Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques

Ludovic Tanguy, Franck Sajous et Nabil Hathout

CLLE-ERSS (CNRS & Université de Toulouse 2)
5, allées Antonio Machado
F - 31058 Toulouse Cedex 9
{ludovic.tanguy, franck.sajous, nabil.hathout}@univ-tlse2.fr

RÉSUMÉ. *Il est possible de construire des modèles distributionnels en ne considérant que la cooccurrence graphique entre les mots, ou bien en utilisant des relations syntaxiques de complexité variable. Si des comparaisons systématiques n'ont jamais pu trancher définitivement en faveur de l'une ou de l'autre, elles ont rarement été menées sur un corpus de taille réduite ou en langue de spécialité. Nous proposons ici une palette d'expériences visant l'observation d'un ensemble de modèles distributionnels construits à partir d'un petit corpus d'articles en français dans le domaine du TAL. Un jeu de données a été spécifiquement conçu pour l'évaluation des différentes configurations. Ces expériences montrent que les modèles qui prennent en compte de façon raisonnable les informations syntaxiques obtiennent globalement de meilleurs résultats.*

ABSTRACT. *Distributional semantics models can be built using simple bag-of-word representation of a word's contexts (window-based) or using more complex syntactic information (syntax-based). Previous studies have compared their relative efficiency without coming to a definitive conclusion, but such examination has never been performed on small and specialised corpora. We have run a set of such comparative experiments based on a collection of French NLP articles and a custom-made gold standard. These experiments show a better global performance of syntax-based models, as long as syntactic information is processed with appropriate care.*

MOTS-CLÉS : *sémantique distributionnelle, corpus spécialisé.*

KEYWORDS: *distributional semantics, specialised corpus.*

1. Introduction

Bien que l'hypothèse harrissienne – les mots dont les distributions sont similaires sont sémantiquement proches – qui sous-tend les recherches sur l'analyse distributionnelle automatique (ADA) soit ancienne (Firth, 1951 ; Harris, 1954), la multitude de travaux et le nombre d'appels à communication récents portant sur cette thématique montrent que ce champ d'investigation n'est pas épuisé. L'ADA connaît actuellement un engouement certain : elle est utilisée pour un ensemble d'applications et de questionnements, allant de la psycholinguistique (Fyshe *et al.*, 2014) jusqu'à la classification de documents (Bullinaria et Levy, 2012). Le principe de base de l'ADA est d'utiliser un corpus pour construire un espace vectoriel dans lequel chaque mot est représenté par les contextes dans lesquels il apparaît. La proximité sémantique de deux mots est alors estimée par leur similarité distributionnelle, établie en mesurant la distance entre leurs représentations dans cet espace vectoriel (Sahlgren, 2006).

Malgré la pertinence des méthodes distributionnelles et la diversité de leurs usages, des questions fondamentales sur leur fonctionnement restent encore ouvertes. Les développements technologiques récents semblent se concentrer sur la sophistication de la mécanique interne de l'ADA à travers, par exemple, la réduction de dimensions (Van de Cruys et Apidianaki, 2011) ou l'échantillonnage négatif (Mikolov *et al.*, 2013) appliqués à des contextes simples tels que cooccurrents de surface ou séquences de type *skipgrams* (*ibid.*) sur de très gros corpus. Les méthodes plus linguistiques utilisant des représentations syntaxiques des contextes qui ont été étudiées par le passé (Grefenstette, 1993 ; Padó et Lapata, 2007, parmi d'autres) semblent actuellement délaissées, le surcoût en termes de calcul, mais aussi de complexité du paramétrage, n'étant pas justifié par une amélioration probante des résultats.

Nous étendons dans cet article l'étude de Fabre *et al.* (2014b) et cherchons à comparer plus finement les méthodes par contextes syntaxiques à celles par contextes graphiques, en les appliquant à nouveau à un corpus spécialisé de taille réduite. L'article propose une analyse de différents paramètres qui déterminent le comportement des modèles distributionnels par une évaluation de 2 592 configurations de systèmes par cooccurrents graphiques ou syntaxiques. Nous rejoignons en cela les travaux de comparaison de Padó et Lapata (2007), Lapesa et Evert (2014) et Kiela et Clark (2014) réalisés sur de gros corpus génériques.

Le fait de s'intéresser à un corpus spécialisé (et de ce fait de taille réduite) rejoint des travaux plus anciens sur l'utilisation de l'ADA pour la constitution de ressources lexicales (Harris *et al.*, 1989 ; Habert et Zweigenbaum, 2002) et répond à des besoins existants. Si de nombreux travaux réalisés sur des collections de documents en langue de spécialité, notamment dans le domaine biomédical, utilisent l'ADA pour l'acquisition et la structuration de connaissances[1] aucune étude de comparaison systématique des modèles distributionnels graphiques *vs* syntaxiques n'a été effectuée à notre connaissance sur des corpus spécialisés, *a fortiori* en langue française. Cette

1. Voir (Cohen et Widdows, 2009) pour une synthèse.

comparaison soulève d'ailleurs des problèmes méthodologiques d'évaluation : quel *gold standard* utiliser ? Comment estimer la qualité des voisinages distributionnels ?

L'une des contributions du travail présenté ici est justement la proposition d'une nouvelle méthode d'évaluation adaptée à notre objet d'étude : l'analyse distributionnelle de corpus spécialisés de taille réduite. Cette méthode n'entre pas dans le cadre des évaluations plus classiques de l'ADA comme l'identification de synonymes sur la base des tests du TOEFL, la corrélation entre similarité distributionnelle et jugements de locuteurs utilisant les jeux de données de Rubenstein et Goodenough (1965), Miller et Charles (1991) ou Finkelstein *et al.* (2002), la classification (*clustering*) comme celles de Almuhareb et Poesio (2004). Notre évaluation repose au contraire sur un jeu de données conçu spécialement pour le corpus étudié : elle est effectuée relativement à un *gold standard* spécifique composé d'un ensemble de paires de mots reliés par des relations sémantiques diversifiées, valuées par un score qui reflète la force de la relation.

Les expériences réalisées montrent que les méthodes utilisant des contextes syntaxiques obtiennent en moyenne des résultats supérieurs à celles fondées sur la simple cooccurrence graphique, sous condition d'une prise en compte raisonnable des informations syntaxiques.

2. Démarche et travaux connexes

La question centrale de l'article rejoint celle de Padó et Lapata (2007), et de Curran et Moens (2002) avant eux : les contextes produits par les analyses syntaxiques permettent-ils d'améliorer les résultats des modèles distributionnels créés à partir de fenêtres graphiques, et dans quelles conditions ? Padó et Lapata (2007) rappellent que les études comparatives ont été souvent peu concluantes : les contextes syntaxiques peuvent, par exemple, être plus performants dans une tâche d'acquisition automatique de thésaurus, tout en dégradant les résultats d'une tâche de recherche d'information. Ces auteurs relèvent par ailleurs que l'utilisation des informations syntaxiques dans ces études est relativement fruste : seules certaines relations directes sont prises en compte, et jamais, par exemple, la relation entre deux actants d'un même verbe. Ils concluent leur étude en montrant la supériorité des méthodes exploitant la syntaxe dans plusieurs tâches. Ces conclusions rejoignent celles de Van der Plas et Bouma (2005), Peirsman *et al.* (2007) et Heylen *et al.* (2008) qui ont mené des analyses sur des corpus journalistiques néerlandais et comparé des méthodes par cooccurrents graphiques et par cooccurrents syntaxiques. Pour ces dernières, chaque relation syntaxique est étudiée séparément. Les trois études concluent en faveur des approches syntaxiques et montrent l'intérêt de considérer l'ensemble de relations de dépendance. Dans un travail similaire, réalisé sur un corpus journalistique portugais, Gamallo Otero (2008) s'intéresse à la relations de co-hyponymie. Il établit que les modèles distributionnels construits à partir de contextes syntaxiques permettent d'identifier cette relation avec une plus grande précision que les modèles fondés sur les fenêtres graphiques. Plus récemment, Kiela et Clark (2014) ont comparé les deux types de méthodes en

analysant de gros corpus (BNC et ukWaC). L'évaluation est réalisée avec les tests de synonymie du TOEFL et quatre autres jeux de données comportant des jugements humains de similarité. Dans cette expérience, ce sont les cooccurrents graphiques qui l'emportent.

3. Données

Nous avons utilisé pour cette étude le corpus TALN (Boudin, 2013), qui comprend 2 millions de mots (62 631 formes et 22 210 lemmes différents). Il se compose de 586 articles des conférences TALN et RECITAL de 2007 à 2013[2]. Il a été étiqueté et analysé syntaxiquement en dépendances par Talismane (Urieli et Tanguy, 2013)[3].

