Évaluation des modèles sémantiques distributionnels : le cas de la dérivation syntaxique

Gabriel Bernier-Colborne Patrick Drouin
Observatoire de linguistique Sens-Texte (OLST), Université de Montréal
C.P. 6128, succ. Centre-Ville, Montréal (QC) Canada, H3C 3J7
{gabriel.bernier-colborne|patrick.drouin}@umontreal.ca

RÉSUMÉ

Nous évaluons deux modèles sémantiques distributionnels au moyen d'un jeu de données représentant quatre types de relations lexicales et analysons l'influence des paramètres des deux modèles. Les résultats indiquent que le modèle qui offre les meilleurs résultats dépend des relations ciblées, et que l'influence des paramètres des deux modèles varie considérablement en fonction de ce facteur. Ils montrent également que ces modèles captent aussi bien la dérivation syntaxique que la synonymie, mais que les configurations qui captent le mieux ces deux types de relations sont très différentes.

ABSTRACT

Evaluation of distributional semantic models : The case of syntactic derivation

Using a dataset representing four kinds of semantic relations, we evaluate two distributional semantic models and analyze the influence of each model's (hyper)parameters. Results indicate that the model which performs best depends on the targeted relations, and that the influence of both models' (hyper)parameters varies considerably with respect to this factor. They also show that distributional models capture syntactic derivation just as well as synonymy, but that the optimal settings for these two kinds of relations are very different.

MOTS-CLÉS : sémantique distributionnelle, sémantique lexicale, évaluation.

KEYWORDS: distributional semantics, lexical semantics, evaluation.

1 Introduction

Dans le cadre d'un travail visant à caractériser l'apport des méthodes distributionnelles à la lexicographie spécialisée, nous avons cherché à déterminer comment le choix et la paramétrisation d'un modèle distributionnel sont influencés par divers facteurs, tels que la langue traitée, la partie du discours des termes utilisés pour interroger le modèle, et les relations lexicales ciblées. Cette étude a été réalisée en évaluant des modèles distributionnels, construits à partir d'un corpus spécialisé, au moyen de données de référence extraites d'un dictionnaire spécialisé, représentant quatre types de relations lexicales paradigmatiques. Les résultats ont notamment mis au jour des différences importantes entre la dérivation syntaxique et d'autres relations paradigmatiques telles que la synonymie en ce qui concerne leur modélisation par l'approche distributionnelle. En effet, les modèles distributionnels captent la dérivation aussi bien que la synonymie, mais les configurations qui captent le mieux ces deux types de relations sont très différentes. En comparant deux modèles distributionnels, nous montrons que celui qui produit les meilleurs résultats dépend des relations ciblées. Puis, en explorant

systématiquement l'espace des (hyper)paramètres des deux modèles, nous montrons comment les paramétrisations optimales varient en fonction de ce facteur. Ce travail fournit ainsi des balises pour l'application de méthodes distributionnelles dans le cadre du travail lexicographique.

2 Travaux reliés

L'analyse et l'optimisation des (hyper)paramètres des modèles sémantiques distributionnels a fait l'objet de nombreux travaux, mais les études traitant systématiquement l'influence des relations lexicales ciblées sur la paramétrisation de ces modèles sont relativement peu nombreuses. Dans le cas des techniques classiques d'analyse distributionnelle (AD), des études réalisées dès les années 1960 auraient montré que certains de ses paramètres, notamment la taille de la fenêtre de contexte, ont une influence sur le genre de relations captées (Moskowich & Caplan, 1978). Plus récemment, des études systématiques de l'influence des paramètres distributionnels ont été réalisées, dont certaines tiennent compte des différentes relations lexicales qu'on peut chercher à identifier. Sahlgren (2006) a comparé deux types de contextes distributionnels (cooccurrents graphiques et segments textuels) et analysé l'influence de quelques paramètres dont la taille des contextes ; les résultats ont montré que la paramétrisation optimale du modèle dépend des relations ciblées (en l'occurrence, la synonymie, l'antonymie ou des relations syntagmatiques). Lapesa *et al.* (2014) ont analysé l'influence des paramètres de l'AD sur sa capacité à capter différentes relations paradigmatiques (synonymie, antonymie et cohyponymie) et syntagmatiques. Notre travail s'inscrit dans ce courant de recherche, se distinguant notamment par la prise en compte d'une gamme plus vaste de relations paradigmatiques.

L'évaluation des modèles sémantiques distributionnels et l'analyse de leurs paramètres sont souvent réalisées au moyen de jeux de données qui ne ciblent pas une relation lexicale particulière (à l'exception de la synonymie) ou qui ne font pas de distinctions entre différentes relations lexicales. En effet, beaucoup des jeux de données couramment utilisés en sémantique distributionnelle [1] représentent des liens de similarité ou de proximité sémantique plus ou moins clairement définis, une exception notable étant le jeu BLESS (Baroni & Lenci, 2011), qui permet d'évaluer la modélisation de relations conceptuelles particulières (cohyponymie, hyperonymie, méronymie, concept-attribut, concept-évènement). Nous utilisons dans ce travail un jeu de données représentant quatre classes particulières de relations lexicales paradigmatiques, qui comprennent une relation qui a reçu relativement peu d'attention dans le domaine de la sémantique distributionnelle, à savoir la dérivation syntaxique (voir Section 3).

