Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille

Cécile Fabre, Nabil Hathout, Franck Sajous, Ludovic Tanguy

Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille

Cécile Fabre Nabil Hathout Franck Sajous Ludovic Tanguy
CLLE/ERSS, CNRS & Université de Toulouse

Résumé. L'analyse distributionnelle sur des corpus spécialisés de taille modeste constitue un objectif applicatif important pour cette famille de méthodes d'extraction des relations sémantiques. Dans ce cadre, nous cherchons à optimiser le calcul distributionnel pour traiter un corpus de 2 millions de mots composé d'articles de la conférence TALN. Notre expertise dans ce champ nous permet de constituer des données d'évaluation adaptées au corpus et à la tâche, et fait de cette configuration expérimentale un lieu idéal pour observer précisément les mécanismes distributionnels à l'œuvre. Un paramétrage précis du calcul distributionnel, depuis l'analyse syntaxique jusqu'aux mesures de proximité sémantique, met en évidence la variété des résultats obtenus, particulièrement selon les catégories grammaticales des mots cibles, et permet de dégager des combinaisons performantes en jouant sur le nombre, la nature et la qualité des contextes pris en compte dans le calcul.

Abstract. Applying distributional semantic models to medium-size specialized corpora is an important objective for the extraction of lexical and terminological ressources. In this context, we seek to optimize the distributional analysis procedure on a 2 million word corpus consisting of NLP conference proceedings. Our expertise in this field allows us to establish a relevant benchmark for the task, thus providing an ideal experimental setup to observe the distributional mechanisms at work. We test several hundred configurations, with parameters ranging from syntactic analysis to similarity measures. This study highlights the variety of the results, particularly according to the POS of the target words, and allows for the identification of the best performing configurations by varying the number, nature and type of the contexts considered.

Mots-clés : Sémantique distributionnelle, analyse syntaxique, corpus spécialisé, évaluation.

Keywords: Distributional semantics, syntactic analysis, specialized corpus, evaluation.

1 Introduction

Les programmes des grandes conférences sur le traitement automatique des langues (TAL) témoignent d'une montée en puissance des recherches sur l'analyse distributionnelle (AD), qui tend à s'imposer comme un mode de représentation et d'exploitation incontournable dans les travaux sur le lexique et la sémantique lexicale (Baroni & Lenci, 2010; Turney & Pantel, 2010). Cet article s'inscrit dans cette lignée, et propose une réponse possible à la tâche exploratoire de l'atelier SemDis 2014, organisé dans le cadre de la conférence TALN, qui sollicite la mise en œuvre de méthodes d'analyse distributionnelle sur un corpus spécialisé d'environ 2 millions de mots, composé d'articles publiés dans les actes des conférences TALN et RÉCITAL.

Cette tâche comporte potentiellement plusieurs difficultés. L'une d'elle est liée à la taille modeste du corpus TALN en comparaison de ceux qui sont habituellement utilisés pour la construction de modèles distributionnels. Par exemple, Baroni & Lenci (2010) ont construit la base DM à partir de ukWaC, un corpus de 2 milliards de mots ; Ferret (2010) considère pour sa part qu'avec 380 millions de mots, AQUAINT 2 est un corpus de taille moyenne. En comparaison, la taille du corpus TALN est de 2 à 3 ordres de magnitude inférieure. L'expérience que nous retraçons ici montre qu'il est possible d'appliquer sur un petit corpus spécialisé les méthodes et les outils généralement destinés à traiter des corpus beaucoup plus volumineux. Nous nous situons de ce fait à mi-chemin entre, d'une part, les méthodes d'extraction et de structuration de terminologie, qui opèrent sur de petits corpus spécialisés et visent la mise au jour de relations conceptuelles spécifiques, et, d'autre part, l'analyse distributionnelle qui traite des corpus de tous types, généralement volumineux, et identifie des relations de proximité sémantique au sens large. Cette démarche amorce en quelque sorte un retour aux sources de l'analyse distributionnelle harrissienne : nous appliquons la méthode distributionnelle à un corpus spécialisé de petite taille sur lequel nous réalisons un ensemble de traitements linguistiques permettant de normaliser les variations dans l'expression

des «dépendances» entre les mots afin de mieux capter les régularités sémantiques qui y sont présentes.

L'application de la démarche distributionnelle au corpus TALN présente par ailleurs plusieurs avantages : nous connaissons parfaitement le domaine et nous trouvons dans des conditions optimales pour l'évaluation et l'analyse des résultats. La taille réduite du corpus nous permet aussi d'étudier plus en détails le comportement de certains des mots-cibles.

L'originalité de ce travail réside d'une part dans l'approche «pragmatique» adoptée pour constituer un référentiel sur lequel les modèles distributionnels sont évalués. Il est en effet essentiel de prendre acte de la diversité des relations sémantiques qui sous-tendent la similarité sémantique (Baroni & Lenci, 2011; Morlane-Hondère & Fabre, 2012) et de la considérer pour ce qu'elle est sans chercher à y retrouver l'inventaire des relations lexicales classiques. Une autre particularité de notre travail concerne l'attention que nous portons à la mise au point des paramètres situés en amont du calcul du similarité. En particulier, nous nous intéressons au traitement et au filtrage des sorties de l'analyseur syntaxique que nous utilisons. Notre objectif est en effet de mieux contrôler les conditions d'utilisation des contextes linguistiques, en jouant sur leur nombre, leur nature et leur fiabilité. De ce fait, si nous avons choisi de recourir à un corpus analysé syntaxiquement, plutôt qu'à un calcul de cooccurrences, c'est pour disposer ainsi d'une meilleure capacité à injecter des connaissances linguistiques dans cette phase du traitement et pour en mesurer l'impact dans les résulats du calcul distributionnel.

Plus généralement, notre effort a porté sur les paramètres de construction des modèles distributionnels. Nous en avons identifié cinq que nous avons testés de manière systématique pour trouver les meilleures configurations. Ces paramètres concernent à la fois l'utilisation des analyses syntaxiques, le filtrage des mots et des relations, leur normalisation et les calculs de similarité. Cette démarche quantitative est complétée par une analyse qualitative des modèles obtenus.

La suite de l'article est organisée comme suit. Nous présentons en section 2 les voisinages de référence que nous avons constitués pour l'évaluation des modèles distributionnels. Nous abordons ensuite en section 3 la construction de ces modèles et les différents paramètres que nous avons testés. Les résultats de ces expériences, leur évaluation et leur analyse sont l'objet de la section 4. Nous présentons enfin en section 5 une courte conclusion et quelques pistes pour des recherches futures.

2 Annotation pour l'évaluation

Le corpus que nous avons traité dans cette expérience est le corpus TALN (Boudin, 2013), fourni dans le cadre de la tâche exploratoire de l'atelier SemDis2014. Ce corpus comprend 586 articles des conférences TALN et RECITAL de 2007 à 2013. Construit par extraction du contenu textuel des articles initialement au format PDF, il compte deux millions de mots [1].