3.1. *Choix des mots cibles*

Dans (Fabre *et al.*, 2014b), un premier jeu d'évaluation a été conçu sur ce même corpus, comportant 15 mots cibles de fréquence moyenne : 5 noms, 5 verbes et 5 adjectifs. Dans la présente étude, la taille de ce jeu initial est doublée en ajoutant 5 mots cibles supplémentaires pour chaque catégorie (soit au total 10 mots cibles par catégorie) afin d'inclure également des mots de haute et de basse fréquence. Le nouveau jeu d'évaluation contient à la fois des termes spécialisés, des termes considérés plus génériques et des mots appartenant à ce que Tutin (2007) appelle « le lexique et la phraséologie transdisciplinaire des écrits scientifiques ». Les 30 mots cibles sélectionnés, listés ci-dessous avec leur fréquence, incluent notamment des items relativement difficiles à caractériser comme *empirique*, *sémantique*, *trait* ou *conduire*.

Adjectifs : sémantique (3 074), important (1 287), complexe (741), temporel (698), correct (622), précis (383), spécialisé (377), significatif (351), empirique (86), computationnel (60).

Noms : méthode (3 816), trait (1 814), élément (1 576), performance (1 315), graphe (1 119), fréquence (952), contrainte (947), sémantique (398), dépendant (96), signification (76).

Verbes : décrire (1 458), évaluer (1 302), extraire (1 165), calculer (1 014), annoter (790), valider (379), caractériser (374), conduire (366), indexer (66), apparier (54).

2. Préparé dans le cadre de l'atelier SemDis2014 (Fabre *et al.*, 2014a), ce corpus est disponible en versions brute et analysée syntaxiquement à l'adresse :
`http://redac.univ-tlse2.fr/corpus/taln.html`

3. Talismane est un analyseur syntaxique en dépendances librement disponible à l'adresse : `http://redac.univ-tlse2.fr/applications/talismane.html`

3.2. *Construction du* gold standard

La liste des « meilleurs » voisins de chacun des mots cibles est constituée en utilisant une *pooling method* directement inspirée de l'évaluation des systèmes de recherche d'information. Pour un mot cible donné, les 3 meilleurs voisins distributionnels générés par chacun des 2 592 systèmes présentés en section 4.2 sont soumis aux juges. Nous sommes conscients du fait que cette méthode ne couvre pas l'ensemble des résultats renvoyés par les systèmes évalués. Cependant, elle permet de réduire le coût de l'annotation tout en examinant les voisins les mieux représentés dans les résultats des différents systèmes et dont le poids sera le plus fort dans l'évaluation (voir section 5.1). Le nombre de candidats examinés par les juges varie de 64 à 445 selon les mots cibles. Un total de 6 091 paires `{mot cible, voisin potentiel}` a ainsi été présenté aux juges auxquels il a été demandé de répondre à la question suivante : *ces deux mots sont-ils sémantiquement proches dans le domaine du TAL ?*

L'annotation, réalisée par qautre juges experts du domaine (dont deux sont des auteurs de l'article), a produit 1 328 paires choisies par au moins un juge, réparties de la manière suivante : 21 % ont été sélectionnées par les quatre juges, 16 % par trois juges, 23 % par deux juges et 40 % par un juge. Par exemple, les voisins choisis par les quatre juges pour l'adjectif *complexe* sont : *aisé*, *ardu*, *compliqué*, *difficile*, *facile*, *polysémique*, *riche*, *simple*, *sophistiqué*, *trivial*, *élaboré*, *élémentaire*, incluant à la fois des synonymes (*compliqué*, *difficile*), des antonymes (*aisé*, *simple*) et des termes dont la présence et la similarité sont totalement dépendantes du domaine (*polysémique*). À l'opposé, le lien sémantique est plus lâche et aucune relation sémantique classique ne peut être identifiée pour les voisins acceptés par un seul annotateur comme *irrégulier*, *abstrait*, *varié*. Ces décisions peuvent découler d'inférences comme le fait qu'un phénomène irrégulier est plus complexe à traiter par des méthodes de TAL. Il semble donc important de prendre en compte cette gradation dans la similarité, dont une estimation satisfaisante est le nombre de juges ayant retenu le voisin.

La valeur moyenne de l'accord inter-annotateurs, calculé pour chaque mot cible avec un kappa de Fleiss est de 0,55 et le coefficient de corrélation de Pearson moyen sur les paires d'annotateurs est de 0,57. Ce score se situe dans la tendance générale pour les annotations de similarité sémantique : voir notamment (Zesch et Gurevych, 2006) pour un comparatif de campagnes avec des scores allant de 0,47 à 0,90. Notre situation est toutefois différente de celles évoquées, le nombre de paires que nous avons annotées étant largement supérieur. Le fait que le corpus soit spécialisé a, semble-t-il, réduit la dispersion des réponses. L'accord fluctue légèrement en fonction de la catégorie du mot cible : l'annotation des adjectifs (κ = 0,59) est plus simple que celle des noms (κ = 0,56) et des verbes (κ = 0,50). Par ailleurs, certains mots sont plus faciles à traiter que d'autres : l'adjectif *complexe*, avec ses nombreux synonymes et antonymes génériques, pose moins de problèmes (κ = 0,70) que les verbes *calculer* (κ = 0,34) et *indexer* (κ = 0,30). L'accord inter-annotateurs est marginalement corrélé à la fréquence des mots cibles, mais il l'est positivement (ρ = 0,40). Cet effet est plus marqué pour les verbes, qui sont notoirement plus polysémiques.

Le jeu d'évaluation que nous venons de présenter reste perfectible par sa couverture et sa fiabilité, mais il n'en demeure pas moins incontournable pour comparer le comportement des méthodes distributionnelles sur le corpus TALN[4]. Son adéquation peut être mise en évidence en comparant par exemple le choix des experts pour les voisins du nom *trait* (*attribut, feature, étiquette, qualia*) à son entrée dans le *Robert des synonymes* : *jet*, *flèche*, *ligne*, *barre*, *dessin*, *rayon*, *figure*, *attribut*, *caractère*, *caractéristique*, *marque*, *signe*, *attaque*, *raillerie*. On voit que les experts ont sélectionné un seul sens du mot (*caractéristique*) en ignorant les acceptions de type *trait de crayon* ou *trait d'humour* et qu'ils ont retenu des termes spécifiques au domaine (*feature* ou *qualia*) ainsi que des termes associés comme *étiquette*.

4. Modèles distributionnels

Nous présentons dans cette section les différents types de contexte et paramètres qui interviennent dans le calcul des mesures de similarité entre mots et les combinaisons de paramètres que nous avons comparées.

4.1. *Types de contexte*

On distingue classiquement en ADA les approches où le contexte d'une occurrence (que l'on appellera « pivot ») est constitué de ses cooccurrents graphiques dans une fenêtre donnée, et celles où ce sont les mots avec lesquels elle entretient des relations syntaxiques qui sont considérés.

4.1.1. *Contextes graphiques*

Bernier-Colborne (2014) a montré que, dans une tâche d'extraction de relations lexico-sémantiques, une fenêtre étroite de 2 à 4 mots donne les meilleurs résultats dès lors que l'on utilise une mesure d'association et non la fréquence absolue. Ce résultat est d'autant plus intéressant qu'il a travaillé sur un corpus spécialisé dans le domaine de l'environnement, de taille comparable à celle du corpus TALN. Il confirme celui de Bullinaria et Levy (2007), obtenu sur un corpus anglais en faisant varier la taille de 1 à 100 millions de mots. Ferret (2010), qui utilise le test du TOEFL étendu et le corpus ACQUAINT-2, obtient de meilleurs résultats avec une fenêtre de longueur 1, sans filtrage des cooccurrents sur la fréquence. Peirsman *et al.* (2007) ont montré qu'élargir la fenêtre conduisait à une baisse conséquente des résultats. Il en va de même pour le filtrage des contextes par leur fréquence. Kiela et Clark (2014) observent que la longueur optimale de la fenêtre dépend de la tâche et de la taille du corpus, la meilleure combinaison globale étant une petite fenêtre avec un grand corpus. Ils montrent également qu'il est inutile de considérer les contextes qui ne figurent pas parmi les 50 000 plus

4. Le jeu de données est disponible à l'adresse :
`http://redac.univ-tlse2.fr/datasets/TAL56-2/`

fréquents. Rappelons que le corpus TALN contient environ 22 000 lemmes différents toutes catégories confondues.

Les approches « non structurées » que nous mettons en œuvre s'appuient sur ces observations : nous considérons comme contextes d'un pivot donné les mots situés à gauche uniquement, à droite uniquement ou apparaissant dans un empan de texte centré sur l'occurrence du mot considéré. Nous avons testé des fenêtres de longueur 1, 3 et 5 (dans l'une des directions ou dans les deux), en conservant tous les mots (*i.e.* sans filtrage sur les catégories). Ces fenêtres peuvent franchir les frontières de phrases, ces dernières étant des *tokens* particuliers faisant partie des contextes. Aucune distinction ni pondération relativement à la distance avec le pivot n'est appliquée. De telles pondérations sont souvent utilisées pour les fenêtres de grande taille, afin de favoriser les mots proches (Sahlgren, 2006). Par ailleurs, le corpus est catégorisé et lemmatisé.