Cette étude comprend également une comparaison entre l'AD et les représentations distribuées apprises au moyen des deux modèles de langue neuronaux implémentés dans l'outil `word2vec`, appelés *continuous bag-of-words* (CBOW) et *continuous skip-gram* (Mikolov *et al.*, 2013a,b). Plusieurs évaluations comparatives de ces deux types de modèles (AD et `word2vec`) ont été réalisées récemment. Baroni *et al.* (2014) ont comparé l'AD et et le modèle CBOW sur plusieurs jeux de données et ont observé que CBOW offrait systématiquement de meilleurs résultats. Par contre, Ferret (2015) a observé le contraire lorsqu'il a comparé deux thésaurus distributionnels construits à partir des représentations rendues disponibles par ces derniers. Levy *et al.* (2015) ont montré que lorsqu'on optimise correctement les (hyper)paramètres de l'AD et du modèle skip-gram, la qualité des résultats qu'on obtient avec ces deux modèles est souvent similaire. Nous ne connaissons aucun travail comparant la capacité de l'AD et de `word2vec` à capter des relations lexicales spécifiques (à l'exception de la synoymie) ; il s'agit là d'une des contributions de ce travail. De plus, nous analysons l'influence des

1. Voir, par exemple, les données utilisées par Bullinaria & Levy (2012), Baroni *et al.* (2014) ou Kiela & Clark (2014).

relations ciblées sur l'optimisation des hyperparamètres de `word2vec`, une question qui n'a pas été étudiée à notre connaissance. En outre, nous comparons ces deux types de modèles distributionnels sur un corpus spécialisé, de taille relativement petite (50M mots).

Soulignons enfin que notre travail est relié aux études traitant l'influence des paramètres distributionnels en corpus spécialisé (Périnet & Hamon, 2014; Fabre *et al.*, 2014; Tanguy *et al.*, 2015, entre autres), notre étude se distinguant de celles-ci notamment par la prise en compte de plusieurs relations lexicales spécifiques et par la comparaison de deux types différents de modèles distributionnels.

3 Ressources

Le corpus utilisé pour construire les modèles distributionnels est le corpus monolingue français PANACEA du domaine de l'environnement [2] (ELRA-W0065). Ce corpus, compilé automatiquement (Prokopidis *et al.*, 2012), est constitué de pages Web totalisant environ 50 millions de mots. Le prétraitement appliqué au corpus comprend l'extraction du texte des documents XML (contenant au moins 50 mots), la normalisation des caractères, la lemmatisation au moyen de TreeTagger (Schmid, 1994) et la mise en minuscules. La lemmatisation, qui consiste à remplacer les variantes flexionnelles d'une même forme lexicale par sa forme canonique, a été appliquée parce qu'elle offre parfois une légère amélioration des résultats de l'approche distributionnelle [3] (Bullinaria & Levy, 2012)

Les données de référence que nous exploitons pour l'évaluation des modèles sont des paires de termes reliés par une relation lexicale paradigmatique. Ces données ont été extraites du DiCoEnviro [4], un dictionnaire spécialisé du domaine de l'environnement. Les paires extraites du dictionnaire ont été filtrées afin de ne retenir que celles constituées de deux termes dont la fréquence dans le corpus était suffisamment élevée pour qu'ils fassent partie des mots-cibles pour lesquels nous avons produit des représentations (voir Section 4). De plus, nous avons seulement retenu les paires constituées de deux termes simples [5] (noms, verbes, adjectifs). Nous avons ainsi extrait 1314 paires de termes reliés, que nous avons regroupées en 4 sous-ensembles correspondant à différents types de relations lexicales :

— **QSYN** (quasi-synonymes) : 357 termes associés à un ou plusieurs synonymes, quasi-synonymes, variantes, cohyponymes ou autres sens voisins, totalisant 689 relations ; p. ex. *écologique*→{*écolo, propre, vert, environnemental, biotique*}.
— **ANTI** (antonymes) : 126 termes associés à un ou plusieurs antonymes, totalisant 173 relations ; p. ex. *absorber* → {*réfléchir, émettre, libérer*}.
— **HYP** (relations hiérarchiques) : 100 termes associés à un ou plusieurs hyponymes ou hyperonymes, totalisant 193 relations ; p. ex. *biocarburant* → {*carburant, biogaz, biodiesel, bioéthanol, éthanol*}.
— **DRV** (dérivés) : 259 termes associés à un [6] dérivé syntaxique (Mel'čuk *et al.*, 1995, p. 133), à savoir un mot ayant le même sens, mais appartenant à une partie du discours différente [7], totalisant 259 relations ; p. ex. *absorber*→{*absorption*}, *combustion*→{*brûler*}.

2. `http://catalog.elra.info/product_info.php?products_id=1186&language=fr`
3. Nous avons effectivement observé que la lemmatisation augmente la qualité des résultats.
4. `http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi`
5. Soulignons qu'une proportion élevée des entrées du DiCoEnviro sont des termes simples, car l'approche lexico-sémantique à la terminologie sur laquelle repose ce dictionnaire ne décrit les termes complexes que si leur sens n'est pas compositionnel (L'Homme, 2005).
6. Un terme peut avoir plusieurs dérivés, mais dans les données que nous avons extraites, aucun terme n'a plus d'un dérivé.
7. Il s'agit souvent, mais pas toujours, de dérivés morphologiques.

Il est important de noter que les QSYN, ANTI et HYP appartiennent à la même partie du discours, tandis que les DRV appartiennent à des parties du discours différentes.

4 Expérience

L'expérience [8] que nous avons réalisée comporte une évaluation comparative de deux modèles sémantiques distributionnels et une exploration systématique de leurs (hyper)paramètres.

Le premier modèle que nous avons utilisé est une méthode distributionnelle classique qu'on appelle généralement *analyse distributionnelle* (AD) en français. L'AD consiste essentiellement à extraire des paires (mot, contexte) d'un corpus, puis à compter et pondérer la fréquence de cooccurrence des mots et des contextes. Les contextes peuvent être définis de différentes façons ; on exploite souvent soit la cooccurrence graphique, soit les relations de dépendance syntaxique pour déterminer les contextes d'une occurrence particulière d'un mot. Dans ce travail, les contextes sont des cooccurrents graphiques, c'est-à-dire les mots apparaissant dans une *fenêtre de contexte* centrée sur une occurrence donnée d'un mot. Ainsi, les contextes sont aussi des mots ; nous appellerons donc *mots-cibles* les mots représentés par le modèle distributionnel et *mots-contextes* les contextes (ou attributs) du modèle.