Nous avons constitué un jeu de données pour l'évaluation du programme d'analyse distributionnelle. On sait que l'évaluation des systèmes d'AD fait difficulté, car leurs résultats sont généralement confontés à des ressources externes (réseaux lexicaux et thésaurus) ou à des tâches (de jugement de synonymie, d'analogie, de proximité sémantique) qui ne permettent d'évaluer que partiellement et indirectement leur qualité, comme le rappellent (Baroni & Lenci, 2011). Le choix du corpus TALN nous a permis de nous affranchir de la nécessité de recourir à des ressources externes, et d'exercer plus directement notre jugement sémantique pour évaluer la qualité des rapprochements sémantiques effectués sur des notions qui relèvent de notre domaine d'expertise.

Nous avons constitué une liste de 15 mots-cibles, en prenant pour point de départ les mots proposés dans le descriptif de la tâche, soit 1 verbe (*calculer*), 2 adjectifs (*précis*, *complexe*) et 5 noms (*fréquence*, *graphe*, *méthode*, *sémantique*, *trait*), que nous avons complétés pour obtenir un ensemble de mots de même effectif selon les 3 catégories, soit 4 verbes et 3 adjectifs de fréquence comparable à celle des mots de la liste initiale. Il s'agit de mots courants dans le corpus : la fréquence moyenne est de 628 occurrences, le mot le moins fréquent, *spécialisé*, a 210 occurrences dans le corpus. La liste de mots-cibles est présentée dans la première colonne du tableau 1.

Nous avons ensuite constitué la liste des meilleurs voisins de chacun de ces 15 mots-cibles. Sur le principe de la *pooling method* pratiquée pour l'évaluation des systèmes de recherche d'information, cette liste a été établie à partir de l'examen d'un sous-ensemble des mots du corpus, correspondant à l'ensemble maximal des voisins distributionnels produits par la méthode que nous présentons dans la section suivante : nous avons réglé au plus bas tous les seuils que nous faisons varier dans l'expérience, de manière à produire la liste la plus large de voisins. Chaque annotateur (chacun des 4 auteurs

1. Le corpus TALN au format texte est disponible à l'adresse : `http://redac.univ-tlse2.fr/corpus/taln/`

Mot-cible	**Accord**	**Exemples** (nb annotateurs)
	adjectifs	
complexe	0,58	*compliqué* (4), *composé* (3), *simple*
correct	0,55	*bon* (4), *pertinent* (4), *valide* (4)
important	0,65	*grand* (4), *majeur* (4), *principal* (4)
précis	0,72	*détaillé* (4), *exhaustif* (4), *fin* (3)
spécialisé	0,73	*juridique* (4), *médical* (4), *spécifique* (3)
	noms	
fréquence	0,58	*nombre* (4), *poids* (4), *probabilité* (4)
graphe	0,55	*réseau* (4), *structure* (4), *treillis* (4)
méthode	0,75	*algorithme* (4), *approche* (4), *procédure* (3)
trait	0,57	*attribut* (4), *caractéristique* (3), *propriété* (3)
sémantique	0,40	*définition* (4), *contenu* (3), *sens* (3)
	verbes	
annoter	0,50	*classer* (4), *étiqueter* (4), *baliser* (3)
calculer	0,47	*construire* (4), *estimer* (4), *évaluer* (4)
décrire	0,57	*détailler* (4), *présenter* (4), *représenter* (4)
évaluer	0,65	*mesurer* (4), *tester* (4), *valider* (4)
extraire	0,58	*acquérir* (4), *identifier* (3), *sélectionner* (3)

TABLE 1: F-mesure de l'accord inter-annotateurs par mot-cible et exemples de voisins sélectionnés

de cet article) avait pour tâche de sélectionner, parmi la liste des voisins distributionnels de chacun des 15 mots-cibles, 10 mots qu'il considérait comme les plus proches sémantiquement de la cible. Nous avons ensuite fait l'union des propositions des 4 annotateurs, en conservant l'information relative au nombre d'annotateurs ayant choisi le mot. La tâche présentait potentiellement deux difficultés susceptibles d'affecter le taux d'accord. Tout d'abord, la consigne était large et ne réduisait pas le jugement de proximité sémantique au repérage de relations lexicales spécifiques, telles la synonymie ou l'hyperonymie. Par ailleurs, la contrainte visant à constituer un ensemble contenant précisément 10 mots pour chaque cible était forte, dans la mesure où elle amenait à exclure certains voisins pertinents, ou à l'inverse (comme pour les mots *sémantique* ou *spécialisé* qui ont peu de très bons voisins) à conserver des mots qui présentaient une proximité sémantique plus faible avec la cible. Malgré ces caractéristiques de la tâche, nous obtenons un score de F-mesure moyen de 0,59. Le tableau 1 montre les variations de ce score selon les mots, l'accord maximum étant obtenu pour certains adjectifs, alors qu'un mot comme *sémantique* donne lieu à un éparpillement plus important des réponses.

Le tableau 1 fournit également pour chaque mot-cible 3 exemples de voisins souvent sélectionnés. On peut constater qu'il s'agit majoritairement de synonymes, mais on trouve aussi des termes plus génériques ou plus spécifiques (*graphe* / *structure*, *juridique* / *spécialisé*), des antonymes (*complexe* / *simple*), ou des voisins correspondant à une relation sémantique plus lâche (*sémantique* / *contenu*). Ces exemples illustrent bien la spécificité des notions employées dans les textes, et par conséquent celle des relations de sens qui s'établissent entre elles. Ainsi, les relations de proximité sémantique entre *fréquence* et *poids*, ou entre *extraire* et *identifier* sont évidentes pour les 4 annotateurs dans le champ considéré, mais elles perdraient de leur pertinence si l'on considérait un autre domaine conceptuel.

3 Méthode

L'analyse distributionnelle consiste à établir une relation de proximité sémantique entre des unités qui apparaissent fréquemment dans les mêmes contextes. Les méthodes d'analyse automatique diffèrent essentiellement par ce que l'on entend par « contexte » et par la manière de mesurer la similitude des contextes d'apparition des unités considérées.

On distingue les approches qui consistent à représenter chaque occurrence d'un mot en corpus par ses cooccurrents graphiques dans une fenêtre donnée, et celles qui représentent le contexte de chaque mot par l'ensemble de ses cooccurrents syntaxiques. La première approche est relativement simple à mettre en œuvre et la seconde, conditionnée par la disponibilité d'un analyseur syntaxique, se révèle plus coûteuse en temps de calcul. Même s'il serait intéressant d'évaluer leurs performances respectives dans le cadre de cette expérience, nous avons fait le choix de ne pas intégrer ce paramètre dans notre étude, en optant pour une méthode basée sur des contextes syntaxiques. Le corpus TALN est traité avec Talismane

(Urieli & Tanguy, 2013), qui produit une analyse syntaxique en dépendances [2]. Ce choix est avant tout motivé par le fait que la phase d'analyse syntaxique nous offre une meilleure marge de manœuvre pour spécifier les caractéristiques linguistiques des contextes qui entrent dans le calcul.