4.1.2. *Dépendances brutes*

L'exploitation des relations de dépendance syntaxique pour construire les contextes peut se faire, à la façon de Kiela et Clark (2014), en utilisant directement les sorties de l'analyseur. Chaque dépendance produit un triplet de la forme $<$`gouverneur`;*relation*;`dépendant`$>$ permettant d'associer au gouverneur le contexte $<$*relation*; `dépendant`$>$ et au dépendant le contexte $<$*relation*$^{-1}$; `gouverneur`$>$. Par exemple, la dépendance sujet dans *« les puces [...] constituent »* en figure 1 produit les deux associations suivantes :

$$< \texttt{constituer}; \textit{suj}; \texttt{puce} > \rightarrow \begin{cases} \texttt{constituer} \leftrightarrow <\textit{suj}; \texttt{puce}> \\ \texttt{puce} \leftrightarrow <\textit{suj}^{-1}; \texttt{constituer}> \end{cases}$$

Toutes les relations fournies par l'analyseur sont utilisées telles quelles, sans modification (cf. tableau 1).

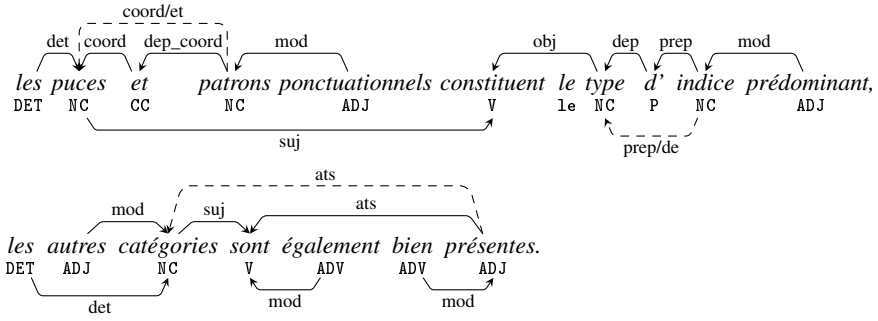

Figure 1 – Dépendances syntaxiques fournies par l'analyseur

4.1.3. *Contextes syntaxiques*

Les informations syntaxiques peuvent donner lieu à une exploitation plus sophistiquée, à l'instar de ce qu'ont proposé Baroni et Lenci (2010). Il est en effet possible de

Relation	**Triplets extraits**
déterminant *(det)*	<NC:puce; *det*; DET:les> <NC:catégorie; *det*; DET:les>
sujet *(suj)*	<V:constituer; *suj*; NC:puce> <V:être; *suj*; NC:catégorie>
objet *(obj)*	<V:constituer; *obj*; NC:type>
modifieur de nom *(nMod)*	<NC:patron; *nMod*; ADJ:ponctuationnel> <NC:indice; *nMod*; ADJ:prédominant> <NC:catégorie; *nMod*; ADJ:autre>
modifieur adverbial *(advMod)*	<V:être; *advMod*; ADV:également> <ADJ:présent; *advMod*; ADV:bien>

Tableau 1 – Dépendances brutes et triplets correspondants

Relation/variante	**Triplets extraits**
coordination *(coord)*	<NC:puce; *coord/et*; NC:patron>
+ fusion des CC *(fusionCC)*	<NC:puce; *coord*; NC:patron>
préposition *(prep)*	<NC:type; *prep/de*; NC:indice>
+ fusion des prépositions *(fusionPrep)*	<NC:type; *prep*; NC:indice>
attribut du sujet *(ats)*	<NC:catégorie; *ats*; ADJ:présent>
attribut → modifieur (*ats→nMod)*	<NC:catégorie; *nMod*; ADJ:présent>

Tableau 2 – Contextes syntaxiques et triplets correspondants

construire des triplets en appliquant des transformations et normalisations au réseau de dépendances fourni par l'analyseur.

Contrairement aux contextes de la section 4.1.2, seules certaines relations syntaxiques sont retenues. Le tableau 2 donne un exemple des triplets syntaxiques extraits de l'analyse présentée en figure 1. Aux relations *suj, obj, nMod* et *advMod* des contextes issus des dépendances directes (la relation *det* disparaissant) s'ajoutent :

1) les relations **préposition** (*prep*) correspondant aux structures N-prép-N, N-prép-V, V-prép-N et V-prép-V, **coordination** (*coord*), **attribut du sujet** (*ats*) et **objet indirect** (*obj2*). Ces relations « non directes », représentées en pointillé dans la figure 1, ne sont pas données explicitement par l'analyseur mais établies en passant respectivement par la préposition, la conjonction de coordination (ou la ponctuation), le verbe attributif et la préposition. La relation *prep* (resp. *coord*) existe en deux variantes, le mot grammatical pouvant être intégré à la relation du triplet (*e.g.* <`type`;*prep/de*;`indice`>) ou ne pas l'être. La deuxième version (*e.g.* <`type`;*prep*;`indice`>) est notée *fusionPrep* (resp. *fusionCC*). La variante *ats→nMod* de la relation *ats* consiste à transformer cette dernière en modifieur de nom classique ;

2) **fermeture transitive de la coordination** (*transCoord*) : si deux mots sont coordonnés à un même troisième, ils deviennent à leur tour coordonnés ;

3) **distribution des relations sur les coordonnés** (*distrCoord*) : un mot coordonné à l'objet (resp. sujet) d'un verbe devient lui-même objet (resp. sujet) de ce verbe. De même, un mot coordonné au modifieur d'un nom devient lui-même modifieur de ce dernier ;

4) **sujets des compléments verbaux sélectionnés par les verbes** (*sujCompVerb*) : certains verbes conjugués (modaux, causatifs, aspectuels, à contrôle, à montée, etc.) se construisent avec un complément verbal à l'infinitif. Dans ce type de structure, le sujet du verbe qui gouverne l'infinitif est attribué à ce dernier ;

5) **sujets des participes présents** (*sujVPR*) : lorsqu'un participe présent modifie un nom, on ajoute une relation sujet entre le nom et le verbe ;

6) recherche des **antécédents des pronoms relatifs** (*antProRel*) : l'antécédent du pronom devient le sujet du verbe gouverné ;

7) **normalisation des passifs** (*normPassifs*) : la relation *sujet* gouvernée par un passif est transformée en *objet* ;

8) **structure argumentale des verbes** : nous combinons les relations *suj*, *obj* et *obj2* pour construire de nouvelles relations. **Sujet-Verbe-Objet** (*svo*) et **Sujet-Objet** (*so*) lient le sujet et l'objet d'un verbe. Comme pour la relation *prep*, *svo* et *so* se différencient par l'inclusion ou non du verbe dans la relation du triplet. De même, **Sujet-Verbe-Objet indirect** (*svo2*) et **Sujet-Objet Indirect** (*so2*) lient le sujet et l'objet indirect d'un verbe. **Objets direct-indirect** (*oo2*) est une relation qui relie les objets direct et indirect d'un même verbe. Une relation réciproque, notée *pred*, reflète le fait qu'un verbe se construit avec un couple (objet direct, objet indirect) donné ;

9) **prise en compte des noms propres** (*inclNPP*) : inclusion des relations syntaxiques gouvernées par ou dépendant d'un nom propre.

En combinant les différentes relations présentées ci-dessus, nous avons défini 4 configurations listées dans le tableau 3. Ces configurations sont cumulatives, *i.e.* les familles 2 à 4 incluent les triplets extraits dans les familles d'indice inférieur et y ajoutent de nouvelles relations syntaxiques ou normalisations. Par exemple, la configuration *Synt3* correspond à la configuration *Synt2* augmentée de la relation de modification adverbiale, la normalisation des passifs, etc.

Synt1	Relations :	*suj, obj, nMod*
Synt2	Relations :	**Synt1** + *coord, prep, ats*
Synt3	Relations :	**Synt2** + *avdMod, inclNPP, sujVPR*
	Normalisations :	*fusionCC, transCoord, ats→nMod, normPassifs, antProRel*
Synt4	Relations :	**Synt3** + *obj2, svo, svo2, oo2, pred, sujCompVerb*

Tableau 3 – Configurations sélectionnées pour les contextes syntaxiques

4.2. *Calcul de similarité*

Les contextes extraits par les méthodes qui viennent d'être présentées servent à calculer une liste ordonnée de voisins distributionnels pour chaque mot du jeu d'éva-

luation. Ce calcul est réalisé en utilisant la bibliothèque *Wordspace* (Evert, 2014). Les paramètres suivants sont pris en compte : filtrage des contextes sur la base de leur distribution, choix d'une mesure d'association entre les mots et leurs contextes, transformation éventuelle de cette mesure par une fonction mathématique, calcul de la similarité entre les mots et ordonnancement et filtrage des voisins de chaque mot. Nous décrivons ici le détail de ces opérations et de leur paramétrage.