Suivant Levy *et al.* (2015), nous notons (w, c) les paires (mot, contexte) extraites du corpus, D l'ensemble des paires extraites, V_w l'ensemble des mots-cibles et V_c l'ensemble des mots-contextes. Une fois la collection D extraite, on calcule pour chaque mot-cible $w \in V_w$ et chaque mot-contexte $c \in V_c$ leur fréquence de cooccurrence $\#(w, c)$ dans D. Ces fréquences sont stockées dans une matrice M, puis pondérées, généralement au moyen d'une mesure d'association telle que l'information mutuelle. Ainsi, chaque valeur M_{ij} indique la force de l'association entre le mot-cible w_i et le mot-contexte c_j. La similarité distributionnelle des mots-cibles est ensuite calculée en comparant les vecteurs-rangées de M au moyen d'une mesure de similarité (ou de distance), la plus courante étant le cosinus de l'angle des vecteurs (Turney & Pantel, 2010).

Quelques détails concernant notre implémentation de ce modèle méritent d'être soulignés. Premièrement, lors de l'extraction de D, nous permettons à la fenêtre de contexte de chevaucher les frontières de phrases ; ainsi, un mot peut être considéré comme contexte (cooccurrent) d'un autre mot même s'il se trouve dans la phrase précédente ou suivante. Deuxièmement, nous avons choisi d'utiliser le même ensemble de mots pour les mots-cibles et les mots-contextes ($V_c = V_w$), plutôt que définir séparément ces deux ensembles. Enfin, les mots hors-vocabulaire (c'est-à-dire qui ne font pas partie des mots-cibles ou des mots-contextes) ne sont pas supprimés du corpus ; ils ne sont simplement pas pris en compte lors de l'extraction de D.

En ce qui concerne les mots-cibles, nous avons utilisé les 10000 formes les plus fréquentes dans le corpus (lemmatisé), en excluant les mots vides, les chaînes contenant tout caractère ne correspondant ni à une lettre ni à un chiffre ni à un trait d'union, ainsi que les formes commençant ou se terminant par un trait d'union. Le mot-cible le moins fréquent a 157 occurrences dans le corpus.

Le deuxième modèle distributionnel que nous avons utilisé est `word2vec`. Les deux architectures neuronales implémentées dans `word2vec` apprennent des représentations distribuées de mots qui peuvent servir à estimer la similarité sémantique de la même façon que les représentations construites par AD. L'AD et les modèles de langue neuronaux reposent sur la même hypothèse, à savoir que les

8. Les programmes Python que nous avons écrits pour réaliser cette expérience, ainsi que les données de référence, sont disponibles à l'adresse `https://github.com/gbcolborne/TALN_2016`.

mots apparaissant dans des contextes similaires ont tendance à avoir des sens similaires (Harris, 1954; Firth, 1957). Bien qu'ils représentent deux approches différentes à l'acquisition de représentations de mots, ces deux types de modèles sémantiques distributionnels exploitent les mêmes genres de contextes, à savoir la cooccurrence graphique ou syntaxique [9].

L'entraînement des modèles neuronaux a été réalisé au moyen du programme `word2vec` [10]. Nous avons utilisé les mêmes mots-cibles que pour l'AD [11]. La valeur par défaut a été utilisée pour tous les hyperparamètres sauf ceux dont nous analysons l'influence (voir la Section 4.1).

En ce qui concerne la mesure d'évaluation, nous utilisons la moyenne des précisions moyennes (en anglais, *mean average precision* ou *MAP*) pour évaluer la qualité de la liste ordonnée de voisins distributionnels qu'on obtient à partir d'un modèle donné pour chaque requête. On obtient ces listes en calculant la similarité entre la représentation de la requête et celle de tous les autres mots-cibles ; la mesure de similarité que nous utilisons est le cosinus de l'angle des vecteurs.

Pour définir le calcul de la MAP, appelons $Q = \{q_1, \ldots, q_n\}$ l'ensemble des requêtes utilisées pour l'évaluation. Pour chaque requête q_i, nous avons un ensemble $T_i = \{t_i^1, \ldots, t_i^{m_i}\}$ de termes reliés à la requête selon les données de référence. La MAP est calculée de la façon suivante : pour chaque requête q_i, on calcule la précision de sa liste ordonnée de voisins à chacun des rangs où se trouvent les termes reliés T_i, puis on calcule la moyenne de ces précisions (*average precision* ou *AP*) ; la MAP est alors la moyenne des précisions moyennes obtenues pour chaque requête $q_i \in Q$. Si on suppose que la fonction $\mathrm{V}(q, t)$ retourne la liste ordonnée des voisins de la requête q jusqu'au rang du terme relié t, on peut alors formuler la précision moyenne (AP) pour la requête q_i de la façon suivante :

$$\mathrm{AP}(q_i) = \frac{1}{|T_i|} \sum_{j=1}^{|T_i|} \frac{|T_i \cap \mathrm{V}(q_i, t_i^j)|}{|\mathrm{V}(q_i, t_i^j)|} \tag{1}$$

et la MAP est la moyenne des précisions moyennes (AP) pour toutes les requêtes $q_i \in Q$.

4.1 Paramétrisations

Afin d'analyser systématiquement l'influence des (hyper)paramètres, nous avons défini un ensemble de valeurs pour chaque paramètre pris en compte, puis pour chaque combinaison possible de ces valeurs, un modèle a été construit et évalué au moyen des données de référence.

Les paramètres de l'AD comprennent les propriétés de la fenêtre de contexte ainsi que la pondération. D'autres paramètres pourraient être pris en compte, tels que le choix d'une technique de réduction de dimension, mais pour limiter le temps de calcul et la complexité des analyses, nous avons choisi d'analyser quatre paramètres fondamentaux de l'AD, dont trois sont reliés à la fenêtre de contexte. Le premier est la taille de la fenêtre, c'est-à-dire le nombre de mots à gauche et à droite du mot-cible qui sont comptés comme cooccurrents. Le deuxième est le type de fenêtre, qui indique si on observe les cooccurrents à gauche (G) ou à droite (D) des occurrences des mot-cibles, ou les deux, ce qui est généralement le cas ; en l'occurrence, on peut prendre la somme des fréquences observées des deux

9. Levy & Goldberg (2014a) ont montré qu'on peut utiliser la cooccurrence syntaxique plutôt que graphique pour entraîner `word2vec`, et que l'influence du type de contexte est semblable à celle qu'on observe dans le cas de l'AD.