Nous décrivons dans les sections 3.1 à 3.5 les différentes étapes du processus d'analyse distributionnelle mis en œuvre dans cette étude en détaillant pour chacune d'elles les facteurs qui entrent en jeu, puis récapitulons en section 3.6 la liste de ces paramètres. La combinaison des différentes valeurs de ces paramètres donne lieu à 720 configurations que nous évaluons en section 4.

3.1 Extraction de triplets syntaxiques

L'analyseur Talismane produit des sorties au format CoNLL. Il indique notamment pour chaque token son lemme, sa catégorie syntaxique, son gouverneur, et le type de relation qui lie ce dernier au token considéré (son dépendant). La première étape de l'analyse distributionnelle consiste, à partir des relations de dépendance, à extraire des triplets syntaxiques de la forme `<gouverneur; relation; dépendant>`. Les relations considérées en première instance sont les suivantes :

– relations sujet (*suj*) ou objet (*obj*) entre un nom (dépendant) et un verbe (gouverneur) ;
– relation modifieur (*mod*) entre un adjectif ou un nom (dépendant) et un autre nom (gouverneur) ;
– relation attribut du sujet (*ats*) entre un adjectif (dépendant) et un nom (gouverneur) ;
– relation préposition (*prép*) dans les constructions de type N-prép-N (ex : *corpus d'apprentissage*), N-prép-V (ex : *phrase à traduire*), V-prép-N (ex : *reposer sur une hypothèse*) et V-prép-V (ex : *choisir d'utiliser*).

Les relations *suj*, *obj* et *mod* correspondent à des dépendances directement fournies par Talismane. Il faut en revanche suivre deux dépendances pour établir les relations *ats* et *prép* en cheminant respectivement par le verbe attributif et la préposition. Cette étape d'extraction fait intervenir un premier paramètre : le seuil sur le score de confiance des dépendances syntaxiques. En tant qu'analyseur probabiliste, Talismane peut produire le score de probabilité de chaque décision prise, et donner ainsi une indication de la confiance à accorder à chaque relation de dépendance. Il a été montré qu'en ne considérant que les relations pour lesquelles le score de confiance est haut, on atteignait des scores de précision plus élevés, au détriment du nombre de relations identifiées (Urieli, 2013, p. 144). Nous avons donc décidé de faire intervenir ce paramètre en envisageant six seuils sur le score de confiance : 0% (toutes les relations de dépendance sont conservées), 70%, 80%, 90%, 95% et 98%. Le nombre de triplets différents extraits passe d'environ 400 000 à 170 000 lorsque l'on fait varier le seuil de 0% à 98%.

3.2 Normalisation et filtrage des triplets syntaxiques

Le second paramètre concerne la normalisation des relations de dépendance : soit les relations extraites à partir des dépendances syntaxiques sont utilisées telles quelles (c'est-à-dire telles que décrites en section 3.1), soit une série d'opérations de transformation de ces relations est effectuée [3] :

1. distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
   Nous illustrons cette première normalisation sur la figure 1. Dans l'extrait 1a, la normalisation permet d'ajouter le triplet `<phrase; mod; correct>` au triplet initial `<phrase; mod; simple>`. Cette normalisation porte sur les coordonnés en position de dépendants syntaxiques. L'extrait 1c illustre un cas où les coordonnés se trouvent en position de gouverneurs : la normalisation permet d'extraire le triplet `<inclure; suj; trait>` en plus de `<reprendre; suj; trait>`.
2. récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. ajout de la relation de coordination (*coord*) à la liste des relations énumérées en 3.1. Cette nouvelle relation permet de construire dans l'exemple 1a le triplet `<simple; coord; correct>` et dans l'exemple 1d le triplet `<apprentissage; coord; méthode>`.
4. transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif (cf. figure 2).
5. conversion de la relation *ats* en *mod*.

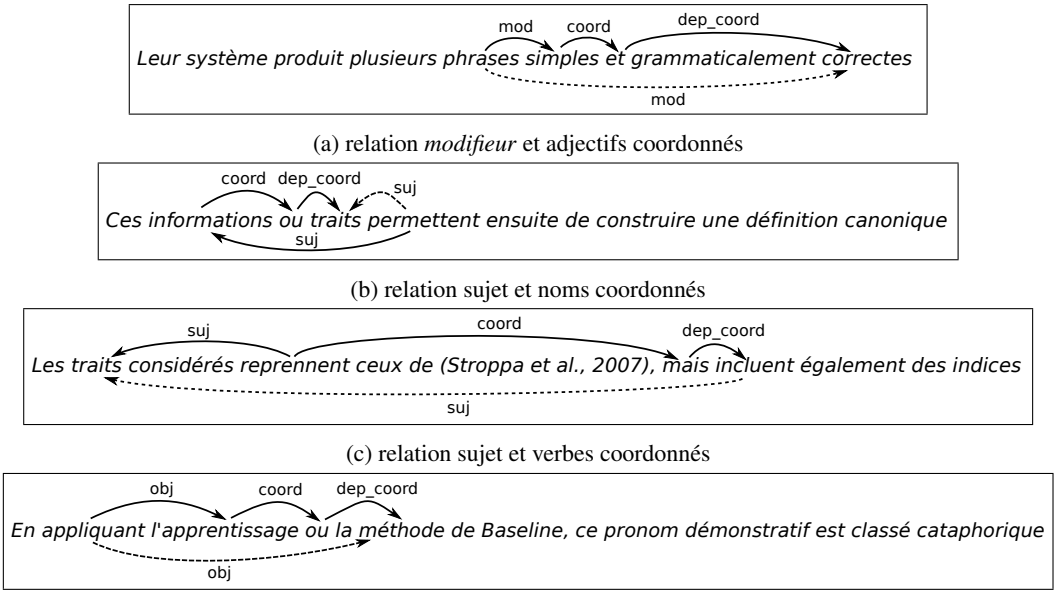
2. Talismane est librement disponible à l'adresse `http://redac.univ-tlse2.fr/applications/talismane.html`
La précision globale du parseur Talismane, pour la configuration utilisée (SVM linéaire, faisceau de largeur 5 avec propagation) est estimée sur des corpus de test à près de 90% pour l'ensemble des dépendances (attachement et labellisation).

3. Nous nous inspirons de l'expérience acquise lors de la construction de bases distributionnelles avec l'outil Upery (Fabre & Bourigault, 2006).