Filtrage par le nombre de contextes différents : un mot n'est retenu que s'il apparaît dans un nombre minimal de contextes différents. Nous filtrons de la même façon les contextes. Le seuil dépend des familles de contextes, les valeurs envisagées étant les suivantes :

– contextes graphiques : $5\times$, $10\times$ et $15\times$ la taille de la fenêtre

– dépendances brutes et contextes syntaxiques : 2 à 10

Le filtrage s'applique itérativement : à chaque fois qu'un mot (resp. contexte) est éliminé, il n'intervient plus dans le décompte des autres contextes (resp. mots).

Mesures d'association entre mots et contextes : plusieurs mesures sont utilisées en ADA pour estimer la force de l'association entre un mot et un contexte. Si la fréquence brute de cooccurrence est la plus simple, on lui préfère habituellement des mesures qui la pondèrent en fonction de la fréquence totale du mot et/ou du contexte. Plus précisément, ces mesures comparent la fréquence de cooccurrence avec les fréquences (relatives) du mot et du contexte, et proposent une valeur qui permet de distinguer les cooccurrences significatives de celles que l'on obtiendrait si la distribution des mots dans le texte était aléatoire. Les mesures envisagées dans cette étude, détaillées dans (Evert, 2007), sont : l'information mutuelle (*MI* dans la librairie *wordspace*), le rapport de vraisemblance (*simple-ll*), le *t-score* et le *z-score*. Il est en outre possible de leur appliquer une transformation simple permettant de pondérer leurs valeurs extrêmes. Trois possibilités ont été envisagées : aucune transformation, racine carrée et logarithme. Le tout produit 12 combinaisons (mesure d'association/transformation).

Mesures de la similarité entre mots : les mesures de similarité entre les vecteurs qui décrivent les contextes des mots sont directement inspirées des distances dans les espaces vectoriels : distance euclidienne, de Manhattan, cosinus, coefficient de corrélation, etc. Le cosinus, rendu populaire par la recherche d'information, est la mesure la plus utilisée et la plus efficace dans la grande majorité des études antérieures. Nous n'avons donc considéré que celle-ci pour nos expérimentations.

Filtrage des voisins distributionnels : une fois calculée la similarité entre tous les couples de mots du corpus, un filtrage supplémentaire est réalisé avant d'extraire les voisins des 30 mots cibles. Tout d'abord, nous avons éliminé les voisins dont la catégorie grammaticale est différente de celle du mot cible considéré : si certains rapprochements intercatégoriels sont pertinents (notamment les liens morphologiques), ils ne nous ont pas paru centraux. Ils n'apparaissent par ailleurs qu'exceptionnellement dans le *gold standard*. Nous avons ajouté une contrainte supplémentaire en imposant un seuil sur le nombre de contextes différents partagés par le mot cible et ses voisins.

Ce seuil est fonction du nombre minimal de contextes différents initialement partagés (nombre de contextes différents + 0, + 5, + 10).

Bilan des configurations testées : nous avons construit pour cette étude un total de 2 592 modèles distributionnels répartis comme indiqué dans le tableau 4.

Types de contextes	Formes des contextes	Contextes différents	Mesures d'association	Transformations	Contextes partagés	Total
Graphiques	9	3	4	3	3	972
Dépendances	1	9	4	3	3	324
Syntaxiques	4	9	4	3	3	1 296

Tableau 4 – Répartition des 2 592 configurations étudiées

Dans ce qui suit, nous ferons référence à une configuration particulière en utilisant une nomenclature du type *Graph_3G_15-z-score-root-20* qui signifie : contextes graphiques (*Graph*) ; fenêtre de taille 3 à gauche (*3G*) ; 15 contextes différents au minimum pour chaque mot pris en compte ; association entre mot et contexte estimée par la racine carrée (*root*) du *z-score* ; seuls les mots qui partagent au moins 20 (15 + 5) contextes différents avec le mot cible sont retenus.

Notre approche expérimentale est très proche de celle de Lapesa et Evert (2014), qui font varier ces paramètres (et quelques autres) pour comparer les approches distributionnelles sur plusieurs jeux de test (TOEFL, classification automatique de noms et corrélation avec des jugements humains de similarité sémantique) en se limitant toutefois aux seuls contextes construits par cooccurrence graphique. Leurs conclusions sont essentiellement les suivantes : les facteurs principaux qui déterminent la qualité d'un modèle sémantique sont la mesure d'association, la transformation et la mesure de similarité. La meilleure configuration globale utilise le log du rapport de vraisemblance avec un cosinus. La taille optimale de la fenêtre de cooccurrence est de 4 mots.

5. Analyse des résultats

La comparaison des 2 592 modèles permet de dégager les configurations et les paramétrages optimaux. Elle permet également d'identifier leurs caractéristiques discriminantes.

5.1. *Méthode*

Pour comparer ces configurations, nous avons extrait pour chaque modèle et pour chacun des 30 mots cibles la liste des 50 mots les plus proches. Rappelons que dans le *gold standard*, chaque voisin a un score compris entre 1 et 4, qui correspond au nombre d'annotateurs ayant déclaré ce voisin sémantiquement proche de la cible.

La comparaison utilise une mesure synthétique qui prend en compte ce score, en favorisant les modèles pour lesquels les voisins distributionnels les plus proches sont ceux qui ont été validés par le plus grand nombre d'annotateurs. Nous avons utilisé le *Normalised Discounted Cumulative Gain* (ci-après NDCG), une mesure utilisée en recherche d'information pour évaluer les systèmes lorsque les documents ciblés ont un score de pertinence associé (Järvelin et Kekäläinen, 2002). Cette mesure est obtenue en additionnant le score des mots renvoyés par le système et en pénalisant les résultats les plus éloignés dans la liste (le score est divisé par le logarithme du rang de chaque mot). Ce calcul est effectué sur les 50 mots renvoyés par le système comme étant les plus proches de la cible. Les formules sont les suivantes :

$$NDCG = \frac{DCG}{DCGI} \qquad DCG = \sum_{i=1}^{50} \frac{score_i}{log_2(i+1)}$$

où $score_i$ est le nombre d'annotateurs qui ont sélectionné le voisin numéro i renvoyé par le système comme un bon voisin du mot cible (ou 0 si le mot n'a pas été sélectionné) et où $DCGI$ est la valeur maximale de DCG obtenue par un modèle qui renverrait tous les voisins du *gold standard*, dans l'ordre décroissant de pertinence, sans aucun bruit. Cette normalisation donne un score entre 0 et 1 et permet de comparer des mots cibles qui n'ont pas les mêmes nombres de voisins pertinents.

5.2. *Vue d'ensemble*

Dans un premier temps, nous comparons les modèles entre eux, globalement, par catégorie de mot cible et par mot cible, en nous concentrant sur les différences entre les types de contexte.

5.2.1. *Meilleures configurations*

Les configurations qui obtiennent les meilleurs scores globalement et pour chaque type de contexte sont identifiées à partir du NDCG moyen pour chaque modèle sur les 30 mots cibles et pour chaque catégorie grammaticale (moyenne sur 10 mots). Elles sont présentées dans le tableau 5.

	Adjectifs		**Noms**	
Contextes	**Config**	**NDCG**	**Config**	**NDCG**
Graphiques	Graph_1GD_10-zscore-log-10	0,539	Graph_3GD_60-ll-log-65	0,615
Dépendances	Dep_4-z-score-root-4	0,539	Dep_5-MI-none-15	0,584
Syntaxiques	Synt3_4-MI-none-4	**0,558**	Synt3_5-ll-root-5	**0,666**
	Verbes		**Global**	
Contextes	**Config**	**NDCG**	**Config**	**NDCG**
Graphiques	Graph_3GD_30-ll-log-40	**0,554**	Graph_3GD_30-ll-log-30	0,559
Dépendances	Dep_2-ll-log-12	0,490	Dep_4-z-score-none-9	0,504
Syntaxiques	Synt3_2-ll-root-2	0,525	Synt4_4-ll-root-4	**0,561**

Tableau 5 – Meilleures configurations par catégorie syntaxique et par type de contexte

Globalement, *Synt4_4-ll-root-4* est la meilleure configuration sur les 30 mots cibles et les modèles syntaxiques l'emportent pour les noms et les adjectifs. Les modèles par cooccurrence graphique ne l'emportent que pour les verbes. Cependant, un test de Wilcoxon par paires indique que les différences entre les meilleures configurations par contextes graphiques (d'une part) et syntaxiques (d'autre part) ne sont pas significatives (au seuil de 0,05). La seule information concluante à ce stade est que les meilleurs modèles par dépendances brutes arrivent systématiquement derrière les meilleurs modèles des deux autres familles, et ce avec une différence significative ($p < 0{,}05$), sauf pour les adjectifs où le meilleur modèle par dépendances brutes arrive *ex aequo* avec la meilleure configuration par cooccurrence graphique.

Hormis l'infériorité notable des contextes par dépendances brutes, l'étude des meilleures configurations ne permet pas de mesurer précisément l'impact des différents paramètres en jeu. Nous avons donc réalisé une analyse globale des 2 592 modèles sur les 30 mots cibles.