10. `https://code.google.com/p/word2vec/`.

11. En fait, le vocabulaire était constitué de tous les mots ayant une fréquence supérieure ou égale à un certain seuil (par défaut, 5 occurrences), mais nous avons seulement conservé les représentations des 10000 mots-cibles.

$log(w,c) = \log(\#(w,c)+1)$
$MI(w,c) = \log_2\left(\frac{\#(w,c)}{\mathbb{E}[\#(w,c)]}\right)$
$MI^k(w,c) = \log_2\left(\frac{\#(w,c)^k}{\mathbb{E}[\#(w,c)]}\right)$
$local\text{-}MI(w,c) = \log\left(\#(w,c)\cdot MI(w,c)+1\right)$
$simple\text{-}LL(w,c) = \log\left(2\cdot\left(\#(w,c)\cdot\log\left(\frac{\#(w,c)}{\mathbb{E}[\#(w,c)]}\right)-(\#(w,c)-\mathbb{E}[\#(w,c)])\right)+1\right)$
$z\text{-}score(w,c) = \sqrt{\frac{\#(w,c)-\mathbb{E}[\#(w,c)]}{\sqrt{\mathbb{E}[\#(w,c)]}}}$
$t\text{-}score(w,c) = \sqrt{\frac{\#(w,c)-\mathbb{E}[\#(w,c)]}{\sqrt{\#(w,c)}}}$
NOTE : $\mathbb{E}[\#(w,c)] = \frac{\#(w)\times\#(c)}{\lvert D\rvert} = \frac{\sum_{w'\in V_w}\#(w',c)\times\sum_{c'\in V_c}\#(w,c')}{\lvert D\rvert}$

TABLE 1: Pondérations utilisées pour l'AD.

côtés (G+D) ou les encoder séparément (G&D), ce qui double la dimension des représentations de mots. Le troisième est la *forme* de la fenêtre, une fonction qui détermine l'incrément ajouté à M_{ij} pour chaque mot-contexte c_j dans une fenêtre de contexte centrée sur une occurrence particulière de w_i. Par exemple, on dit que la fenêtre est *rectangulaire* si l'incrément est 1 pour tous les contextes ; la fenêtre est dite *triangulaire* si l'incrément est inversement proportionnel à la distance entre w_i et c_j (1 pour les contextes directement adjacents à w_i, $\frac{1}{2}$ pour les suivants, et ainsi de suite).

Le dernier paramètre que nous avons pris en compte est la pondération, une fonction appliquée aux fréquences de cooccurrence dans M. Les pondérations que nous avons testées sont présentées dans la Table 1, les formules étant basées sur celles fournies par Evert (2007, ch. 4) (excepté pour la première pondération, qui est une simple transformation logarithmique), mais l'espérance de la fréquence de cooccurrence étant définie d'une façon légèrement différente [12]. Soulignons qu'une transformation est appliquée à certaines pondérations : nous prenons le logarithme dans le cas de *simple-LL* et la racine carrée dans le cas du *z-score*, suivant Lapesa *et al.* (2014) ; ainsi, la formule de *simple-LL* est en fait le logarithme de la pondération *simple-LL* telle que la définit Evert (2007). Nous appliquons également une transformation aux pondérations *local-MI* et *t-score* (log et racine carrée respectivement). De plus, nous imposons une contrainte de non-négativité : pour toutes les pondérations [13], nous remplaçons le résultat par 0 s'il est négatif.

Nous avons testé toutes les combinaisons possibles des valeurs suivantes de ces quatre paramètres :

— Type de fenêtre de contexte : G+D ou G&D.
— Forme de la fenêtre : rectangulaire ou triangulaire.
— Taille de la fenêtre : 1, 2, 3, 4, 5, 6, 7, 8, 9 ou 10.
— Pondération : *log*, *MI*, MI^2, MI^3, *local-MI*, *simple-LL*, *z-score* ou *t-score*.

Dans le cas de `word2vec` (W2V), nous avons analysé les 5 hyperparamètres qui ont une influence importante sur la qualité des résultats selon la documentation de cet outil [14]. Le premier est l'architecture du modèle ; deux architectures neuronales ont été implémentées dans `word2vec` : *continuous bag-of-words* (CBOW) et *continuous skip-gram*. Le deuxième est l'algorithme d'entraînement : quelle que soit l'architecture, le modèle peut être entraîné au moyen d'un *softmax hiérarchique* ou par

12. Notre définition est celle sur laquelle repose la définition donnée par Levy *et al.* (2015) pour l'information mutuelle.
13. Dans le cas de *simple-LL*, nous remplaçons le résultat par 0 si $\#(w,c) < \mathbb{E}[\#(w,c)]$, suivant Evert (2007, p. 21).
14. `https://code.google.com/p/word2vec/`.

échantillonnage d'exemples négatifs. Le troisième, le sous-échantillonnage de mots fréquents, est une fonction qui supprime du corpus des occurrences de mots dont la fréquence relative est supérieure à un seuil t, la probabilité qu'une occurrence soit supprimée augmentant en fonction de la fréquence du mot ; on recommande pour le seuil t d'utiliser une valeur entre 10^{-5} et 10^{-3}. Les deux derniers hyperparamètres sont la taille de la fenêtre de contexte et la dimension des représentations de mots. Pour chacun de ces hyperparamètres, nous avons testé soit toutes les valeurs possibles, soit des valeurs recommandées dans la documentation ou utilisées dans d'autres travaux. Nous avons testé toutes les combinaisons possibles des valeurs suivantes des 5 hyperparamètres :

— Architecture : CBOW ou skip-gram.
— Algorithme d'entraînement : softmax hiérarchique ou échantillonnage d'exemples négatifs (en l'occurrence, 5 ou 10 exemples).
— Sous-échantillonnage : seuil faible ($t = 10^{-5}$), seuil élevé ($t = 10^{-3}$) ou aucun sous-échantillonnage.
— Taille de la fenêtre de contexte : 1, 2, 3, 4, 5, 6, 7, 8, 9 ou 10.
— Dimension des représentations : 100 ou 300.