6. regroupement des relations *prép*. L'extraction par défaut produit, à partir des syntagmes *méthode d'apprentissage* et *méthode pour l'apprentissage* les triplets `<méthode; prép_de; apprentissage>` et `<méthode; prép_pour; apprentissage>`. L'extraction « normalisée » omet la préposition à l'origine de la relation et produit, à partir des deux syntagmes précédents, l'unique triplet `<méthode; prép; apprentissage>` (avec une fréquence de 2).

Les opérations 1 à 4 sont de nature à produire davantage de triplets syntaxiques, en établissant des relations qui ne sont pas explicitement fournies par l'analyseur. Les opérations 5 et 6 quant à elles rassemblent une information potentiellement dispersée. Par exemple, dans les extraits *« ce mode d'évaluation sous-estime une **réalité** linguistique **complexe** »* et *« le **jugement** par les humains devient alors encore plus **complexe** »*, il semble peu pertinent de considérer que la nature de la relation qu'entretiennent *réalité* et *complexe* est différente de celle qu'entretiennent *jugement* et *complexe*.
Le bénéfice que l'on peut espérer tirer de la normalisation de la relation *prép* va moins de soi. On sait en effet que la sémantique de chaque préposition a un rôle à jouer dans la représentation distributionnelle de certains mots-cibles (par exemple, pour la construction de certaines classes de noms), et que la génération de triplets différents pour chaque préposition peut empêcher des regroupements abusifs. Néanmoins, notre hypothèse est que la distinction des prépositions peut entraîner une dispersion de l'information lorsque l'on traite une faible masse de données, en empêchant l'établissement de liens de voisinage pertinents. Si une étude plus détaillée serait nécessaire pour estimer dans quelle mesure le regroupement des relations *prép* est souhaitable, celui-ci a clairement un impact dès lors que l'on augmente la valeur de certains seuils comme le nombre d'occurrences des triplets, ou le nombre de contextes syntaxiques partagés par deux voisins (cf. section 3.5).

Les triplets syntaxiques produits, avec ou sans normalisations, sont filtrés sur leur fréquence. Le seuil de ce filtrage (fixé à deux ou cinq occurrences minimum dans le corpus) constitue le troisième paramètre de notre processus d'analyse.



(a) relation *modifieur* et adjectifs coordonnés

(b) relation sujet et noms coordonnés

(c) relation sujet et verbes coordonnés

(d) relation objet et noms coordonnés

FIGURE 1: Normalisation des dépendances sur les gouverneurs et les dépendants coordonnés

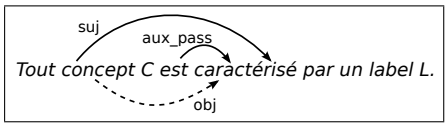


FIGURE 2: Normalisation de la relation sujet avec un gouverneur passif

3.3 Association entre mots-cibles et contextes syntaxiques

Chaque triplet `<gouv; rel; dép>` issu des étapes décrites en sections 3.1 et 3.2 donne lieu à deux associations entre un mot-cible et un contexte syntaxique : le lemme `gouv` (resp. `dép`) est associé au contexte `<rel; dép>` (resp. `<gouv; rel>`).

Par exemple, à partir du triplet `<N:phrase; mod; ADJ:simple>`, on retient que la cible `N:phrase` apparaît avec le contexte syntaxique `<mod; ADJ:simple>` et que `ADJ:simple` apparaît avec le contexte `<N:phrase; mod>`. Chaque mot-cible peut alors être représenté par un vecteur de contextes syntaxiques dont les coordonnées sont une pondération attribuée au couple mot-cible/contexte. Les pondérations testées sont deux mesures d'association habituellement utilisées en analyse distributionnelle : l'information mutuelle (IM) et le t-score (cf. fig. 3). Ces mesures, en relativisant la fréquence d'un couple mot/contexte par rapport à la fréquence du mot et à celle du contexte, reflètent la spécificité de chaque contexte pour la cible considérée. Si un mot apparaît fréquemment dans un contexte donné, mais que ce mot apparaît également très fréquemment avec d'autres contextes et que, de plus, le contexte considéré apparaît très fréquemment avec d'autres mots, la mesure d'association sera faible.

$$IM(l,c) = log_2\left(\frac{N \times f(l,c)}{f(l)f(c)}\right)$$

(a) Information mutuelle

$$t\text{-}score(l,c) = \frac{f(l,c) - \frac{f(l)f(c)}{N}}{\sqrt{f(l,c)}}$$

(b) t-score

$f(l)$: nombre d'occurrences du lemme l
$f(c)$: nombre d'occurrences du contexte syntaxique c
$f(l,c)$: nombre de cooccurrences de l et c
N : nombre total d'occurrences de tous les triplets

FIGURE 3: Mesures d'association utilisées pour la pondération des contextes syntaxiques

3.4 Calculs de similarité

Pour une paire de mots-cibles donnée, deux mesures de similarité (notées *cosIM* et *cosTS*) sont obtenues en calculant le cosinus de leurs vecteurs de contextes. Nous testons également le Jaccard, qui n'est pas basé sur la fréquence d'apparition des lemmes et des triplets syntaxiques, mais sur la *productivité* des lemmes (*i.e.* le nombre de contextes *différents* avec lesquels un lemme apparaît). Le détail de ces mesures est donné en figure 4.

$$cos(l_1,l_2) = \frac{\sum_i p_{1i}p_{2i}}{\sqrt{\sum_i {p_{1i}}^2 \sum_i {p_{2i}}^2}}$$

(a) Cosinus

$$jacc(l_1,l_2) = \frac{|C(l_1) \cap C(l_2)|}{|C(l_1) \cup C(l_2)|}$$

(b) Jaccard

p_{ji} : pondération (IM ou t-score) du contexte i pour le lemme l_j
$C(l_i)$: ensemble des contextes dans lesquels le lemme l_i apparaît

FIGURE 4: Mesures de similarité utilisées pour le calcul de voisinage

3.5 Filtrage sur le nombre de contextes syntaxiques partagés

Pour un mot-cible donné, ses voisins sont donc ordonnés par une des trois mesures de similarité présentées ci-dessus (*cosIM*, *cosTS*, *Jaccard*). La dernière étape consiste à filtrer les couples partageant trop peu de contextes syntaxiques. En effet, des contextes marginaux tendent à rapprocher des paires de mots non pertinentes pour l'établissement d'une relation de voisinage. Par exemple, *contraception* arrive au 3^e rang des voisins du nom *sémantique* (après *optimalité* et *outillage*), alors que *syntaxe* apparaît seulement au 26^e rang. Le rapprochement de *sémantique* et *contraception* est dû à leur unique contexte commun `<question; prép(_de)>` : *contraception* apparaît deux fois dans le corpus (comme exemple d'alignement de textes français-japonais), et uniquement avec ce contexte. L'information mutuelle et le t-score entre *contraception* et son unique contexte `<question; prép(_de)>` sont donc très élevés. En imposant au moins deux contextes syntaxiques partagés, le mot *contraception*, comme d'autres, n'est plus considéré comme voisin de *sémantique* et *syntaxe* remonte en 8^e position. Nous avons testé des configurations avec un seuil variant de 1 à 10 contextes partagés.