5.2.2. *Variation suivant les catégories des mots cibles*

La figure 2 montre la variation de NDCG pour chaque type de contexte sur l'ensemble des mots et pour chaque catégorie de mot cible. On voit que les variations au sein des familles de configuration sont assez importantes. Si les valeurs maximales sont proches (sauf pour les contextes par dépendances), les valeurs centrales indiquent que les contextes syntaxiques dominent les contextes par dépendances brutes et les contextes graphiques. Les écarts sont plus marqués pour les noms avec des résultats globalement meilleurs que pour les deux autres catégories. Ils sont moins nets pour les adjectifs. Il semble donc au vu de ces scores que les contextes syntaxiques obtiennent globalement les meilleurs résultats.

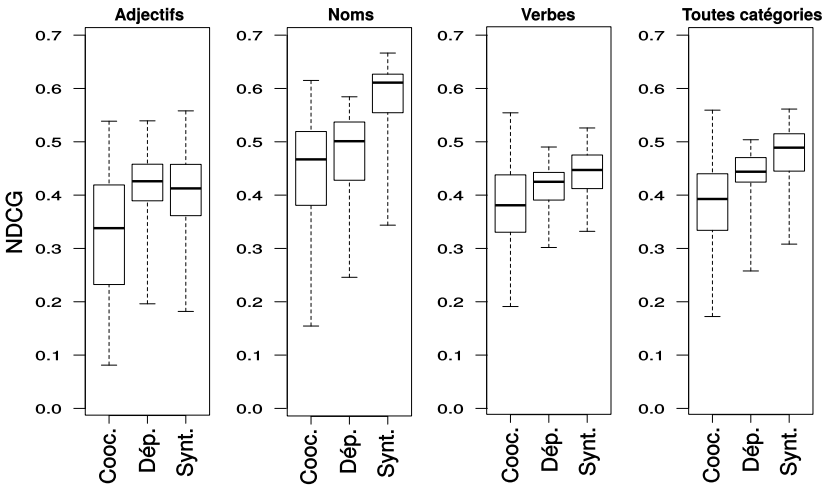

Figure 2 – Variation du NDCG en fonction des catégories et des types de contexte

5.2.3. *Variation suivant les mots cibles*

Les résultats pour chacun des 30 mots cibles font également apparaître d'importantes variations, comme l'illustre la figure 3 qui présente la moyenne sur tous les modèles. Ces variations dépassent les frontières catégorielles, même si les valeurs les plus élevées sont atteintes pour les noms. Malgré l'étendue des boîtes à moustaches, le comportement des différents systèmes est en fait très stable : sur les 30 mots cibles, le coefficient de corrélation moyen entre deux modèles est de 72,2 % (ρ de Spearman). Les différents modèles rencontrent les mêmes difficultés face aux mêmes mots cibles. Signalons que Peirsman *et al.* (2007) ont observé une corrélation similaire, de 70 %, entre approches syntaxiques et graphiques.

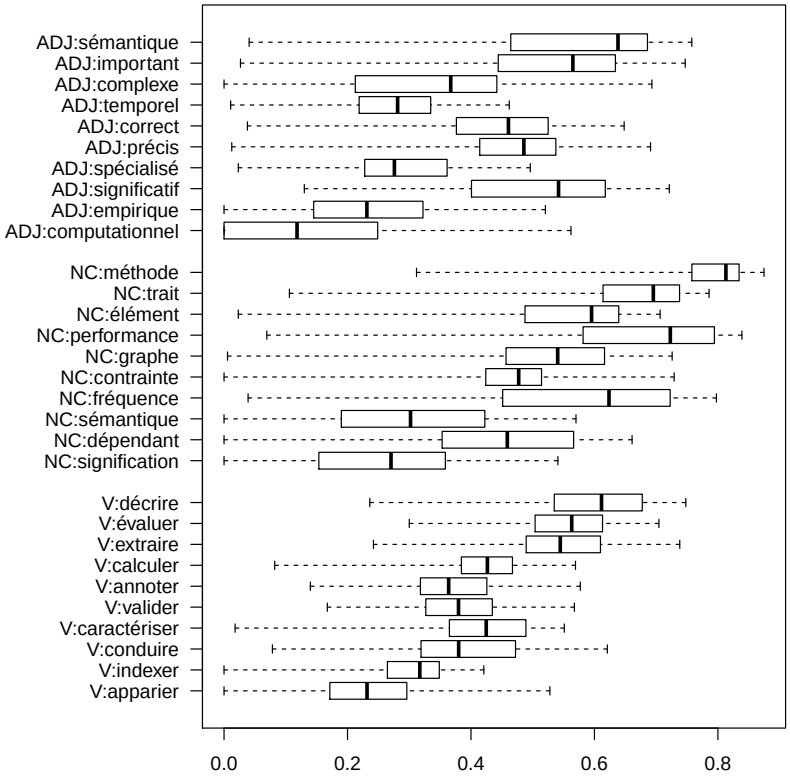

Figure 3 – Variation du NDCG moyen en fonction des mots cibles

D'autre part, il apparaît que les scores varient avec la fréquence du mot cible (dans la figure 3, les mots cibles de chaque catégorie sont rangés par fréquence décroissante), comme on le verra dans la section suivante. Les scores nuls atteints par certains systèmes correspondent à des configurations trop restreintes (notamment en termes de seuils) pour produire le moindre voisin.

5.3. *Étude des paramètres*

L'impact des différents paramètres de chaque type de modèle a été mesuré afin d'identifier ceux qui ont un effet significatif sur les résultats.

5.3.1. *Analyse globale*

Nous avons calculé, à partir des scores de NDCG pour chaque mot cible et pour chaque configuration, une régression linéaire multiple sur les caractéristiques suivantes, en prenant en compte leurs interactions deux à deux : le type de contexte (parmi les trois envisagés), la catégorie du mot cible, la fréquence du mot cible dans le corpus, le nombre de contextes minimal, la mesure d'association, la transformation de la mesure et le nombre minimal de contextes communs.

Le paramètre le plus important, au vu des coefficients de détermination (R^2), est de très loin la fréquence du mot cible, qui explique à elle seule 31 % de la variance globale. De fait, le coefficient de corrélation linéaire entre la fréquence et le NDCG est de 0,6 : les mots les plus fréquents sont ceux pour lesquels les modèles obtiennent les meilleures performances. Le deuxième paramètre par ordre d'importance est la catégorie du mot cible (9 % de la variance expliquée) : les résultats pour les noms sont supérieurs à ceux pour les adjectifs, eux-mêmes meilleurs que pour les verbes. Vient ensuite le type de contexte (5 %) avec les variations vues en figure 2. Les autres paramètres sont négligeables à ce stade ; nous avons opté, comme Lapesa et Evert (2014), pour un seuil arbitraire de 5 % sur le coefficient R^2, les valeurs-p étant toutes infinitésimales lorsque l'on travaille sur de tels effectifs. Le modèle linéaire global a un taux de résidus de 43 % (R^2 ajusté de 57 %) : une grande partie de la variance ne s'explique que par le paramétrage spécifique de chaque type de modèle (fenêtre graphique, type de contexte syntaxique, etc.) et par les variations d'un mot cible à un autre.

Si l'on ne considère pour chaque configuration que la moyenne des scores de NDCG sur les 30 mots cibles (*i.e.* si on masque les spécificités de chaque mot cible), le modèle linéaire atteint une meilleure qualité de représentation de la variation (R^2 ajusté de 72 %) et indique que le facteur principal est le type de contexte (33 %), bien devant les paramètres liés au calcul de similarité. Le premier de ces paramètres (18 %) est l'interaction entre la mesure d'association et la transformation qui lui est appliquée. La figure 4 montre que les scores moyens, calculés sur l'ensemble des configurations et sur les 30 mots cibles, varient lorsqu'une transformation est appliquée à une mesure. On remarque notamment la nécessité d'appliquer une transformation quelle qu'elle soit au rapport de vraisemblance (*simple-ll*) sous peine de voir chuter dramatiquement l'efficacité du calcul de similarité. On note par ailleurs que toute transformation logarithmique est bénéfique, ou au pire sans effet pour l'information mutuelle. Ces observations rejoignent celles de Lapesa et Evert (2014). Le reste des variations d'une combinaison à l'autre n'est pas décisif, comme on l'observe dans le palmarès des meilleures configurations (cf. tableau 5). L'interaction entre ces mesures et la fréquence ou la catégorie du mot cible n'est pas significative.