Soulignons que la taille de fenêtre est également un paramètre de l'AD. En revanche, les architectures et les algorithmes d'entraînement décrits ci-dessus sont propres à `word2vec`. Le sous-échantillonnage de mots fréquents pourrait être appliqué lors du calcul de l'AD, ainsi qu'une technique de lissage utilisée lors de l'échantillonnage d'exemples négatifs (Levy *et al.*, 2015). Donc, il aurait été possible d'observer l'influence de ces deux techniques dans les deux modèles (ainsi que l'influence de la forme de la fenêtre dans le cas de W2V, par exemple), mais nous avons plutôt choisi de nous pencher sur un ensemble restreint d'(hyper)paramètres typiques pour chacun des deux modèles. D'ailleurs, dans le cas des (hyper)paramètres qui s'appliquent aux deux modèles, il est probable que leur influence serait semblable dans les deux modèles, comme le suggèrent nos résultats en ce qui concerne la taille de fenêtre (voir ci-dessous).

5 Résultats

Nous comparons d'abord la MAP atteinte par les deux modèles distributionnels sur l'ensemble des données de référence (toutes relations confondues). La Figure 1a, qui présente la dispersion de la MAP en fonction du modèle utilisé, montre que l'AD atteint une MAP plus élevée que W2V lorsqu'on explore systématiquement l'espace des (hyper)paramètres des deux modèles ; en effet, l'AD atteint 0.4107 tandis que la MAP maximale de W2V est 0.3819. Si on choisissait une paramétrisation au hasard, la MAP qu'on obtiendrait, en moyenne, avec l'AD (0.3283) est également plus élevée que celle qu'on obtiendrait avec W2V (0.3065).

La Figure 1b présente la dispersion de la MAP en fonction de la relation lexicale ciblée, les deux modèles confondus. Cette figure montre qu'on atteint une MAP beaucoup plus élevée sur les QSYN et les DRV que sur les ANTI ou les HYP, ce qui signifie que les termes reliés qu'on souhaite repérer se trouvent plus près du début des listes ordonnées de voisins distributionnels, en moyenne, lorsqu'on cherche des quasi-synonymes ou des dérivés que lorsqu'on cherche des antonymes, des hyponymes ou des hyperonymes. Cette figure montre également qu'on arrive à modéliser les DRV aussi bien que les QSYN, mais que la dispersion de la MAP est beaucoup plus élevée sur les DRV, ce qui indique que la qualité des résultats qu'on obtient pour cette relation est beaucoup plus sensible au choix et à la paramétrisation du modèle que pour les autres relations.

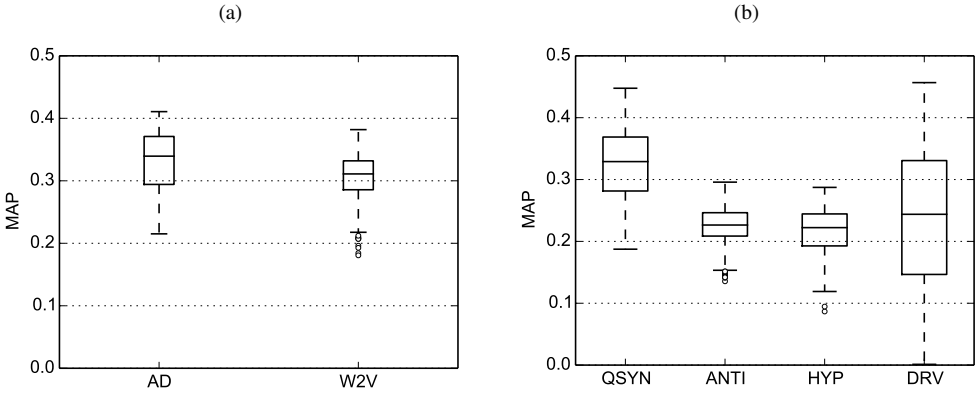

FIGURE 1: Dispersion de la MAP en fonction **(a)** du modèle et **(b)** de la relation.

On peut vérifier dans quelle mesure le choix du modèle détermine la qualité des résultats qu'on obtient pour chaque relation en observant les MAP moyenne et maximale de chaque modèle sur les 4 types de relations lexicales. Ces données, qui sont présentées dans la Table 2, indiquent que l'AD offre de meilleurs résultats que W2V sur les QSYN, les ANTI et les HYP, mais que W2V atteint une MAP plus élevée sur les DRV, l'écart entre les deux modèles étant particulièrement important sur cette relation (plus de 6 points). Ces résultats indiquent que W2V modélise mieux la similarité des mots qui ont le même sens, mais des comportements syntaxiques différents, tandis que l'AD capte mieux la similarité syntaxique. Cette différence serait éventuellement attribuable à la réduction de dimension opérée par W2V [15]. Soulignons qu'un test de Wilcoxon (Hull, 1993) a indiqué que les différences de MAP entre les meilleurs modèles AD et W2V sur chaque jeu de données sont toutes significatives ($p < 0.05$).

Relation	AD	W2V
QSYN	**0.4476** (0.3366)	0.4147 (0.3128)
ANTI	**0.2958** (0.2308)	0.2810 (0.2214)
HYP	**0.2873** (0.2347)	0.2653 (0.2024)
DRV	0.3960 (0.2358)	**0.4567** (0.2311)
TOUTES	**0.4107** (0.3283)	0.3819 (0.3065)

TABLE 2: MAP maximale (et moyenne entre parenthèses) des deux modèles en fonction de la relation.

