Traitement automatique des langues

Sémantique distributionnelle

sous la direction de
Cécile Fabre
Alessandro Lenci

Vol. 56- n°2/ 2015

ATALA

Sémantique distributionnelle

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
56

n°2
2015

Sémantique distributionnelle

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS, de l'Université Paris VII et de l'Université de Provence

ISSN 1965-0906

http ://www.atala.org/-Revue-TAL-

Traitement automatique des langues

Traitement automatique des langues

Volume 56 – n°2/2015

SEMANTIQUE DISTRIBUTIONNELLE

Table des matières

Distributional Semantics Today
Introduction to the special issue

Cécile Fabre* — **Alessandro Lenci****

* *University of Toulouse, CLLE-ERSS*
** *University of Pisa, Computational Linguistics Laboratory*

ABSTRACT. *This introduction to the special issue of the* TAL *journal on distributional semantics provides an overview of the current topics of this field and gives a brief summary of the contributions.*

RÉSUMÉ. *Cette introduction au numéro spécial de la revue* TAL *consacré à la sémantique distributionnelle propose un panorama des thèmes de recherche actuels dans ce champ et fournit un résumé succinct des contributions acceptées.*

KEYWORDS: *Distributional Semantic Models, vector-space models, corpus-based semantics, semantic proximity, semantic relations, evaluation of distributional ressources.*

MOTS-CLÉS : *Sémantique distributionnelle, modèles vectoriels, proximité sémantique, relations sémantiques, évaluation de ressources distributionnelles.*

1. Introduction

Distributional Semantic Models (DSMs) have been the focus of considerable research over the past twenty years. The use of distributional information extracted from corpora to compute semantic similarity between words has become a very common method in NLP. Its popularity is easily explained. In distributional models, the meaning of words is estimated from the statistical analysis of their contexts, in a bottom-up fashion: requiring no sources of knowledge other than corpus-derived information about word distribution in contexts, it provides access to semantic content on the basis of an elementary principle which states that semantic proximity can be inferred from proximity of distribution. It gives access to meaning in usage, as it emerges from word occurrences in texts. Distributional semantics based on vector space models has benefited from the availability of massive amounts of textual data and increased computational power, allowing for the application of these methods on a large scale. Today, the field has reached maturity: many experiments have been carried out on different languages, several survey articles have helped to consolidate the concepts and procedures used for distributional computations (Turney and Pantel, 2010; Baroni and Lenci, 2010), various distributional models and evaluation data are now available. Still, many issues remain open to gain a better understanding of the type of information that is induced by these methods and to extend their use to new applications and new linguistic phenomena.

In recent years, much research effort has focused on optimization methods to handle massive corpora and on the adjustment of the many parameters that are likely to have impact on the quality and nature of semantic relations. A second important issue relates to the relevance of distributional semantic information for a large number of tasks and applications. Finally, in the last few years, research has also focused on a better understanding of the semantic information that is conveyed by these models. Before presenting the papers that appear in this special issue of the *TAL* journal dedicated to distributional semantics, this introduction provides an overview on these different topics.

2. Principles and Methodology of the Construction of DSMs

2.1. *The Distributional Hypothesis*

Distributional semantics is grounded on the Distributional Hypothesis: *similarity of meaning correlates with similarity of distribution*. Zellig Harris is usually referred to as the theoretical and methodological source of research for distributional semantic models (Harris, 1954). In fact, he considered distributional method the only viable scientific approach to the study of linguistic meaning. In his later works, he designed a method to classify words on the basis of the contexts they share in a given corpus, through the careful collection and analysis of dependency relations involving operators and arguments (Harris, 1991). What was clearly asserted in Harris' original method was the fact that such inductive semantic classifications reflected the

use of words in specific corpora. The approach was set in the context of the theory of sublanguages, based on the assumption that only corpora from restricted domains could guarantee the possibility to build up clear-cut semantic categories (Habert and Zweigenbaum, 2002).

Since the 1960s, several implementation of the Distributional Hypothesis have been carried out for the automatic constructions of thesauri for machine translation and information retrieval (Grefenstette, 1994). A crucial contribution to distributional semantics has indeed come from the vector space model in information retrieval (Salton *et al.*, 1975), resulting in successive improvements to the original methodology with respect to the nature of data and the mathematical formalization, thereby boosting its spread in computational linguistics. In the last twenty years, the possibility to apply the method on a much larger scale to huge corpora has imposed the distributional approach as the default approach to semantics in NLP.

2.2. *Design of Distributional Semantic Models*

In DSMs, words are represented as vectors built from their distribution in contexts, and similarity between words is approximated in terms of geometric distance between their vectors. The standard organization of DS systems is usually described as a four-step method (Turney and Pantel, 2010): for each target word, contexts are collected and counted and a co-occurrence method is generated; raw frequencies are then usually transformed into significance scores that are more suitable to reflect the importance of the contexts; the resulting matrix tends to be very large and sparse, requiring techniques to limit the number of dimensions. Finally, a similarity score is computed between the vector rows, using various similarity measures. DSMs have many design options, due to the variety of parameters that can be set up at each step of the process and may affect the results and performances of the system.

2.2.1. *Parameters*

A corpus-based semantic model reflects the semantic behaviour of words in use. It is thus by definition highly dependent on the type of corpus that is being analyzed. There has been a clear shift from the treatment of middle-sized specialized corpora for the acquisition of distributional thesauri in the 90's (Grefenstette, 1994; Nazarenko *et al.*, 2001), to the compilation of corpora as large as possible, often heterogeneous in genre and domain. Newspaper and encyclopedic articles (Peirsman and Geeraerts, 2009), balanced reference corpora such as the BNC (Sadrzadeh and Grefenstette, 2011), very large corpora obtained from the web (Agirre *et al.*, 2009), or any combination of the former (Baroni and Lenci, 2010) have been used. The trend to use huge corpora is mainly motivated by the joint need of increasing the coverage of distributional lexical resources while reducing data-sparseness, which is known to negatively affect the performance of DSMs.

The definition of contexts is another crucial parameter in the implementation of the systems. Three types of linguistic environments have been considered (Peirsman and Geeraerts, 2009): in document-based models, as in *Latent Semantic Analysis* (LSA) (Landauer and Dumais, 1997), words are similar if they appear in the same documents or in the same paragraphs; word-based models consider a "bag-of-word" window of collocates around the target words (Lund and Burgess, 1997; Sahlgren, 2008; Ferret, 2013); syntax-based models are closer to Harris' approach as they compare words on the basis of their dependency relations (Curran, 2004; Padó and Lapata, 2007; Baroni and Lenci, 2010). Word-based models have an additional parameter represented by the window size (from a few words to an entire paragraph), while syntax-based models need to specify the type of dependency relations that are selected as contexts (Baroni and Lenci, 2010; Peirsman *et al.*, 2007). Some experiments suggest that syntax-based models tend to identify distributional neighbors that are taxonomically related, mainly co-hyponyms, whereas word-based models are more oriented towards identifying associative relations (Van de Cruys, 2008; Peirsman *et al.*, 2007; Levy and Goldberg, 2014). However, the question whether syntactic contexts provide a real advantage over "bag-of words" models is still open. On the other hand, a more dramatic difference exists with respect to document-based models, which are strongly oriented towards neighbors belonging to loosely defined semantic topics or domains (Sahlgren, 2006).

Other parameters have received particular attention: weighting scores and similarity measures. A wide range of setting exists for both parameters (Curran, 2004; Bullinaria and Levy, 2007), but nowadays the most common practice is to use Positive Pointwise Mutual Information as weighting scheme and cosine as similarity measure, which are typically credited for granting the best performances across a wide range of tasks (Turney and Pantel, 2010).

Vectors in the co-occurrence matrix provide an *explicit* representation (Levy and Goldberg, 2014) of the lexeme distribution in contexts. Each vector dimension in fact represents a specific context in which the target word has been observed. Explicit co-occurrence vectors are huge and sparse. Techniques are therefore used to reduce their dimension and limit computational complexity. The most common approach consists in mapping the original sparse matrix into a low-dimensional dense matrix with methods such as Singular Value Decomposition (Landauer and Dumais, 1997), Non-Negative Matrix Factorization (Van de Cruys, 2010), and Latent Dirichlet Allocation (Blei *et al.*, 2003). Crucially, the dimensions of the reduced vectors no longer correspond to explicit contexts, but rather to "latent" semantic dimensions implicit in the original distributional data. Matrix reduction techniques smooth unseen data, remove noise and exploit redundancies and correlations between the linguistic contexts, thereby improving the quality of the resulting semantic space (Turney and Pantel, 2010). A popular as well as effective alternative to matrix reduction is Random Indexing (Sahlgren, 2006): instead of reducing a previously constructed matrix, low-dimensional representations are incrementally built by assigning each word a random vector that is summed to the vectors of the co-occurring words.

Much research has been dedicated to the investigation of the impact of some or all these parameters on the performance of DSMs systems in a variety of tasks. The most recent and comprehensive studies are those of Lapesa and Evert (2014) and Kiela and Clark (2014). They investigate a very large set of parameters, including type of corpus, use of stemming and lemmatization, type of contexts (dependency vs co-occurrence, direction and size of the window), weighting scores, similarity measures, dimensionality reduction techniques. These experiments provide a very useful presentation of the best configurations according to the type of semantic task.

2.2.2. *Count vs. Prediction Models*

The DSMs we have just described use a *count-based* approach to build distributional representations: corpus co-occurrences are first counted, then weighted and finally optionally reduced to dense vectors. Recently, a new family of *prediction-based* DSMs has appeared: neural network algorithms directly create dense, low-dimensional word representations by learning to optimally predict the contexts of a target word (Mikolov *et al.*, 2013a). These representations are also typically referred to as *embeddings*, because words are embedded into a low-dimensional linear space of latent features. Various types of "linguistic regularities" have been claimed to be identifiable by embeddings (Mikolov *et al.*, 2013b). For instance, the fact that *king* and *queen* have the same gender relation as *man* and *woman* is represented in their embeddings offsets, so that the vector of one word (e.g. $queen$) can be recovered by the representations of the other words by simple vector arithmetics (i.e., $king - man + woman$). Moreover, prediction-based models have been shown to outperform count-based ones in various semantic tests (Baroni *et al.*, 2014)

Despite their increasing popularity, the question whether embeddings are really a breakthrough with respect to more traditional methods is far from being set. For instance, the same linguistic regularities captured by embeddings are also captured by explicit count-based models (Levy and Goldberg, 2014). When parameters of the latter are carefully tuned, no significant difference is observed in the performance between count and prediction-based models (Levy *et al.*, 2015). It is possible that future research will able to show some clear advantage for embeddings, but for the time being the two approaches do not substantially differ for the type of semantic aspects they are able to address. They are just alternative ways to build distributional representations.

3. Evaluation of DSMs

The classical dichotomy between *intrinsic* and *extrinsic* modes of evaluation in NLP applies to DSMs as well. Intrinsic evaluations aim at measuring the quality of the resource in itself, by confronting it with human evaluation or with similar semantic resources that can be used as gold standards. Extrinsic evaluations measure the specific contribution of the resource to enhance the performance of a system in which it is integrated.

The intrinsic evaluation of the DSMs has been conducted through the comparison to various lexical resources, such as the TOEFL synonym detection task (Landauer and Dumais, 1997), specialized thesauri (Grefenstette, 1994), wordnets (Curran and Moens, 2002; Padró *et al.*, 2014; Anguiano and Denis, 2011), synonym dictionaries (Van der Plas *et al.*, 2011). Intrinsic evaluation of DSMs is a complex issue for various reasons. First of all, DSMs capture a very broad notion of semantic proximity (cf. section below). Therefore, there is an inevitable mismatch between DSMs results and resources that focus on specific, classical lexical relations, such as synonymy dictionaries, thesauri and wordnets. A second kind of potential mismatch is due to the fact that DSMs results reflect the specificities of the corpus and as such they can identify potentially relevant semantic relations and yet missing in general-purpose resources. It is indeed difficult, perhaps impossible, to assess the validity of a semantic relation out of context (Muller *et al.*, 2014). In order to overcome such limitations, a number of resources specifically geared towards DSM evaluation have been developed, mostly for English. One of the most popular gold standard is WordSim-353 (Finkelstein *et al.*, 2002), with 353 word pairs rated by human judgments. A multilingual version of this dataset has also been recently released (Leviant and Reichart, 2015).

Regarding extrinsic evaluation, the use of distributional features is useful each time there is a need to compute similarities between words or longer stretches of text. Several experiments have been dedicated to the use of distributional resources in information retrieval to compute query similarity (Alfonseca *et al.*, 2009; Claveau and Kijak, 2015). In the lexical substitution task (McCarthy and Navigli, 2007), a DSM is used to compute potential substitutes before the disambiguation process (Fabre *et al.*, 2014). Distributional similarity is also used as a cue to determine the predominant sense of a word in a corpus (McCarthy *et al.*, 2007). DSMs have proved efficient in even more complex NLP applications such as textual entailment or summarization (Cheung and Penn, 2013). Word embeddings have also been successfully used to improve Semantic Role Labeling and Named Entity Recognition (Collobert and Weston, 2008).

4. The challenges for DSMs

Critics have been regularly addressed to DSMs, even by researchers involved in the field: the bottom-up approach to meaning pursued by distributional semantics is very practical in terms of processing, but it is an open issue whether statistical co-occurrences alone are enough to address deeper semantic questions or just provide a shallow proxy of lexical meaning (Sahlgren, 2008; Lenci, 2008; Koller, 2015).

The Distributional Hypothesis is a claim about semantic similarity, which DSMs measure with proximity in vector spaces. However, semantic similarity is itself a very vague notion, ranging from similarity between words to similarity between relations (Turney, 2006; Baroni and Lenci, 2010; Turney, 2013). It is also necessary to distinguish semantic similarity *stricto sensu* (also called *attributional similarity*), as a relation between words sharing similar semantic features, such as as *car* and *van*, from

the *semantic relatedness* of words that are strongly associated, like *car* and *wheel* (Budanitsky and Hirst, 2006; Agirre *et al.*, 2009). These two types of similarities have very different semantic properties. Yet, they are hardly distinguished by DSMs. Even gold standards like WordSim-353 are populated with semantically related pairs (Agirre *et al.*, 2009). In order to address this issue, the dataset SimLex-999 has been recently developed in order to specifically evaluate DSMs' ability to capture semantic similarity rather than semantic relatedness (Hill *et al.*, forthcoming).

An additional problem is that both semantic similarity and semantic relatedness are cover terms for very different types of lexical relations. For instance, both synonyms, co-hyponyms and even antonyms can be said to be semantically similar because they share a high number of features. Semantic relatedness includes meronymy, locative relations, up to topical and other non-classical relations (Morris and Hirst, 2004). This large and graded notion of relatedness is both useful and problematic for NLP applications, because it is very difficult to draw a clear limit between relevant and non-relevant associates (Sahlgren, 2008; Ferret, 2013). In general, the distributional neighbors identified by DSMs have very different semantic relations with the target, suggesting that DSM provide quite a coarse-grained representation of lexical meaning. The BLESS (Baroni and Lenci, 2011) and the most recent EVALution (Santus *et al.*, 2015) datasets have specifically been designed to test the ability of DSMs to discriminate different types of relations, which represents an important area of research in distributional semantics (Van der Plas and Tiedemann, 2006; Lenci and Benotto, 2012; Santus *et al.*, 2014; The Pham *et al.*, 2015).

One important issue it to determine what type of semantic information can be grasped on the basis of contextual properties, and what part of the meaning of words remains unreachable without complementary knowledge. Recent works focus on this question: Gupta *et al.* (2015) show that referential information is accessible, while results from Herbelot and Ganesalingam (2013) suggest that informativeness (discriminating between more or less contentful words) is difficult to assess on the basis of distributional information. Zarcone *et al.* (2015) show that not only the argument thematic fit to a predicate but also semantic type constraints can be approximated by DSMs to model complement coercion.

In a similar perspective, recent works propose to connect formal and distributional semantics (Guevara, 2011; Grefenstette, 2013), so as to combine the capacity of DSMs to provide semantic representations of word meanings (Erk, 2013; Boleda and Erk, 2015) and the capacity of formal models to account for semantics at the level of complex structures. Compositionality issues have been the focus of many research studies: until recently, most works on DSMs have been concerned with words in isolation, but in the last few years research has been conducted on the extension of these models to process larger semantic units such as phrases and sentences. Two approaches can be considered. The first one consists in taking into account phrases in addition to words as the basic processing units, as did Baldwin *et al.* (2003) in the LSA framework. In this issue, the paper by Périnet and Hamon follows this orientation. Yet this remains a very minority approach, because of the sparsity of data when one con-

siders the distribution of complex units. The second option has generated a large bulk of research. It consists in modeling semantic compositionality within a distributional framework, under the assumption that semantic information about phrases can be computed by combining information about its components. The work by Mitchell and Lapata (2010) proposes a thorough account and evaluation of the combination functions that can be used. Very recently, a task has been proposed on compositional semantics at SemEval (Marelli *et al.*, 2014). Different types of units have been examined, such as Adjective-Noun (Baroni and Zamparelli, 2010),Verb-Noun (Mitchell and Lapata, 2010), sentences. Another area of research concerns the integration of extralinguistic features to complement distributional information with other sources of information, in multimodal models (Bruni *et al.*, 2014).

5. Conclusions and presentation of the papers

Distributional semantics is a young paradigm, but despite its short history we can reliably state that it has been able to gain a large credibility in NLP community and beyond, with increasing interest in cognitive and linguistic research. As shown in this short review, the variety of DSMs is expanding fast, but even more importantly, we have been gaining a much deeper understanding of the effects of their various parameters. The number of semantic tasks that are now addressed by these models has constantly increased, going well beyond the original application of the Distributional Hypothesis to synonym identification. Of course lots of challenges still lie ahead. Under many respects, DSMs still provide a very coarse-grained representation of meaning, and their actual limits and potentialities need to be explored. All this makes distributional semantics a very lively and fascinating research field, as confirmed by the contributions in this special issue.

The four papers published in this issue address a large proportion of the topics we have just listed, such as parameter tuning, evaluation, compositionality or processing of larger lexical units. It is interesting to note that they depart from the dominant trend that consider huge corpora to build distributional models, as three papers out of four are concerned with the treatment of specialized corpora for knowledge acquisition. Two papers are dealing with complex terms, but in a different perspective. Périnet and Hamon ("Analyse distributionnelle appliquée aux textes de spécialité") apply the distributional method to complex terms and propose a solution to normalize the contexts to deal with the problem of sparsity of data. Daille and Hazem ("Méthode semi-compositionnelle pour l'extraction de synonymes des termes complexes") use distributional semantics to generate synonyms of multi-word expressions by leveraging the compositionality properties of these terms. The two other papers focus on the evaluation and improvement of distributional models. Tanguy, Sajous and Hathout's experiment ("Évaluation sur-mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques") is also based on the treatment of a specialized corpus. They use a specifically designed evaluation dataset to define the best parameters for their distributional model, focus-

ing on the contribution of accurate syntactic information. Ferret's paper ("'Combiner différents critères pour améliorer les thésaurus distributionnels") proposes a way to improve a distributional thesaurus by using a bootstrapping method based on the automatic selection of positive and negative examples of semantic neighbors. The selection procedure takes advantage of the symmetry of the semantic relations, and of the compositionality of compounds.

Acknowledgements

We want to thank the *TAL* journal editors and committee as well as the specific scientific committee. We are particularly grateful to the reviewers for their time and effort to improve this special issue.

Specific scientific committee :

– Marianna Apidianaki – LIMSI, Orsay
– Marco Baroni – CIMeC, Trento
– Ann Bertels – ILT, K.U. Leuven
– Romaric Besançon – CEA, Gif-sur-Yvette
– Yves Bestgen – UCL/CECL, Louvain-La-Neuve
– Gemma Boleda – Université Pompeu Fabra, Barcelone
– Marie Candito – ALPAGE, Paris
– Georgiana Dinu – CIMeC, Trento
– Olivier Ferret – CEA, Gif-sur-Yvette
– Andre Freitas – DERI, National University of Ireland, Galway
– Gregory Grefenstette – INRIA, Saclay
– Thierry Hamon – LIMSI, Paris
– Aurélie Herbelot – Institut für Linguistik, Potsdam
– Mai Ho-Dac – CLLE-ERSS, Toulouse
– Guillaume Jacquet – European Commission, JRC, Ispra
– Olivier Kraif – LIDILEM, Grenoble
– François Morlane-Hondère – LIMSI, Paris
– Yves Peirsman – Leuven
– Laurent Prévot – LPL, Aix-Marseille
– Benoît Sagot – ALPAGE, Paris
– Magnus Sahlgren – Gavagai, Inc., Sweden
– Franck Sajous – CLLE-ERSS, Toulouse
– Sabine Schulte im Walde – IMS, Stuttgart
– Peter Turney – National Research Council Canada, Ottawa
– Tim Van de Cruys – IRIT, Toulouse

6. References

Agirre E., Alfonseca E., Hall K., Kravalova J., Paşca M., Soroa A., "A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches", *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 19-27, 2009.

Alfonseca E., Hall K., Hartmann S., "Large-scale computation of distributional similarities for queries", *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Association for Computational Linguistics, p. 29-32, 2009.

Anguiano E. H., Denis P., "FreDist: Automatic construction of distributional thesauri for French", *Actes de la 18e conférence sur le traitement automatique des langues naturelles – TALN*, p. 119-124, 2011.

Baldwin T., Bannard C., Tanaka T., Widdows D., "An empirical model of multiword expression decomposability", *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, Association for Computational Linguistics, p. 89-96, 2003.

Baroni M., Dinu G., Kruszewski G., "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, p. 238-247, 2014.

Baroni M., Lenci A., "Distributional memory: A general framework for corpus-based semantics", *Computational Linguistics*, vol. 36, no 4, p. 673-721, 2010.

Baroni M., Lenci A., "How we BLESSed distributional semantic evaluation", *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Association for Computational Linguistics, p. 1-10, 2011.

Baroni M., Zamparelli R., "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space", *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 1183-1193, 2010.

Blei D. M., Ng A. Y., Jordan M. I., "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol. 3, p. 993-1022, 2003.

Boleda G., Erk K., "Distributional semantic features as semantic primitives – or not", *AAAI Spring Symposium on Knowledge Representation and Reasoning*, Stanford University, USA, 2015.

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

Budanitsky A., Hirst G., "Evaluating wordnet-based measures of lexical semantic relatedness", *Computational Linguistics*, vol. 32, no 1, p. 13-47, 2006.

Bullinaria J., Levy J. P., "Extracting semantic representations from word co-occurrence statistics: A computational study", *Behavior Research Methods*, vol. 39, p. 510-526, 2007.

Cheung J. C. K., Penn G., "Probabilistic Domain Modelling With Contextualized Distributional Semantic Vectors.", *Association for Computational Linguistics (ACL)*, p. 392-401, 2013.

Claveau V., Kijak E., "Thésaurus distributionnels pour la recherche d'information et vice-versa", *Actes de la 13e Conférence en Recherche d'Information et Applications (CORIA)*, 2015.

Collobert R., Weston J., "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning", *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, p. 160-167, 2008.

Curran J. R., *From distributional to semantic similarity*, PhD thesis, University of Edinburgh, 2004.

Curran J. R., Moens M., "Improvements in automatic thesaurus extraction", *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, Association for Computational Linguistics, p. 59-66, 2002.

Erk K., "Towards a semantics for distributional representations", *Proceedings of the 10th International Conference on Computational Semantics (IWCS-2013)*, 2013.

Fabre C., Hathout N., Ho-Dac L.-M., Morlane-Hondère F., Muller P., Sajous F., Tanguy L., Van de Cruys T., "Présentation de l'atelier SemDis 2014: sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés", *Actes de la conférence Traitement Automatique du Langage Naturel*, Marseille, France, p. 196-205, 2014.

Ferret O., "Identifying Bad Semantic Neighbors for Improving Distributional Thesauri", *51st Annual Meeting of the Association for Computational Linguistics–ACL 2013*, p. 561-571, 2013.

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

Grefenstette E., "Towards a formal distributional semantics: Simulating logical calculi with tensors", *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, Atlanta, USA, p. 1-10, 2013.

Grefenstette G., *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers, Norwell, MA, USA, 1994.

Guevara E., "Computing semantic compositionality in distributional semantics", *Proceedings of the Ninth International Conference on Computational Semantics*, Association for Computational Linguistics, p. 135-144, 2011.

Gupta A., Boleda G., Baroni M., Padó S., "Mapping conceptual features to referential properties", *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

Habert B., Zweigenbaum P., "Contextual acquisition of information categories", *The Legacy of Zellig Harris: Language and information into the 21st century*, vol. 2, n° 203, p. 139-159, 2002.