3.6 Paramètres et configurations obtenues

Le tableau 3 récapitule les paramètres que nous faisons varier pour construire les différentes bases de voisins. Selon les combinaisons de paramètres, le nombre de lemmes ayant des voisins varie de 10 963 (1,4 million de couples) pour les configurations `0_norm_2_*_1` (pas de filtrage sur le score de confiance des dépendances, normalisation des relations, fréquence minimale de 2 triplets, et pas de seuil sur le nombre de contextes partagés, quelle que soit la mesure de similarité choisie) à 279 (3 638 couples) pour les configurations `98_norm_5_*_10` faisant intervenir les filtres les plus sévères.

Paramètre	nb valeurs	valeurs
seuil sur le score de confiance des dépendances syntaxiques	6	{0, 70, 80, 90, 95, 98}
normalisation des relations	2	avec ou sans : {norm, nonorm}
seuil sur le nombre d'occurrences des triplets	2	{2, 5}
mesure de similarité	3	{cosIM, cosTS, Jaccard}
seuil sur le nombre de contextes partagés	10	[1-10]

TABLE 2: Paramètres de calcul des voisins

4 Analyse des résultats

Nous avons effectué une extraction des 20 premiers voisins distributionnels pour chacun des 15 mots-cibles présentés en section 2 et pour les 720 configurations différentes envisagées pour la méthode. La première analyse de ces données vise à examiner le rôle des différents paramètres, en cherchant à mesurer l'impact de chacun d'eux sur les résultats et à dégager les configurations optimales au vu de l'annotation manuelle. La seconde série d'observations concerne le fonctionnement détaillé de ces configurations optimales pour les différents mots et les catégories grammaticales, afin de mieux appréhender les mécanismes distributionnels à l'œuvre dans ce corpus spécialisé.

4.1 Méthode de comparaison

Afin de comparer les différentes configurations, nous devons prendre en compte pour chacune, et pour chaque mot-cible :

– l'ordre dans lequel les voisins sont classés, en suivant la mesure de similarité de cette configuration (rang, de 1 à 20) ;
– le nombre d'annotateurs qui ont choisi ce mot comme étant un voisin pertinent du mot-cible (pertinence, notée de 0 à 4).

Nous nous trouvons donc dans une situation similaire à celle de la comparaison de systèmes de recherche d'information pour lesquels le jugement de pertinence des réponses est mesuré sur une échelle. Les mesures classiques de rappel et de précision ne considèrent qu'un jugement binaire et sont ainsi moins bien adaptées à notre cas. Nous avons donc utilisé la mesure du *Normalised Discounted Cumulated Gain* ou $nDCG$ (Järvelin & Kekäläinen, 2002). Cette mesure est obtenue en additionnant le score de pertinence des mots renvoyés par le système, mais en pénalisant les résultats les plus éloignés dans la liste en divisant ce score de pertinence par le logarithme du rang de chaque mot. Autrement dit, pour obtenir un haut score pour cette mesure, le système doit renvoyer en premier les voisins déclarés pertinents par le plus grand nombre d'annotateurs.

Le détail de cette mesure est plus précisément :

$$nDCG = \frac{DCG}{DCGI}$$

où

$$DCG = \sum_{i=1}^{20} annot_i / log_2(i+1)$$

$annot_i$ étant le nombre d'annotateurs qui ont sélectionné comme un bon voisin du mot cible le voisin numéro i renvoyé par le système.

$DCGI$ est la valeur maximale de DCG, obtenue par un système qui renverrait tous les mots dans l'ordre décroissant de pertinence. Cette normalisation permet ainsi d'obtenir des valeurs comparables pour des mots dont le nombre de voisins pertinents varie (comme c'est notre cas), et comprises entre 0 et 1 comme les autres mesures classiques d'évaluation en RI [4]. Nous avons donc pu sur cette base calculer des scores moyens de $nDCG$ à travers les différents mots-cibles.

4. Nous avons aussi calculé les scores de précision, rappel et F-mesure à différents points dans les listes de résultats, donc sans prendre en compte le nombre d'annotateurs pour définir la pertinence d'un voisin distributionnel. Les conclusions présentées par la suite restent globalement valables également pour ces différents scores.

4.2 Impact des différents paramètres

Dans un premier temps, nous avons cherché à identifier le rôle global de chacun des paramètres sélectionnés (voir tableau 3). Pour ce faire, nous avons mesuré le score de $nDCG$ pour chaque valeur de paramètre, en faisant une double moyenne : sur les 15 mots-cibles et sur l'ensemble des configurations concernées par cette valeur. Le tableau 3 donne la valeur moyenne, la valeur maximale et l'écart-type du $nDCG$.

Paramètre	Moyenne	Max	Écart-type
Score global	0,446	0,917	0,234
Score de confiance			
0%	**0,473**	0,917	0,233
70%	0,466	0,916	0,228
80%	0,464	0,901	0,231
90%	0,453	0,893	0,231
95%	0,428	0,898	0,230
98%	0,391	0,891	0,239
Normalisation			
Avec	**0,448**	0,905	0,234
Sans	0,443	0,917	0,233
Seuil de fréquence des triplets			
2	**0,500**	0,891	0,211
5	0,391	0,917	0,242

Paramètre	Moyenne	Max	Écart-type
Mesure de similarité			
Cosinus IM	**0,521**	0,872	0,245
Cosinus t-score	0,389	0,917	0,233
Jaccard	0,427	0,792	0,202
Seuil sur les contextes partagés			
1	0,385	0,871	0,251
2	0,438	0,876	0,216
3	0,466	0,889	0,204
4	**0,474**	0,891	0,206
5	0,467	0,898	0,224
6	0,464	0,905	0,232
7	0,456	0,905	0,238
8	0,448	0,916	0,245
9	0,434	0,917	0,250
10	0,426	0,917	0,251

TABLE 3: Scores $nDCG$ moyens et maximaux pour chaque valeur des paramètres

La première ligne du tableau donne la valeur moyenne sur l'ensemble des 720 configurations, et sert de référentiel pour chaque paramètre. Les conclusions suivantes peuvent être tirées pour chaque paramètre pris individuellement (la valeur moyenne la plus élevée a été indiquée en gras dans le tableau) :

- score de confiance : le score de $nDCG$ diminue de façon monotone avec ce seuil, il semble donc préférable de ne pas filtrer les triplets en fonction de la confiance estimée par l'analyseur syntaxique ;
- normalisation : la normalisation des triplets syntaxiques extraits apporte un léger gain par rapport à l'utilisation des contextes « bruts » ;
- seuil de fréquence des triplets : l'augmentation de ce seuil de 2 à 5 fait baisser sensiblement la performance globale du système ;
- mesure de similarité : c'est la similarité cosinus basée sur les scores d'information mutuelle qui donne les meilleurs résultats, suivie d'assez loin par le Jaccard. Le cosinus sur les t-scores donne globalement de très mauvais résultats, mais les valeurs maximales atteintes sont étonnamment supérieures aux autres méthodes ;
- seuil sur les contextes partagés : un minimum de 4 contextes syntaxiques différents est la valeur optimale à ce stade.

