 10 (23), 146-162.

HEYLEN K. *et al.* (2014), "Monitoring Polysemy: Word Space Models as a Tool for Large-Scale Lexical Semantic Analysis", *Lingua: International Review of General Linguistics* 157, 153-172.

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 et UNCLH*, Avignon, France, 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.

LANDAUER T. K. & DUMAIS S. T. (1997), "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge", *Psychological Review* 104 (2), 211-240.

LUND K. & BURGESS C. (1996), "Producing high-dimensional semantic spaces from lexical co-occurrence", *Behavior Research Methods, Instrumentation and Computers* 28, 203-208.

MORARDO M. & VILLEMONTE DE LA CLERGERIE É. (2013), « Vers un environnement de production et de validation de ressources lexicales sémantiques », *Actes du 20e Traitement Automatique des Langues Naturelles (TALN2013), Atelier SemDis 2013*, Les Sables-d'Olonne, France, 167-180.

MORLANE-HONDÈRE F. (2013), « Utiliser une base distributionnelle pour filtrer un dictionnaire de synonymes », *Actes du 20e Traitement Automatique des Langues Naturelles (TALN2013), Atelier SemDis 2013*, Les Sables-d'Olonne, France, 112-125.

NAVIGLI R. (2012), "A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches", *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, Spindleruv Mlyn, Czech Republic, 115-129.

PEIRSMAN Y., GEERAERTS D. & SPEELMAN D. (2010), "The automatic identification of lexical variation between language varieties", *Natural Language Engineering* 16 (4), 469-491.

PEREK F. (2014), "Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, USA, 309-314.

PÉRINET A. & HAMON T. (2014), « Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité », *Actes des 12e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2014)*, Paris, France, 507-518.

SAGI E., KAUFMANN S. & CLARK B. (2009), "Semantic density analysis: Comparing word meaning across time and phonetic space", *Proceedings of the EACL 2009 Workshop on GEMS*, Athens, Greece, 104-111.

SAHLGREN M. (2006), *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*, Ph.D. Thesis, Department of Linguistics, Stockholm University.

SINCLAIR J. (1991), *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

THANOPOULOS A., FAKOTAKIS N. & KOKKINAKIS G. (2002), "Comparative Evaluation of Collocation Extraction Metrics", *Proceedings of the 3rd Language Resources Evaluation Conference*, Las Palmas, Spain, 620-625.

TURNEY P. D. & PANTEL P. (2010), "From Frequency to Meaning: Vector Space Models of Semantics", *Journal of Artificial Intelligence Research* 37 (1), 141-188.

VENABLES W. N. & RIPLEY B. D. (2002[4]), *Modern Applied Statistics with S*, New York: Springer-Verlag.

WEAVER W. (1955), "Translation", *in* W. N. Locke & A. D. Booth (eds), *Machine Translation of Languages*, Cambridge (MA): The MIT Press, 15-23.

WIECHMANN D. (2008), "On the Computation of Collostruction Strength", *Corpus Linguistics and Linguistic Theory* 4 (2), 253-290.