Harris Z. S., "Distributional structure", *Word*, vol. 10, n° 2-3, p. 146-162, 1954.

Harris Z. S., *A Theory of Language and Information: A Mathematical Approach*, Clarendon Press, Oxford, 1991.

Herbelot A., Ganesalingam M., "Measuring semantic content in distributional vectors", *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, p. 440-445, 2013.

Hill F., Reichart R., Korhonen A., "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation", *Computational Linguistics*, forthcoming.

Kiela D., Clark S., "A systematic study of semantic vector space model parameters", *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL*, p. 21-30, 2014.

Koller A., "Top-down questions for distributional semantics", *Presentation at the Workshop on formal and distributional semantics*, Toulouse, 2015.

Landauer T. K., Dumais S. T., "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.", *Psychological review*, vol. 104, n° 2, p. 211, 1997.

Lapesa G., Evert S., "A large scale evaluation of distributional semantic models: Parameters, interactions and model selection", *Transactions of the Association for Computational Linguistics*, vol. 2, p. 531-545, 2014.

Lenci A., "Distributional semantics in linguistic and cognitive research", *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics*, vol. 20, n° 1, p. 1-31, 2008.

Lenci A., Benotto G., "Identifying hypernyms in distributional semantic spaces", **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, Association for Computational Linguistics, Montréal, Canada, p. 75-79, 7-8 June, 2012.

Leviant I., Reichart R., "Judgment Language Matters: Multilingual Vector Space Models for Judgment Language Aware Lexical Semantics", *ArXiv e-prints*, August, 2015.

Levy O., Goldberg Y., "Linguistic Regularities in Sparse and Explicit Word Representations", *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)*, Baltimore, Maryland, USA, p. 171-180, 2014.

Levy O., Goldberg Y., Dagan I., "Improving Distributional Similarity with Lessons Learned from Word Embeddings", *Transactions of the ACL*, vol. 3, p. 211-225, 2015.

Lund C., Burgess K., "Modelling parsing constraints with high-dimensional context space", *Language and cognitive processes*, vol. 12, n° 2-3, p. 177-210, 1997.

Marelli M., Bentivogli L., Baroni M., Bernardi R., Menini S., Zamparelli R., "Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment", *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, p. 1-8, 2014.

McCarthy D., Koeling R., Weeds J., Carroll J., "Unsupervised acquisition of predominant word senses", *Computational Linguistics*, vol. 33, n° 4, p. 553-590, 2007.

McCarthy D., Navigli R., "Semeval-2007 task 10: English lexical substitution task", *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval)*, p. 48-53, 2007.

Mikolov T., Chen K., Corrado G., Dean J., "Efficient Estimation of Word Representations in Vector Space", *In Proceedings of Workshop at ICLR 2013*, p. 1-12, 2013a.

Mikolov T., Yih W.-t., Zweig G., "Linguistic Regularities in Continuous Space Word Representations", *In Proceedings of NAACL-HLT 2013, Atlanta, Georgia*, p. 746-751, 2013b.

Mitchell J., Lapata M., "Composition in Distributional Models of Semantics", *Cognitive Science*, vol. 34, n° 8, p. 1388-1439, 2010.

Morris J., Hirst G., "Non-classical lexical semantic relations", *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, Association for Computational Linguistics, p. 46-51, 2004.

Muller P., Fabre C., Adam C., "Predicting the relevance of distributional semantic similarity with contextual information", *52nd Annual Meeting of the Association for Computational Linguistics-ACL 2014*, p. 479-488, 2014.

Nazarenko A., Zweigenbaum P., Habert B., Bouaud J., "Corpus-based Extension of a Terminological Semantic Lexicon", *In Recent Advances in Computational Terminology*, John Benjamins, p. 327-351, 2001.

Padó S., Lapata M., "Dependency-based construction of semantic space models", *Computational Linguistics*, vol. 33, n° 2, p. 161-199, 2007.

Padró M., Idiart M., Ramisch C., Villavicencio A., "Nothing like Good Old Frequency: Studying Context Filters for Distributional Thesauri", *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 419-424, 2014.

Peirsman Y., Geeraerts D., "Predicting strong associations on the basis of corpus data", *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 648-656, 2009.

Peirsman Y., Heylen K., Speelman D., "Finding semantically related words in Dutch. Cooccurrences versus syntactic contexts", *Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models: Beyond Words and Doc- uments*, p. 9-16, 2007.

Sadrzadeh M., Grefenstette E., "A compositional distributional semantics, two concrete constructions, and some experimental evaluations", *Quantum Interaction*, Springer, p. 35-47, 2011.

Sahlgren M., The Word-Space Model, PhD thesis, University of Stockholm, 2006.

Sahlgren M., "The distributional hypothesis", *Italian Journal of Linguistics*, vol. 20, n° 1, p. 33-54, 2008.

Salton G., Wong A., Yang C.-S., "A vector space model for automatic indexing", *Communications of the ACM*, vol. 18, n° 11, p. 613-620, 1975.

Santus E., Lenci A., Lu Q., Schulte im Walde S., "Chasing Hypernyms in Vector Spaces with Entropy", *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, Association for Computational Linguistics, Gothenburg, Sweden, p. 38-42, April, 2014.

Santus E., Yung F., Lenci A., Huang C.-R., "EVALution 1.0: an Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models", *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, Association for Computational Linguistics, Beijing, China, p. 64-69, July, 2015.

The Pham N., Lazaridou A., Baroni M., "A multitask Objective to inject Lexical Contrast into Distributional Semantics", *53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Beijing, China, p. 21-26, 2015.

Turney P. D., "Similarity of semantic relations", *Computational Linguistics*, vol. 32, n° 3, p. 379-416, 2006.

Turney P. D., "Distributional semantics beyond words: Supervised learning of analogy and paraphrase", *Transactions of the Association for Computational Linguistics (TACL)*, vol. 1, p. 353-366, 2013.

Turney P. D., Pantel P., "From frequency to meaning: Vector space models of semantics", *Journal of artificial intelligence research*, vol. 37, n° 1, p. 141-188, 2010.

Van de Cruys T., "A comparison of bag of words and syntax-based approaches for word categorization", *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, p. 47-54, 2008.

Van de Cruys T., "A non-negative tensor factorization model for selectional preference induction", *Natural Language Engineering*, vol. 16, n° 04, p. 417-437, October, 2010.

Van der Plas L., Tiedemann J., "Finding synonyms using automatic word alignment and measures of distributional similarity", *Proceedings of the COLING/ACL on Main conference poster sessions*, Association for Computational Linguistics, p. 866-873, 2006.

Van der Plas L., Tiedemann J., Manguin J.-L., "Synonym acquisition across domains and languages", *Advances in Distributed Agent-Based Retrieval Tools*, Springer, p. 41-57, 2011.

Zarcone A., Padó S., Lenci A., "Same same but different: Type and typicality in a distributional model of complement coercion", *Proceedings of the NetWordS Final Conference on Word Knowledge and Word Usage*, p. 91-94, 2015.

Réordonnancer des thésaurus distributionnels en combinant différents critères

Olivier Ferret

CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
91191 Gif-sur-Yvette Cedex, France
olivier.ferret@cea.fr

RÉSUMÉ. *Dans cet article, nous proposons une méthode pour améliorer les thésaurus distributionnels grâce à un mécanisme d'amorçage : un ensemble d'exemples positifs et négatifs de mots sémantiquement similaires sont sélectionnés de façon non supervisée et utilisés pour entraîner un classifieur supervisé. Celui-ci est ensuite appliqué pour réordonner les voisins sémantiques du thésaurus utilisé pour la sélection des exemples. Nous montrons comment les relations entre les constituants de noms composés similaires peuvent être utilisées pour réaliser une telle sélection et comment conjuguer ce critère, soit de façon précoce, soit de façon tardive, à un critère déjà expérimenté touchant à la symétrie des relations sémantiques. Nous évaluons l'intérêt de ces propositions sur un large ensemble de noms en anglais couvrant un vaste spectre de fréquences. Cet article est une version étendue de (Ferret, 2013 ; Ferret, 2015a).*

ABSTRACT. *In this article, we propose a method for improving distributional thesauri based on a bootstrapping mechanism: a set of positive and negative examples of semantically similar words are selected in an unsupervised way and used for training a supervised classifier. This classifier is then applied for reranking the semantic neighbors of the thesaurus used for example selection. We show how the relations between the mono-terms of similar nominal compounds can be used for performing this selection and how to associate this criterion, either by early fusion or late fusion, with an already tested criterion based on the symmetry of semantic relations. We evaluate the interest of the proposed procedure for a large set of English nouns with various frequencies. This article is an extended version of (Ferret, 2013 ; Ferret, 2015a).*

MOTS-CLÉS : *sémantique lexicale, similarité sémantique, thésaurus distributionnel.*

KEYWORDS: *lexical semantics semantic similarity distributional thesaurus.*

TAL. Volume 56 – n° 2/2015, pages 21 à 49

1. Introduction

Les ressources de nature distributionnelle sont utilisées dans un ensemble de tâches de plus en plus important, allant de l'analyse syntaxique (Henestroza Anguiano et Candito, 2012) à l'extraction de relations (Min *et al.*, 2012). Le travail sur lequel se focalise cet article concerne plus spécifiquement les thésaurus distributionnels, qui associent à un mot un ensemble de voisins dits sémantiques, généralement ordonnés selon l'ordre décroissant de leur similarité avec ce mot, à l'image des exemples donnés par le tableau 1. À la suite de Grefenstette (1994), la façon la plus répandue de construire de tels thésaurus à partir d'un corpus est de caractériser chaque mot du corpus par l'ensemble de ses contextes d'occurrence et d'évaluer le niveau de similarité de deux mots en fonction d'une mesure de similarité reposant sur les contextes qu'ils partagent. Cette mesure permet alors de sélectionner les plus proches voisins d'un mot. Ce schéma général se retrouve sous diverses variantes dans des travaux comme (Lin, 1998), (Curran et Moens, 2002), (Weeds, 2003) ou (Heylen *et al.*, 2008).

Au-delà du problème spécifique de la construction de thésaurus, cette façon d'aborder le problème de la similarité sémantique des mots est caractéristique de la mise en œuvre traditionnelle de l'approche distributionnelle. Cette mise en œuvre a fait depuis quelque temps l'objet de nombreux développements. Une partie d'entre eux se sont attachés à améliorer l'approche de Grefenstette (1994), mais sans la changer en profondeur. Ces travaux se focalisent principalement sur la pondération des éléments constituant les contextes distributionnels, à l'instar de Broda *et al.* (2009), qui transforment les poids au sein des contextes en rangs, ou de Zhitomirsky-Geffet et Dagan (2009), repris et étendus par Yamamoto et Asakura (2010), qui proposent une méthode fondée sur l'amorçage pour modifier les poids des éléments des contextes. Kazama *et al.* (2010) ont pour leur part adopté un point de vue bayésien pour aborder la question. D'autres travaux ont envisagé des changements plus radicaux. Les modèles à base d'exemples (Erk et Padó, 2010) ou de prototypes multiples (Reisinger et Mooney, 2010), dans lesquels la représentation d'un mot est fondée sur un ensemble d'exemples caractéristiques au lieu d'une agrégation de contextes d'occurrence, en sont une manifestation. Les méthodes s'appuyant sur la construction de représentations lexicales distribuées en sont une autre, que ce soit par le biais de techniques de factorisation de matrice comme l'analyse sémantique latente (Landauer et Dumais, 1997) ou la factorisation de matrice non négative (Van de Cruys, 2010), de modèles probabilistes fondés sur la mise en évidence de facteurs latents, prenant la forme de sens en utilisant l'allocation de Dirichlet latente (Dinu et Lapata, 2010) ou de classes sémantiques avec un modèle de Markov caché (Grave *et al.*, 2014), de méthodes fondées sur la notion de hachage comme le Random Indexing (Kanerva *et al.*, 2000) ou plus récemment de celles issues du Deep Learning pour la construction de représentations de type *word embedding* (Huang *et al.*, 2012 ; Mikolov *et al.*, 2013) ou du modèle GloVe de Pennington *et al.* (2014).

En dehors des avancées réalisées globalement dans le champ de la sémantique distributionnelle, certains travaux se concentrent sur des voies d'amélioration plus spécifiques aux thésaurus distributionnels. Même s'ils ne traitent pas explicitement de

cette notion de thésaurus, Zhitomirsky-Geffet et Dagan (2009) et Yamamoto et Asakura (2010), déjà mentionnés ci-dessus, relèvent de cette problématique. Ils s'appuient en effet sur un mécanisme d'amorçage dont la première étape consiste à trouver des voisins sémantiques selon une approche comparable à Grefenstette (1994), le résultat ne constituant rien d'autre qu'un thésaurus distributionnel. Ces voisins sont utilisés dans un second temps pour repondérer les éléments constitutifs des contextes distributionnels et aboutir ainsi à une version améliorée du thésaurus initial. Une telle forme d'amorçage se retrouve également au niveau de Ferret (2012). Dans ce cas, le thésaurus initial est à la base de la sélection non supervisée d'exemples positifs et négatifs de mots sémantiquement liés, exemples servant ensuite à entraîner un classifieur permettant de réordonnancer le thésaurus initial. Dans le cas de Ferret (2012), cette sélection s'appuie sur l'exploitation des relations de symétrie existant au niveau du thésaurus initial. Claveau *et al.* (2014) proposent, quant à eux, plusieurs façons de généraliser cette idée d'exploitation des relations à l'échelle du thésaurus pour améliorer celui-ci.

Dans cet article, après avoir défini la méthode de construction et d'évaluation des thésaurus distributionnels que nous utilisons et analysé en détail les caractéristiques de ces thésaurus, nous examinerons comment les principes développés par Ferret (2012) peuvent être conjugués à la sélection non supervisée d'exemples de mots sémantiquement liés fondée sur les mots composés (Ferret, 2013 ; Ferret, 2015a) pour améliorer les thésaurus distributionnels. Nous présenterons, en outre, deux modes de conjugaison des informations apportées par les deux critères de sélection des exemples, l'un correspondant à une fusion dite précoce, réalisée au niveau des ensembles d'apprentissage, l'autre à une fusion dite tardive, opérant au niveau des thésaurus réordonnancés selon chacun des critères.

2. Construire et évaluer un thésaurus distributionnel

2.1. *Paramètres distributionnels*

L'utilisation de l'amorçage implique dans notre cas de construire un thésaurus initial dont la qualité, au moins pour un sous-ensemble de celui-ci, soit suffisamment élevée pour servir de marchepied à une amélioration plus globale. Une telle construction dépend d'un ensemble suffisamment large de paramètres pour qu'il soit en pratique impossible de mener une optimisation globale pour fixer la valeur de ceux-ci. Même les travaux les plus récents menés pour explorer cet espace de paramètres (Bullinaria et Levy, 2012 ; Kiela et Clark, 2014 ; Lapesa et Evert, 2014) font un certain nombre d'impasses et surtout, utilisent pour faire leurs évaluations des jeux de test souvent de petite taille constitués majoritairement de mots ayant des fréquences assez élevées. Des jeux de test tels que WordSim 353 (Gabrilovich et Markovitch, 2007) ou les quatre-vingts questions du TOEFL (Landauer et Dumais, 1997) en sont des exemples typiques. Le fait d'en conjuguer plusieurs de différents types, comme le font les travaux cités, vise à pallier ces insuffisances, mais il n'est pas évident que cumuler des résultats biaisés permette d'aboutir à des conclusions plus solides.

Dans notre cas, nous avons fait le choix de mener des évaluations à large échelle pour éviter certains de ces biais, tout en restant bien sûr prisonnier de l'impossibilité de tester toutes les valeurs de paramètres. Compte tenu de notre objectif final – la construction de thésaurus – nous avons adopté un test de type TOEFL, option la plus proche de cette optique, mais en utilisant le jeu de test WordNet-based Synonymy Test (WBST) proposé par Freitag *et al.* (2005), comportant 9 887 questions pour les noms. Ferret (2010) s'est attaché à la sélection des meilleurs paramètres distributionnels dans ce cadre. Nous reprenons ici les conclusions de ce travail.

Bien que notre langue cible soit l'anglais, nous avons ainsi choisi de limiter le niveau des traitements linguistiques appliqués au corpus source de nos données distributionnelles à l'étiquetage morphosyntaxique et à la lemmatisation, de manière à faciliter la transposition du travail à des langues moins dotées. Cette approche apparaît à cet égard comme un compromis raisonnable entre l'approche de Freitag *et al.* (2005), dans laquelle aucune normalisation n'est faite, et l'approche assez largement répandue consistant à utiliser un analyseur syntaxique, à l'instar de Curran et Moens (2002). Plus précisément, nous nous sommes appuyé sur l'outil *TreeTagger* (Schmid, 1994) pour assurer le prétraitement du corpus AQUAINT-2 (Voorhees et Graff, 2008) qui est à la base de ce travail. Ce corpus, que l'on peut qualifier de taille moyenne avec ses 380 millions de mots environ, est composé d'articles de journaux.

Les expérimentations menées par Ferret (2010) ont abouti par ailleurs aux choix suivants concernant les paramètres de construction des contextes distributionnels et d'évaluation de leur similarité :

– contextes distributionnels constitués de cooccurrents graphiques : noms, verbes et adjectifs collectés grâce à une fenêtre de taille fixe centrée sur chaque occurrence du mot cible ;
– taille de la fenêtre = 3 (un mot plein à droite et un mot plein à gauche du mot cible), c'est-à-dire des cooccurrents de très courte portée ;
– filtrage minimal des contextes : suppression des seuls cooccurrents de fréquence égale à 1 ;
– fonction de pondération des cooccurrents dans les contextes = *information mutuelle ponctuelle* entre le mot cible et son cooccurrent, restreinte aux valeurs positives (*Positive Pointwise Mutual Information*) ;
– mesure de similarité entre contextes, pour évaluer la similarité sémantique de deux mots = mesure *cosinus*.

Un filtre fréquentiel est en outre appliqué à la fois aux mots cibles et à leurs cooccurrents : seuls les mots de fréquence supérieure à 10 sont considérés, limite un peu arbitraire en dessous de laquelle nous considérons la pauvreté des contextes distributionnels comme trop importante pour que leur comparaison soit significative.

Les paramètres que nous avons ainsi sélectionnés, en particulier pour ce qui est de la taille de la fenêtre, de la fonction de pondération des cooccurrents et de la mesure de similarité entre contextes, sont très directement en phase avec ceux faisant consensus parmi les travaux récents réalisés avec des corpus de taille importante et pour des éva-

abnormality	defect [0,30], disorder [0,23], deformity [0,22], mutation [0,21], prolapse [0,21], anomaly [0,21] ...
agreement	accord [0,44], deal [0,41], pact [0,38], treaty [0,36], negotiation [0,35], proposal [0,32], arrangement [0,30] ...
cabdriver	waterworks [0,23], toolmaker [0,22], weaponeer [0,17], valkyry [0,17], wang [0,17], amusement-park [0,17] ...
machination	hollowness [0,15], share-price [0,12], clockmaker [0,12], huguenot [0,12], wrangling [0,12], alternation [0,12] ...

Tableau 1. *Premiers voisins de quelques entrées du thésaurus distributionnel A2ST*

luations portant également sur la similarité sémantique au sens large du terme (Kiela et Clark, 2014 ; Baroni *et al.*, 2014 ; Levy *et al.*, 2015). Le jeu de test WBST que nous avons utilisé est très clairement centré sur la similarité sémantique, par opposition à la proximité sémantique, ce qui explique en particulier le choix d'une fenêtre de petite taille. Il faut néanmoins remarquer que l'idée généralement acceptée d'une association des petites fenêtres à la similarité sémantique et des plus larges fenêtres à la proximité sémantique semble à relativiser dans le cas des grands corpus : les jeux de test utilisés dans bon nombre des travaux évoqués, qui mettent tous en avant les meilleures performances des fenêtres de petite taille, incluent aussi bien des relations de similarité sémantique que de proximité sémantique, les secondes étant même parfois plus nombreuses que les premières. Nous avons d'ailleurs fait le même constat pour les thésaurus distributionnels dans le cadre d'expérimentations non présentées ici.

2.2. *Construction et évaluation du thésaurus*

Disposant d'une mesure permettant d'évaluer la similarité sémantique d'un couple de mots, la procédure de construction d'un thésaurus est simple : les voisins sémantiques d'un mot sont trouvés en recherchant les N plus proches voisins de ce mot selon la mesure de similarité considérée. Plus précisément, cette recherche consiste à appliquer cette mesure de similarité entre le mot cible et tous les autres mots du vocabulaire considéré ayant la même catégorie morphosyntaxique (ici, les noms). Finalement, tous ces mots sont triés suivant leur valeur de similarité et seuls les N plus proches voisins, N étant égal à 100 dans nos expérimentations, sont conservés en tant que voisins sémantiques.

Nous avons appliqué la procédure de construction décrite à l'ensemble des noms du corpus AQUAINT-2 de fréquence strictement supérieure à 10, soit 26 210 noms. Au final, le thésaurus résultat, appelé A2ST, est constitué de 25 988 entrées, la différence s'expliquant par les cas où le contexte distributionnel d'un nom ne comporte aucun élément commun avec celui d'un autre nom. La limite fixée des 100 voisins est atteinte par 99,8 % d'entre elles. Le tableau 1 donne les premiers voisins associés à

type de relation	% des relations	
c	9,0	c : co-hyponymie (= H + h)
c + h	6,5	h : hyponymie
H + c	5,8	H : hyperonymie
H + c + h	3,8	s : synonymie
s	3,4	r1 + r2 : composition des relations r1 et r2
h	3,0	
H	2,8	

Tableau 2. *Relations sémantiques les plus fréquentes au sein du thésaurus Moby caractérisées en fonction de WordNet*

quatre entrées en illustrant le fait que ces voisins peuvent être très pertinents, au regard de la notion de similarité sémantique, pour certaines entrées, comme *abnormality* ou *agreement* ici, et beaucoup moins pour d'autres, comme dans le cas de *cabdriver* ou *machination*.

L'évaluation de l'ensemble du thésaurus demande néanmoins une procédure plus formelle et plus automatique. Comme dans le cas des paramètres distributionnels, cette évaluation est de nature intrinsèque et à large échelle : les voisins du thésaurus sont comparés à des ressources de référence. Nous avons plus spécifiquement adopté deux ressources complémentaires de large couverture : les synonymes de WordNet [W] (Miller, 1990), dans sa version 3.0, et le thésaurus Moby [M] (Ward, 1996). Les premiers sont représentatifs de la similarité sémantique tandis que le second recouvre un spectre plus large de relations sémantiques que l'on peut en première analyse regrouper sous la notion de proximité sémantique. Nous avons également créé une ressource fusionnant WordNet et le thésaurus Moby [WM]. En reprenant Ferret (2015b), le tableau 2 donne un aperçu des relations présentes dans Moby en les caractérisant en fonction de WordNet, soit comme des relations élémentaires de WordNet (synonymie, hyponymie, etc.), soit comme des relations composées d'une suite de ces relations élémentaires. Il montre ainsi que les relations les plus fréquentes sont composées mais également que Moby recèle une grande diversité de types de relations puisque les sept types de relations les plus fréquents ne couvrent que 34,3 % de ses relations.

Notre but étant d'abord d'évaluer le thésaurus construit et non la capacité de celui-ci à reconstituer les ressources de référence, nous avons filtré ces ressources en éliminant en leur sein, aussi bien au niveau des entrées que des mots qui leur sont liés, les termes ne faisant pas partie du vocabulaire des noms simples retenus pour construire nos données distributionnelles. Au final, seules 14 670 entrées du thésaurus, intersection entre les noms de ce dernier et ceux de WordNet[1], ont été utilisées

1. Il est à noter que tous les noms présents dans WordNet ne sont pas nécessairement associés à un synset, ce qui explique la différence entre le nombre d'entrées du thésaurus considéré et le nombre d'entrées évaluées par rapport à WordNet.

fréq.	réf.	#mots éval	#syn./ mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
	W	10 473	2,9	24,6	8,2	9,8	11,7	5,1	3,4	0,7
toutes	M	9 216	50,0	9,5	6,7	3,2	24,1	16,4	13,0	4,8
14 670	WM	12 243	38,7	9,8	7,7	5,6	22,5	14,1	10,8	3,8
	W	5 889	3,3	29,4	11,8	13,5	17,4	7,5	4,9	1,0
hautes	M	5 751	60,5	11,2	9,4	4,6	35,9	24,2	18,9	6,8
7 335	WM	6 754	52,6	11,4	11,1	7,4	36,4	22,8	17,5	6,0
	W	4 584	2,3	16,0	3,7	5,1	4,2	2,0	1,4	0,4
basses	M	3 465	32,5	4,4	2,3	0,9	4,4	3,4	3,1	1,4
7 335	WM	5 489	21,6	5,1	3,6	3,4	5,5	3,3	2,7	1,1

Tableau 3. *Évaluation du thésaurus distributionnel A2ST*

pour l'évaluation présentée dans le tableau 3. La troisième colonne de ce même tableau donne le nombre effectif de noms pour lesquels l'évaluation a été réalisée pour chaque ressource, chaque entrée retenue pour l'évaluation n'apparaissant pas dans chaque ressource de référence. La quatrième colonne correspond pour sa part au nombre moyen de voisins de référence (appelés synonymes par facilité de langage) à trouver dans chaque ressource pour chacune de leurs entrées faisant partie du vocabulaire AQUAINT-2.