Bien évidemment, les différents paramètres ne sont pas indépendants, et la configuration optimale n'est pas nécessairement celle qui correspond à la combinaison des valeurs identifiées précédemment. De fait, sur l'ensemble des 15 mots, le paramétrage optimal (avec un $nDCG$ moyen de 0,659) est : `0_norm_2_cosIM_3`. C'est-à-dire : pas de filtrage sur les relations de dépendance, normalisation des contextes, élimination des triplets ayant une fréquence inférieure à 2, tri par similarité cosinus sur les valeurs d'information mutuelle, élimination des voisins ayant moins de 3 contextes syntaxiques partagés avec la cible.

4.3 Variation par catégorie du mot-cible

Nous allons maintenant nous pencher sur les variations entre les trois catégories grammaticales possibles pour les mots-cibles (nom, verbe et adjectif). La table 4 donne les valeurs maximales et moyennes sur l'ensemble des 720 configurations envisagées pour chaque catégorie.

On peut voir que les scores sont assez proches pour les noms et les verbes. Les adjectifs, quant à eux, semblent nettement plus difficiles à traiter et ont un score moyen largement inférieur.

$nDCG$	Adjectifs	Noms	Verbes	Toutes catégories
Maximum	0,827	0,917	0,872	0,917
Moyenne	0,311	0,533	0,493	0,446
Écart-type	0,212	0,221	0,206	0,234

TABLE 4: Valeurs du $nDCG$ pour les 720 configurations envisagées

En calculant, pour chacune des configurations, le score $nDCG$ sur les 5 mots de chaque catégorie, il est possible d'identifier le paramétrage optimal pour cette catégorie.

Pour les **verbes**, le meilleur système est : `0_norm_2_cosIM_3`. Il s'agit du système qui a obtenu les meilleures performances globales.

Pour les **noms**, le meilleur système est : `80_norm_2_cosIM_7`. Il semble donc préférable pour les noms de restreindre les données exploitées par la méthode, tant sur la confiance de l'analyseur syntaxique (minimum de 80%, ce qui correspond globalement à rejeter 10% des dépendances syntaxiques) que sur le nombre de contextes partagés par les voisins (7 contextes différents au moins). Une hypothèse à ce stade pourrait être que les noms sont impliqués dans une plus grande variété de contextes syntaxiques, qu'il devient alors nécessaire de filtrer pour faire émerger les voisins les plus pertinents.

Pour les **adjectifs**, le meilleur système est : `0_norm_2_cosIM_1`. Autrement dit, la principale différence avec le meilleur système global est de ne pas exiger un nombre minimal de contextes partagés pour les adjectifs. Là encore, on peut expliquer cette différence par la faible variété de contextes syntaxiques des adjectifs : on les trouve principalement en relation de modifieur et éventuellement en coordination avec d'autres adjectifs. De ce fait, le filtrage des contextes semble trop pénalisant.

4.4 Variation par mot-cible

Si l'on regarde le comportement pour chaque mot-cible, on peut voir dans la table 5 les scores maximum et moyens (sur les 720 configurations envisagées). Nous avons également reproduit les scores de F-mesure de l'accord inter-annotateurs du tableau 1 (voir section 2).

Mot-cible	**Maximum**	**Moyenne**	**Accord**
complexe	0,620	0,194	0,58
correct	0,773	0,343	0,55
important	0,827	0,527	0,65
précis	0,748	0,285	0,72
spécialisé	0,454	0,208	0,73
Tous les adjectifs	0,591	0,311	0,65
fréquence	0,776	0,587	0,58
graphe	0,760	0,547	0,55
méthode	0,917	0,729	0,75
trait	0,802	0,565	0,57
sémantique	0,649	0,237	0,40
Tous les noms	0,733	0,533	0,57
annoter	0,607	0,355	0,50
calculer	0,815	0,545	0,47
décrire	0,816	0,504	0,57
évaluer	0,872	0,677	0,65
extraire	0,793	0,383	0,58
Tous les verbes	0,761	0,493	0,55

TABLE 5: Comparaison des scores $nDCG$ et de l'accord inter-annotateurs par mot et par catégorie

Il apparaît que les scores ne sont pas corrélés, autrement dit que les mots qui semblent aisément analysables par les

humains ne sont pas ceux pour lesquels les systèmes obtiennent de bons scores. Le coefficient de corrélation entre les $nDCG$ maximum et l'accord inter-annotateurs est nul, et il est très faiblement positif ($r = 0,13$) lorsque l'on considère le $nDCG$ moyen. Cette constatation est assez surprenante, et on aurait attendu une liaison positive forte.

Il se trouve que cette liaison existe pour les noms ($r = 0,96$), est moins marquée pour les verbes ($r = 0,47$) mais est même négative pour les adjectifs ($r = -0,24$). On voit donc bien ici que c'est le comportement des adjectifs qui est le plus contre-intuitif : s'il s'agit de la catégorie la plus « facile » pour l'annotation humaine, nous avons vu que c'est celle qui a posé le plus de problèmes aux systèmes. Ces constatations, ainsi que le faible nombre d'individus envisagés nous amènent à recourir à un examen qualitatif des résultats de l'analyse distributionnelle.

4.5 Étude détaillée de quelques résultats

Nous allons ici examiner plus en détails les résultats obtenus pour quelques mots parmi ceux étudiés. Plus précisément, nous allons examiner deux noms (*méthode* et *sémantique*) et deux adjectifs (*complexe* et *spécialisé*).

4.5.1 De la *méthode* avant tout, la *sémantique* finira bien par émerger

Nous présentons ici les résultats du meilleur système global (configuration `0_norm_2_cosIM_3`) pour les deux noms *méthode* (table 6) et *sémantique* (table 7). Nous avons indiqué pour chacun d'eux les 20 premiers voisins renvoyés par le système, et leur score de pertinence résultant de l'annotation manuelle.