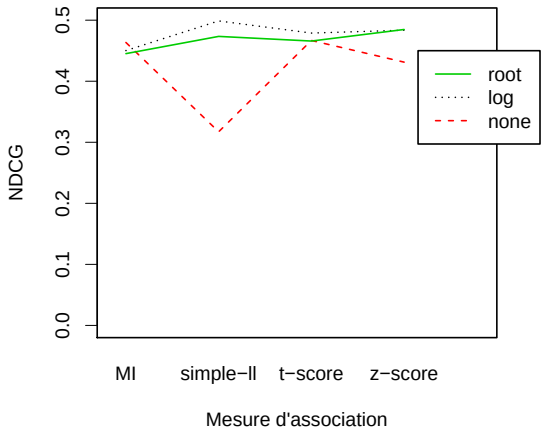

Figure 4 – Variation du NDCG moyen en fonction de la mesure d'association et de la transformation appliquée

Dans ce qui suit, nous étudions plus finement le fonctionnement de chacune des trois familles de contextes pour mesurer l'impact des différentes caractéristiques sur leurs résultats.

5.3.2. *Impact du paramétrage des contextes graphiques*

Les modèles fondés sur les cooccurrents graphiques sont décrits par les paramètres suivants : taille et direction de la fenêtre, seuil sur le nombre de contextes différents, mesure d'association et transformation et seuil sur les contextes partagés.

Un modèle de régression linéaire sur ces paramètres et leurs interactions deux à deux pour prédire le NDCG moyen sur les 30 mots cibles obtient un taux de résidus faible (18 %). Une analyse de variance permet d'identifier les principaux facteurs (*i.e.* expliquant plus de 5 % de la variance) :

1) l'interaction entre la mesure d'association et la transformation (21 %), avec la même tendance qu'en figure 4 ;

2) la direction de la fenêtre (14 %) : une fenêtre bidirectionnelle est globalement meilleure qu'une fenêtre droite ou gauche seulement ;

3) la mesure d'association (8 %), avec une légère préférence pour le *t-score* ;

4) la taille de la fenêtre (6 %), avec une préférence pour les fenêtres de 3 mots.

Si l'on détaille cette analyse en fonction des catégories grammaticales du mot cible, on retrouve globalement les mêmes tendances, avec les différences suivantes :

– pour les adjectifs, la direction de la fenêtre est le facteur critique : une fenêtre unilatérale droite entraîne une baisse importante des performances. Il est clair que, la plupart des adjectifs épithètes étant postposés, négliger leur contexte gauche est une erreur fatale. En revanche, comme la plupart des noms qualifiés sont immédiatement à gauche de l'adjectif, une fenêtre de taille 1 n'est absolument pas pénalisante à condition qu'elle englobe la partie gauche ;

– pour les noms, la taille de la fenêtre est primordiale, les fenêtres de taille 1 étant très pénalisantes. On peut supposer que de telles fenêtres empêchent notamment l'accès aux compléments et aux gouverneurs à cause des éventuelles prépositions ou déterminants ;

– pour les verbes, peu de facteurs discriminants émergent. La direction de la fenêtre est ici aussi importante, avec une préférence pour les contextes droits. Ceci semble donc indiquer que le rapprochement distributionnel des verbes est plus efficace en utilisant leurs objets et leurs compléments indirects plutôt que leurs sujets.

Aucune des analyses ne semble indiquer que les différents seuils sur le nombre minimal de contextes aient un effet significatif sur les résultats. La corrélation est légèrement négative pour le seuil de filtrage initial, et nulle pour celui du nombre de contextes partagés.

5.3.3. *Impact du paramétrage des dépendances brutes*

Les modèles fondés sur les dépendances brutes ont peu de paramètres : le seuil sur le nombre de contextes différents, la mesure d'association et sa transformation, et le seuil pour les contextes partagés. En procédant avec la même méthode, nous observons que le paramètre principal qui explique les variations de performance est l'interaction entre la mesure d'association et la transformation qu'on lui applique. On y retrouve exactement les mêmes tendances que précédemment.

5.3.4. *Impact du paramétrage des contextes syntaxiques*

L'étude des paramètres des contextes syntaxiques est plus complexe. Comme indiqué en section 4.1.3, nous avons choisi de regrouper les différents triplets syntaxiques en quatre configurations de base (tableau 3). Dans un premier temps, nous avons étudié l'impact de la configuration (*Synt1* à *Synt4*) et des paramètres du calcul de similarité en procédant comme pour les deux familles précédentes.

Sur l'ensemble des mots cibles, le modèle linéaire obtenu ne laisse que 3 % de résidus. Ses principaux facteurs sont :

– la configuration (39 %) : *Synt3* et *Synt4* sont largement supérieures aux deux autres, *Synt1* étant de loin la plus faible ;

– l'interaction entre la mesure d'association et la transformation (36 %), avec les mêmes tendances que pour les contextes graphiques ;

– le seuil de filtrage initial des mots en fonction du nombre de contextes (5 %) : la valeur optimale est de 3 ou 4 contextes différents ;

– le seuil de filtrage sur les contextes partagés (5 %) : il semble préférable de ne pas réaliser ce type de filtrage.

On retrouve les mêmes tendances générales quand on observe les détails par catégorie de mot cible. Il semble donc que le choix des triplets syntaxiques constitue le facteur principal. Nous explorons plus en profondeur cet aspect dans ce qui suit.

5.3.5. *Impact des différentes relations syntaxiques*

À partir des 4 configurations principales (*Synt1* à *Synt4*), nous avons produit un nouvel ensemble de 58 configurations dérivées par ajout ou suppression de relations syntaxiques et de normalisations. Pour chacune des configurations de départ, nous avons sélectionné les paramètres de calcul de similarité permettant d'obtenir globalement les meilleurs résultats pour chaque catégorie de mot cible. Ces paramètres étant fixés, nous comparons le score obtenu par la configuration de base et ses variantes.

Nous avons tout d'abord évalué les relations principales en supprimant tour à tour de *Synt1* chacune des 3 relations qui la composent. Dans un deuxième temps, nous ajoutons tour à tour à *Synt1* différentes relations syntaxiques ou normalisations. Nous avons ensuite testé sur *Synt2* plusieurs normalisations de relations syntaxiques absentes de *Synt1*. Nous ne détaillons pas ci-dessous les variantes testées sur *Synt3* et *Synt4* car elles ne produisent que des différences mineures : le nombre de relations présentes dans ces configurations rend l'ajout ou la suppression de l'une d'entre elles imperceptible. Les résultats obtenus sont les suivants :

– *nMod :* la relation *modifieur de nom* est essentielle aux adjectifs. La supprimer de *Synt1* rend impossible le calcul de voisinage pour cette catégorie car c'est la seule relation à laquelle ils participent dans cette configuration. Sa suppression a également un impact significativement négatif pour les noms (− 11 %)[5] ;

– *obj :* la relation *objet* est essentielle aux verbes (− 13 %). Sa suppression fait aussi baisser les noms de 4 %, mais cette différence n'est pas significative ;

– *suj :* la contribution de la relation *sujet* n'est significative ni pour les noms, ni pour les verbes. Notons que la pertinence de la relation *objet* et la faible qualité de la relation *sujet* pour les noms avaient déjà été observées sur le néerlandais par Peirsman *et al.* (2007) et Heylen *et al.* (2008). Elles rejoignent les observations de Fabre (2010, p. 54), et nos remarques sur l'importance d'une fenêtre à droite pour les cooccurrents graphiques des verbes (cf. section 5.3.2) ;

– *advMod :* la relation *modification adverbiale* est bénéfique aux adjectifs (+ 5 % lorsqu'elle est ajoutée), sans que ce résultat soit significatif. Cette relation fait très légèrement chuter le score pour les noms. Son ajout à *Synt2* améliore significativement les résultats pour les adjectifs (+ 4,2 %) et donne la meilleure configuration pour cette catégorie syntaxique ;

– *inclNPP :* la prise en compte des noms propres a un impact très légèrement négatif sur l'ensemble des configurations que nous avons testées ;

5. Nous utilisons le test de Wilcoxon par paires au seuil de 0,05.

– *prep :* ajouter la relation *préposition* à *Synt1* améliore significativement les résultats pour les noms (+ 5,7 %) et les verbes (+ 7,1 %). Le fait d'ignorer la préposition qui lie un pivot et son contexte (*fusionPrep*) n'a pas d'effet significatif ;

– *coord :* le repérage des mots coordonnés améliore également les résultats pour toutes les catégories syntaxiques, mais la différence n'est jamais significative. Concernant les normalisations opérées sur cette relation, la fermeture transitive et la distribution des relations sur les coordonnés dégradent légèrement les résultats. Cette observation rejoint une nouvelle fois celles de Peirsman *et al.* (2007) et Heylen *et al.* (2008) qui mentionnent le traitement problématique de la coordination, notamment dans le cas des énumérations ou de la coordination à longue distance ;

– *autres relations et normalisations :* l'ajout de l'attribut du sujet (*ats*) à *Synt1* n'est pas significatif. Sa transformation en modifieur de nom (*ats→nMod*) dans *Synt2* dégrade significativement les résultats pour les adjectifs (− 2 %). Aucune des autres opérations de normalisation testées (recherche de l'antécédent des pronoms relatifs, normalisation des passifs, ajout du sujet des participes présents), prise séparément, n'est probante.