Les voisins étant ordonnés pour chaque entrée du thésaurus, il est possible de faire le parallèle entre la recherche de voisins sémantiques et la recherche de documents en recherche d'information et de réutiliser ainsi les métriques d'évaluation classiquement utilisées pour cette dernière en faisant jouer aux entrées du thésaurus le rôle de requêtes et aux autres noms celui de documents. Les six dernières colonnes du tableau 3 donnent ainsi les résultats en pourcentage pour les métriques suivantes : la R-précision (R-préc.) est la précision obtenue en se limitant aux R premiers voisins, R étant le nombre de synonymes dans la ressource de référence pour l'entrée considérée ; la MAP (*Mean Average Precision*) est la moyenne des précisions pour chacun des rangs auxquels un synonyme de référence a été identifié ; enfin, sont données les précisions pour différents seuils de nombre de voisins sémantiques examinés (précision pour les 1, 5, 10 et 100 premiers voisins). Ces métriques sont complétées par la donnée à la cinquième colonne du tableau du pourcentage des voisins de référence figurant parmi les 100 voisins sémantiques de chaque entrée de notre thésaurus distributionnel. La fréquence des mots, en relation avec la taille des corpus, étant une donnée importante des approches distributionnelles, les résultats globaux sont différenciés suivant deux tranches fréquentielles de même effectif (7 335 mots chacune) : *hautes* pour les mots de fréquence $>$ à la fréquence médiane (249) et *basses* pour les autres.

L'analyse du tableau 3 conduit à faire trois grandes observations. Tout d'abord, malgré leurs performances intéressantes sur un test de similarité sémantique à large

couverture et adapté à l'application visée, les paramètres distributionnels sélectionnés n'obtiennent dans l'absolu que des résultats assez modestes lorsqu'ils sont appliqués au problème de la construction d'un thésaurus distributionnel. Cette faiblesse est observable aussi bien au niveau du taux de rappel des voisins de référence – environ 25 % pour WordNet et 10 % pour le thésaurus Moby – qu'au niveau de leur rang parmi les voisins retenus : la R-précision générale dépasse à peine 8 % dans le meilleur des cas, en l'occurrence WordNet. Ce constat a une portée plus générale que notre travail spécifique dans la mesure où ces paramètres peuvent être considérés comme classiques.

La deuxième observation est que cette faiblesse générale recouvre des différences importantes suivant la fréquence des mots. On observe ainsi une corrélation claire entre le niveau des résultats et la fréquence des mots dans le corpus de constitution des données distributionnelles : plus cette fréquence est élevée, plus la qualité des voisins sémantiques est élevée, à la fois en termes de quantité et de rang. Le phénomène conduit d'ailleurs à considérer que pour les basses fréquences, les résultats obtenus sont difficilement exploitables d'un point de vue applicatif. Ce niveau de résultats est assez aisément compréhensible : un mot cible avec peu d'occurrences ne peut avoir qu'un contexte distributionnel très pauvre. Ainsi, deux mots de faible fréquence risquent fort de n'avoir aucun élément en commun au niveau de leurs contextes et, dans le cas de la comparaison d'un mot de faible fréquence avec un mot de fréquence plus élevée, la taille très réduite de l'intersection de leurs contextes rend cette comparaison très sensible à des cooccurrences non significatives sur le plan sémantique. Même si cette constatation plaide en faveur de l'accroissement de la taille des corpus, ce que l'on observe de fait actuellement, des mots de faible fréquence existeront toujours dans ces corpus en vertu de la loi de Zipf. Par ailleurs, il existe de nombreuses situations où de très larges corpus ne sont pas disponibles.

La dernière observation suscitée par le tableau 3 est que le profil des ressources de référence considérées a aussi son importance quant aux résultats obtenus. WordNet fournit un nombre restreint de synonymes stricts pour chaque nom (2,9 en moyenne) tandis que le thésaurus Moby contient pour chaque entrée un nombre beaucoup plus important de mots sémantiquement proches (50 en moyenne). Cette différence de profil explique en particulier que si les valeurs pour Moby sont assez élevées par rapport à celles pour WordNet pour des précisions à un rang faible – P@1 = 35,9 pour les hautes fréquences par exemple, à comparer à 17,4 – le rapport s'inverse dans les mêmes conditions pour la MAP : 13,5 pour WordNet et 4,6 pour Moby. Intuitivement, les premiers voisins du thésaurus ont dans le cas de Moby plus de chances de figurer dans un ensemble de voisins de référence plus large et couvrant un plus grand nombre de types de relations sémantiques mais, en contrepartie, il est plus difficile pour le thésaurus de couvrir ce large ensemble de voisins de façon significative.

2.3. *Mise en perspective*

Aborder la problématique de la similarité sémantique fondée sur des bases distributionnelles par le biais des thésaurus n'est pas la façon de faire la plus répandue

jeu de test	réf.	#mots éval	#syn./ mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
	W	59	5,2	23,6	9,3	9,6	20,3	8,1	5,6	1,2
70	M	64	103,2	11,2	11,4	5,1	54,7	43,1	33,3	11,6
(69)	WM	65	103,1	11,2	12,0	5,8	55,4	43,4	33,5	11,6
	W	247	4,7	27,2	12,3	12,7	19,8	8,4	5,8	1,3
300	M	253	97,6	11,1	11,5	5,6	53,0	37,0	29,0	10,8
(296)	WM	269	93,0	11,2	12,2	6,7	52,0	36,0	28,2	10,4

Tableau 4. *Évaluation des voisins du thésaurus A2ST pour le jeu de test de 70 mots de Curran et Moens (2002) et de 300 mots de Curran (2003)*

à l'heure actuelle. En conséquence, les méthodologies d'évaluation dans cette sphère sont moins uniformes que pour l'évaluation de la stricte similarité sémantique, réalisée grâce à des jeux de test de type WordSim 353. Les comparaisons sont donc aussi plus difficiles. Dans le cas du néerlandais par exemple, Van der Plas et Bouma (2004) et Van de Cruys (2010) ont ainsi adopté la version néerlandaise d'EuroWordNet comme référence, assez comparable à WordNet, mais en s'appuyant sur sa structure hiérarchique : au lieu de simplement se fonder sur l'appartenance à une liste de synonymes présents dans EuroWordNet, la pertinence sémantique d'un voisin par rapport à une entrée est définie en calculant la mesure de Wu et Palmer (1994) entre cette entrée et le voisin. Cette méthode présente l'avantage d'intégrer de façon cohérente des types de relations différents dans une même mesure. Elle est cependant moins directement interprétable que l'option que nous avons adoptée. Pantel *et al.* (2009) s'inscrivent, pour leur part, dans un cadre plus applicatif en s'intéressant à la notion d'ensemble d'entités (*Entity Sets*), sous-tendue par une gamme de relations très étendue et se focalisant beaucoup sur des entités nommées.

Le travail de Curran et Moens (2002) est en revanche plus directement comparable au nôtre. Il met en œuvre diverses mesures de similarité fondées sur des cooccurrences syntaxiques qui sont ensuite évaluées du point de vue de l'extraction de voisins sémantiques en adoptant comme référence la fusion des thésaurus Roget, Moby et Macquarie. Cette évaluation porte sur 70 noms choisis au hasard dans WordNet en respectant une diversité de fréquences et de degrés de spécificité. Parmi les différentes mesures testées, la meilleure performance obtenue (Dice† + T-test) est une précision au rang 1 de 76 %, au rang 5 de 52 % et au rang 10 de 45 % pour 70 noms, à comparer avec 41,3 %, 28,0 % et 21,9 % dans notre cas en se restreignant aux 3 732 noms de fréquence $>$ à 1 000[2].

2. Nous reprenons ici les chiffres de Ferret (2010), qui effectue un découpage plus fin en trois tranches fréquentielles à peu près de même taille, les fréquences $>$ à 1 000 constituant la tranche supérieure et les fréquences $\leq$ 100, la tranche inférieure.

méthode	#mots éval	#syn./ mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
A2ST	12 243	38,7	9,8	7,7	5,6	22,5	14,1	10,8	3,8
A2ST-SYNT	11 887	39,4	13,2	10,7	7,9	29,4	18,9	14,6	5,2
[Lin, 98]	9 823	44,5	12,7	11,6	8,1	36,1	23,7	18,2	5,6
[Huang *et al.*, 12]	10 537	42,6	3,8	1,9	0,8	7,1	5,0	4,0	1,6
[Mikolov *et al.*, 13]	12 326	38,6	6,2	5,5	4,2	16,3	9,5	7,0	2,4
ESA	7 756	44,3	7,0	6,9	5,1	13,2	9,1	7,3	3,1
[Baroni *et al.*, 14]-C	12 052	39,3	13,6	12,5	9,8	31,9	19,6	15,2	5,3
[Baroni *et al.*, 14]-P	12 052	39,3	11,3	10,9	8,5	30,3	18,4	13,8	4,4

Tableau 5. *Évaluation de thésaurus construits selon des méthodes différentes*

Une explication de cette différence pourrait être l'utilisation de cooccurrents syntaxiques par Curran et Moens (2002) alors que nous nous contentons de cooccurrents graphiques. Néanmoins, comme le montre la ligne A2ST-SYNT du tableau 5, la référence étant [WM], l'utilisation de cooccurrents syntaxiques sur le corpus AQUAINT-2 ne suffit pas à expliquer la différence avec Curran et Moens (2002). Même les meilleurs résultats pour ce thésaurus A2ST – P@1 = 44,0 %, P@5 = 31,5 % et P@10 = 25,2 % pour 3 727 noms de fréquence $>$ à 1 000 avec [M] comme référence – sont encore assez éloignés des chiffres de Curran et Moens (2002). Deux autres facteurs sont aussi à considérer. Tout d'abord, le niveau de richesse des références utilisées est très différent. Pour 3 732 noms de fréquence $>$ à 1 000, le thésaurus Moby fournit en moyenne 69 mots sémantiquement liés dans notre cas tandis que pour les 70 noms de Curran et Moens (2002), ce nombre monte à 331. Or, ce facteur a une grande influence sur les résultats ainsi que nous l'avons illustré ci-dessus où le passage d'une moyenne de 2,9 synonymes par entrée pour WordNet à 50 mots liés pour Moby s'accompagne d'une montée de la précision au rang 5 de 5,1 % à 16,4 %. Le même phénomène, observé aussi entre WordNet et Moby, explique que le rappel soit dans notre cas supérieur, avec 11,4 %, à celui de Curran et Moens (2002), égal à 8,3 %.

Le second facteur est néanmoins aussi important. Bien que les 70 noms du jeu de test aient été en principe sélectionnés de manière équilibrée en termes notamment de fréquence, on constate en pratique que sur les 69 présents dans notre thésaurus, 65 figurent parmi les entrées de fréquence $>$ à 1 000 et aucun parmi les mots de fréquence $\leq$ à 100. La première ligne du tableau 4 montre par ailleurs que les performances de notre thésaurus obtenus pour ces 69 entrées sont bien supérieures aux performances de l'ensemble de nos entrées de fréquence $>$ à 1 000. Ce constat s'applique également à un jeu de test plus large constitué de 300 noms et utilisé par Curran (2003). Parmi les 296 noms faisant partie de notre thésaurus, 244 figurent parmi les entrées de fréquence $>$ à 1 000 tandis que 3 seulement ont une fréquence $\leq$ à 100. Encore une fois, les

performances pour ce jeu de test étendu sont plus élevées que celles obtenues pour l'ensemble de nos entrées de fréquence $>$ à 1 000.

Pour achever cette mise en perspective, le tableau 5 compare les résultats de notre thésaurus (A2ST) avec ceux de plusieurs autres thésaurus en utilisant [WM] comme référence et les mêmes entrées que pour l'évaluation d'A2ST. A2ST-SYNT est le thésaurus que nous avons produit dans les mêmes conditions qu'A2ST en nous contentant de remplacer les cooccurrents graphiques par des cooccurrents syntaxiques obtenus grâce à l'analyseur MINIPAR (Lin, 1994). Comme nous l'avons indiqué précédemment, et en cohérence avec Curran et Moens (2002) et Heylen *et al.* (2008), cette substitution a un effet très clairement positif sur les résultats et ce, pour toute la gamme des fréquences. La seule restriction à noter est un petit rétrécissement du nombre des entrées pour lesquelles des voisins sont trouvés. [Lin, 98] est le thésaurus mis à disposition par Lin[3], construit comme A2ST-SYNT grâce à des cooccurrents syntaxiques obtenus par l'analyseur MINIPAR. L'évaluation de ce thésaurus donne de meilleurs résultats que pour A2ST-SYNT, ce qui peut s'expliquer par deux facteurs : d'une part, le corpus utilisé par Lin, d'une taille de 1,5 milliard de mots, est beaucoup plus important que le corpus AQUAINT-2, d'autre part, du fait des entrées disponibles, l'évaluation du corpus de Lin a été réalisée sur un plus petit ensemble d'entrées (seulement 1 510 pour les fréquences $\leq$ à 100 à comparer à 3 687 pour A2ST), en moyenne de plus forte fréquence comme le montre le nombre plus élevé de synonymes par entrée.

Les lignes restantes du tableau 5 correspondent à des thésaurus que nous avons construits en suivant le même processus que pour A2ST mais en utilisant des représentations différentes, toujours exprimées sous la forme vectorielle, en lieu et place des contextes distributionnels classiques (l'exception étant [Baroni *et al.*, 14]-C, fondé sur des contextes distributionnels classiques mais non construits par nos soins). Dans chacun des cas, nous appliquons la même mesure, en l'occurrence la mesure *cosinus*, aux représentations associées aux mots afin d'évaluer la similarité de ceux-ci. Dans ce cadre, [Huang *et al.*, 12] et [Mikolov *et al.*, 13] renvoient à deux approches récentes évoquées en introduction et fondées sur la construction de représentations distribuées de mots par des réseaux de neurones. Dans le cas de [Huang *et al.*, 12], nous avons utilisé les représentations construites à partir de Wikipédia[4] fournies par les auteurs tandis que dans le cas de [Mikolov *et al.*, 13], nous avons calculé ces représentations à partir du corpus AQUAINT-2 grâce au logiciel *word2vec*[5] en utilisant les meilleurs paramètres sélectionnés par Mikolov *et al.* (2013)[6]. Dans les deux cas, les résultats sont significativement inférieurs à ceux d'A2ST, avec un niveau particulièrement bas pour [Huang *et al.*, 12] qui peut s'expliquer au moins en partie par la différence de corpus. Ces résultats suggèrent néanmoins que l'utilisation de ce type de représentations distribuées n'est pas encore une option intéressante pour la construction de thésaurus distributionnels, un peu à contresens de Baroni *et al.* (2014) mais plus compatible avec

3. `http://webdocs.cs.ualberta.ca/~lindek/Downloads/sim.tgz`
4. `http://nlp.stanford.edu/~socherr/ACL2012_wordVectorsTextFile.zip`
5. `http://code.google.com/p/word2vec`
6. word2vec -cbow 0 -size 600 -window 10 -negative 0 -hs 0 -sample 1e-5

Hill *et al.* (2014). Ce constat est renforcé par les deux dernières lignes du tableau 5, qui donnent les résultats des thésaurus construits avec les vecteurs de contexte mis à disposition par Baroni *et al.* (2014)[7] : [Baroni *et al.*, 14]-P correspond à des vecteurs construits grâce au modèle CBOW de Mikolov *et al.* (2013) tandis que [Baroni *et al.*, 14]-C correspond à des vecteurs de cooccurrents obtenus de façon classique par une fenêtre graphique. Le niveau des résultats obtenus, rivalisant et même dépassant pour bon nombre de mesures les résultats d'A2ST-SYNT et ceux du thésaurus de Lin, confirme l'observation faite à propos du thésaurus de Lin de la grande importance de la taille du corpus initial sur les résultats, égale à 2,8 milliards de mots dans le cas de Baroni *et al.* (2014). Mais l'observation la plus importante est ici la supériorité de [Baroni *et al.*, 14]-C par rapport à [Baroni *et al.*, 14]-P, ce qui vient limiter le constat général fait par Baroni *et al.* (2014) de la supériorité des modèles neuronaux par rapport aux approches traditionnelles. Ce constat n'est visiblement pas vérifié dans le cas des thésaurus distributionnels.

Enfin, nous donnons également l'évaluation d'un thésaurus fondé sur l'approche ESA proposée par Gabrilovich et Markovitch (2007). Dans ce cas, les traits sur lesquels s'appuie le calcul de la similarité entre deux mots sont constitués de concepts Wikipédia, chaque concept correspondant en pratique à un article de cette encyclopédie. Pour construire notre thésaurus, nous avons exploité les données constituées par Popescu et Grefenstette (2011)[8]. Bien que l'ensemble des entrées soit ici plus limité que pour les autres thésaurus, il est suffisamment important pour se rendre compte qu'il existe une certaine variabilité de la performance de l'approche ESA, qui est plutôt bonne sur le test WordSim 353 et moins intéressante pour la construction d'un thésaurus, se situant assez proche de celle de Mikolov *et al.* (2013).

3. Utiliser l'amorçage pour améliorer un thésaurus distributionnel

Les analyses faites ci-dessus ont montré que les performances de l'approche distributionnelle restent globalement modestes quant à la qualité des thésaurus qu'elle produit et que ce constat n'est pas propre au cadre que nous avons adopté, même si certains facteurs, comme l'utilisation de cooccurrents syntaxiques ou l'augmentation de la taille des corpus, permettent d'améliorer en partie la situation. Il est, dès lors, légitime de chercher à améliorer ces thésaurus. Nous nous sommes plus spécifiquement concentré sur des méthodes non supervisées ou très faiblement supervisées. Dans cette section, nous proposons un cadre d'amélioration se fondant sur l'amorçage.

3.1. *Principes*

L'évaluation de notre thésaurus distributionnel initial, A2ST, montre que les voisins sémantiques obtenus sont significativement meilleurs pour certaines entrées que

7. `http://clic.cimec.unitn.it/composes/semantic-vectors.html`
8. Nous remercions Adrian Popescu pour avoir mis à notre disposition ces données.

pour d'autres. Une telle configuration est *a priori* favorable à un mécanisme de type amorçage dans la mesure où il est envisageable de s'appuyer sur les résultats des « bonnes » entrées pour obtenir une amélioration plus globale. Zhitomirsky-Geffet et Dagan (2009) ont déjà fait appel à l'amorçage dans un contexte proche du nôtre, l'acquisition de relations d'implication textuelle entre mots. Cependant, des expérimentations rapportées par Ferret (2010) ont montré que la transposition de cette approche à notre problème n'était pas concluante. Ainsi, au lieu d'utiliser les résultats d'une mesure de similarité initiale pour modifier directement les poids des éléments constitutifs des contextes distributionnels, nous avons adopté une approche plus indirecte, exploitant les résultats de Hagiwara (2008).

Hagiwara (2008) a en effet montré qu'il est possible d'entraîner et d'appliquer avec un bon niveau de performance un classifieur statistique, en l'occurrence de type machine à vecteurs de support (SVM), pour décider si deux mots sont ou ne sont pas synonymes, au sens large du terme. Par ailleurs, ce travail montre également que la valeur de la fonction de décision caractérisant les SVM, dont on n'utilise que le signe dans le cas d'une classification binaire, peut jouer pour l'ordonnancement des voisins sémantiques le même rôle que la valeur d'une mesure de similarité telle que celle définie à la section 2.1.

À la différence de Hagiwara (2008), nous ne faisons volontairement pas l'hypothèse de l'accès possible à un ensemble d'exemples et de contre-exemples étiquetés manuellement pour réaliser l'entraînement d'un tel classifieur. Le nombre de ces exemples dans (Hagiwara, 2008), 2 148 pour les positifs et 13 855 pour les négatifs, est en effet très important. En revanche, les voisins sémantiques de notre thésaurus initial peuvent être exploités pour construire un tel ensemble. La mesure de similarité à laquelle est adossé ce thésaurus n'offre pas de critère évident pour discriminer les mots sémantiquement liés[9]. Cependant, elle peut être utilisée plus indirectement pour sélectionner un ensemble d'exemples et de contre-exemples de façon non supervisée en minimisant le nombre d'erreurs. Ces erreurs correspondent à des exemples considérés comme positifs mais en réalité négatifs, et d'exemples considérés comme négatifs mais en fait positifs. Dans cette optique, nous proposons d'entraîner un classifieur SVM grâce à ces ensembles et de l'appliquer ensuite pour réordonner les voisins sémantiques obtenus précédemment. L'ensemble de la démarche peut être résumée par la procédure suivante, que l'on peut rapprocher dans une certaine mesure de la notion d'auto-apprentissage (*self-training*) :

– construction d'un thésaurus distributionnel ;
– sélection non supervisée d'un ensemble d'exemples et de contre-exemples de mots sémantiquement similaires au sein de ce thésaurus au moyen d'heuristiques ;
– entraînement d'un classifieur statistique à partir des exemples sélectionnés ;

9. Fixer pour ce faire un seuil sur les valeurs de similarité produit de mauvais résultats du fait de la variabilité de ces valeurs d'une entrée à l'autre. Ce constat a motivé notre choix d'utiliser un SVM en classification plutôt qu'en régression.

– application du classifieur entraîné au réordonnancement des voisins du thésaurus initial.

Le point clé de l'amélioration des résultats par ce moyen est de sélectionner de façon non supervisée un nombre suffisant d'exemples et de contre-exemples en minimisant les erreurs propres à une telle sélection. Dans la section 3.3, nous proposons d'associer deux méthodes faibles, à la fois au sens de la productivité et de la validité des résultats, pour accomplir cette tâche.

3.2. *Représentation des exemples*

Avant de présenter plus en détail ce processus de sélection, il convient de préciser la nature des exemples et des contre-exemples. Nous reprenons de ce point de vue la conception développée par Hagiwara (2008) : un exemple est constitué d'un couple de mots considérés comme synonymes ou plus généralement sémantiquement liés ; un contre-exemple est formé d'un couple de mots entre lesquels un tel lien sémantique n'existe pas. La représentation de ces couples pour un classifieur de type SVM s'effectue en associant leurs représentations distributionnelles. Cette association s'effectue pour chaque couple (M_1, M_2) en sommant le poids des cooccurrents communs aux mots M_1 et M_2. Les cooccurrents de M_x non présents dans M_y se voient attribuer un poids nul. Chaque exemple ou contre-exemple a donc la même forme que la représentation distributionnelle d'un mot, c'est-à-dire un vecteur de mots pondérés.

3.3. *Sélection des exemples et des contre-exemples*

Du point de vue de la sélection des exemples et des contre-exemples de mots sémantiquement liés, le tableau 3 offre une image claire : trouver des exemples est beaucoup plus problématique que trouver des contre-exemples dans la mesure où le nombre de mots sémantiquement liés à une entrée du thésaurus diminue très fortement dès que l'on considère ses voisins de rang un peu élevé. Dans les expérimentations de la section 4, nous avons ainsi construit nos contre-exemples à partir de nos exemples en créant pour chaque exemple (A,B) deux contre-exemples de la forme : (A, *voisin de rang 10 de A*) et (B, *voisin de rang 10 de B*). Le choix d'un rang supérieur garantirait un nombre plus faible de faux contre-exemples (*i.e.* couples de synonymes) et donc *a priori*, de meilleurs résultats. En pratique, l'utilisation de voisins du mot cible de rang assez faible conduit à une performance supérieure, sans doute parce que ceux-ci sont plus utiles en termes de discrimination, étant plus proches de la zone de transition entre exemples et contre-exemples. Nous avons par ailleurs constaté expérimentalement que le rapport entre contre-exemples et exemples dans (Hagiwara, 2008), égal à 6,5 et donc fortement déséquilibré en faveur des contre-exemples, n'était pas nécessaire dans notre situation et pouvait se ramener à 2.