Rang	Mot	Pertinence
1	approche	4
2	technique	4
3	système	1
4	algorithme	4
5	stratégie	4
6	modèle	1
7	outil	1
8	méthodologie	4
9	processus	3
10	module	0
11	procédure	4
12	mesure	0
13	étape	1
14	analyseur	0
15	classifieur	1
16	règle	1
17	ressource	0
18	travail	0
19	critère	0
20	résultat	0

TABLE 6: Résultats de la meilleure configuration globale pour le nom *méthode* ($nDCG = 0,89$)

Rang	Mot	Pertinence
1	propriété	0
2	signification	2
3	ambiguïté	1
4	nature	1
5	polysémie	0
6	syntaxe	3
7	aspect	0
8	définition	4
9	idée	0
10	diversité	0
11	notion	3
12	comportement	0
13	représentation	2
14	diacritique	0
15	caractéristique	1
16	distribution	1
17	délimitation	0
18	fermeture	0
19	structure	0
20	spécificité	1

TABLE 7: Résultats de la meilleure configuration globale pour le nom *sémantique* ($nDCG = 0,30$)

Méthode. Comme on peut le voir dans la table 6, les voisins pertinents de *méthode* ont pratiquement tous été retrouvés (il ne manque qu'*heuristique* et *stratégie* parmi ceux qui ont été sélectionnés par deux annotateurs ou plus) et sont placés dans le haut de la liste. De plus, certains des résultats déclarés non pertinents sont tout de même acceptables (*module, mesure, analyseur*).

Les principaux contextes syntaxiques de *méthode* dans le corpus sont `<proposer; obj>`, `<permettre; suj>`, `<présenter; obj>` et `<prep(de); apprentissage>`. On voit bien à la fois la variété des relations syntaxiques impliquées et l'unité

sémantique (il s'agit bien ici des méthodes de TAL que les articles du corpus présentent, un grand nombre d'entre elles étant des méthodes d'apprentissage).

Sémantique. Pour le nom *sémantique* (table 7), les résultats sont moins probants, et la délimitation est floue pour chaque voisin envisagé. Plusieurs voisins fortement pertinents manquent à l'appel : *contenu*, *sens*, *connaissance* et *valeur*. Par contre, la pertinence des voisins renvoyés en haut de liste est difficile à interpréter (*propriété, polysémie, aspect, idée, diversité*, etc.). Cette difficulté à cerner la zone de sens autour de *sémantique* avait bien été perçue lors de la phase d'annotation, comme le montre le faible taux d'accord et les commentaires des annotateurs.

Les contextes syntaxiques de *sémantique* sont par exemple : `<mod; lexical>`, `<mod; compositionnel>`, `<prep(de); Montague>`, `<prep(de); mot>`, `<prep(de); phrase>`, `<prep(de); texte>`. On constate une moins grande variété syntaxique que pour *méthode* (très peu de contextes verbaux notamment) mais surtout une plus grande dispersion.

4.5.2 Le problème des adjectifs : une tâche *complexe*, même pour un corpus *spécialisé*

Comme on l'a vu dans les résultats globaux, le cas des adjectifs est très différent de celui des noms et des verbes : malgré un accord inter-annotateurs très élevé, les différentes configurations peinent à faire émerger les voisins déclarés pertinents. De plus, au sein des adjectifs, il ne semble pas y avoir de lien entre leur traitement par les annotateurs et par les approches distributionnelles. Nous allons voir plus en détails les résultats concernant les adjectifs *complexe* et *spécialisé*, produits par la même configuration que précédemment.

Rang	Mot	Pertinence
1	distinct	0
2	fastidieux	0
3	multimots	2
4	particulier	0
5	simple	3
6	composé	3
7	incomplet	0
8	typique	0
9	long	0
10	considéré	0
11	trivial	3
12	extrait	0
13	délicat	4
14	monosémique	0
15	spécifique	1
16	compliqué	4
17	classique	0
18	coûteux	3
19	fondamental	0
20	visé	0

TABLE 8: Résultats de la meilleure configuration globale pour l'adjectif *complexe* ($nDCG = 0,38$)

Rang	Mot	Pertinence
1	cible	0
2	biomédical	4
3	généraliste	3
4	juridique	4
5	considéré	0
6	multilingue	0
7	analysé	0
8	médical	4
9	anglais	0
10	bilingue	0
11	technique	4
12	structuré	0
13	monolingue	0
14	volumineux	0
15	japonais	0
16	orthographié	0
17	existant	0
18	annoté	0
19	vietnamien	0
20	source	0

TABLE 9: Résultats de la meilleure configuration globale pour l'adjectif *spécialisé* ($nDCG = 0,39$)

***Complexe*.** Pour l'adjectif *Complexe* (table 8), plusieurs voisins pertinents manquent à l'appel : *difficile*, *élémentaire*, *dérivé*.

Plusieurs des adjectifs non pertinents sont d'une nature particulière : ils n'expriment pas une propriété caractéristique du référent, en d'autres termes ce ne sont pas des adjectifs qualificatifs typiques. C'est le cas d'adjectifs comme *distinct*, *particulier*, *visé*, *considéré*, qui sont généralement utilisés sur le plan rhétorique, pour structurer le discours ("deux X

distincts", "ce X particulier", etc.). De fait, les contextes qu'ils partagent avec *complexe* correspondent à des noms exprimant des notions très générales : *problème*, *format*, *besoin*, *genre*, *tâche*, *modèle*, *configuration*, *phénomène*, etc. On a donc affaire à des adjectifs qui peuvent modifier une très large gamme de noms, ce qui peut expliquer le décalage par rapport à l'annotation humaine, focalisée sur des emplois plus spécifiques de l'adjectif (complexité des traitements et des résultats). À l'inverse, l'observation des contextes de *complexe* dans le corpus montre la prédominance du nom *terme* (127 occurrences de *terme complexe*). C'est ce contexte et ses quelques variantes sémantiques (*mot*, *phrase*, *morphologie*) qui ont permis de faire émerger des voisins plus spécifiques comme *multimots* et *composé*.

Spécialisé. Pour l'adjectif *spécialisé* (table 9), les principaux voisins manquants sont : *spécifique*, *général*, *générique*, *quelconque* et *savant*.

Les principaux contextes sur lesquels s'appuient les voisins distributionnels sont des noms correspondant aux différents types de données langagières : *langue*, *terme*, *document*, *domaine*, *texte*, *corpus*, *discours*, *lexique*. Mais malgré cette homogénéité les adjectifs renvoyés correspondent à des qualifications très disparates de ces données (*cible*, *multilingue*, *anglais*, *structuré*, etc.) qui ont apparemment « noyé » les voisins les plus pertinents. Les adjectifs sous-spécifiés sont moins nombreux que pour l'adjectif *complexe* : seuls *considéré* et *existant* semblent avoir ce statut.