5.1 Influence des paramètres de l'AD

Dans cette section, nous analysons l'influence des paramètres de l'AD et vérifions dans quelle mesure la paramétrisation optimale varie en fonction des relations lexicales que l'on souhaite identifier. Nous

15. Levy & Goldberg (2014b) ont montré que `word2vec` (plus précisément, le modèle skip-gram entraîné par échantillonnage d'exemples négatifs) factorise implicitement une matrice de cooccurrence comme celles qu'on obtient au moyen de l'AD. Par ailleurs, l'influence des techniques de réduction de dimension appliquées aux modèles distributionnels a été abordée dans plusieurs travaux (Bullinaria & Levy, 2012; Lapesa *et al.*, 2014; Levy *et al.*, 2015, entre autres).

analysons l'influence d'un paramètre en observant la MAP moyenne pour chaque valeur du paramètre, c'est-à-dire la MAP moyenne de toutes les paramétrisations où ce paramètre a cette valeur. Nous observons dans ce cas les valeurs moyennes plutôt que maximales afin de déterminer pour chaque paramètre la valeur qui produit les meilleurs résultats en moyenne, quelles que soient les valeurs utilisées pour les autres paramètres ; dans le cadre de cet article, nous ne traitons pas les interactions entre les paramètres.

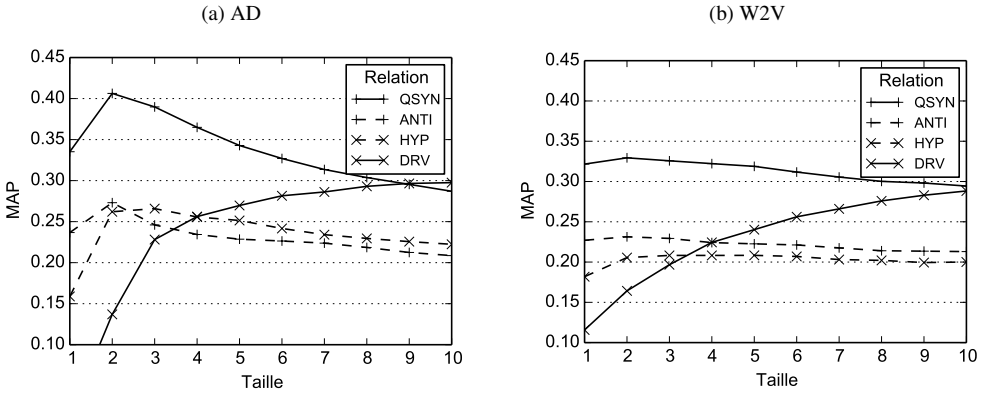

FIGURE 2: MAP moyenne en fonction de la taille de fenêtre.

La Figure 2a illustre l'influence de la taille de fenêtre sur la MAP, en fonction de la relation lexicale ciblée. Cette figure montre que la taille optimale de la fenêtre de contexte est beaucoup plus élevée pour les DRV que pour les autres relations. En effet, la taille optimale est 2 pour les QSYN et les ANTI et 3 pour les HYP, mais dans le cas des DRV, la taille optimale est 10 (ou plus [16]). Ainsi, si on utilise une fenêtre de 2 mots, on modélise bien les quasi-synonymes, mais on obtiendrait une MAP beaucoup plus élevée sur les dérivés en utilisant une fenêtre plus large.

Paramètre	Valeur	QSYN	ANTI	HYP	DRV	TOUTES
Type de fenêtre	G&D	**0.3495**	**0.2312**	**0.2369**	0.2040	0.3268
	G+D	0.3236	0.2305	0.2325	**0.2676**	**0.3297**
Forme de fenêtre	Rect.	0.3065	0.2118	0.2292	**0.2565**	0.3112
	Tri.	**0.3666**	**0.2499**	**0.2402**	0.2151	**0.3453**

TABLE 3: MAP moyenne en fonction du type de fenêtre et de sa forme.

L'influence du type de fenêtre est illustrée dans la Table 3. Comme dans le cas de la taille de fenêtre, la valeur optimale de ce paramètre est différente pour les DRV que pour les autres relations. En effet, la fenêtre G&D produit les meilleurs résultats, en moyenne, pour les QSYN, les ANTI et les HYP, mais on obtient une MAP beaucoup plus élevée sur les DRV avec une fenêtre G+D. Nous proposons que cette différence peut être expliquée de la façon suivante. Beaucoup des DRV dans notre jeu de données sont constitués d'un nom et d'un verbe ayant le même sens et partageant des actants sémantiques ; p. ex. les termes *disparaître* et *disparition* ont un patient, dont les réalisations

16. Il faudrait tester des tailles plus élevées pour vérifier si la moyenne continue à augmenter, mais soulignons que la MAP maximale a été atteinte avec une taille de 8 pour l'AD et de 9 pour W2V.

comprennent des termes tels que *écosystème* et *espèce*. Si la réalisation de cet actant a tendance à apparaître à gauche de l'un des deux termes, mais à droite de l'autre, on ne modélisera pas bien le fait que *disparaître* et *disparition* partagent un actant si les fréquences de cooccurrence observées à gauche et à droite de ces deux termes sont encodées séparément, c'est-à-dire si on utilise une fenêtre G&D. Cela expliquerait vraisemblablement le fait que la fenêtre G+D produit de meilleurs résultats pour les DRV.

En ce qui concerne la forme de la fenêtre de contexte, on observe également que la valeur optimale de ce paramètre est différente pour les DRV que pour les autres relations lexicales, comme le montre la Table 3. En effet, la MAP qu'on obtient sur les DRV est plus élevée, en moyenne, lorsqu'on utilise une fenêtre rectangulaire, alors que c'est la fenêtre triangulaire qui produit les meilleurs résultats pour les autres relations. Comme le montrait la Figure 2a, on modélise mieux les DRV en utilisant une fenêtre de contexte large plutôt qu'étroite, donc il est plus important de prendre en compte des contextes éloignés dans le cas des DRV. Or, une fenêtre rectangulaire accorde plus de poids aux contextes éloignés qu'une fenêtre triangulaire, la différence étant d'autant plus importante que la distance est élevée. Cela expliquerait éventuellement pourquoi la fenêtre rectangulaire produit de meilleurs résultats pour les DRV, du moins en partie.