Pour la sélection des exemples, le tableau 3 impose un double constat : trouver un voisin sémantiquement proche est d'autant plus probable que la fréquence de l'entrée

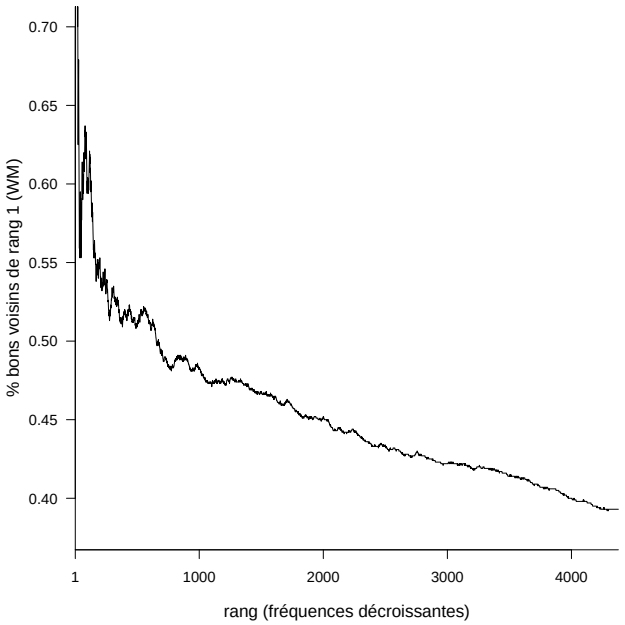

Figure 1. *Proportion de bon voisins de rang 1 selon les fréquentes décroissantes de leur entrée*

du thésaurus considérée est élevée et que le rang du voisin est faible. La forme extrême de cette logique conduirait à retenir comme exemples tous les couples de mots (*entrée de haute fréquence*, *voisin de rang 1*), ce qui donne un large nombre d'exemples – 7 335 – mais un taux d'erreur (*i.e.* nombre de couples de mots non liés sémantiquement) également élevé – 63,6 % dans le cas le plus favorable (référence WM). Comme le montre la figure 1, qui donne la proportion de bons voisins au rang 1 selon les fréquences décroissantes des entrées de forte fréquence, il n'existe pas de critère évident permettant de fixer un seuil. À titre indicatif, prendre les 2 148 premières entrées en termes de fréquence (nombre d'exemples positifs de Hagiwara (2008)) conduit à un taux d'erreur encore égal à 55,7 % et ce taux ne descend en dessous de 50 % qu'avec un nombre d'entrées égal à 654.

Nous avons donc proposé une approche plus sélective pour choisir nos exemples parmi les entrées fréquentes du thésaurus afin d'aboutir à une solution plus équilibrée entre le nombre d'exemples et leur taux d'erreur. Cette approche associe deux méthodes de sélection non supervisées produisant chacune un nombre limité d'exemples mais avec un meilleur taux d'erreur.

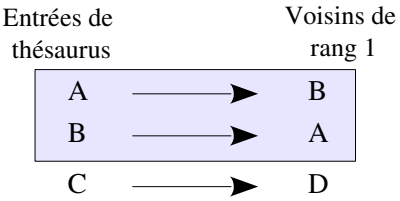

Figure 2. *Sélection d'exemples reposant sur la symétrie des relations sémantiques*

3.4. *Sélection des exemples fondée sur les relations de symétrie dans le thésaurus*

Notre première méthode de sélection d'exemples de mots sémantiquement similaires a été introduite par Ferret (2012). Elle est fondée sur l'hypothèse que les relations de similarité sémantique sont symétriques, ce qui est strictement vrai dans le cas des synonymes de WordNet mais l'est moins pour les mots liés de Moby. En accord avec cette hypothèse, nous avons considéré que si une entrée A du thésaurus initial a pour voisin un mot B, ce voisin a d'autant plus de chances d'être sémantiquement similaire à A que A est lui-même un voisin de B en tant qu'entrée du thésaurus, situation qu'illustre la figure 2. Plus précisément, les résultats du tableau 3 nous ont conduit à limiter l'application de ce principe aux voisins de rang 1 et aux entrées de haute fréquence, dont les voisins sont eux-mêmes généralement des noms de haute fréquence. Nous avons donc appliqué ce principe aux 7 335 entrées dites de haute fréquence du thésaurus, obtenant des cas de symétrie entre entrées et voisins de rang 1 pour 1 592 entrées. Un ensemble de 796 exemples de mots sémantiquement similaires ont finalement été produits puisque les couples (A, B) et (B, A) représentent un même exemple.

3.5. *Sélection des exemples fondée sur les mots composés*

La seconde méthode que nous proposons pour la sélection de couples de mots sémantiquement similaires repose sur l'hypothèse que les mono-termes de deux mots composés sémantiquement similaires occupant dans ces deux composés le même rôle syntaxique sont eux-mêmes susceptibles d'être sémantiquement similaires. Par exemple, le fait que les noms composés *movie_director* et *film_director* soient considérés comme similaires et que les têtes syntaxiques de ces deux mots composés soient identiques conduit à valider la similarité sémantique observée entre *film* et *movie* dans le thésaurus initial.

3.5.1. *Construction d'un thésaurus distributionnel de noms composés*

Le point de départ de cette hypothèse étant la similarité sémantique des mots composés, nous avons commencé par construire un thésaurus distributionnel de noms composés pour l'anglais, à l'image du thésaurus de la section 2 pour les noms simples.

Cette construction a été réalisée à partir du même corpus et avec les mêmes paramètres que pour les mono-termes, à l'exception, bien entendu, de l'ajout d'une étape dans le prétraitement linguistique des documents du corpus pour l'identification des noms composés. Cette identification a été réalisée en deux étapes : un ensemble de noms composés ont d'abord été extraits du corpus AQUAINT-2 sur la base d'un nombre limité de patrons morphosyntaxiques ; les plus fréquents de ces composés ont ensuite été utilisés comme référence dans un processus d'indexation contrôlée.

La première étape a été mise en œuvre grâce à l'outil *mwetoolkit* (Ramisch *et al.*, 2010), qui permet d'extraire efficacement des mots composés d'un corpus à partir du résultat d'un étiqueteur morphosyntaxique, l'étiqueteur *TreeTagger* dans notre cas, en s'appuyant sur un ensemble de patrons morphosyntaxiques. Nous nous sommes limité aux trois patrons de noms composés suivants[10] :

Patrons	**Exemples**
NN NN	village chief, league team, cruise ship, oil producer, movie director
JJ NN	medical information, commercial right, educational program
NN IN NN	sense of duty, director of photography, fall in oil[11]

Un ensemble de 3 246 401 noms composés ont ainsi été extraits du corpus AQUAINT-2 parmi lesquels seuls les 30 121 termes de fréquence supérieure à 100 ont été retenus pour des raisons à la fois de fiabilité et de limitation du vocabulaire pour la construction du thésaurus. L'identification de ces termes de référence dans les textes a ensuite été réalisée en appliquant la stratégie de l'appariement maximal à la sortie lemmatisée du *TreeTagger*. Finalement, des contextes distributionnels constitués à la fois de mots simples et de termes complexes ont été construits suivant les principes de la section 2 et des voisins ont été trouvés pour 29 174 noms composés.

réf.	**#mots éval.**	**#syn. /mot**	**rappel**	**R-préc.**	**MAP**	**P@1**	**P@5**	**P@10**	**P@100**
W	608	1,2	82,0	41,5	50,0	43,4	14,3	8,0	1,0
M	241	2,3	38,0	9,0	12,2	11,2	6,5	4,2	0,9
WM	813	1,6	63,5	32,7	39,5	34,9	12,3	7,1	1,0

Tableau 6. *Évaluation du thésaurus distributionnel de noms composés*

Le tableau 6 donne les résultats de l'évaluation des voisins sémantiques trouvés en prenant comme précédemment en tant que référence WordNet, le thésaurus Moby et la fusion des deux. Le premier constat pouvant être fait est la proportion très faible, par rapport aux mono-termes, d'entrées ayant pu être évaluées : seulement 2,8 % des en-

10. Avec NN : nom (y compris NNS pour les formes plurielles), JJ : adjectif et IN : préposition.
11. Comme on peut le constater avec ce dernier exemple, l'extraction des termes n'est pas parfaite. Même si *fall in oil* correspond à un syntagme nominal correct, il est probable que le terme à extraire soit dans ce cas plus long, comme dans *fall in oil prices*.

trées, à comparer à 83,5 % des entrées pour les mono-termes[12]. De ce fait, les résultats de cette évaluation doivent être considérés avec prudence, même si le nombre d'entrées évaluées est globalement plus élevé que le nombre d'entrées considérées dans beaucoup d'évaluations standard : 70 pour Curran et Moens (2002) ou 353 pour Gabrilovich et Markovitch (2007). Cette prudence est particulièrement de mise pour les mots liés de Moby : les résultats, à l'exception du rappel, sont très significativement inférieurs à ceux obtenus avec les mono-termes mais le nombre d'entrées évaluées – 241 – est aussi faible. À l'inverse, les performances obtenues pour les synonymes de WordNet sont très nettement supérieures sur tous les plans à celles caractérisant les mono-termes, ces résultats étant obtenus pour un nombre d'entrées – 608 – nettement supérieur. Cette différence ne s'expliquant pas par un biais concernant la fréquence des entrées évaluées vis-à-vis respectivement de WordNet et de Moby, il semble donc que le comportement des noms composés soit, du point de vue des similarités distributionnelles, l'inverse de celui des noms simples, favorisant les relations sémantiques paradigmatiques par rapport aux relations syntagmatiques. La plus faible ambiguïté sémantique des noms composés serait une explication possible de ce phénomène qui tend à être confirmé par l'évaluation de plus large ampleur de Ferret (2014).

3.5.2. *Sélection d'exemples à partir de noms composés*

La sélection d'exemples de mots simples sémantiquement similaires à partir de noms composés s'appuie sur la structure syntaxique de ces noms composés. Compte tenu des patrons utilisés pour l'extraction des termes, cette structure prend la forme de l'un des trois grands schémas suivants :

<nom>$_{expansion}$ <nom>$_{tête}$

<adjectif>$_{expansion}$ <nom>$_{tête}$

<nom>$_{tête}$ <préposition> <nom>$_{expansion}$

Chaque nom composé C_i a ainsi été représenté sous la forme d'un couple de noms (T_i, E_i), dans lequel T_i représente la tête syntaxique de C_i et E_i, son expansion, au sens des grammaires de dépendance. Conformément au principe sous-tendant notre méthode de sélection, si un nom composé (T_2, E_2) est un voisin sémantique d'un nom composé (T_1, E_1) (au plus, son $c^{ième}$ voisin), il est probable que T_1 et T_2 ou E_1 et E_2 soient sémantiquement similaires[13]. Comme le montre le tableau 6, notre thésaurus distributionnel de noms composés est cependant loin d'être parfait. Pour limiter les erreurs, nous avons ajouté des contraintes sur l'appariement des constituants des noms composés similaires en nous appuyant sur la similarité distributionnelle de ces constituants. Au final, nous sélectionnons des exemples de noms simples sémantiquement similaires (couples de noms après $\rightarrow$) en appliquant les trois règles suivantes,

12. Il faut néanmoins noter que dans le cas des mono-termes, les entrées considérées pour l'évaluation sont des noms présents dans WordNet. Si l'on considère toutes les entrées du thésaurus, la couverture n'est plus que de 47 %, ce qui reste toutefois très supérieur à 2,8 %.

13. La similarité des expansions ne nous intéresse pas ici lorsque ce sont des adjectifs.

dans lesquelles $E_1 = E_2$ signifie que E_1 et E_2 sont identiques et $T_1 \equiv T_2$ signifie que T_2 est au plus le $n^{ième}$ voisin de T_1 dans notre thésaurus de noms simples :

(1) $T_1 \equiv T_2$ et $E_1 = E_2 \rightarrow (T_1, T_2)$
(*crash*, *accident*) issu de *car_crash* et *car_accident*

(2) $E_1 \equiv E_2$ et $T_1 = T_2 \rightarrow (E_1, E_2)$
(*ocean*, *sea*) de *ocean_floor* et *sea_floor* ; (*jail*, *prison*) de *prison_cell* et *jail_cell*

(3) $E_1 \equiv E_2$ et $T_1 \equiv T_2 \rightarrow (T_1, T_2), (E_1, E_2)$
(*increase*, *rise*) et (*salary*, *pay*) de *salary_increase* et *pay_rise*

4. Expérimentations et évaluation

4.1. *Sélection des exemples de mots sémantiquement similaires*

Le tableau 7 fait une synthèse des résultats de nos deux méthodes de sélection de mots sémantiquement similaires en donnant le pourcentage des couples sélectionnés trouvés dans chacune de nos ressources (W, M et WM) ainsi que la taille de chaque ensemble d'exemples. Dans le cas de la seconde méthode, ces mesures sont également déclinées au niveau de chacune des trois règles de sélection. Les chiffres donnés entre crochets représentent, quant à eux, les pourcentages d'erreurs parmi les contre-exemples. Ces résultats ont été obtenus en fixant expérimentalement la taille du voisinage considéré pour les entrées à 3 pour les noms composés (c) et à 1 pour les noms simples (n). En outre, ces trois règles de sélection ont été appliquées avec l'ensemble des entrées du thésaurus des noms composés et les entrées du thésaurus des noms simples dites de haute fréquence. Les valeurs des paramètres c et n ne résultent pas d'une optimisation particulière mais répondent plutôt à une logique induite des évaluations réalisées : pour les mono-termes, seul le premier voisin est retenu du fait de la faiblesse des résultats alors que pour les multi-termes, le voisinage peut être légèrement élargi du fait d'une meilleure fiabilité des voisins. Il est à noter, par ailleurs, que l'association de deux ensembles d'exemples sélectionnés par des méthodes différentes rend les résultats plus stables vis-à-vis des valeurs de c et n.

L'évaluation de la seconde méthode de sélection montre d'abord que la règle (3), qui est *a priori* la moins fiable des trois, ne produit effectivement qu'un petit nombre d'exemples tendant à dégrader les résultats. De ce fait, seule la combinaison des règles (1) et (2) a ensuite été utilisée. Cette évaluation montre en outre que les têtes de deux noms composés sémantiquement liés ont davantage tendance à être elles-mêmes similaires si leurs expansions sont égales, que n'ont tendance à être similaires des expansions de deux noms composés dont les têtes sont égales. Ce résultat peut se comprendre si l'on fait l'hypothèse, *a priori* fondée, que la tête d'un composé est plus représentative du sens de ce composé que son expansion. Plus globalement, le tableau 7 laisse apparaître que la première méthode de sélection est supérieure à la seconde mais que leur association produit un compromis intéressant entre le nombre d'exemples –

méthode	**W**		**M**		**WM**		**# exemples**
symétrie	36,6	[2,0]	55,5	[14,4]	59,7	[12,4]	796
règle (1)	19,3	[2,5]	56,1	[16,6]	56,9	[16,1]	921
règle (2)	16,2	[1,5]	42,4	[16,0]	44,7	[14,7]	308
règle (3)	13,5	[1,4]	45,9	[17,8]	46,2	[16,9]	40
règles (1,2)	*17,8*	[2,5]	*52,2*	[16,8]	*53,0*	[16,1]	*1 115*
règles (1,2,3)	17,6	[2,5]	51,7	[16,6]	52,4	[15,9]	1 131
symétrie + *règles (1,2)*	*23,5*	[2,3]	*52,5*	[16,3]	*54,3*	[15,0]	*1 710*
symétrie + règles (1,2,3)	23,3	[2,1]	52,1	[15,7]	53,9	[14,5]	1 725

Tableau 7. *Évaluation de la qualité des exemples sélectionnés par rapport aux ressources de référence (% bons exemples [% mauvais contre-exemples])*

1 710 – et son taux d'erreur – 45,7 % avec WM comme référence. Cette complémentarité est également illustrée par le faible nombre d'exemples – 201 – qu'elles partagent.

4.2. *Mise en œuvre du réordonnancement des voisins*

La mise en œuvre effective de notre approche de réordonnancement des voisins sémantiques nécessite de fixer un certain nombre de paramètres liés aux SVM. De même que Hagiwara (2008), nous avons adopté un noyau RBF et une stratégie de type recherche en grille (*grid search*) pour l'optimisation du paramètre γ fixant la largeur de la fonction gaussienne du noyau RBF et du paramètre C d'ajustement entre la taille de la marge et le taux d'erreur. Cette optimisation a été réalisée pour chaque ensemble d'apprentissage considéré en se fondant sur la mesure de précision calculée dans le cadre d'une validation croisée divisant ces ensembles en cinq parties. Chaque modèle SVM correspondant a été construit en utilisant l'outil LIBSVM (Chang et Lin, 2001) puis appliqué à la totalité des 14 670 noms cibles de notre évaluation initiale. Plus précisément, pour chaque nom cible NC, une représentation d'exemple a été construite pour chaque couple (NC, voisin de NC) et a été soumise au modèle SVM considéré en mode classification. L'ensemble de ces voisins ont ensuite été réordonnés suivant la valeur de la fonction de décision ainsi calculée pour chaque voisin.

4.3. *Évaluation*

Le tableau 8 donne les résultats globaux du réordonnancement réalisé sur la base des exemples sélectionnés par chacune des deux méthodes présentées (*symétrie* et *composés*) et leur combinaison (*sym.+comp.*) tandis que la figure 3 en donne une vision plus détaillée en fonction des tranches fréquentielles pour la seule combinaison *sym.+comp.* Cette dernière correspond à ce que nous avons appelé dans l'introduction *fusion précoce*. Chacun de ces trois thésaurus a été évalué selon les mêmes principes

qu'à la section 2.2. La valeur de chaque mesure se voit associer sa différence avec la valeur correspondante pour le thésaurus initial dans le tableau 3. Enfin, comme l'évaluation s'applique au résultat d'un réordonnancement, les mesures de rappel et de précision au rang le plus lointain ne changent pas et ne sont pas rappelées.

méthode	réf.	R-préc.	MAP	P@1	P@5	P@10
symétrie	W	7,8 $(-0,4)$	9,4 $(-0,4)$	11,2 $(-0,5)$ ‡	5,0 $(-0,1)$ ‡	3,3 $(-0,1)$ ‡
	M	7,1 $(+0,4)$	3,4 $(+0,2)$	27,3 $(+3,2)$	17,6 $(+1,2)$	13,7 $(+0,7)$
	WM	8,0 $(+0,3)$	5,7 $(+0,1)$	24,6 $(+2,1)$	14,9 $(+0,8)$	11,4 $(+0,6)$
composés	W	7,2 $(-1,0)$	8,8 $(-1,0)$	10,4 $(-1,3)$	4,6 $(-0,5)$	3,1 $(-0,3)$
	M	7,1 $(+0,4)$	3,3 $(+0,1)$	26,8 $(+2,7)$	17,4 $(+1,0)$	13,5 $(+0,5)$
	WM	7,8 $(+0,1)$	5,5 $(-0,1)$	24,0 $(+1,5)$	14,6 $(+0,5)$	11,2 $(+0,4)$
sym.+comp.	W	7,9 $(-0,3)$ ‡	9,5 $(-0,3)$ ‡	11,5 $(-0,2)$ ‡	5,1 $(+0,0)$ ‡	3,4 $(+0,0)$ ‡
	M	7,2 $(+0,5)$	3,5 $(+0,3)$	27,9 $(+3,8)$	18,1 $(+1,7)$	14,1 $(+1,1)$
	WM	8,0 $(+0,3)$	5,8 $(+0,2)$	25,3 $(+2,8)$	15,3 $(+1,2)$	11,7 $(+0,9)$

Tableau 8. *Réordonnancement des voisins sémantiques de toutes les entrées du thésaurus initial pour chaque méthode de sélection d'exemples et leur combinaison*[14]

La tendance générale est claire : le processus de réordonnancement conduit à une amélioration significative des résultats globaux pour toutes les méthodes dans le cas des références M et WM, la seule exception étant une très légère diminution de la MAP pour la référence WM dans le cas de la méthode dite *composés*. Parallèlement, une diminution des résultats est observée pour la référence W, jugée statistiquement non significative. En d'autres termes, par rapport au thésaurus initial, la procédure de réordonnancement tend à favoriser les mots similaires au détriment des synonymes. Cette tendance n'est pas surprenante compte tenu du principe de ce réordonnancement : les premiers sont en effet mieux représentés que les seconds dans les exemples sélectionnés du fait même de leur meilleure représentation au niveau global. Les modèles SVM appris ne font en l'occurrence qu'amplifier un état de fait déjà présent initialement. Ce biais est particulièrement fort pour la méthode de sélection fondée sur les noms composés, comme l'illustre le tableau 8. Cependant, les résultats du tableau 8 montrent clairement l'intérêt de l'association des deux méthodes de sélection : d'un côté, la sélection fondée sur la symétrie des relations vient rééquilibrer ce biais au bénéfice des résultats globaux, de l'autre côté, les exemples apportés par la méthode fondée sur les mots composés élargissent l'ensemble d'apprentissage de celle fondée sur la symétrie dont la plus grande qualité des résultats ne suffit pas totalement à compenser la taille restreinte.

L'analyse des résultats donnés par la figure 3 en termes de fréquence des mots met en évidence une seconde grande tendance : l'amélioration produite par le réordonnancement est d'autant plus sensible que la fréquence de l'entrée du thésaurus est faible.

14. La significativité statistique des différences a été évaluée grâce à un test de Wilcoxon avec un seuil de 0,05, les échantillons étant appariés. Seules les différences suivies du signe ‡ sont considérées comme non significatives.

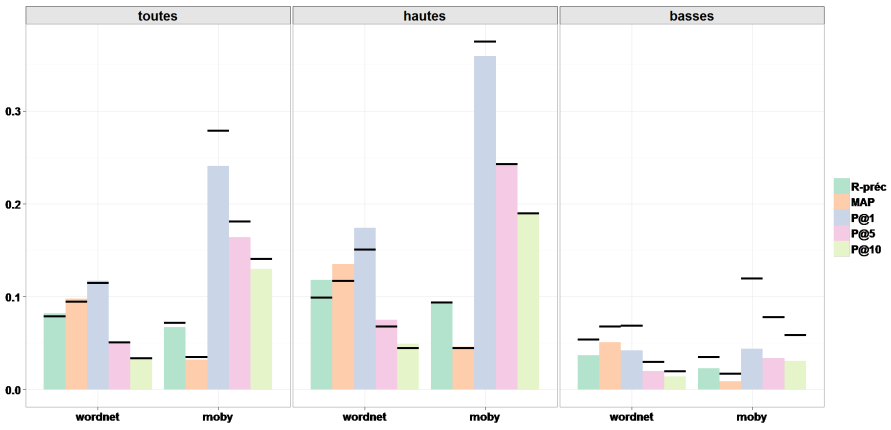

Figure 3. *Résultats du réordonnancement du thésaurus initial suivant les tranches fréquentielles. L'histogramme représente les valeurs pour le thésaurus initial et les barres* ***noires****, les valeurs pour le réordonnancement conjuguant l'utilisation de la symétrie des relations et celle des mots composés.*

Ainsi, pour les noms de faible fréquence, cette amélioration s'observe quelle que soit la référence tandis que pour les noms de forte fréquence, la variation est négative pour certaines références et mesures et positive pour d'autres. Ce constat montre que le réordonnancement tend ainsi à rééquilibrer le thésaurus initial, très fortement biaisé vers les fortes fréquences. L'évaluation de ces trois thésaurus confirme, par ailleurs, les résultats du tableau 7 à propos de chaque ensemble d'exemples sélectionnés : le thésaurus construit à partir des exemples de la première méthode de sélection est meilleur que celui construit à partir des exemples de la seconde méthode de sélection et les deux sont nettement dépassés par le thésaurus construit à partir de la fusion des deux ensembles d'exemples.

Enfin, le tableau 9 illustre pour une entrée spécifique du thésaurus initial, en l'occurrence le mot *esteem*, l'impact du réordonnancement fondé sur les deux méthodes de sélection d'exemples. Ce tableau donne d'abord pour cette entrée ses synonymes dans **WordNet** et les premiers mots qui lui sont liés dans Moby. Il fait ensuite apparaître que dans notre thésaurus initial, les deux premiers voisins de cette entrée présents dans une de nos deux ressources de référence sont les mots *admiration*, au rang 3, et le mot *respect*, au rang 7. Le réordonnancement améliore significativement la situation puisque ces deux mots deviennent les deux premiers voisins tandis que le troisième synonyme donné par WordNet passe du rang 22 au rang 12. Par ailleurs, le nombre de voisins présents parmi les 14 premiers mots liés de Moby passe de 3 à 6.