4.6 Premier bilan

On a donc pu voir dans ces différentes analyses que la complexité des mécanismes distributionnels entraîne une grande variété des résultats, à travers les paramétrages, les catégories des mots-cibles et les mots-cibles eux-mêmes. Si nous avons pu dégager de grandes tendances concernant le paramétrage, et identifier un sous-ensemble de configurations performantes, il ne faut pas oublier que celles-ci dépendent bien entendu des caractéristiques du corpus, et ne sont pas à priori réutilisables telles quelles sur des corpus de plus grande taille, de genre différent ou moins homogènes. Certaines de ces tendances semblent rejoindre les conclusions de travaux antérieurs (Ferret, 2010), à savoir la préférence pour le cosinus et l'information mutuelle comme bases du calcul de la similarité.

En ce qui concerne les différentes catégories, les résultats pour les noms et les verbes sont encourageants : non seulement la précision globale obtenue par ces méthodes est tout à fait acceptable, mais surtout les variations entre les mots semblent aller dans le même sens que la difficulté ressentie par les annotateurs pour interpréter le sens des mots-cibles. Pour ces deux catégories, nous avons mesuré pour chaque mot la concentration des contextes syntaxiques (à savoir l'entropie normalisée des contextes dans le corpus), et observé une corrélation négative significative (sur les seuls 10 mots envisagés) avec l'accord inter-annotateurs ($r = -0,72$, $p < 0,05$). Il est donc plus facile pour le système comme pour les juges de cerner le voisinage sémantique de mots dont l'usage est régulier dans le corpus.

Le cas des adjectifs est très différent des deux autres catégories. Non seulement les résultats globaux sont nettement inférieurs, même pour les meilleures configurations, mais de plus la corrélation précédente avec la distribution des contextes n'est pas observée, et il est difficile de comprendre quels sont les phénomènes qui bloquent un traitement efficace de ce type de mots. Les deux seules pistes à ce stade sont la présence massive d'adjectifs sous-spécifiés et la pauvreté des types de contextes syntaxiques impliquant les adjectifs : il y a notamment très peu de syntagmes prépositionnels à tête adjectivale. Il devrait donc être possible d'améliorer les résultats pour les adjectifs en utilisant des contextes élargis ou en cherchant à filtrer plus sévèrement les adjectifs sous-spécifiés.

Enfin, l'observation précise des résultats permet de confirmer la variété des relations sémantiques identifiées par la méthode distributionnelle. On avait vu (section 2) que les annotateurs ne s'étaient pas limités à la seule synonymie, et les autres relations sont aussi bien (ou mal) identifiées qu'elle par les systèmes étudiés. On retrouve bien dans les résultats positifs des cas d'antonymie (*simple* et *trivial* pour *complexe*, *généraliste* pour *spécialisé*), d'hyponymie (*médical*, *juridique*, *technique*, etc. pour *spécialisé*, *algorithme* et *classifieur* pour *méthode*) et de co-hyponymie (*syntaxe* pour *sémantique*).

Cette variété est en tout cas une confirmation de la difficulté à évaluer efficacement les sorties d'une telle méthode, et du fait que nous avons mis toutes les chances de notre côté en effectuant une annotation experte à la fois sur les données et les méthodes.

5 Conclusion et perspectives

En résumé, nous avons déployé un système d'analyse distributionnelle sur le corpus des articles de TALN en faisant varier un ensemble de paramètres et en comparant les résultats avec une annotation manuelle pour 15 mots. En nous basant sur une comparaison purement quantitative, nous avons montré que les 720 combinaisons distinctes pour les 5 paramètres donnaient des résultats très variables. Cette variabilité concerne aussi bien les paramètres liés à l'exploitation des sorties de l'analyseur syntaxique que les filtrages et les choix de mesures à effectuer pour calculer la similarité distributionelle. Ceci confirme donc qu'il est important d'accorder une attention égale à chaque aspect de la chaîne de traitement, et qu'il nous reste encore de grandes marges de progression sur ce terrain.

Nous souhaitons maintenant étudier plus finement les contributions relatives des différentes relations syntaxiques impliquées, élargir la gamme des opérations de normalisation et observer leurs interactions, en suivant un protocole similaire à celui utilisé ici. Une piste intéressante concerne la difficulté d'identifier des voisins pertinents pour les adjectifs. Une des pistes consisterait alors à diversifier les contextes syntaxiques de ceux-ci pour permettre un rapprochement plus efficace. Pour les autres catégories (noms et verbes) il semble par contre que la difficulté de traitement soit directement liée à celle qu'ont rencontrée les annotateurs.

Si le choix d'utiliser d'emblée un analyseur syntaxique nous semble justifié par la taille du corpus et surtout par les possibilités d'interaction qu'il permet sur le plan de la caractérisation linguistique des contextes, nous souhaitons tout de même comparer les résultats obtenus à ceux que fournirait une méthode se basant sur la cooccurrence de surface.

Pour revenir sur les données utilisées pour l'évaluation, il est clair que la taille du corpus TALN le situe tout de même confortablement dans la zone d'application de ce type de méthodes. De même, nous avons choisi des mots-cibles dont la fréquence est suffisamment élevée pour garantir un rendement minimal même après filtrage. On pourrait donc envisager de contraindre ces deux aspects, surtout en appliquant la méthode à des mots de basses voire très basses fréquences, comme le font Périnet & Hamon (2013) dans le domaine culinaire.

De plus, nous nous sommes limités ici au cas classique de l'étude des mots simples, alors qu'un tel corpus spécialisé appelle bien évidemment la prise en compte des nombreuses unités polylexicales qui le peuplent, tant comme cibles que comme éléments de contexte.

Références

BARONI M. & LENCI A. (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 673–721.

BARONI M. & LENCI A. (2011). How we BLESSed distributional semantic evaluation. *Proceedings of the GEMS 2011, Workshop on GEometrical Models of Natural Language Semantics*, p. 1–10.

BOUDIN F. (2013). TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 507–514, Les Sables d'Olonne, France.

FABRE C. & BOURIGAULT D. (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In *Actes de la 13econférence sur le Traitement Automatique de la Langue Naturelle (TALN 2006)*, Leuven, Belgique.

FERRET O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *7th International Conference on Language Resources and Evaluation (LREC'10)*, p. 3338–3343, Malta.

JÄRVELIN K. & KEKÄLÄINEN J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, **20**(4), 422–446.

MORLANE-HONDÈRE F. & FABRE C. (2012). Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. In *Actes du 3e Congrès Mondial de Linguistique Français (CMLF 2012)*, p. 1001–1015, Lyon.

PÉRINET A. & HAMON T. (2013). Hybrid acquisition of semantic relations based on context normalization in distributional analysis. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA2013)*, p. 113–122.

TURNEY P. & PANTEL P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, **37**(1), 141–188.

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat, Université de Toulouse II le Mirail.

URIELI A. & TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 188–201, Les Sables d'Olonne, France.