Ainsi les enseignements que l'on peut tirer de ces expérimentations, pour le corpus et le jeu d'évaluation utilisés ici, sont qu'il est essentiel de sélectionner un noyau de relations pertinentes pour chaque catégorie syntaxique particulière :

– pour les adjectifs, modifieur de nom et modification adverbiale ;

– pour les noms, modifieur de nom et préposition ;

– pour les verbes, objet direct et préposition.

Les traitements plus fins et plus complexes n'apportent aucun gain substantiel.

5.4. *Bilan des observations*

Si l'on résume les observations effectuées sur ces données, il semblerait que les contextes syntaxiques dépassent les deux autres types de configurations. Certes, il est toujours possible d'obtenir un niveau équivalent avec une méthode graphique en utilisant un paramétrage optimal, mais à défaut d'une telle optimisation on peut voir (spécialement en figure 2) que les méthodes syntaxiques atteignent globalement des scores supérieurs.

Au sein des différentes possibilités offertes par l'analyse syntaxique, on voit également clairement qu'une utilisation directe des dépendances brutes, comme elle a été faite dans plusieurs études, notamment (Kiela et Clark, 2014), n'offre que peu d'avantages par rapport aux cooccurrents graphiques et produit des résultats inférieurs à ceux obtenus en sélectionnant les relations de dépendance à prendre en compte.

Parmi le grand nombre de choix possibles pour configurer un modèle distributionnel fondé sur les contextes syntaxiques, il apparaît que le facteur principal est bien la nature de ces contextes. Nous pouvons dégager au vu des modèles examinés ici un

noyau minimal constitué des relations et des normalisations correspondant au niveau *Synt3*. Les transformations plus sophistiquées et les normalisations complexes au-delà de ce noyau n'apportent en définitive qu'un gain très faible, rarement mesurable. Nous avons également montré que ces paramètres pouvaient varier en efficacité en fonction de la catégorie du mot cible.

Si les modèles à base de contextes graphiques sont moins paramétrables, nous avons tout de même pu mettre en évidence la sensibilité à la géométrie de la fenêtre utilisée, notamment pour certaines catégories de mots : importance du contexte gauche des adjectifs, besoin d'élargir la fenêtre pour les noms et préférence pour une fenêtre droite pour les verbes.

En ce qui concerne la « mécanique interne » de la méthode distributionnelle, le paramètre principal concerne l'utilisation des mesures d'association et leur transformation, mais la variation semble essentiellement due à des configurations déficientes (notamment *simple-ll* brut), et les conclusions à cet égard sont les mêmes quelle que soit la famille de contextes utilisée.

Nous avons enfin confirmé une forte corrélation entre tous les modèles qui ont un même comportement face aux pivots, notamment une facilité pour traiter les mots fréquents et les noms.

5.5. *Analyses qualitatives*

Dans cette dernière partie, nous nous intéressons plus en détail à la nature des voisins distributionnels identifiés par ces méthodes, afin de voir si ces faibles différences en termes de score traduisent des différences qualitatives importantes. Pour ce faire, nous avons tout d'abord sélectionné un sous-ensemble de modèles, afin de réduire les coûts du calcul. Le jeu que nous avons examiné a été construit de la façon suivante : à partir des 2 592 modèles initiaux et de ceux qui ont été définis pour tester les variations dans les triplets syntaxiques comme décrit en 4.1.3, soit un total de 20 880 modèles, nous avons sélectionné ceux qui apparaissent comme les meilleurs selon un des critères suivants : le meilleur modèle global, le meilleur modèle pour une catégorie de mot cible, le meilleur modèle pour un mot cible, le meilleur modèle d'une famille de contextes (graphiques, dépendances, syntaxiques) pour une catégorie et le meilleur modèle d'une famille pour un mot cible.

Étant donné les recouvrements, 106 modèles différents ont ainsi été sélectionnés. Ce nombre réduit de configurations nous a permis d'observer plus en détail leur comportement. Tous ces modèles obtiennent des scores NDCG relativement élevés pour chaque mot, et ont sur cette base une très forte corrélation (ρ de Spearman de 0,76). Autrement dit, à l'échelle de notre *gold standard*, ils ont des comportements très proches.

Nous nous sommes donc intéressés aux résultats de chacun de ces modèles, indépendamment du *gold standard*, en nous fondant sur les voisins distributionnels

eux-mêmes. Pour ce faire, nous avons utilisé la mesure *Rank-Biased Overlap* (ci-après RBO) qui permet de comparer deux listes ordonnées quelconques (Webber *et al.*, 2010). Cette mesure a notamment été utilisée pour comparer les résultats de deux moteurs de recherche face à une collection ouverte (comme le Web). Elle est définie comme suit :

$$RBO(A,B) = \frac{1-p}{p} \sum_{d=1}^{50} p^d \frac{|A_{1:d} \cap B_{1:d}|}{d}$$

où $A_{1:d}$ est l'ensemble des d premiers voisins de A et p est le biais qui pénalise les recouvrements dans les rangs inférieurs. Nous avons utilisé ici la valeur recommandée de p = 0,98. Intuitivement, cette mesure calcule à chaque rang (de 1 à 50) le recouvrement entre les deux listes, en le pondérant par rapport au rang à la manière de ce qui est fait pour NDCG. La valeur obtenue en seuillant ces recouvrements pondérés est ensuite normalisée pour obtenir une valeur entre 0 (différence totale, donc intersection vide entre les deux listes jusqu'au rang 50) et 1 (listes identiques jusqu'au rang 50). Nous avons calculé cette mesure de similarité entre chacun de nos 106 modèles en faisant la moyenne de leur RBO au rang 50 pour les 30 mots cibles.

Sur la base de la matrice de similarité ainsi obtenue, les différentes familles de modèles se répartissent comme illustré en figure 5, où nous avons utilisé une projection de Sammon (Sammon, 1969) pour calculer un espace de dimension 2 dans lequel la distance entre les vecteurs est la plus proche de celle qu'ils ont dans l'espace initial.

Même si les deux axes de cette représentation n'ont aucune signification particulière, on voit clairement que les modèles syntaxiques sont assez nettement séparés des modèles graphiques. Les contextes par dépendances brutes forment une catégorie intermédiaire entre les deux autres, mais plus proches des modèles par cooccurrence graphique. Une classification hiérarchique ascendante sur cette même mesure de similarité (non représentée ici) montre qu'à de rares exceptions près, les deux principaux types de modèles sont bien séparables.

Autrement dit, même si les modèles obtiennent des scores très similaires, ils produisent des voisinages différents : ils sélectionnent des voisins différents ou ils les ordonnent de façon distincte. Il devrait donc être possible d'identifier, s'ils existent, les voisins de chaque mot cible qui sont préférentiellement renvoyés par chaque type de méthode.

6. Conclusion

Cet article présente un ensemble de modèles distributionnels construits sur un corpus spécialisé et évalués sur un jeu de données conçu spécialement pour cette étude. Nous avons testé un ensemble de modèles, comparé leurs performances et étudié les différents facteurs qui les caractérisent, à savoir les types de contexte et les paramètres qui interviennent dans le calcul de similarité. Les performances ont été évaluées de manière globale, par catégorie syntaxique des mots cibles et par mot cible.

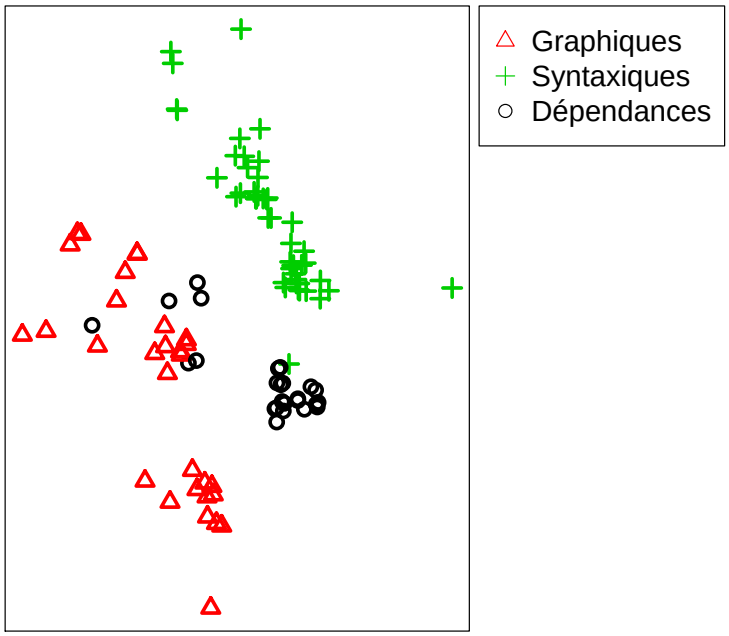

Figure 5 – Projection de Sammon des 106 meilleurs modèles sur la base de leur similarité RBO

Les modèles construits à partir de contextes qui utilisent des dépendances syntaxiques brutes obtiennent systématiquement des résultats inférieurs à ceux qui font un usage plus fin des informations syntaxiques. À l'inverse, les transformations et normalisations syntaxiques complexes atteignent assez rapidement un plafond, et nous avons ainsi pu dégager un noyau efficace de traitements pour exploiter les sorties d'un analyseur syntaxique.