WordNet	respect, admiration, regard
Moby	admiration, appreciation, acceptance, dignity, regard, respect, account, adherence, consideration, estimate, estimation, fame, greatness, homage, honor, prestige, prominence, reverence, veneration + 74 mots liés supplémentaires
initial	cordiality, gratitude, **admiration**, comradeship, back-scratching, perplexity, **respect**, ruination, appreciation, neighbourliness, trust, empathy, suffragette, goodwill ...
après réordonnancement	**respect**, **admiration**, trust, recognition, gratitude, confidence, affection, understanding, solidarity, dignity, appreciation, **regard**, sympathy, acceptance ...

Tableau 9. *Impact du réordonnancement pour l'entrée* esteem

4.4. *Fusion des thésaurus*

Les performances de l'approche *sym.+comp.* présentées à la section précédente illustrent l'intérêt de combiner des approches reposant sur des critères différents. Les travaux sur la fusion de données font généralement la distinction entre fusion précoce et fusion tardive (Atrey *et al.*, 2010), la première opérant au niveau des représentations, la seconde au niveau des résultats. L'approche précédente *sym.+comp.* peut être considérée comme relevant d'une fusion précoce dans la mesure où la fusion, ici des ensembles d'apprentissage, s'opère en amont du processus de réordonnancement. À l'instar de Curran (2002), nous avons également testé une fusion tardive des deux méthodes de réordonnancement de thésaurus. Chaque thésaurus résultat donnant pour chacune de ses entrées une liste de voisins ordonnés selon l'ordre décroissant de leur proximité avec leur entrée, la solution la plus évidente est de procéder pour chaque entrée à une fusion de la liste des voisins issue de chacun des thésaurus résultats en adoptant une méthode classique de vote. Le tableau 10 donne les résultats que nous avons obtenus avec quatre de ces méthodes. Trois d'entre elles, *Borda*[15], *Condorcet*[16] (Nuray et Can, 2006) et *Reciprocal Rank Fusion* (RRF, avec le paramètre $k = 60$ de (Cormack *et al.*, 2009))[17], s'appuient uniquement sur les rangs tandis que *CombSum*,

15. La méthode *Borda* attribue à chaque voisin de chaque liste à fusionner un poids égal à : taille de la liste − rang du voisin − 1. Le poids final de chaque voisin au sein de l'union des listes est donné par la somme de ses poids dans chaque liste.

16. La méthode *Condorcet* fusionne les listes de voisins en s'assurant que dans le résultat de la fusion, pour un voisin A de rang i et un voisin B de rang j avec $i < j$, le rang de A dans chacune des listes fusionnées est inférieur au rang de B pour une majorité de ces listes.

17. RRF classe les voisins selon l'ordre croissant du score $\sum_l \frac{1}{k+rang(v,l)}$ où $rang(v, l)$ est le rang du voisin v dans la liste l.

Thésaurus	**R-préc.**	**MAP**	**P@1**	**P@5**	**P@10**
initial	7,7	5,6	22,5	14,1	10,8
symétrie	+ 0,3	+ 0,1	+ 2,1	+ 0,8	+ 0,6
composés	+ 0,1	+ 0,0	+ 2,0	+ 0,9	+ 0,6
sym.+ comp.	+ 0,3	+ 0,2	+ 2,8	+ 1,2	+ 0,9
RRF	+ 0,7	+ 0,6	+ 3,7	+ 1,9	+ 1,4
Borda	+ 0,7	+ 0,5	+ 3,6	+ 1,7	+ 1,3
Condorcet	+ 0,5	+ 0,4	+ 3,4	+ 1,6	+ 1,2
CombSum	**+ 0,9**	**+ 0,8**	**+ 4,7**	**+ 2,2**	**+ 1,5**

Tableau 10. *Comparaison des différentes méthodes de fusion avec la référence [WM]*

utilisée ici avec une normalisation des valeurs selon Lee (1997)[18], exploite les valeurs de similarité. Trois thésaurus sont ainsi fusionnés : le thésaurus initial, le thésaurus réordonné grâce au critère de symétrie et celui réordonné grâce aux mots composés.

Outre les résultats pour ces quatre méthodes de fusion, donnés pour la seule référence [WM] pour des raisons de place, le tableau 10 rappelle les résultats pour les thésaurus fusionnés. Ces résultats, de même que ceux issus des fusions, sont donnés en différence de valeur par rapport au thésaurus initial. Un premier constat d'évidence s'impose : les méthodes de fusion permettent toutes de dépasser les résultats de chacun des quatre thésaurus fusionnés. Les gains en termes de R-précision et de MAP apparaissent modestes, mais la référence étant [WM], le nombre de voisins de référence est important, ce qui a un impact direct sur ces deux mesures. En revanche, les gains sont nettement plus substantiels concernant la précision aux rangs 1, 5 et 10. Dans une optique applicative, cette tendance est la plus importante : seuls les voisins des tout premiers rangs sont en effet utilisés dans un tel contexte. Parmi l'ensemble des méthodes de fusion, *CombSum* se détache clairement pour toutes les mesures, l'effet étant particulièrement notable pour la précision au rang 1. L'utilisation des valeurs de similarité, dont la normalisation est indispensable dans le cas présent, s'avère donc supérieure à celle des rangs. Parmi les méthodes exploitant les rangs, *RRF* est la meilleure option, de façon similaire aux constatations de Cormack *et al.* (2009) dans le domaine de la recherche d'information.

18. *CombSum* attribue un score à chaque voisin, après fusion des listes, égal à la somme des valeurs de similarité de ce voisin dans chacune des listes. La normalisation des valeurs de similarité $sim(v)$ est, quant à elle, donnée par $\frac{sim(v)-min_{sim}}{max_{sim}-min_{sim}}$.

5. Conclusions et perspectives

Dans cet article, nous avons présenté une méthode fondée sur l'amorçage pour améliorer un thésaurus distributionnel. Plus précisément, cette méthode se fonde sur le réordonnancement des voisins sémantiques de ce thésaurus par le biais d'un classifieur SVM. Ce classifieur est entraîné à partir d'un ensemble d'exemples et de contre-exemples sélectionnés de façon non supervisée en appliquant deux critères faibles fondés sur la similarité distributionnelle. L'un exploite la symétrie des relations sémantiques tandis que l'autre s'appuie sur l'appariement des constituants de noms composés similaires. Nous avons plus particulièrement testé deux méthodes de fusion des résultats de ces deux critères. L'une, qualifiée de précoce, consiste à regrouper les ensembles d'apprentissage produits par les deux critères. Les améliorations apportées par ce biais sont plus particulièrement notables pour les noms de fréquence faible et pour des mots similaires plutôt que pour de stricts synonymes. L'autre méthode, dite tardive, fusionne les listes de voisins réordonnancés produits par chacun des critères et permet d'obtenir ainsi des résultats nettement supérieurs à la fusion précoce.

Au-delà de leur analyse précise, les résultats obtenus doivent être replacés dans un contexte plus large sur les possibilités offertes pour améliorer les thésaurus distributionnels. Comme nous l'avons vu à la section 2.3, certains paramètres de base intervenant dans la construction des thésaurus distributionnels ont une incidence sensible sur la qualité de ceux-ci, en particulier la taille du corpus utilisé et la nature des contextes distributionnels. Une façon évidente d'améliorer la qualité des thésaurus est ainsi d'augmenter la taille de leur corpus source, tendance que l'on observe clairement dans les travaux récents où les corpus sont rarement inférieurs au milliard de mots. Cette approche est concevable en domaine général où les corpus de grandes tailles ne sont pas rares. Elle est plus difficile à mettre en œuvre dans beaucoup de domaines spécialisés, même si ce n'est pas le cas de tous.

Concernant la nature des contextes distributionnels, l'utilisation de cooccurrents syntaxiques apporte également un plus indéniable. Elle se heurte à la disponibilité d'un analyseur syntaxique et aux temps de traitement qu'il engendre mais le tableau 5 illustre assez bien le fait que les cooccurrents syntaxiques permettent de compenser une taille de corpus plus faible, ce qui module un peu le problème du temps de traitement. Dans le registre des cooccurrents graphiques, Claveau *et al.* (2014) montrent que la directionnalité des cooccurrents, conjuguée dans le cas présent à une fonction de pondération et à une mesure de similarité issues de la recherche d'information[19], a une incidence importante sur les résultats, ce que Curran (2003) avait déjà noté à une moindre échelle. Enfin, des travaux récents suggèrent des gains intéressants en lien avec d'autres aspects des contextes distributionnels, comme la sélection des éléments constituant les contextes et la normalisation de leur poids (Polajnar et Clark, 2014) ou l'adoption d'une variante de la fonction de pondération PPMI compensant sa sensibilité aux faibles fréquences (Levy *et al.*, 2015).

19. Les expériences menées depuis confirment ce gain pour le couple *cosinus* – PPMI.

Comparés aux différences de performance imputables à tous ces facteurs, les gains obtenus par les méthodes d'amorçage proposées sont en apparence assez limités. Mais plusieurs points sont à prendre en considération pour juger de leur intérêt. Tout d'abord, les principes d'amorçage utilisés peuvent être appliqués indépendamment du mode de représentation et de constitution des données distributionnelles. Les améliorations issues de ces données et celles issues des méthodes proposées ici sont donc complémentaires, voire susceptibles de synergie : *a priori*, plus la qualité du thésaurus initial est élevée et plus l'effet d'amorçage doit être lui-même important, de meilleurs exemples devant conduire à un meilleur classifieur et donc *in fine*, à un meilleur réordonnancement des voisins. C'est un effet qu'il nous reste néanmoins à confirmer.

La seconde dimension à prendre en compte est la différenciation des résultats obtenus en fonction des tranches fréquentielles. Si le gain au niveau global reste modeste, il est en revanche beaucoup plus significatif au niveau des entrées de basse fréquence ainsi que l'illustre la figure 3. Or, ces entrées représentent la moitié du vocabulaire du corpus. L'effet plutôt négatif du réordonnancement au niveau des hautes fréquences, sauf dans certains cas avec Moby comme référence, se comprend d'ailleurs assez bien en considérant que les exemples sélectionnés en sont issus, ce qui rend la difficulté d'une amélioration possible beaucoup plus grande. Sur le plan applicatif, une façon directe d'élever la performance globale est donc de construire un thésaurus hybride composé des voisins du thésaurus initial pour les entrées de haute fréquence et des voisins du thésaurus réordonnancé pour les entrées de basse fréquence.

Nous envisageons, par ailleurs, plusieurs pistes d'extension de ce travail. En premier lieu, nous souhaitons élargir les critères de sélection non supervisée d'exemples. Alors que les techniques de sélection expérimentées reposent toutes deux sur des thésaurus distributionnels, des critères s'attachant aux occurrences des mots et à leur environnement plutôt qu'à une représentation distributionnelle sont également envisageables, comme l'utilisation de patrons linguistiques classiques d'extraction de synonymes ou l'exploitation des chaînes de coréférence, à l'instar de Adel et Schütze (2014) mais en exploitant un système de résolution des coréférences n'intégrant pas déjà la connaissance que l'on cherche à extraire. Au-delà d'une approche purement non supervisée de sélection d'exemples, il serait également intéressant d'étudier dans quelle mesure un ensemble très restreint d'exemples fournis manuellement peut aider ou non à améliorer significativement les résultats. Enfin, sur un autre plan, l'évaluation menée, fondée sur la comparaison avec des ressources de référence, pourrait être complétée avec profit par une évaluation extrinsèque permettant de juger de l'impact des améliorations du thésaurus distributionnel sur une tâche à laquelle il contribue, à l'image de Claveau et Kijak (2015) dans le champ de la recherche d'information pour l'expansion de requêtes. Nous serions, pour notre part, intéressé par une application à la segmentation thématique, dans le prolongement de Adam et Morlane-Hondère (2009).

6. Bibliographie

Adam C., Morlane-Hondère F., « Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique », *RECITAL'09*, Senlis, France, 2009.

Adel H., Schütze H., « Using Mined Coreference Chains as a Resource for a Semantic Task », *EMNLP 2014*, Doha, Qatar, p. 1447-1452, 2014.

Atrey P. K., Hossain M. A., El Saddik A., Kankanhalli M. S., « Multimodal fusion for multimedia analysis : a survey », *Multimedia Systems*, vol. 16, n° 6, p. 345-379, 2010.

Baroni M., Dinu G., Kruszewski G., « Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors », *ACL 2014*, Baltimore, Maryland, USA, p. 238-247, 2014.

Broda B., Piasecki M., Szpakowicz S., « Rank-Based Transformation in Measuring Semantic Relatedness », *Canadian AI 2009*, p. 187-190, 2009.

Bullinaria J. A., Levy J. P., « Extracting semantic representations from word co-occurrence statistics : stop-lists, stemming, and SVD », *Behavior Research Methods*, vol. 44, n° 3, p. 890-907, 2012.

Chang C.-C., Lin C.-J., *LIBSVM : a library for support vector machines*, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 2001.

Claveau V., Kijak E., « Thésaurus distributionnels pour la recherche d'information et vice-versa », *CORIA 2015*, Paris, France, 2015.

Claveau V., Kijak E., Ferret O., « Improving distributional thesauri by exploring the graph of neighbors », *COLING 2014*, Dublin, Ireland, p. 709-720, 2014.

Cormack G. V., Clarke C. L. A., Buettcher S., « Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods », *SIGIR'09*, p. 758-759, 2009.

Curran J., « Ensemble Methods for Automatic Thesaurus Extraction », *EMNLP 2002*, p. 222-229, 2002.

Curran J., Moens M., « Improvements in automatic thesaurus extraction », *Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, Philadelphia, USA, p. 59-66, 2002.

Curran J. R., From Distributional to Semantic Similarity, PhD thesis, University of Edinburgh, 2003.

Dinu G., Lapata M., « Measuring Distributional Similarity in Context », *EMNLP 2010*, Cambridge, MA, USA, p. 1162-1172, 2010.

Erk K., Padó S., « Exemplar-Based Models for Word Meaning in Context », *ACL 2010, short paper*, Uppsala, Sweden, p. 92-97, 2010.

Ferret O., « Similarité sémantique et extraction de synonymes à partir de corpus », *TALN 2010*, Montréal, Canada, 2010.

Ferret O., « Combining Bootstrapping and Feature Selection for Improving a Distributional Thesaurus », *ECAI 2012*, Montpellier, France, p. 336-341, 2012.

Ferret O., « Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel », *TALN 2013*, Les Sables d'Olonne, France, p. 48-61, 2013.

Ferret O., « Compounds and distributional thesauri », *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, p. 2979-2984, 2014.

Ferret O., « Early and Late Combinations of Criteria for Reranking Distributional Thesauri », *ACL-IJCNLP 2015, short paper session*, Beijing, China, p. 470-476, 2015a.

Ferret O., « Typing relations in distributional thesauri », *in* N. Gala, R. Rapp, G. Bel (eds), *Language Production, Cognition, and the Lexicon*, vol. 48 of *Text, Speech and Language Technology*, Springer, p. 113-134, 2015b.

Freitag D., Blume M., Byrnes J., Chow E., Kapadia S., Rohwer R., Wang Z., « New experiments in distributional representations of synonymy », *CoNLL 2005*, p. 25-32, 2005.

Gabrilovich E., Markovitch S., « Computing semantic relatedness using wikipedia-based explicit semantic analysis », *IJCAI 2007*, Hyderabad, India, p. 6-12, 2007.

Grave E., Obozinski G., Bach F., « A Markovian approach to distributional semantics with application to semantic compositionality », *COLING 2014*, p. 1447-1456, 2014.

Grefenstette G., *Explorations in automatic thesaurus discovery*, Kluwer, 1994.

Hagiwara M., « A Supervised Learning Approach to Automatic Synonym Identification Based on Distributional Features », *ACL-08 : HLT, student session*, p. 1-6, 2008.

Henestroza Anguiano E., Candito M., « Probabilistic Lexical Generalization for French Dependency Parsing », *SP-Sem-MRL 2012 workshop*, Jeju, Republic of Korea, p. 1-11, 2012.

Heylen K., Peirsman Y., Geeraerts D., Speelman D., « Modelling Word Similarity : An Evaluation of Automatic Synonymy Extraction Algorithms », *LREC'08*, 2008.

Hill F., Reichart R., Korhonen A., « SimLex-999 : Evaluating Semantic Models with (Genuine) Similarity Estimation », *CoRR*, 2014.

Huang E. H., Socher R., Manning C. D., Ng A. Y., « Improving word representations via global context and multiple word prototypes », *ACL'12*, p. 873-882, 2012.

Kanerva P., Kristoferson J., Holst A., « Random Indexing of Text Samples for Latent Semantic Analysis », *CogSci 2000*, Lawrence Erlbaum, p. 103-6, 2000.

Kazama J., De Saeger S., Kuroda K., Murata M., Torisawa K., « A Bayesian Method for Robust Estimation of Distributional Similarities », *ACL 2010*, p. 247-256, 2010.

Kiela D., Clark S., « A Systematic Study of Semantic Vector Space Model Parameters », 2^{nd} *CVSC workshop*, Gothenburg, Sweden, p. 21-30, 2014.

Landauer T. K., Dumais S. T., « A solution to Plato's problem : the latent semantic analysis theory of acquisition, induction, and representation of knowledge », *Psychological review*, vol. 104, n° 2, p. 211-240, 1997.

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

Lee J. H., « Analyses of Multiple Evidence Combination », *SIGIR'97*, ACM, p. 267-276, 1997.

Levy O., Goldberg Y., Dagan I., « Improving Distributional Similarity with Lessons Learned from Word Embeddings », *TALC*, vol. 3, p. 211-225, 2015.

Lin D., « PRINCIPAR : An efficient, broad-coverage, principle-based parser », *COLING'94*, Kyoto, Japan, p. 42-48, 1994.

Lin D., « Automatic retrieval and clustering of similar words », *ACL-COLING'98*, Montréal, Canada, p. 768-774, 1998.

Mikolov T., Yih W.-t., Zweig G., « Linguistic Regularities in Continuous Space Word Representations », *NAACL HLT 2013*, Atlanta, Georgia, p. 746-751, 2013.

Miller G. A., « WordNet : An On-Line Lexical Database », *International Journal of Lexicography*, 1990.

Min B., Shi S., Grishman R., Lin C.-Y., « Ensemble Semantics for Large-scale Unsupervised Relation Extraction », *EMNLP-CoNLL 2012*, Jeju Island, Korea, p. 1027-1037, July, 2012.

Nuray R., Can F., « Automatic Ranking of Information Retrieval Systems Using Data Fusion », *Information Processing and Management*, vol. 42, n° 3, p. 595-614, 2006.

Pantel P., Crestan E., Borkovsky A., Popescu A.-M., Vyas V., « Web-Scale Distributional Similarity and Entity Set Expansion », *EMNLP 2009*, Singapore, p. 938-947, 2009.

Pennington J., Socher R., Manning C., « Glove : Global Vectors for Word Representation », *EMNLP 2014*, Doha, Qatar, p. 1532-1543, 2014.

Polajnar T., Clark S., « Improving Distributional Semantic Vectors through Context Selection and Normalisation », *EACL 2014*, Gothenburg, Sweden, p. 230-238, 2014.

Popescu A., Grefenstette G., « Social Media Driven Image Retrieval », *1^{st} ACM International Conference on Multimedia Retrieval (ICMR'11)*, ACM, Trento, Italy, p. 1-8, 2011.

Ramisch C., Villavicencio A., Boitet C., « mwetoolkit : a Framework for Multiword Expression Identification », *LREC'10*, Valetta, Malta, p. 662-669, 2010.

Reisinger J., Mooney R. J., « Multi-Prototype Vector-Space Models of Word Meaning », *HLT-NAACL 2010*, Los Angeles, California, USA, p. 109-117, June, 2010.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *International Conference on New Methods in Language Processing*, Manchester, UK, p. 44-49, 1994.

Van de Cruys T., Mining for Meaning. The Extraction of Lexico-semantic Knowledge from Text, PhD thesis, University of Groningen, The Netherlands, 2010.

Van der Plas L., Bouma G., « Syntactic Contexts for Finding Semantically Related Words », *CLIN 2004*, Leiden, Netherlands, 2004.

Voorhees E., Graff D., *AQUAINT-2 Information-Retrieval Text Research Collection*, `https://catalog.ldc.upenn.edu/LDC2008T25`. 2008.

Ward G., *Moby Thesaurus*, Moby Project, 1996.

Weeds J., Measures and Applications of Lexical Distributional Similarity, PhD thesis, Department of Informatics, University of Sussex, 2003.

Wu Z., Palmer M., « Verbs semantics and lexical selection », *ACL'94*, Las Cruces, New Mexico, USA, p. 133-138, 1994.

Yamamoto K., Asakura T., « Even Unassociated Features Can Improve Lexical Distributional Similarity », *NLPIX 2010 workshop*, Beijing, China, p. 32-39, 2010.

Zhitomirsky-Geffet M., Dagan I., « Bootstrapping Distributional Feature Vector Quality », *Computational Linguistics*, vol. 35, n° 3, p. 435-461, 2009.

Méthode semi-compositionnelle pour l'extraction de synonymes des termes complexes

Amir Hazem — Béatrice Daille

Université de Nantes, LINA UMR CNRS 6241
2 rue de la Houssinière, BP 92208
F-44322 Nantes cedex 3
{amir.hazem,beatrice.daille}@univ-nantes.fr

RÉSUMÉ. *L'extraction de synonymes et des mots sémantiquement liés est une tâche utile en recherche d'information et en traitement automatique des langues. L'analyse distributionnelle a fourni un cadre théorique et opérationnel pour la détection de synonymes en corpus qui a principalement été exploité pour la découverte des synonymes de mots simples relevant de la langue générale. Dans cet article, nous nous intéressons à la découverte de synonymes de phrasèmes nominaux relevant de domaines de spécialités. Nous proposons une méthode semi-compositionnelle non supervisée qui mêle analyse compositionnelle et analyse distributionnelle. Nous montrons que cette méthode permet d'identifier nombre de termes complexes synonymes non découverts par la méthode état de l'art fondée sur une analyse compositionnelle seule, tout en étant beaucoup plus précise que la méthode exploitant la seule analyse distributionnelle.*

ABSTRACT. *Automatic synonyms and semantically related word extraction is a challenging task, useful in many NLP applications such as question answering, search query expansion, text summarization, etc. While different studies addressed the task of word synonym extraction, only a few investigations tackled the problem of acquiring synonyms of multi-word terms (MWT) from specialized corpora. To extract pairs of synonyms of multi-word terms, we propose in this paper an unsupervised semi-compositional method that makes use of distributional semantics and exploit the compositional property shared by most MWT. We show that our method outperforms significantly the state-of-the-art.*

MOTS-CLÉS : *synonymes, termes complexes, compositionnalité, sémantique distributionnelle, méthode non supervisée.*

KEYWORDS: *synonyms, multi-word terms, compositionality, distributional semantics, unsupervised method.*

1. Introduction

L'extraction de synonymes et des mots sémantiquement liés est une tâche utile en recherche d'information et en traitement automatique des langues. L'analyse distributionnelle a fourni un cadre théorique et opérationnel pour la détection de synonymes en corpus qui a principalement été exploité pour la découverte des synonymes de mots simples relevant de la langue générale. La synonymie des termes est un phénomène couramment rencontré dans les textes. En revanche, les synonymes des termes ne figurent généralement pas dans une ressource dictionnairique (Kremer *et al.*, 2014) à l'exception de quelques sigles et acronymes. Dans cet article, nous nous intéressons à la découverte de synonymes de phrasèmes nominaux relevant de domaines de spécialités, et en particulier aux synonymes de termes complexes où l'un des lexèmes du terme est substitué par un de ses synonymes comme *énergie éolienne* ↔ *courant éolien* dans le domaine de spécialité des énergies renouvelables. Le principe de compositionnalité sémantique est utilisé pour générer les synonymes d'un terme complexe : deux termes complexes sont synonymes si l'un des composants est substitué par l'un de ses synonymes ; l'existence du terme complexe généré est vérifiée en corpus. Cette méthode est identique à celle de Hamon et Nazarenko (2001) mais diffère sur la fourniture des synonymes des composants des termes complexes. Hamon et Nazarenko (2001) s'appuient sur un dictionnaire de langue générale. Nous proposons d'exploiter plutôt les mots sémantiquement liés identifiés par une analyse distributionnelle un peu bridée. Nous démontrons que cette méthode appelée *semi-compositionnelle* permet d'extraire des synonymes de termes complexes avec une bonne précision là où l'approche à base de dictionnaire échoue. Des expérimentations sont faites sur trois langues et deux domaines de spécialités.