Pondération	QSYN	ANTI	HYP	DRV	TOUTES
aucune	0.1817	0.1455	0.1078	0.1318	0.1739
log	0.2890	0.2017	0.2178	0.1930	0.2733
MI	0.3199	0.2321	0.2362	**0.2916**	0.3367
MI^2	0.3190	0.2069	0.2321	0.1967	0.3023
MI^3	0.3217	0.2119	0.2278	0.1982	0.3017
local-MI	0.3692	0.2486	0.2425	0.2416	0.3533
simple-LL	**0.3710**	**0.2516**	**0.2457**	0.2540	**0.3621**
t-score	0.3566	0.2491	0.2440	0.2436	0.3457
z-score	0.3460	0.2448	0.2316	0.2678	0.3510

TABLE 4: MAP moyenne en fonction de la pondération.

Enfin, la Table 4 illustre l'influence de la pondération en fonction de la relation ciblée ; nous avons inclus dans cette table les résultats qu'on obtient sans pondération, qui montrent l'importance d'appliquer une pondération quelconque aux fréquences de cooccurrence, mais il est important de noter que ces résultats ne sont pas pris en compte dans le reste de l'analyse présentée dans cet article. On observe dans la Table 4 que la valeur optimale de ce paramètre est encore une fois différente pour les DRV que pour les autres relations. La pondération optimale pour les DRV est l'information mutuelle (*MI*), tandis que c'est *simple-LL* qui produit les meilleurs résultats pour les autres relations, en moyenne. L'information mutuelle accorde plus de poids aux mots-contextes dont la fréquence est faible que d'autres pondérations telles que *simple-LL* [17], et cela semble avoir un effet favorable dans le cas des DRV, mais pas pour les autres relations.

Ainsi, pour tous les paramètres de l'AD que nous avons pris en compte, la valeur qui produit les meilleurs résultats sur les DRV est différente de la valeur optimale pour les QSYN et les autres relations liant des mots de la même partie du discours. L'analyse présentée ci-dessus (qui ne tient

17. Cette propriété de l'information mutuelle peut être atténuée au moyen du lissage de la distribution des contextes (Levy *et al.*, 2015).

pas compte des interactions entre les paramètres) suggère que la paramétrisation optimale pour les DRV est une fenêtre G+D rectangulaire de 10 mots (ou plus) et la pondération *MI*, tandis que la paramétrisation optimale pour les autres relations est une fenêtre G&D triangulaire de 2 ou 3 mots et la pondération *simple-LL*.

5.2 Influence des hyperparamètres de word2vec

Nous analysons dans cette section l'influence des hyperparamètres de W2V comme nous l'avons fait pour les paramètres de l'AD à la section précédente.

La Figure 2b présentée ci-dessus illustre l'influence de la taille de fenêtre sur la MAP de W2V, en fonction de la relation ciblée. Cette figure montre que la taille de fenêtre optimale est 2 pour les QSYN et les ANTI, la MAP diminuant lentement à mesure qu'on élargit la fenêtre. Pour les HYP, la MAP plafonne lorsque la taille de fenêtre atteint 3. En revanche, la MAP qu'on obtient sur les DRV augmente rapidement à mesure que la taille de fenêtre augmente, et on ne semble pas avoir atteint un maximum même avec une taille de 10 mots. Comme dans le cas de l'AD, si on utilise la taille optimale pour les QSYN, la MAP qu'on obtient sur les DRV est beaucoup plus faible que si on utilisait une fenêtre plus large. Les résultats qu'on obtient sur les QSYN semblent être moins sensibles à la taille de fenêtre que dans le cas de l'AD, de sorte que la MAP ne baisse pas beaucoup si on utilise une fenêtre plus large afin d'améliorer les résultats sur les DRV. Cela est peut-être lié aux propriétés de la fenêtre de contexte implémentée dans `word2vec` : celle-ci a une taille *dynamique* (elle est tirée au hasard entre 1 et la taille précisée par l'utilisateur pour chaque occurrence de mot) et sa forme est définie par la fonction $f(d) = m - (d-1)/m$, où d est la distance entre un cooccurrent donné et le mot-cible, et m est la taille de la fenêtre (Levy *et al.*, 2015).

Hyperparamètre	Valeur	QSYN	ANTI	HYP	DRV	TOUTES
Architecture	Skip-gram	0.3008	**0.2248**	**0.2048**	**0.2732**	**0.3128**
	CBOW	**0.3247**	0.2180	0.1999	0.1889	0.3003
Seuil (sous-échantillonnage)	Aucun	0.3357	**0.2328**	0.1998	0.1065	0.2828
	Faible	0.2588	0.2051	0.1915	**0.3717**	0.3121
	Élevé	**0.3438**	0.2263	**0.2158**	0.2150	**0.3247**

TABLE 5: MAP moyenne en fonction de l'architecture et du seuil pour le sous-échantillonnage.

L'influence de l'architecture est présentée dans la Table 5. Ces résultats indiquent que l'architecture CBOW produit les meilleurs résultats, en moyenne, sur les QSYN ; en revanche, l'architecture skip-gram produit des résultats beaucoup meilleurs sur les DRV et, dans une moindre mesure, sur les ANTI et les HYP.

Un autre hyperparamètre dont la valeur optimale varie selon la relation ciblée est le seuil (t) pour le sous-échantillonnage des mots fréquents. Les données présentées dans la Table 5 montrent que pour les QSYN et les HYP, la qualité des résultats augmente légèrement si on applique le sous-échantillonnage avec un seuil élevé ($t = 10^{-3}$), mais diminue si on utilise un seuil plus faible ($t = 10^{-5}$), particulièrement dans le cas des QSYN. En revanche, dans le cas des DRV, le seuil faible donne de meilleurs résultats que le seuil élevé ; de plus, la MAP qu'on obtient sur les DRV semble extrêmement sensible à la valeur de cet hyperparamètre, car on observe un gain de presque 250% en

appliquant le sous-échantillonnage (avec seuil faible). Pour ce qui est des ANTI, on n'améliore pas la qualité des résultats en appliquant le sous-échantillonnage.