Si les meilleurs modèles par cooccurrence graphique peuvent rivaliser avec les modèles syntaxiques, leur paramétrage nécessite un soin particulier, susceptible de varier d'une catégorie de mot cible à une autre. Nous avons néanmoins dégagé des tendances générales pour les combinaisons de paramètres, notamment pour le calcul de l'association entre mots et contextes. Enfin, les sorties de ces modèles sont qualitativement différentes de celles des modèles syntaxiques.

La supériorité des modèles syntaxiques non triviaux pourrait être en partie expliquée par la taille réduite du corpus, les modèles pauvres en information ayant besoin d'un volume de texte plus important. Ceci semble confirmé par une étude préliminaire (non présentée ici) où nous avons utilisé des sous-ensembles de notre corpus et observé que les écarts se creusaient entre ces méthodes, toujours au profit des modèles syntaxiques. Mais nous ne pouvons prétendre à ce stade à la généralisation des résultats obtenus ici sur un corpus particulier et avec un jeu d'évaluation *ad hoc*.

Il nous semble cependant important d'insister sur le fait que ces familles de méthodes produisent des voisins différents même si les différences sont marginales pour les scores globaux. Cet aspect, ainsi que les variations importantes repérées en fonction de la catégorie et de la fréquence du mot cible nous amènent à envisager pour la suite des études plus qualitatives, facilitées dans notre cas par notre connaissance du corpus et de son contenu terminologique.

Remerciements

Nous tenons à remercier Cécile Fabre et Lydia-Mai Ho-Dac pour leur travail dans la conception du jeu d'évaluation et pour l'ensemble des interactions que nous avons eues au cours de ce travail. Nous remercions également Florian Boudin pour la constitution du corpus TALN et l'ATALA pour en avoir autorisé l'usage.

7. Bibliographie

Almuhareb A., Poesio M., « Attribute-Based and Value-Based Clustering: An Evaluation. », *Proceedings of EMNLP*, p. 158-165, 2004.

Baroni M., Lenci A., « Distributional Memory: A General Framework for Corpus-Based Semantics », *Computational Linguistics*, vol. 36, n° 4, p. 673-721, 2010.

Bernier-Colborne G., « Analyse distributionnelle de corpus spécialisés pour l'identification de relations lexico-sémantiques », *Actes de l'atelier SemDis à TALN'2014*, Marseille, p. 238-251, 2014.

Boudin F., « TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue », *Actes de TALN'2013*, Les Sables d'Olonne, p. 507-514, 2013.

Bullinaria J. A., Levy J. P., « Extracting Semantic Representations from Word Co-occurrence statistics: A Computational Study », *Behavior Research Methods*, vol. 39, n° 3, p. 510-526, 2007.

Bullinaria J. A., Levy J. P., « Extracting Semantic Representations from Word Co-occurrence Statistics: stop-lists, stemming, and SVD », *Behavior Research Methods*, vol. 44, n° 3, p. 890-907, 2012.

Cohen T., Widdows D., « Empirical Distributional Semantics: Methods and Biomedical Applications », *Journal of biomedical informatics*, vol. 42, n° 2, p. 390–405, 2009.

Curran J. R., Moens M., « Improvements in Automatic Thesaurus Extraction », *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, p. 59-66, 2002.

Evert S., « Corpora and Collocations », *in* A. Lüdeling, M. Kytö (eds), *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, chapter 58, 2007.

Evert S., « Distributional Semantics in R with the Wordspace Package », *Proceedings of COLING 2014, System Demonstrations*, Dublin, p. 110-114, 2014.

Fabre C., Affinités syntaxiques et sémantiques entre les mots : Apports mutuels de la linguistique et du TAL, Habilitation à diriger des recherches, Université de Toulouse, 2010.

Fabre C., Hathout N., Ho-Dac L.-M., Morlane-Hondère F., Muller P., Sajous F., Tanguy L., Van de Cruys T., « Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés », *Actes de l'atelier SemDis, TALN'2014*, Marseille, p. 196-205, 2014a.

Fabre C., Hathout N., Sajous F., Tanguy L., « Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille », *Actes de l'atelier SemDis, TALN'2014*, Marseille, p. 266-279, 2014b.

Ferret O., « Testing Semantic Similarity Measures for Extracting Synonyms from a Corpus », *Proceedings of LREC'10*, Malta, p. 3338–3343, 2010.

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G., Ruppin E., « Placing Search in Context: The Concept Revisited », *ACM Transactions on Information Systems*, vol. 20, n° 1, p. 116-131, 2002.

Firth J. R., « Modes of Meaning », *Papers in linguistics 1934-1951 (1957)*, Oxford University Press, 1951.

Fyshe A., Talukdar P. P., Murphy B., Mitchell T. M., « Interpretable Semantic Vectors from a Joint Model of Brain- and Text- Based Meaning », *Proceedings of ACL 2014*, Baltimore, Maryland, p. 489-499, 2014.

Gamallo Otero P., « Comparing window and syntax based strategies for semantic extraction », *Computational Processing of the Portuguese Language. PROPOR 2008*, vol. 5190 of *LNAI*, Springer, p. 41-50, 2008.

Grefenstette G., « Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches », *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, Columbus, Ohio, 1993.

Habert B., Zweigenbaum P., « Contextual Aquisition 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*, vol. 2: computability of language and computer applications, John Benjamins Publishing Company, Amsterdam / Philadelphia, chapter 8, p. 203-231, 2002.

Harris Z., « Distributional Structure », *Word*, vol. 10, n° 2-3, p. 146-162, 1954. Traduction française dans *Langages* (20) 1970.

Harris Z. S., Gottfried M., Ryckman T., Mattick P., Daladier A., Harris T. N., Harris S., *The form of information in science: analysis of an immunology sublanguage*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.

Heylen K., Peirsman Y., Geeraerts D., Speelman D., « Modelling Word Similarity: an Evaluation of Automatic Synonymy Extraction Algorithms. », *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may, 2008.

Järvelin K., Kekäläinen J., « Cumulated Gain-Based evaluation of IR techniques », *ACM Transactions on Information Systems (TOIS)*, vol. 20, n° 4, p. 422-446, 2002.

Kiela D., Clark S., « A Systematic Study of Semantic Vector Space Model Parameters », *Proceedings of the 2nd CVSC Workshop*, p. 21-30, 2014.

Lapesa G., Evert S., « A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection », *Transactions of the ACL*, vol. 2, p. 531-545, 2014.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient Estimation of Word Representations in Vector Space », *CoRR*, 2013.

Miller G. A., Charles W. G., « Contextual correlates of semantic similarity », *Language and cognitive processes*, vol. 6, n° 1, p. 1-28, 1991.

Padó S., Lapata M., « Dependency-based Construction of Semantic Space Models », *Computational Linguistics*, vol. 33, n° 2, p. 161-199, 2007.

Peirsman Y., Heylen K., Speelman D., « Finding Semantically Related Words in Dutch. Co-occurrences versus Syntactic Contexts », *Proceedings of the CoSMO workshop*, Roskilde, Danemark, p. 9-16, 2007.

Rubenstein H., Goodenough J. B., « Contextual Correlates of Synonymy », *Communications of the ACM*, vol. 8, n° 10, p. 627-633, 1965.

Sahlgren M., The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces, Phd thesis, Stockholm University, 2006.

Sammon J. W., « A nonlinear mapping for data structure analysis », *IEEE Transactions on Computers*, vol. 18, p. 401–409, 1969.

Tutin A., « Autour du lexique et de la phraséologie des écrits scientifiques », *Revue Française de Linguistique Appliquée Lexique et écrits scientifiques*, vol. XII, n° 2, p. 5-14, 2007.

Urieli A., Tanguy L., « L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane », *Actes de TALN'2013*, Les Sables d'Olonne, p. 188-201, 2013.

Van de Cruys T., Apidianaki M., « Latent Semantic Word Sense Induction and Disambiguation », *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, p. 1476-1485, 2011.

Van der Plas L., Bouma G., « Syntactic Contexts for Finding Semantically Related Words », *in* T. Van der Wouden, M. Poß, H. Reckman, C. Cremers (eds), *Computational Linguistics in the Netherlands 2004: Selected papers from the fifteenth CLIN meeting*, vol. 4 of *LOT Occasional Series*, Utrecht University, 2005.

Webber W., Moffat A., Zobel J., « A Similarity Measure for Indefinite Rankings », *ACM Transactions on Information Systems*, vol. 28, n° 4, p. 20, 2010.

Zesch T., Gurevych I., « Automatically creating datasets for measures of semantic relatedness », *COLING/ACL 2006 Workshop on Linguistic Distances*, Sydney, Australia, p. 16-24, 2006.