La suite de cet article est organisée de la manière suivante. La section 2 fait un tour d'horizon des principales méthodes employées pour la détection automatique de synonymes dans les textes. La section 3 décrit le principe de compositionnalité et de synonymie. La section 4 présente l'approche semi-compositionnelle utilisée pour la génération de synonymes de termes complexes et les filtrages en corpus opérés. La section 5 décrit les différentes ressources linguistiques utilisées dans nos expériences, et en particulier les listes de référence exploitées. La section 6 évalue la contribution de divers paramètres de l'analyse distributionnelle vis-à-vis de la qualité des termes complexes synonymes extraits. Les sections 7 et 8 détaillent les acquis et les points restant à améliorer.

2. Identification automatique des synonymes

L'identification automatique de synonymes, et plus généralement des mots sémantiquement liés, est possible grâce à trois grandes familles d'approches, à savoir : (i) les approches distributionnelles qui effectuent une analyse distributionnelle en corpus dans un cadre purement monolingue, (ii) les approches qui exploitent les énoncés définitoires en corpus ou présents dans des ressources lexicales, et (iii) les approches d'alignements lexicaux multilingues qui s'appuient sur des textes traduits considérés

comme des réservoirs de synonymes. Une exploitation par combinaison des approches de ces trois familles permet, elle aussi, d'améliorer la qualité des synonymes extraits (Wu et Zhou, 2003).

L'analyse distributionnelle est la méthode la plus populaire pour la découverte de synonymes en corpus. Elle est fondée sur la conception contextualiste du sens d'un mot : « On reconnaît un mot à ses fréquentations »[1]. Le sens d'un mot est défini par l'ensemble des contextes dans lequel il apparaît. Deux mots, ou plus largement deux unités lexicales, partageant des contextes similaires, vont être sémantiquement liés (Harris, 1954). Plus les contextes seront similaires plus les mots le seront également. Plusieurs études ont exploité l'analyse distributionnelle automatique pour la détection de synonymes en corpus (Hindle, 1990 ; Grefenstette, 1994 ; Lin, 1998 ; Hagiwara, 2008 ; Ferret, 2010 ; Ferret, 2013). Elles diffèrent selon la définition du contexte adoptée, les méthodes pour comparer ces contextes et l'ordonnancement de ces mots identifiés comme sémantiquement liés. Lin (1998), par exemple, a introduit l'idée que les mots partageant le plus de relations syntaxiques étaient les plus favorables à être en relation de synonymie. Cette idée a été reprise et étendue au chemin syntaxique (Hagiwara, 2008) dans le but de pallier le manque de relations de dépendance syntaxique directes. Lorsqu'un nombre important de mots sémantiquement liés sont détectés, Ferret (2013) améliore leur ordonnancement en détectant et en déclassant les mots les plus ambigus.

Les énoncés définitoires sont des réservoirs à synonymes. Ceux-ci sont détectés dans les textes à l'aide de patrons morphosyntaxiques ou extraits de ressources lexicographiques telles que les dictionnaires de langue générale ou les dictionnaires terminologiques. Les patrons lexico-syntaxiques ont montré leur efficacité pour la détection de relations sémantiques telles que l'hyperonymie ou la méronymie. Dans le cas de la synonymie, il existe peu de constructions syntaxiques ou partiellement lexicalisées la dénotant. Blondel et Senellart (2002), par exemple, ont adopté l'hypothèse selon laquelle les synonymes partagent plusieurs mots en commun dans leurs définitions. Ainsi, des graphes d'énoncés définitoires sont construits à partir d'un dictionnaire. Chaque mot du dictionnaire représente un sommet du graphe, et deux mots u et v seront reliés entre eux par un arc, si et seulement si le mot u apparaît dans la définition du mot v ou inversement. Les synonymes sont extraits en fonction de la similarité entre les sommets des graphes.

Se positionner dans un contexte multilingue peut aussi favoriser l'extraction de synonymes. Plusieurs méthodes exploitent des corpus multilingues sous l'hypothèse que les mots qui ont des contextes de traduction similaires auront tendance à être en relation sémantique (Wu et Zhou, 2003 ; Van der Plas et Tiedemann, 2006). Partant de l'observation que les approches fondées sur la similarité distributionnelle étaient incapables de distinguer les synonymes des autres relations sémantiques, telles que l'hyperonymie ou l'antonymie, et qu'il y avait plus de chances de trouver des synonymes d'un mot en le caractérisant par ses traductions que par son contexte distri-

1. « *You shall know a word by the company it keeps.* »(Firth, 1957).

butionnel, Van der Plas et Tiedemann (2006) ont utilisé l'alignement multilingue des mots à partir de corpus parallèles dans plusieurs langues afin d'extraire les synonymes. L'hypothèse est qu'un mot n'est jamais traduit par son antonyme, son hyperonyme ou son co-hyponyme. Le mot anglais *apple*, par exemple, ne serait jamais traduit en français par *fruit* ou par *poire*. Ainsi, chaque mot est caractérisé par un vecteur de ses traductions dans dix langues cibles. Cette méthode qui est une extension de celle proposée par Wu et Zhou (2003) concernant les corpus parallèles, a produit de meilleurs résultats que l'approche distributionnelle monolingue exploitant les relations de dépendance syntaxique avec une augmentation moyenne de 7 % de F-mesure.

Certains travaux ont montré l'utilité de combiner plusieurs ressources pour améliorer la qualité des synonymes extraits. Wu et Zhou (2003), par exemple, combinent plusieurs ressources dont un dictionnaire monolingue, un corpus parallèle bilingue et un large corpus monolingue. Dans un premier temps, chaque ressource est utilisée individuellement pour extraire les synonymes. L'approche utilisant un dictionnaire monolingue est fondée sur la méthode de Blondel et Senellart (2002) citée plus haut. L'approche utilisant un corpus parallèle, quant à elle, est fondée sur la traduction d'un mot pour exprimer son sens. Ainsi, chaque mot est représenté par le vecteur de ses traductions accompagnées de leur probabilité de traduction. L'extraction des synonymes se fait en mesurant la similarité entre les vecteurs traduits selon la mesure du cosinus. Enfin, l'approche utilisant un large corpus monolingue est fondée sur l'hypothèse distributionnelle selon laquelle les synonymes auront tendance à apparaître dans les mêmes contextes lexicaux. Chaque mot est associé aux mots avec lesquels il est en relations de dépendance syntaxique. Un contexte est représenté par un triplet <mot1, type de relation, mot2>, où mot2 est appelé *attribut*. Ainsi, un mot sera caractérisé par un vecteur contenant l'ensemble de ses attributs. Deux mots seront synonymes s'ils partagent des attributs similaires. Par la suite, les trois approches sont combinées de telle sorte que la similarité entre deux mots sera la somme pondérée de la similarité renvoyée par chaque approche individuelle.

Quelle que soit la méthode utilisée, celle-ci produit une liste de mots potentiellement synonymes pour une unité lexicale donnée. Parmi ces candidats, il n'est pas évident d'établir une distinction claire et précise entre la catégorie des synonymes et les autres catégories de relations sémantiques (Lin *et al.*, 2003 ; Van der Plas et Tiedemann, 2006). Concernant l'approche distributionnelle, Resnik (1993 :18)[2] déclare que « l'information capturée en utilisant les méthodes distributionnelles n'apparaît comme ni vraiment syntaxique, ni purement sémantique ». La sémantique reliant les éléments lexicaux ne relève pas des seules synonymie ou quasi-synonymie mais aussi des autres relations sémantiques classiques telles que l'antonymie, l'hyperonymie, la co-hyponymie, la méronymie, et les relations sémantiques non classiques telles que les relations action-agent (Morris et Hirst 2004). Morlane-Hondère (2013) a réalisé une étude exhaustive et approfondie des relations sémantiques générées à l'aide d'une ana-

2. « *It would seem that the information captured using distributional methods is not precisely syntactic, nor purely semantic - in some sense the only word that appears is* distributional. »

lyse distributionnelle automatique pour le français dans le domaine général et confirme le vaste ensemble de relations sémantiques qu'elle met à jour.

Toutes ces approches ont été appliquées à l'extraction de synonymes de termes simples. L'extraction de synonymes de termes complexes composés de plusieurs unités lexicales, quant à elle, a été très peu étudiée alors que les synonymes de termes complexes sont de loin les plus fréquents quand il s'agit de langues de spécialités, particulièrement les langues romanes. Seuls Hamon et Nazarenko (2001) se sont intéressés à cette tâche et la méthode qu'ils ont proposée fait office de référence.

3. Compositionnalité et synonymie

Une définition générale de la compositionnalité admise par tous est celle proposée par Parnee *et al.* (1990) : une expression est compositionnelle si son sens est exprimé en fonction du sens de ses parties en respectant les règles syntaxiques de combinaison[3].

Hamon et Nazarenko (2001) font l'hypothèse que les synonymes des termes complexes sont compositionnels si leurs parties sont des synonymes. Ils ont défini trois règles pour détecter les relations de synonymie. Étant donné les candidats termes complexes $CCT_1 = (T_1, E_1)$ et $CCT_2 = (T_2, E_2)$ où T_1 (respectivement T_2) correspond à la tête du terme complexe et E_1 (respectivement E_2) correspond à son expansion et $syn(CT_1, CT_2)$ une relation de synonymie entre les candidats termes CT_1 et CT_2, les règles d'inférences suivantes sont utilisées :

– $R_1 : T_1 = T_2 \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$
– $R_2 : E_1 = E_2 \wedge syn(T_1, T_2) \supset syn(CCT_1, CCT_2)$
– $R_3 : syn(T_1, T_2) \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$

La règle R_1 signifie que les têtes syntagmatiques sont identiques et les expansions sont des synonymes (*collecteur général*/*collecteur commun*). Les synonymes des constituants lexicaux d'un terme complexe comme *général* dans cet exemple sont fournis par un dictionnaire de synonymes de la langue générale.

Kraft (2007) note que les expressions et leurs parties sont habituellement ambiguës et qu'il est difficile de leur assigner un seul sens. Nous illustrons cette remarque par l'analyse des synonymes des termes complexes qui sont présents dans les banques de données terminologiques. En examinant les parties des synonymes des termes du domaine des énergies renouvelables partageant au moins une partie commune, nous rencontrons des relations diverses :

3. « *A compound expression is composional if its meaning is a function of the meaning of the parts and of the syntactic rule by which they are combined.* »

– synonyme : *energy output*/*energy production* donnés par Termium[4] où *output*/*production* sont synonymes ;

– hyperonyme : *turbine noise*/*turbine sound* donnés par le GDT (*Le grand dictionnaire terminologique*)[5] où *sound* est un hyperonyme de *noise* ou encore *implantación de las máquinas*/*implantación de aerogeneradores* donnés par le Lexique panlatin où *máquina* 'machine' est un hyperonyme de *aerogenerador* 'aérogénérateur' ;

– indéfini : *nuclear plant*/*nuclear station* donnés par Termium où *plant* et *station* ne sont pas reliés sémantiquement ou encore *arbre lent*/*arbre primaire* donnés par Terminalf sans relation sémantique entre *lent* et *primaire*.

Notre hypothèse est que la sémantique distributionnelle qui permet d'identifier les mots en relation sémantique dans un domaine de spécialité doit aider à découvrir d'autres synonymes de termes complexes. Pour intégrer l'analyse distributionnelle au sein de la méthode compositionnelle, nous avons besoin d'adapter la méthode proposée dans (Hamon et Nazarenko, 2001).

4. Méthode semi-compositionnelle

Notre méthode s'inspire des travaux de Morin et Daille (2012) qui s'étaient intéressés à l'amélioration de l'alignement de termes complexes à partir de corpus bilingues comparables en combinant une approche compositionnelle avec une approche distributionnelle. Nous utilisons la même approche pour la tâche d'extraction de synonymes de termes complexes (TC). Nous partons de l'hypothèse qu'un terme complexe et ses synonymes adoptent le principe de compositionnalité. Par exemple, le synonyme d'*énergie renouvelable* peut être obtenu par : (i) la décomposition en parties du TC, puis (ii) la détection des mots en relation sémantique avec *énergie* et/ou *renouvelable* en utilisant la méthode distributionnelle, et enfin (iii) la recomposition des parties des termes candidats et le filtrage des termes recomposés à l'aide de listes construites à partir du corpus monolingue spécialisé. Notre méthode semi-compositionnelle diffère en deux points de la méthode proposée par Hamon et Nazarenko (2001), à savoir : (i) la manière d'extraire les synonymes des termes simples, et (ii) la longueur des termes complexes traités.

4.1. *Approche distributionnelle*

Contrairement à Hamon et Nazarenko (2001) qui utilisent un dictionnaire de langue générale pour déterminer les synonymes de chaque partie du terme complexe, notre approche exploite la palette des relations sémantiques des mots fournie par une analyse distributionnelle. L'approche distributionnelle propose des synonymes et des relations sémantiques en contexte, ce qui n'est pas le cas des dictionnaires de langue

4. www.btb.termiumplus.gc.ca
5. www.oqlf.gouv.qc.ca/

générale et surtout des lexiques recensant les vocabulaires spécialisés. De plus, et comme cité précédemment, les parties d'un terme complexe et de son synonyme ne sont pas obligatoirement en relation de synonymie et peuvent être reliées par d'autres relations comme la quasi-synonymie, l'hyperonymie ou l'hyponymie. Ainsi, l'utilisation d'un dictionnaire de synonymes de la langue générale limite l'identification de synonymes de termes complexes au seul cas où un lien de synonymie est attesté entre deux constituants de deux termes complexes.

Nous adoptons l'hypothèse classique de l'analyse distributionnelle qui dit que deux mots sont en relation sémantique s'ils partagent les mêmes contextes lexicaux. Ainsi, pour identifier ces liens sémantiques, nous modélisons le contexte des mots à l'aide de vecteurs, appelés *vecteurs de contexte*. Étant donné un corpus, nous calculons le contexte de chaque mot modélisé dans un vecteur, puis nous mesurons la similarité entre tous les vecteurs de contexte construits. Un score élevé de similarité entre deux mots induit une similarité sémantique forte.

Le vecteur de contexte $v_{w_i^s}$ d'un mot source donné w_i^s[6] par exemple, contiendra tous les mots qui apparaissent avec w_i^s dans une fenêtre contextuelle de n mots autour de lui. Soit $occ(w_i^s, w_j^s)$ la valeur de cooccurrence de w_i^s avec w_j^s qui est un mot appartenant à son contexte. Une mesure d'association, comme l'information mutuelle (notée IM) (Fano, 1961), le taux de vraisemblance (noté TV) (Dunning, 1993) ou le *discounted odds-ratio* (noté DOR) (Evert, 2008), est utilisée pour mesurer la corrélation entre w_i^s et w_j^s ainsi que tous les autres mots de son contexte. Enfin, une mesure de similarité est utilisée également pour chaque mot candidat w_i^t en fonction de son vecteur de contexte, $v_{w_i^s}^t$. Plusieurs mesures de similarité peuvent être utilisées. La plus populaire est la mesure du cosinus (Salton et Lesk, 1968) (notée COS) ou la mesure du Jaccard pondéré (notée JAC) (Grefenstette, 1994). Les candidats en relation sémantique avec w_i^s seront les mots candidats, ordonnés en fonction de leur score de similarité.

Ci-dessous la table de contingence et les mesures d'association et de similarité :

	j	$\neg j$
i	$a = occ(i, j)$	$b = occ(i, \neg j)$
$\neg i$	$c = occ(\neg i, j)$	$d = occ(\neg i, \neg j)$

Tableau 1. *Table de contingence*

$$\begin{aligned} TV(i,j) = \quad & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & +(N)\log(N) - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) \\ & -(c+d)\log(c+d) \end{aligned} \qquad [1]$$

6. L'exposant s de w_i^s fait référence au corpus source, et l'indice i au i-ième mot de ce corpus.

avec $N = a + b + c + d$.

$$\text{IM}(i,j) = log\frac{a}{(a+b)(a+c)} \quad [2]$$

$$DOR(i,j) = \log\frac{(a+\frac{1}{2}) \times (d+\frac{1}{2})}{(b+\frac{1}{2}) \times (c+\frac{1}{2})} \quad [3]$$

$$Cosinus_{v_l}^{v_k} = \frac{\sum_t assoc_t^l \, assoc_t^k}{\sqrt{\sum_t assoc_t^{l\,2}}\sqrt{\sum_t assoc_t^{k\,2}}} \quad [4]$$

$$Jaccard_{v_l}^{v_k} = \frac{\sum_t min(assoc_t^l, assoc_t^k)}{\sum_t max(assoc_t^l, assoc_t^k)} \quad [5]$$

avec $assoc_t^l$ par exemple qui fait référence à une mesure d'association donnée (TV, IM ou DOR) entre les deux mots t et l.

4.2. *Longueur des termes complexes*

À la différence de la méthode de Hamon et Nazarenko (2001), notre approche semi-compositionnelle ne se limite pas à des termes complexes de longueur 2. Elle peut être appliquée à des termes complexes de n'importe quelle longueur à partir du moment où les termes complexes suivent les règles R_1 et R_2. Nous proposons de généraliser ces règles à la fois en acceptant une plus large gamme de relations sémantiques et en autorisant les substitutions sémantiques aux termes complexes de n'importe quelle longueur. Nous étendons les règles R_1 et R_2 en remplaçant les relations synonymiques $syn(CCT_1, CCT_2)$ par les relations sémantiques $sem(CCT_1, CCT_2)$. La règle R_1^G correspond à la généralisation de la règle R_1 (respectivement la règle R_2^G correspond à la généralisation de la règle R_2) et T_1, T_2, E_1, E_2 sont des termes complexes. De plus, nous supprimons la règle R_3 en nous appuyant sur les résultats obtenus par Hamon et Nazarenko (2001) où ils ont montré que cette règle était peu productive et extrêmement bruitée. Nous obtenons donc les règles suivantes :

– $R_1^G : T_1 = T_2 \wedge sem(E_1, E_2) \supset sem(CCT_1, CCT_2)$
– $R_2^G : E_1 = E_2 \wedge sem(T_1, T_2) \supset sem(CCT_1, CCT_2)$

Le tableau 2 illustre quelques exemples de termes complexes et de leurs synonymes en français, en anglais et en espagnol dans le domaine des énergies renouvelables. De même, le tableau 3 illustre quelques exemples de termes complexes

et de leurs synonymes en français et en anglais dans le domaine du cancer du sein. Ces termes et leurs variantes synonymiques ont été extraits de ressources terminologiques. La plupart des exemples de synonymes correspondent bien à nos règles où l'un des composants du terme complexe reste inchangé. Quelques exemples :

En : dans le domaine de l'énergie éolienne *wind turbine*/*wind machine*, *power supply*/*energy supply*, dans le domaine médical *invasive carcinoma*/*infiltrating carcinoma*, *adjuvant therapy*/*adjuvant treatment*.

Fr : dans le domaine de l'énergie éolienne *énergie renouvelable*/*énergie durable*, dans le domaine médical *curage du ganglion*/*ablation du ganglion*.

Pour chaque terme complexe (TC), nous fixons alternativement sa partie gauche pour extraire les mots en relation sémantique avec sa partie droite et sa partie droite pour extraire les mots en relation sémantique avec sa partie gauche. Ceci correspond aux règles R_1^G et R_2^G.

L'inconvénient de cette méthode est l'impossibilité de traiter les synonymes qui ne suivent pas les règles citées plus haut, par exemple : le terme complexe *moulin à vent* et son synonyme : *éolienne*.

Nous filtrons ensuite les termes complexes candidats obtenus en les comparant avec deux listes extraites du corpus : une liste de n-grammes et une liste de termes complexes candidats.

4.3. *Filtrage*

Le filtrage consiste à comparer la liste des termes synonymes candidats obtenue à l'aide de la méthode semi-compositionnelle à des listes extraites du corpus. Le but est de s'assurer de la correction syntagmatique des candidats, l'analyse distributionnelle par définition s'exonère de cette contrainte.

4.3.1. *Filtrage par n-grammes*

Une liste de n-grammes est construite automatiquement à partir du corpus préalablement lemmatisé. Partant du principe de filtrage des suites de mots improbables dans le corpus, une méthode simple consiste à collecter tous les n-grammes apparaissant dans le corpus. Tout n-gramme est considéré comme un terme candidat potentiel. Les termes synonymes candidats qui ne sont pas des n-grammes sont éliminés.

4.3.2. *Filtrage par extracteur terminologique*

Les termes synonymes candidats sont comparés à une liste de termes candidats extraits par un extracteur de termes. Un terme synonyme candidat non proposé par l'extracteur est éliminé. Nous avons utilisé TermSuite[7] (Rocheteau et Daille, 2011).

7. `logiciels.lina.univ-nantes.fr/redmine/projects/termsuite`

TermSuite détecte les occurrences de termes simples et complexes en exploitant leurs structures morphosyntaxiques caractéristiques. Il normalise et regroupe les différentes variantes des candidats termes. Ceux-ci sont ensuite ordonnés en fonction de leur spécificité. Les hapax et les termes avec une spécificité équivalente à celle d'un hapax sont éliminés. Voici un extrait de la liste produite par TermSuite où le premier chiffre indique le rang, T s'il s'agit d'un terme, V s'il s'agit d'une variante, suivi de la forme fléchie du terme ou de la variante la plus fréquente.

```
20 T tower
21 T wind farm
21 V wind energy farm
21 V wind power farms
21 V wind turbine farms
```

Le terme *wind farm* fait partie de la liste des termes candidats extraits du corpus des énergies éoliennes (cf. section 5). Il est le 21^e candidat et est accompagné de trois variantes terminologiques.

5. Ressources expérimentales

Dans cette section nous détaillons les corpus et les listes de référence utilisées dans nos expériences ainsi que les paramètres des différentes approches.

5.1. *Corpus*

Les expériences ont été menées sur deux corpus comparables spécialisés.

Le premier corpus comparable relève du domaine des énergies renouvelables et est disponible en trois langues : le français, l'anglais et l'espagnol. Les corpus ont été collectés à partir des pages Web en utilisant le *crawler* BABOUK (De Groc, 2011). BABOUK prend en entrée une liste de termes spécifiques à un domaine, appelée *liste d'amorces de termes*, et trouve des textes sur le Web qui traitent du domaine spécialisé dénoté par ces termes. Lors de la première itération (de collection des pages Web), cette liste d'amorces de termes est étendue en s'appuyant sur les nouveaux termes se trouvant dans les pages Web collectées. Pour élargir la recherche, les amorces de termes sont combinées de manière aléatoire pour former une requête. Cette dernière est ensuite soumise à un moteur de recherche qui va retourner les n meilleures pages qui correspondent à cette recherche. Ensuite, des filtres définis choisissent les pages qui sont riches en terminologies spécifiques au domaine et ignorent les pages qui ne sont pas propres, après conversion au format texte, celles qui sont de très petite taille ou de très grande taille, etc. Certains corpus monolingues ont été étendus avec des documents recueillis manuellement sur le Web afin d'atteindre la taille minimale fixée

Synonymes des termes anglais

wind turbine	wind machine
power supply	energy supply
power plant	electricity plant
savonius model	savonius type
energy output	energy production
sea wind farm	offshore wind farm
wind farm	wind power plant
wind turbine	aeroturbine

Synonymes des termes français

énergie renouvelable	énergie durable
centrale électrique	centrale éolienne
unité de stockage	dispositif de stockage
arbre primaire	arbre lent
force du vent	vitesse du vent
éolienne	moulin à vent

Synonymes des termes espagnols

ángulo de paso	ángulo de calaje
extremo de la pala	punta de la pala
mapa de vientos	mapa eólico
coeficiente de potencia	coeficiente de rendimiento
implantación de las máquinas	implantación de aerogeneradores
aerogenerador	torre eólica

Tableau 2. *Exemples de synonymes anglais, français et espagnols de termes complexes extraits de ressources terminologiques du domaine des énergies renouvelables*

à 300 000 mots. La collecte a été effectuée en 2010. Le tableau 4 résume la taille des corpus français, anglais et espagnol en nombre de mots et d'articles[8].