Ainsi, l'influence de 3 des hyperparamètres que nous avons pris en compte varie considérablement en fonction du type de relation lexicale ciblée. En ce qui concerne l'algorithme d'entraînement et la dimension des représentations, la valeur optimale de ces deux hyperparamètres ne varie pas beaucoup en fonction de ce facteur. L'échantillonnage d'exemples négatifs produit de meilleurs résultats, en moyenne, que le softmax hiérarchique, et on obtient généralement une légère augmentation de la MAP en utilisant 10 exemples plutôt que 5. Quant à la dimension des représentations, une dimension de 300 produit de meilleurs résultats qu'une dimension de 100, excepté sur les ANTI, qu'on modélise légèrement mieux avec une dimension de 100.

6 Conclusion

Nous avons présenté dans cet article une partie des résultats d'un travail visant à caractériser l'apport des méthodes distributionnelles à la lexicographie spécialisée, en analysant l'influence des relations lexicales ciblées sur le choix et la paramétrisation du modèle distributionnel. Nous avons montré que les modèles distributionnels captent la dérivation syntaxique aussi bien que la (quasi-)synonymie, et mieux que d'autres relations paradigmatiques, mais que les modèles qui offrent les meilleurs résultats pour ces deux relations sont très différents. D'abord, les résultats que nous avons obtenus en comparant une technique classique d'analyse distributionnelle et un modèle de langue neuronal indiquent que ce dernier modélise mieux la dérivation syntaxique, mais que l'analyse distributionnelle offre de meilleurs résultats pour les relations paradigmatiques reliant deux mots appartenant à la même partie du discours (quasi-synonymie, antonymie, hyperonymie et hyponymie). Nous avons montré que la paramétrisation optimale des deux modèles dépend également de la relation ciblée, les différences les plus importantes étant celles entre la dérivation syntaxique et les autres relations que nous avons prises en compte. Nous avons ainsi fourni des balises pour l'utilisation de méthodes distributionnelles en lexicographie. Par ailleurs, bien que nous avons seulement traité le cas du français dans cet article, nous avons observé des tendances très similaires en anglais.

Soulignons qu'il serait intéressant de vérifier si on peut augmenter la qualité des résultats en combinant différents modèles, une possibilité que nous avons explorée dans un autre travail (Bernier-Colborne & Drouin, 2016). Une autre possibilité consisterait à identifier d'abord des vecteurs correspondant à chacune des relations ciblées, comme le font Mikolov *et al.* (2013c) pour résoudre des analogies syntaxiques et sémantiques, puis à exploiter ces vecteurs pour identifier des paires de termes participant à des relations particulières.

Remerciements

Ce travail a bénéficié du soutien financier du Conseil de recherches en sciences humaines (CRSH) du Canada. Nous remercions également les relecteurs anonymes ainsi que Vincent Claveau pour leurs commentaires utiles.

Références

BARONI M., DINU G. & KRUSZEWSKI G. (2014). Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 238–247, Baltimore, Maryland : ACL.

BARONI M. & LENCI A. (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, p. 1–10 : Association for Computational Linguistics.

BERNIER-COLBORNE G. & DROUIN P. (2016). Combiner des modèles sémantiques distributionnels pour mieux détecter les termes évoquant le même cadre sémantique. In *Actes de la 23ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Paris.

BULLINARIA J. A. & LEVY J. P. (2012). Extracting semantic representations from word co-occurrence statistics : stop-lists, stemming, and SVD. *Behavior research methods*, **44**(3), 890–907.

EVERT S. (2007). Corpora and collocations (extended manuscript). In A. LÜDELING & M. KYTÖ, Eds., *Corpus Linguistics. An International Handbook*, volume 2. Berlin/New York : Walter de Gruyter.

FABRE C., HATHOUT N., SAJOUS F. & TANGUY L. (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *Actes de l'atelier SemDis, TALN 2014 (Traitement automatique des langues naturelles)*, p. 266–279, Marseille : ATALA LPL.

FERRET O. (2015). Réordonnancer des thésaurus distributionnels en combinant différents critères. *Traitement automatique des langues*, **56**(2), 21–49.

FIRTH J. R. (1957). A synopsis of linguistic theory 1930–1955. In THE PHILOLOGICAL SOCIETY, Ed., *Studies in Linguistic Analysis*, p. 1–32. Oxford : Blackwell.

HARRIS Z. S. (1954). Distributional structure. *Word*, **10**(2–3), 146–162.

HULL D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, p. 329–338, New York, NY, USA : ACM.

KIELA D. & CLARK S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, p. 21–30 : Association for Computational Linguistics.

LAPESA G., EVERT S. & SCHULTE IM WALDE S. (2014). Contrasting syntagmatic and paradigmatic relations : Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, p. 160–170, Dublin, Ireland : ACL/DCU.

LEVY O. & GOLDBERG Y. (2014a). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, p. 302–308.

LEVY O. & GOLDBERG Y. (2014b). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, p. 2177–2185.

LEVY O., GOLDBERG Y. & DAGAN I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, **3**, 211–225.

L'HOMME M.-C. (2005). Sur la notion de "terme". *Meta*, **50**(4), 1112–1132.

MEL'ČUK I. A., CLAS A. & POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Duculot.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26 (NIPS)*, p. 3111–3119. Curran Associates, Inc.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 746–751, Atlanta, Georgia : Association for Computational Linguistics.

MOSKOWICH W. & CAPLAN R. (1978). Distributive-statistical text analysis : A new tool for semantic and stylistic research. In G. ALTMANN, Ed., *Glottometrika*, p. 107–153. Bochum : Studienverlag Dr. N. Brockmeyer.

PROKOPIDIS P., PAPAVASSILIOU V., TORAL A., RIERA M. P., FRONTINI F., RUBINO F. & THURMAIR G. (2012). *Final Report on the Corpus Acquisition & Annotation subsystem and its components*. Rapport interne WP-4.5, PANACEA Project.

PÉRINET A. & HAMON T. (2014). Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité. In *Actes des 12es Journées d'analyse statistique des données textuelles (JADT 2014)*, p. 507–518.

SAHLGREN M. (2006). *The word-space model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

TANGUY L., SAJOUS F. & HATHOUT N. (2015). Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques. *Traitement automatique des langues*, **56**(2), 103–127.

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