Le deuxième corpus comparable relève du domaine du cancer du sein. Il est disponible en français et en anglais. La collection des textes a été construite à partir de publications scientifiques collectées sur le portail *Elsevier*[9] et sur *Google Scholar*[10]. Les documents ont été collectés de manière à ce qu'ils répondent aux critères de comparabilité suivants :

8. Le corpus des énergies renouvelables est téléchargeable à l'URL : `www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html`
9. `www.elsevier.com`
10. `scholar.google.com`

Synonymes des termes anglais

anti-oestrogens therapy	anti-oestrogens treatment
invasive carcinoma	infiltrating carcinoma
tumour tissue	cancerous tissue
adjuvant therapy	adjuvant treatment
complete breast prosthesis	full breast prosthesis
invasive ductal carcinoma	invasive ductal cancer

Synonymes des termes français

radiographie mammaire	radiographie du sein
cellule de la tumeur	cellule cancéreuse
curage du ganglion	ablation du ganglion
reconstruction du sein	reconstruction mammaire
reconstruction mammaire différée	reconstruction mammaire secondaire
reconstruction mammaire différée	reconstruction du sein différée

Tableau 3. *Exemples de synonymes anglais et français de termes complexes dans le domaine du cancer du sein extraits de Termium*

	Fr	**En**	**Es**
Nb. de mots	313 943	314 549	453 953
Nb. d'articles	11	28	46

Tableau 4. *Caractéristiques des corpus du domaine des énergies renouvelables*

(a) contenir le terme clé *cancer du sein* pour le français et son équivalent *breast cancer* pour l'anglais ;

(b) être publiés dans la période 2001-2008.

Le tableau 5 précise les caractéristiques du corpus du cancer du sein.

	Fr	**En**
Nb. de mots	267 180	198 244
Nb. de documents	78	70

Tableau 5. *Caractéristiques des corpus du domaine du cancer du sein*

Tous les corpus sont prétraités linguistiquement en utilisant la tokénisation, l'analyse morphosyntaxique et la lemmatisation à l'aide de TreeTagger.

5.2. *Listes de référence*

Les listes de référence ont été construites à partir de diverses ressources terminologiques. Seules ont été retenues les ressources qui recensaient dans leurs fiches des termes synonymes. Dans de telles bases, les synonymes ne sont pas systématiquement présents, et pour les fiches en comprenant, la nature de la synonymie est diverse. Nombre de synonymes proposés dans ces ressources sont des mots liés par d'autres types de relations sémantiques, comme la quasi-synonymie ou l'hyperonymie. De nombreuses variantes de termes complexes comme les abréviations, les constructions syntaxiques concurrentes comme N P N A/N A P N ou N P N/N Ar avec Ar adjectif relationnel pour le français, les réductions lexicales y sont listées. Toutes ces variantes de termes complexes ont été écartées de nos listes car elles peuvent être détectées à l'aide d'une analyse syntagmatique.

Concernant le domaine des énergies renouvelables, les termes complexes français ont été sélectionnés à partir de Terminalf[11]et ses 84 fiches terminologiques. Les synonymes sont indiqués sous le champ `forme concurrente`. Il existe 56 termes qui acceptent entre 1 et 3 formes concurrentes soit 76 synonymes. À ces 76 synonymes, nous avons ajouté 14 autres synonymes ou quasi-synonymes déduits à la lecture des fiches terminologiques, en particulier dans les champs `définition` ou `genre`. Le contenu du champ `genre` fait souvent référence à un hyperonyme qui peut être utilisé comme une variante synonymique en contexte. N'ont pas été considérées comme des synonymes, les formes concurrentes qui sont des variantes de réduction portant :

– sur les éléments fonctionnels comme *rotor Darrieus* et *rotor de Darrieus* ;

– sur les composants lexicaux comme la variante *traînée* pour le terme *traînée aérodynamique* ou encore la variante *coefficient de puissance* pour le terme *coefficient de puissance du rotor*.

Pour l'espagnol, nous avons collecté 64 termes complexes et leurs synonymes à partir du Lexique panlatin de l'énergie éolienne[12]. Ce lexique contient 300 entrées en 8 langues. Certains termes comprennent des synonymes disponibles dans deux zones géographiques différentes : l'Espagne et le Mexique. Par exemple, le terme espagnol *aerogenerador de dos palas* (*éolienne bipale*) aura l'équivalent suivant en espagnol mexicain *aerogenerador de doble aspa*. Nous avons considéré ces variantes terminologiques géographiques comme des synonymes. En comparaison de Terminalf qui ne contient que des termes simples ou des termes complexes composés de deux unités lexicales, le Lexique panlatin contient aussi de nombreux termes de longueur supérieure à deux unités lexicales pleines comme *aerogenerador de eje vertical con geometría variable* (*éolienne à axe vertical et géométrie variable*). Les synonymes recensés contiennent de nombreuses variantes que nous n'avons pas incluses dans nos listes :

11. `terminalf.scicog.fr`
12. `www.realiter.net/wp-content/uploads/2013/06/pan-energie-power.pdf`

– les variantes morphologiques portant sur les composants du terme complexe comme les équivalences groupe prépositionnel et adjectif *efecto de aceleración*/*efecto acelerador* avec *acelerador* dérivé du nom *aceleración* ;

– les variantes de réduction ou d'augmentation de terme complexe que la réduction ou l'augmentation porte sur l'un des composants autonomes du terme comme *efecto de pérdida aerodinámica*/*efecto de entrada en pérdida aerodinámica* ou sur l'un des composants morphologiques comme *acoplamiento dinámico*/*acoplamiento aerodinámico* ;

– les variantes de substitution des prépositions comme *enfriamiento del aire*/*enfriamiento por aire* ;

– les constructions syntaxiques concurrentes où les composants en position d'expansion du terme complexe ont été inversés : *aerogenerador para vientos de baja velocidad*/*aerogenerador para bajas velocidades de viento*.

Les termes anglais ont été sélectionnés à partir du glossaire en ligne (Gipe, 2004) et de la banque de données terminologiques Termium[13]. Nous avons collecté 84 termes complexes et leurs synonymes en écartant les variantes qui sont moins diversifiées que pour le français et l'espagnol. Les variantes prédominantes sont les variantes de réduction lexicale comme *wind power plant*/*wind plant*.

Après projection dans les corpus du domaine de l'éolien respectifs de chaque langue, nous avons obtenu 34 termes français, 20 termes anglais et 26 termes complexes espagnol associés à leurs synonymes.

Nous avons procédé de manière semblable pour constituer les synonymes de termes complexes du domaine du cancer du sein. Nous avons principalement extrait les synonymes de Termium, là encore en supprimant les mêmes types de variantes que pour le domaine de l'éolien. Après projection dans le corpus du cancer du sein dans chaque langue, nous avons obtenu 20 termes français et 16 termes anglais associés à leurs synonymes.

La taille réduite des listes de référence que ce soit pour le domaine de l'éolien ou du cancer du sein peut s'expliquer par le fait que les termes suivent le principe de monosémie et mononymie comme le rappellent Bowker et Hawkins (2006), p. 83 : « un terme correspond à un concept, et un concept n'est désigné que par un seul terme »[14]. Une autre raison est que les corpus de spécialité étant souvent de petite taille, ceci induit un nombre limité de termes spécialisés et de variantes synonymiques. Enfin, nombre de variantes synonymiques sont contextuelles, et il est difficile pour un terminologue de prédire et de détecter toutes les variantes synonymiques qui peuvent être produites.

13. `www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng`

14. « *A term should be applied to a single concept, and a concept should be designed by only one term. So, synonyms of terms are rare phenomena.* »

5.3. *Méthode à base de dictionnaire*

Nous avons utilisé la méthode proposée par Hamon et Nazarenko (2001). Pour extraire les synonymes des termes simples en français nous avons utilisé le dictionnaire en ligne DES[15]. Il contient 49 168 entrées et 201 511 relations de synonymie relevant de la langue générale. Pour l'anglais, le dictionnaire de synonymes a été construit en utilisant la base de données lexicale WordNet[16] qui contient environ 117 000 entrées (synsets). La principale relation présente dans WordNet est la synonymie.

5.4. *Paramètres de l'approche distributionnelle*

Nous avons besoin de fixer trois paramètres concernant l'approche distributionnelle, à savoir :

1) la taille de la fenêtre contextuelle qui sert à construire le vecteur de contexte (Morin *et al.*, 2007 ; Gamallo Otero, 2008) ;

2) la mesure d'association (le taux de vraisemblance (TV) (Dunning, 1993), l'information mutuelle (IM) (Fano, 1961), le *discounted odds-ratio* (DOR) (Evert, 2008)) qui sert à mesurer la force de la relation entre les mots ;

3) la mesure de similarité (l'indice du Jaccard pondéré (JAC) (Grefenstette, 1994), le cosinus (COS) (Salton et Lesk, 1968)) qui sert à mesurer la similarité entre les vecteurs de contexte des mots.

Pour construire le vecteur de contexte, nous avons choisi une taille de fenêtre égale à 7, c'est-à-dire 3 mots précédant et trois mots suivant le mot à caractériser. Comme mesures d'association nous avons utilisé IM, TV et DOR et comme mesures de similarité nous avons utilisé COS et JAC.

Pour évaluer les différentes combinaisons de paramètres de l'analyse distributionnelle, nous avons adopté la mesure de la MAP (*Mean Average Precision*) (Manning *et al.*, 2008) ainsi que la précision aux TOP1, TOP5 et TOP10 qui examine le premier candidat, puis les ensembles des 5 et 10 premiers candidats.

$$MAP = \frac{1}{|W|} \sum_{i=1}^{|W|} \frac{1}{Rang_i} \qquad [6]$$

où $|W|$ correspond à la taille de la liste d'évaluation et $Rang_i$ correspond au rang du candidat synonyme correct i.

Si les TOP1, TOP5 et TOP10 renvoient la précision des systèmes aux premiers rangs, la MAP, quant à elle, renvoie la précision moyenne globale des systèmes. Ainsi,

15. www.crisco.unicaen.fr/des/synonyms
16. wordnetweb.princeton.edu/perl/webwn/

un système A ayant une meilleure précision au TOP1 qu'un système B peut avoir une moins bonne MAP, si dans la suite du classement, ses synonymes sont moins bien classés que ceux du système B.

6. Résultats

Dans cette section, nous présentons les résultats des expériences menées sur le corpus des énergies renouvelables pour le français, l'anglais et l'espagnol, et sur le corpus du cancer du sein pour le français et l'anglais. La méthode de Hamon et Nazarenko (2001) est notée *référence* et notre approche est notée *Semi-Comp*. Les candidats renvoyés par Semi-Comp sont ordonnés de trois manières différentes :

– le rang par score d'association où les termes complexes candidats sont ordonnés par rapport à la valeur décroissante du score d'association du composant substitué (Semi-Comp (assoc)) ;

– le rang par effectif où les termes complexes candidats sont ordonnés par rapport à leur fréquence d'occurrence dans le corpus (Semi-Comp (effec)) ;

– le rang par rapport de fréquence où les termes complexes candidats sont ordonnés en fonction de leur rapport de fréquence avec le terme complexe dont les synonymes sont recherchés (Semi-Comp (rdf)).

Nous étudions deux manières de filtrer les candidats termes synonymes, à savoir un filtrage à l'aide d'une liste de n-grammes, noté *sans patrons syntaxiques* (illustré dans la deuxième colonne de chaque tableau de résultats) et un filtrage par une liste de termes complexes candidats, noté *avec patrons syntaxiques* (illustré dans la troisième colonne de chaque tableau de résultats). Enfin, nous expérimentons trois configurations correspondant à trois combinaisons impliquant chacune une mesure d'association et une mesure de similarité différentes. Ces trois combinaisons ne représentent qu'un sous-ensemble de la combinatoire entre mesures d'association et de similarité mais sont classiquement utilisées en analyse distributionnelle et ont montré leur efficacité comparativement à d'autres combinaisons. Les configurations sont indiquées sur l'axe des ordonnées des tableaux. Par exemple, la légende IM-COS signifie que l'approche distributionnelle utilisée dans Semi-Comp exploite l'information mutuelle comme mesure d'association et le cosinus comme mesure de similarité.

Pour chaque configuration, les meilleurs résultats au TOP1, TOP5, TOP10 et MAP sont indiqués en gras. Le tableau 6 donne les résultats de l'approche de référence ainsi que ceux de l'approche semi-compositionnelle sur le corpus français des énergies renouvelables. La première remarque concerne l'approche de référence où les résultats obtenus sont très faibles en comparaison avec l'approche semi-compositionnelle quelle que soit la configuration choisie. Les meilleurs résultats sont obtenus avec l'approche Semi-Comp (assoc) avec la configuration IM-COS associée à un filtrage par patrons syntaxiques et donnent une précision au TOP1 de 26,4 %, au TOP5 de 55,8 % et une MAP de 38,5 %. La meilleure précision au TOP5 de 55,8 % est partagée par toutes les configurations avec Semi-Comp (effec) quel que soit le filtrage. La meilleure

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	0,25	-	-	-	0,25
TV-JAC	Semi-Comp (assoc)	17,6	52,9	52,9	31,7	23,5	52,9	58,8	38,2
	Semi-Comp (effec)	14,7	**55,8**	64,7	31,2	17,6	**55,8**	64,7	34,8
	Semi-Comp (rdf)	5,88	23,5	35,2	14,9	5,88	47,0	52,9	23,2
IM-COS	Semi-Comp (assoc)	17,6	47,0	52,9	29,4	**26,4**	**55,8**	61,7	**38,5**
	Semi-Comp (effec)	14,7	**55,8**	64,7	31,2	17,6	**55,8**	64,7	34,8
	Semi-Comp (rdf)	5,88	23,5	35,2	14,9	5,88	47,0	52,9	23,2
DOR-COS	Semi-Comp (assoc)	14,7	44,1	47,0	26,9	23,5	52,9	**70,5**	37,9
	Semi-Comp (effec)	14,7	**55,8**	64,7	31,2	17,6	**55,8**	64,7	34,8
	Semi-Comp (rdf)	5,88	23,5	35,2	14,9	5,88	47,0	52,9	23,2

Tableau 6. *Résultats des expériences sur le corpus français des énergies renouvelables*

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	3,63	-	-	-	3,63
TV-JAC	Semi-Comp (assoc)	20,0	55,0	65,0	36,1	15,0	60,0	70,0	38,2
	Semi-Comp (effec)	45,0	80,0	90,0	59,2	**45,0**	**80,0**	**90,0**	**61,1**
	Semi-Comp (rdf)	20,0	40,0	40,0	28,2	15,0	40,0	45,0	26,7
IM-COS	Semi-Comp (assoc)	15,0	55,0	65,0	32,8	15,0	60,0	70,0	35,8
	Semi-Comp (effec)	**45,0**	**80,0**	**90,0**	59,2	**45,0**	**80,0**	**90,0**	**61,1**
	Semi-Comp (rdf)	20,0	40,0	40,0	28,2	15,0	40,0	45,0	26,7
DOR-COS	Semi-Comp (assoc)	15,0	40,0	55,0	27,2	10,0	40,0	60,0	26,8
	Semi-Comp (effec)	**45,0**	**80,0**	**90,0**	60,0	**45,0**	**80,0**	**90,0**	**61,1**
	Semi-Comp (rdf)	20,0	40,0	40,0	28,3	15,0	40,0	45,0	26,7

Tableau 7. *Résultats des expériences sur le corpus anglais des énergies renouvelables*

précision au TOP10 de 70,5 % est fournie par Semi-Comp (assoc) avec la configuration DOR-COS associée à un filtrage par patrons syntaxiques.

Le tableau 7 donne les résultats de l'approche de référence ainsi que ceux de l'approche semi-compositionnelle sur le corpus anglais des énergies renouvelables. Nous constatons que les résultats de l'approche de référence sont de nouveau très faibles

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	8,09	-	-	-	8,09
TV-JAC	Semi-Comp (assoc)	7,69	38,4	42,3	19,7	11,5	50,0	61,5	29,8
	Semi-Comp (effec)	15,3	34,6	42,3	24,8	15,3	38,4	46,1	25,7
	Semi-Comp (rdf)	3,84	11,5	15,3	7,90	11,5	30,7	42,3	23,1
IM-COS	Semi-Comp (assoc)	15,3	38,4	42,3	25,3	**23,0**	**57,6**	**65,3**	**37,0**
	Semi-Comp (effec)	15,3	34,6	42,3	24,8	15,3	38,4	46,1	25,7
	Semi-Comp (rdf)	3,84	11,5	15,3	7,90	11,5	30,7	42,3	22,5
DOR-COS	Semi-Comp (assoc)	0	30,7	42,3	14,4	19,2	53,8	**65,3**	32,6
	Semi-Comp (effec)	15,3	34,6	42,3	24,9	15,3	38,4	50,0	25,8
	Semi-Comp (rdf)	3,84	11,5	15,3	7,90	11,5	30,7	42,3	22,5

Tableau 8. *Résultats des expériences sur le corpus espagnol des énergies renouvelables*

en comparaison de ceux de l'approche semi-compositionnelle. Les meilleurs résultats sont obtenus avec Semi-Comp (effec) associé à un filtrage par patrons syntaxiques et donnent une précision au TOP1 de 45 %, au TOP5 de 80 %, au TOP10 de 90 % et une MAP de 61,1 %. La configuration n'influence pas les résultats et le filtrage ne l'influence qu'à la marge.

Les résultats de la dernière expérience dans le domaine des énergies renouvelables et qui concernent le corpus espagnol sont représentés dans le tableau 8. Pour le corpus espagnol comme pour le corpus français, c'est le classement par score d'association Semi-Comp (assoc) exploitant la configuration IM-COS associée à un filtrage par patrons syntaxiques qui donne les meilleurs résultats avec une précision au TOP1 de 23 %, au TOP5 de 57,6 %, au TOP10 de 65,3 % et une MAP de 37 %. La meilleure précision au TOP10 de 65,3 % est obtenue par Semi-Comp (assoc) avec la configuration DOR-COS associée à un filtrage par patrons syntaxiques.

D'une manière générale, l'approche Semi-Comp donne toujours de meilleurs résultats que l'approche de référence quelle que soit la configuration choisie. Concernant la manière de classer les candidats, les classements par score d'association et par effectif sont à privilégier et à associer à un filtrage par patrons syntaxiques. Les trois configurations testées donnent des résultats très proches qui varient selon la langue et la méthode de classement choisies. Le classement par effectif montre une grande stabilité vis-à-vis de la configuration choisie et ceux pour les trois langues.

Les tableaux 9 et 10 fournissent les résultats de l'approche de référence ainsi que ceux de l'approche semi-compositionnelle sur le corpus du cancer du sein, respectivement pour le français et l'anglais.

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	4,92	-	-	-	4,92
TV-JAC	Semi-Comp (assoc)	5,26	15,7	47,3	13,9	26,3	47,3	57,8	34,7
	Semi-Comp (effec)	**31,5**	**42,1**	**73,6**	40,3	31,5	52,6	63,1	39,9
	Semi-Comp (rdf)	0	15,7	31,5	8,6	15,7	47,3	47,3	30,3
IM-COS	Semi-Comp (assoc)	10,5	36,8	52,6	19,9	21,0	**52,6**	57,8	35,7
	Semi-Comp (effec)	**31,5**	42,1	**73,6**	40,3	**31,5**	**52,6**	63,1	39,9
	Semi-Comp (rdf)	0	15,7	31,5	8,6	15,7	47,3	47,3	30,3
DOR-COS	Semi-Comp (assoc)	15,7	42,1	52,6	27,1	26,3	57,8	57,8	38,5
	Semi-Comp (effec)	**31.5**	42,1	**73,6**	**40,4**	**31,5**	**52,6**	63,1	39,9
	Semi-Comp (rdf)	0	15,7	31,5	8,6	15,7	47,3	47,3	30,3

Tableau 9. *Résultats des expériences sur le corpus français du cancer du sein*

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	7,03	-	-	-	7,03
TV-JAC	Semi-Comp (assoc)	6,66	20,0	26,6	13,3	0	26,6	40,0	16,7
	Semi-Comp (effec)	20,0	46,6	46,6	29,0	**26,6**	**53,3**	**53,3**	**38,1**
	Semi-Comp (rdf)	0	6,66	6,66	3,0	0	20,0	33,3	9,0
IM-COS	Semi-Comp(assoc)	6,66	20,0	26,6	12,6	6,66	26,6	53,3	18,1
	Semi-Comp (effec)	20,0	46,6	46,6	29,0	**26,6**	**53,3**	**53,3**	**38,1**
	Semi-Comp (rdf)	0	6,66	6,66	3,0	0	20,0	33,3	9,0
DOR-COS	Semi-Comp (assoc)	6,66	13,3	20,0	11,0	6,66	20,0	33,3	15,7
	Semi-Comp (effec)	20,0	40,0	40,0	25,7	**26,6**	46,6	46,6	34,8
	Semi-Comp (rdf)	0	6,66	6,66	3,0	0	20,0	33,3	9,0

Tableau 10. *Résultats des expériences sur le corpus anglais du cancer du sein*

À nouveau, l'approche semi-compositionnelle donne de meilleurs résultats que l'approche de référence. Pour le français, nous constatons que Semi-Comp (effec) en utilisant la configuration DOR-COS associée à un filtrage sans patrons syntaxiques donne les meilleurs résultats avec une précision pour le TOP1 de 31,5 %, pour le TOP10 de 73,6 % et une MAP de 40,4 %. Néanmoins, la précision au TOP1 de 31,5 % est obtenue par tous les classements quels que soient la configuration et le filtrage, et la précision pour le TOP5 de 52,6 % est proposée par Semi-Comp (effec) pour le filtrage avec patrons syntaxiques et à 42,1 % quelle que soit la configuration sans patrons syn-

taxiques. L'amélioration de la MAP pour le filtrage sans patrons syntaxiques apparaît entre les rangs 5 à 10 uniquement, ce qui en atténue l'intérêt.

Pour l'anglais, c'est encore Semi-Comp (effec) en utilisant les configurations TV-JAC et IM-COS associées à un filtrage avec patrons syntaxiques qui donne les meilleurs résultats avec une précision au TOP1 de 26,6 %, au TOP5 de 53,3 %, au TOP10 de 53,3 % et une MAP de 38,1 %.

Les résultats de Semi-Comp qui utilisent le classement par rapport de fréquence sont très faibles et confirment l'inefficacité de ce type de classement. La méthode semi-compositionnelle fournissant les meilleurs résultats au TOP1 en moyenne, un classement par effectif et un filtrage des candidats par patrons syntaxiques, ne propose qu'une fois sur quatre un synonyme. En revanche, la méthode semi-compositionnelle adoptant un classement par effectif, la configuration IM-COS et un filtrage des candidats par patrons syntaxiques donne toujours les meilleurs résultats au TOP5 quels que soit le domaine ou la langue. Le score le plus faible de précision au TOP5 est de 52,6 % pour le français et le plus élevé est de 80 % pour l'anglais dans le domaine des énergies renouvelables. Cela signifie que l'examen du TOP5 fournit en moyenne entre 2 à 3 synonymes. Le dépouillement des résultats fournis par la méthode semi-compositionnelle peut donc à moindre effort aider à recenser les variantes synonymiques contextuelles.

Certains résultats quantitatifs nous ont donné envie de les examiner plus finement. Le filtrage à l'aide de patrons syntaxiques fournit généralement de meilleurs résultats que le filtrage par n-grammes. Ce résultat est vérifié pour l'anglais dans le domaine du cancer du sein. Par exemple le filtrage par patrons syntaxiques propose en rang 2 le couple *nipple prosthesis*/*nipple reconstruction* et en rang 12 le couple *os biopsy*/*bone biopsy* qui n'apparaissent pas avec le filtrage par n-grammes. Le filtrage par patrons syntaxiques donne aussi généralement de meilleurs rangs aux couples synonymes, comme *breast tissue*/*mammary tissue*, rang 7 avec filtrage par patrons syntaxiques, et au rang 29 avec filtrage par les n-grammes. En revanche, dans le domaine des énergies renouvelables, le filtrage par patrons syntaxiques est équivalent à celui proposé par les n-grammes. Les mêmes couples sont proposés au rang 1 comme *wind generator*/*wind turbine*, *lattice construction*/*lattice tower*, *drive train*/*power train*.

Bien entendu, certains couples de synonymes sont filtrés par les patrons syntaxiques alors qu'ils ne le sont pas avec les n-grammes. Il s'agit pour l'anglais de termes comportant un participe présent comme *infiltrating carcinoma* dans le domaine du cancer du sein. Le patron $V_{\text{participe présent}}$ N produit trop de candidats termes incorrects et ne fait pas partie des patrons de termes complexes anglais. Des remarques similaires peuvent être faites dans d'autres langues. Pour le français, les couples *protocole radiothérapie*/*protocole cmf* et *protocole chimiothérapie*/*protocole cmf* sont proposés en rang 1 et 2 avec le filtrage par n-grammes et sont absents du filtrage par patrons syntaxiques, sans doute à cause d'une erreur de lemmatisation du corpus.

Pour l'espagnol et le domaine des énergies renouvelables, le meilleur classement est obtenu très nettement par le classement par association contrairement aux résultats

pour le français et l'anglais dans les deux domaines étudiés qui privilégie un classement par effectif. La liste de référence de l'espagnol comportait des spécificités par rapport aux autres listes de référence : termes complexes de longueur plus importante, variantes géographiques d'espagnol. Plusieurs candidats dans les premiers rangs sont partagés par les deux classements comme *torre eólica*[ESP]*turbina eólica*[ESP] ou *flujo de viento*[ESP]/*corriente de viento*[ESP]. Le fait d'avoir inclus des variantes géographiques Espagne [ESP] et Mexique [MEX] ne pose pas de problèmes particuliers. Leurs occurrences dans le corpus apparaissent en haut des deux classements comme *energía eólica*[MEX]/*potencia eólica*[ESP] et *mapa de los vientos*[ESP]/*mapa eólico*[MEX]. Tous les synonymes de rang 1 fournis par le classement par effectif et par le classement par association sont des termes synonymes proposés par le Lexique panlatin. Les différences portent sur les positions de certains couples du Lexique panlatin qui apparaissent plus loin dans le classement par effectif que dans le classement par association, comme par exemple *implantación máquina*[ESP]/*implantación aerogeneradores*[ESP], *implantación aerogeneradores*[ESP]/*instalación aerogeneradores*[ESP] en rang 8 proposé pour le classement par l'effectif et en rang 1 pour le classement par association. Un couple discutable de notre liste de référence puisqu'il n'apparaît ni dans le Lexique panlatin, ni dans Termium : *abrigo de torre*/*sombra de torre* apparaît en rang 1 pour le classement par association et en rang 22 pour le classement par effectif. La taille des listes étant réduite, ces quelques faits expliquent la différence des résultats.

En conclusion, les résultats obtenus sur les deux corpus confirment l'efficacité et l'opérationabilité pour la mise à jour de ressources terminologiques de l'approche semi-compositionnelle pour l'extraction de termes complexes reliés sémantiquement. Les variantes synonymiques extraites relèvent de causes différentes : dénominations concurrentes géographiques comme pour l'espagnol et variantes contextuelles. L'approche semi-compositionnelle n'est pas sensible à ces différentes variantes.

7. Discussion

Peu de travaux se sont intéressés à l'identification de synonymes de termes complexes. À notre connaissance, seuls Hamon et Nazarenko (2001) se sont attelés à cette tâche en exploitant le principe de compositionnalité des termes complexes. Les faibles résultats obtenus par l'approche de référence peuvent s'expliquer de deux façons : la première est que dans le domaine de spécialité les dictionnaires de synonymes sont rares ou non disponibles. La seconde est que le fait de s'appuyer sur un dictionnaire de langue générale peut conduire à l'extraction de termes complexes inadaptés et/ou hors domaine comme cela a été montré dans nos expériences.

La principale contribution de notre approche est l'utilisation de l'analyse distributionnelle à la place des dictionnaires pour identifier les synonymes et les mots sémantiquement liés des composants d'un terme complexe. L'analyse distributionnelle permet d'identifier des mots en relation sémantique qui sont utilisés pour construire des synonymes de termes complexes en exploitant, là encore, le principe de compositionnalité.

Nous avons aussi identifié le fait que les synonymes des termes complexes ne sont pas toujours composés de synonymes de leurs parties, et que l'utilisation de mots sémantiquement liés était plus souhaitable pour cette tâche. Notre approche peut être appliquée à n'importe quel terme complexe qui suit les règles R_1^G et R_2^G et pour n'importe quelle catégorie de variantes synonymiques, géographiques ou contextuelles. Le classement des candidats synonymes par leur effectif associé à un filtrage par patrons syntaxiques fournit les meilleurs résultats.

Si nos différentes expériences ont montré que le classement des cinq premiers candidats synonymes comprenait la moitié des synonymes corrects, il reste une marge de progression. Outre l'identification des synonymes de termes complexes, et sachant que l'analyse distributionnelle peut renvoyer des termes simples antonymes (*chaud/froid*) ou contrastifs (*blanc/noir*), il est légitime de s'intéresser à l'extraction de termes complexes antonymes ou contrastifs *via* l'approche semi-compositionnelle. Pour ce faire, nous avons construit une liste d'antonymes de termes complexes incluant des termes constrastifs pour le corpus du cancer du sein et nous avons mené une expérience supplémentaire pour identifier les antonymes des termes complexes qui suivent les règles R_1^G et R_2^G. Nous considérons que deux termes complexes sont antonymes si leurs parties le sont. Les listes de référence pour l'anglais et le français dans le domaine médical contiennent respectivement 12 et 9 paires de termes complexes antonymes. Ces listes ont été encore plus difficiles à construire que pour les termes synonymes. Les antonymes étaient absents de toutes les ressources terminologiques que nous avions utilisées pour la construction les listes de synonymes. Nous avons consulté le *Trésor de la langue française*, et pour les antonymes listés comme *artificiel/naturel* cherché des termes complexes ayant au moins un élément en commun et apparaissant dans le corpus comme *ménopause artificielle/ménopause naturelle*. Nous nous sommes limités au corpus du domaine médical. Pour l'anglais, nous avons traduit les couples d'antonymes français et vérifié leur appartenance au corpus et inclus quelques antonymes marqués lexicalement comme *malignant tissue/non-malignant tissue*.

Comme le montrent les résultats des tableaux 11 et 12, l'approche semi-compositionnelle est aussi capable d'identifier les antonymes des termes complexes. Dans la majorité des cas, c'est le classement par score d'association qui donne les meilleurs résultats avec une MAP de 40,6 % pour le français (Semi-Comp(assoc) avec DOR-COS avec filtrage par patrons syntaxiques) et 27,6 % pour l'anglais (Semi-Comp (assoc) avec TV-JAC avec filtrage sans patrons syntaxiques). Le classement par effectif comme celui par rapport de fréquence donnent de très faibles résultats.

Une première interprétation de ces résultats est de dire qu'en utilisant l'analyse distributionnelle, les termes simples antonymes sont mieux classés que les termes simples synonymes si l'on s'appuie sur un classement par score d'association. Ainsi, un classement par effectif favoriserait plutôt l'identification des synonymes de termes complexes et un classement par score d'association l'extraction d'antonymes de termes complexes. Ce résultat conforte l'observation faite par les linguistes Morlane-Hondère (2013) que les contextes partagés par un terme et son antonyme sont peu nombreux et diffèrent des contextes partagés par un terme et ses synonymes. Le score d'associa-

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
TV-JAC	Semi-Comp (assoc)	**37,5**	37,5	37,5	39,7	25,0	**50,0**	**62,5**	37,2
	Semi-Comp (effec)	0	0	25,0	3,4	0	12,5	37,5	7,4
	Semi-Comp (rdf)	0	0	0	2,0	0	25,0	37,5	8,4
IM-COS	Semi-Comp(assoc)	25,0	**50,0**	50,0	34,5	37,5	**37,5**	50,0	40,0
	Semi-Comp (effec)	0	0	25,0	3,4	0	12,5	37,5	7,4
	Semi-Comp (rdf)	0	0	0	2,0	0	25,0	37,5	8,4
DOR-COS	Semi-Comp (assoc)	12,5	37,5	50,0	27,8	**37,5**	37,5	**62,5**	**40,6**
	Semi-Comp (effec)	0	0	25,0	3,8	0	12,5	37,5	7,6
	Semi-Comp (rdf)	0	0	0	2,0	0	25,0	37,5	8,4

Tableau 11. *Résultats des expériences pour l'identification des antonymes sur le corpus français du cancer du sein*

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
TV-JAC	Semi-Comp (assoc)	**25,0**	**25,0**	33,3	**27,6**	16,6	**25,0**	**41,6**	23,6
	Semi-Comp (effec)	8,33	8,33	16,6	10,8	8,33	16,6	25,0	12,8
	Semi-Comp (rdf)	0	8,33	16,6	6,1	8,33	16,6	16,6	12,8
IM-COS	Semi-Comp(assoc)	16,6	16,6	16,6	17,8	16,6	16,6	25,0	18,1
	Semi-Comp (effec)	8,33	8,33	16,6	10,8	8,33	16,6	25,0	12,8
	Semi-Comp (rdf)	0	8,33	16,6	6,1	8,33	16,6	16,6	12,8
DOR-COS	Semi-Comp (assoc)	16,6	16,6	25,0	18,8	16,6	16,6	25,0	18,9
	Semi-Comp (effec)	8,33	8,33	16,6	10,8	8,33	16,6	25,0	12,8
	Semi-Comp (rdf)	0	8,33	16,6	6,1	8,33	16,6	16,6	12,8

Tableau 12. *Résultats des expériences pour l'identification des antonymes sur le corpus anglais du cancer du sein*

tion mettrait en lumière le caractère remarquable au sens collocationnel des contextes partagés par un terme et son antonyme et le classement par effectif privilégierait l'importance du nombre de contextes partagés par un terme et son synonyme.

8. Conclusion

Nous avons présenté dans cet article l'approche semi-compositionnelle pour l'extraction de synonymes de termes complexes. Fondée sur le principe de distributionnalité et de compositionnalité, notre approche a montré des gains significatifs en comparaison avec l'approche état de l'art. Si plus d'expériences sont sûrement nécessaires, les résultats encourageants de l'approche semi-compositionnelle confirment l'intérêt de combiner les analyses distributionnelle et compositionnelle pour l'identification de termes complexes synonymes. Dans des travaux futurs, nous nous attellerons à l'extraction de synonymes de termes complexes qui ne sont pas compositionnels et inversement, nous explorerons l'extraction de synonymes de termes simples. Perinet et Hamon (2013) proposent une méthode hybride pour l'acquisition de relations sémantiques fondées sur une normalisation contextuelle qui pourrait donner de bons résultats pour l'extraction de termes synonymes tout comme une extraction des contextes par une analyse syntaxique.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet TERMITH (`www.atilf.fr/ressources/termith`) a bénéficié d'une aide de l'Agence nationale de la recherche portant la référence ANR-2-CORD-0029.

9. Bibliographie

Blondel V. D., Senellart P., « Automatic Extraction of Synonyms in a Dictionary », 2002.

Bowker L., Hawkins S., « Variation in the organization of medical terms - Exploring some motivations of term choice », *Terminology*, vol. 12, n° 1, p. 79-110, 2006.

De Groc C., « Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction », *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT'11, IEEE Computer Society, Washington, DC, USA, p. 497-498, 2011.

Dunning T., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, vol. 19, n° 1, p. 61-74, 1993.

Evert S., *Corpus Linguistics. An International Handbook*, vol. 2, De Gruyter Mouton, chapter Corpora and collocations, p. 1212-1248, 2008.

Fano R. M., *Transmission of Information : A Statistical Theory of Communications*, MIT Press, Cambridge, MA, USA, 1961.

Ferret O., « Testing Semantic Similarity Measures for Extracting Synonyms from a Corpus. », *in* N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds), *LREC*, 2010.

Ferret O., « Identifying bad semantic neighbors for improving distributional thesauri », *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, p. 561-571, 2013.

Firth J. R., « A synopsis of linguistic theory 1930-1955 », *in* J. R. Firth, W. Haas, M. A. K. Halliday (eds), *Studies in Linguistic Analysis*, Special volume of the Philological Society, Blackwell, Oxford, p. 1-32, 1957.

Gamallo Otero P., « Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora », *Proceedings of LREC 2008 Workshop on Comparable Corpora (LREC'08)*, Marrakech, Morroco, p. 19-26, 2008.

Gipe P., *Wind power : renewable energy for home, farm, and business*, Chelsea Green Pub. Co., 2004.

Grefenstette G., *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publisher, Boston, MA, USA, 1994.

Hagiwara M., « A Supervised Learning Approach to Automatic Synonym Identification Based on Distributional Features », *Proceedings of the ACL-08 : HLT Student Research Workshop*, Association for Computational Linguistics, Columbus, Ohio, p. 1-6, June, 2008.

Hamon T., Nazarenko A., « Detection of synonymy links between terms : experiment and results », *Recent Advances in Computational Terminology*, John Benjamins, p. 185-208, 2001.

Harris Z. S., « Distributional structure. », *Word*, 1954.

Hindle D., « Noun Classification from Predicate-Argument Structures. », *ACL*, p. 268-275, 1990.

Kraft M., « Compositionality : The very Idea », *Research on Language and Computation*, vol. 5, n° 3, p. 287-308, 2007.

Kremer G., Erk K., Padó S., Thater S., « What Substitutes Tell Us - Analysis of an "All-Words" Lexical Substitution Corpus », *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Gothenburg, Sweden, p. 540-549, April, 2014.

Lin D., « Automatic retrieval and clustering of similar words », *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 768-774, 1998.

Lin D., Zhao S., Qin L., Zhou M., « Identifying Synonyms among Distributionally Similar Words », *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 2003.

Manning D. C., Raghavan P., Schütze H., *Introduction to information retrieval*, Cambridge University Press, 2008.

Morin E., Daille B., « Revising the Compositional Method for Terminology Acquisition from Comparable Corpora », *24th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers, Coling'12*, Mumbai, India, p. 1797-1810, 2012.

Morin E., Daille B., Takeuchi K., Kageura K., « Bilingual Terminology Mining – Using Brain, not brawn comparable corpora », *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, p. 664-671, 2007.

Morlane-Hondère F., Une approche linguistique de l'évaluation des ressources extraites par l'analyse distributionnelle automatique, PhD thesis, Université Toulouse II Le Mirail, 2013.

Morris J., Hirst G., « Non-classical lexical semantic relations », *HLT-NAACL Worlshop on Computational Lexical semantics (CLS'04)*, ACL, p. 46-51, 2004.

Parnee B. H., Ter Meulen A., Wall R. E., *Mathematical Methods in Linguistics*, vol. 30 of *Studies in Linguistics and Philosophy*, Kluwer Academic Publishers, Dordrecht, 1990.

Périnet A., Hamon T., « Hybrid acquisition of semantic relations based on context normalization in distributional analysis », *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA2013)*, Paris Nord, France, p. 113-122, October, 2013.

Resnik P., Selection and Information : A class-based approach to lexical relationships, PhD thesis, University of Pennsylvania, 1993.

Rocheteau J., Daille B., « TTC TermSuite - A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora », *Proceedings of the IJCNLP 2011 System Demonstrations*, Asian Federation of Natural Language Processing, Chiang Mai, Thailand, p. 9-12, November, 2011.

Salton G., Lesk M. E., « Computer evaluation of indexing and text processing », *Journal of the Association for Computational Machinery*, vol. 15, n° 1, p. 8-36, 1968.

Van der Plas L., Tiedemann J., « Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity », *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics ACL'06*, Sydney, Australia, 2006.

Wu H., Zhou M., « Optimizing synonym extraction using monolingual and bilingual resources », *In Proceedings of the second international workshop on Paraphrasing*, p. 72, 2003.

Analyse distributionnelle appliquée aux textes de spécialité

Réduction de la dispersion des données par abstraction des contextes

Amandine Périnet* — **Thierry Hamon*,****

* *Université Paris 13, Sorbonne Paris Cité, France*
amandine.perinet@edu.univ-paris13.fr
** *LIMSI-CNRS, BP133, Orsay, France*
hamon@limsi.fr

RÉSUMÉ. *Les modèles vectoriels utilisés pour l'analyse distributionnelle souffrent de la dispersion des données dans la matrice des contextes et du nombre important de dimensions de cette matrice. Ces limitations rendent difficile leur application aux corpus de spécialité, et les termes ne sont habituellement pas pris en compte alors qu'ils sont essentiels. Dans cet article, nous proposons une adaptation de l'analyse distributionnelle afin de pouvoir l'utiliser efficacement sur des textes de spécialité. L'approche proposée réalise une abstraction des contextes distributionnels pour réduire la dispersion des données et ainsi améliorer la qualité des regroupements tout en y incluant les termes. Nous avons évalué notre approche sur deux corpus médicaux. L'analyse des résultats montre que tout en permettant la prise en compte des termes dans l'analyse distributionnelle, l'abstraction des contextes, notamment grâce à l'inclusion lexicale, permet d'obtenir des regroupements sémantiques de meilleure qualité et plus homogènes.*

ABSTRACT. *The vector space models used for the distributional analysis suffer from data sparseness in the context matrix and from the great number of dimensions of the matrix. These limitations make its use on domain-specific corpora difficult and terms are usually not considered while they are essential. In this paper, we propose an adaptation of the distributional analysis in order to apply it efficiently on domain-specific corpora. The proposed approach performs an abstraction of the distributional contexts in order to reduce data sparseness and thus improve the quality of the distributional classes, and also to include terms in these classes. We evaluated our approach on two medical corpora. The results analysis shows that context abstraction, especially thanks to the lexical inclusion, leads to more semantically homogeneous classes with a better quality. The approach also achieves to take terms into account in the classes.*

MOTS-CLÉS : *analyse distributionnelle, corpus de spécialité, dispersion des données.*

KEYWORDS: *distributional analysis, domain-specific corpora, data sparseness.*

TAL. Volume 56 – n° 2/2015, pages 77 à 102

1. Introduction

Les relations entre termes jouent un rôle prépondérant dans les terminologies pour définir le sens des termes, mais aussi lors de l'utilisation des ressources terminologiques dans des applications en langue de spécialité pour augmenter leur couverture (Bodenreider *et al.*, 2002 ; McCray *et al.*, 2002) ou permettre leur adaptation (Cohen et Demner-Fushman, 2013). De nombreuses méthodes ont été proposées pour acquérir des relations sémantiques entre termes (Nastase *et al.*, 2013). Parmi celles-ci, l'analyse distributionnelle conduit au regroupement de mots sémantiquement proches en tenant compte des contextes qu'ils partagent : plus le nombre de contextes communs est élevé, plus les deux mots cibles sont sémantiquement proches (Harris, 1954 ; Firth, 1957). L'acquisition de relations sémantiques est vue comme un problème de regroupement de mots sémantiquement proches fondé sur des informations statistiques liées aux contextes et des mesures de similarité sémantique. À partir de ces regroupements, il est ainsi possible d'obtenir un nombre important de relations sémantiques qui ne sont, cependant, pas typées : il peut s'agir aussi bien de relations classiques telles que la synonymie ou l'hyperonymie, que de relations propres au domaine (Morlane-Hondère, 2013).

La mise en œuvre de l'analyse distributionnelle dans un processus automatique s'appuie généralement sur une représentation vectorielle des mots du corpus (Curran, 2004 ; Sahlgren, 2006). Un mot cible est alors un point dans un espace à n-dimensions où chaque dimension correspond à des contextes possibles, et où la valeur associée aux dimensions est le nombre d'occurrences du contexte correspondant. Le vecteur de chaque mot cible représente donc les informations contextuelles mais aussi des données statistiques distributionnelles comme le nombre de contextes et le nombre d'occurrences des contextes partagés (Turney et Pantel, 2010 ; Lund et Burgess, 1996). La similarité sémantique entre deux mots cibles est ainsi définie comme une proximité dans cet espace, calculée à l'aide du cosinus de l'angle par exemple. Les modèles vectoriels ont l'avantage de permettre une quantification facile de la proximité sémantique entre deux mots. Cependant, ils souffrent d'un problème de dispersion des données, car ils s'appuient sur des espaces aux très grandes dimensions et sur la redondance des contextes partagés alors qu'il s'agit d'évènements souvent rares (Chatterjee et Mohan, 2008). Ainsi, en considérant les espaces vectoriels comme des matrices de contexte, où les lignes sont les mots cibles du texte et les colonnes sont les contextes, on dispose généralement de matrices creuses ou éparses où beaucoup d'éléments sont à zéro car peu de contextes sont associés à un mot cible (Turney et Pantel, 2010).

Lorsqu'il s'agit de corpus de spécialité, ce problème de dispersion des données est accentué par des tailles de corpus beaucoup plus petites, un faible nombre d'occurrences du vocabulaire et un nombre de contextes partagés plus faible. Or, quand l'analyse distributionnelle est utilisée sur des corpus de spécialité, il est essentiel de prendre en compte les termes simples et complexes, à la fois dans les mots cibles, c'est-à-dire les mots regroupés, et dans les contextes des mots cibles (pour le calcul distributionnel). Ceci est généralement difficile à réaliser. Dans le cadre de traitement

de corpus monolingues, très peu de travaux existent. En revanche, pour l'extraction de lexiques bilingues à partir de corpus comparables, plusieurs travaux ont recours à l'analyse distributionnelle et portent un intérêt particulier aux termes complexes ((Daille et Morin, 2005), (Déjean *et al.*, 2002) et (Zweigenbaum et Habert, 2006) pour un aperçu général). Pour faire face au problème du faible nombre d'occurrences, Morin et Hazem (2014) utilisent un modèle de régression en amont de l'analyse distributionnelle. Ce modèle, entraîné sur des corpus de petite et de grande taille, leur permet de prédire le nombre d'occurrences de chaque contexte de manière à rendre ces valeurs plus fiables sur de nouveaux textes. Notre problématique est proche de ces travaux, qui s'appuient également sur les travaux fondateurs de Grefenstette (1994).

Ainsi, en raison de leur très faible nombre d'occurrences, les termes complexes se retrouvent généralement écartés du calcul de similarité, et les mots du corpus sont généralement considérés indépendamment du fait qu'il s'agisse de termes ou non.

Dans cet article, nous nous intéressons à l'adaptation d'une méthode d'analyse distributionnelle en prenant en compte les termes simples et complexes. Cette adaptation nous amène ainsi à aborder le problème de la dispersion des données et la réduction de la dimension de l'espace vectoriel dans le contexte des textes de spécialité. Nous émettons l'hypothèse que la suppression de la variation terminologique dans les contextes[1] ne dégrade pas trop leur sémantique, et permette d'améliorer la qualité des regroupements tout en prenant en compte les termes complexes. Pour cela, nous réalisons une abstraction des contextes à l'aide de relations sémantiques acquises automatiquement à partir de nos corpus de travail. Ainsi, les relations calculées par trois méthodes d'acquisition de relations d'hyperonymie (l'inclusion lexicale, les patrons lexico-syntaxiques et la variation terminologique) et d'une méthode d'acquisition de relations de synonymie sont utilisées pour généraliser ou normaliser les contextes distributionnels. Cette étape d'abstraction des contextes permet de réduire leur diversité ; le nombre de contextes différents est ainsi réduit.

Dans la suite de cet article, nous revenons en détail, à la section 2, sur le problème de la dispersion des données et nous exposons les solutions proposées dans les travaux précédents. La section 3 est consacrée à la description générale de la méthode distributionnelle mise en œuvre. Le processus d'abstraction des contextes est exposé à la section 4. Après avoir présenté le matériel utilisé (section 5), nous montrons les expériences réalisées et les résultats obtenus à la section 6.

2. Réduction de la dispersion des données

Les modèles vectoriels sont limités par la dispersion des données dans la matrice des contextes : beaucoup d'éléments de la matrice sont à zéro car généralement peu de contextes sont associés à un mot cible. On dispose alors d'une matrice de très faible densité, considérée comme creuse (Turney et Pantel, 2010). Sahlgren (2006) constate

1. Nous entendons par *contexte* une unité lexicale qui apparaît dans le voisinage du mot cible.

à travers ses expériences que plus de 99 % des entrées d'une matrice sont égales à zéro. Cet inconvénient est dû notamment à la distribution des mots dans le corpus (Baroni, 2009) : quelle que soit la taille du corpus, la plupart des mots ont un faible nombre d'occurrences, et un nombre de contextes très limité au regard du nombre de mots dans le corpus. La dispersion des données touche à la fois les corpus de langue générale, habituellement très volumineux (Weeds et Weir, 2005 ; van der Plas, 2008), et les textes de spécialité, souvent de plus petite taille et caractérisés par un vocabulaire avec un plus petit nombre d'occurrences. Ainsi, même dans un gros corpus tel que le BNC (100 millions de mots), moins de 14 % des mots ont un nombre d'occurrences de 20 ou plus (Baroni, 2009). Comme conséquence, les méthodes fondées sur l'analyse distributionnelle obtiennent de meilleures performances lorsque beaucoup d'informations sont disponibles, et notamment sur ces corpus volumineux, caractérisés par des nombres d'occurrences des mots du vocabulaire plus élevés. La réduction de la dispersion des données devient donc un enjeu majeur dans le cas des corpus de spécialité.

Il existe une forte corrélation entre la densité de la matrice et la performance du modèle vectoriel. Ainsi, même s'il est difficile de saisir la structure sémantique sous-jacente des matrices creuses, plus une matrice est creuse, moins le modèle vectoriel est performant sur la tâche donnée. Ce rapport est équivalent à celui liant le nombre d'occurrences des mots et la qualité des vecteurs (Bullinaria et Levy, 2007) : plus le nombre d'occurrences des mots est élevé, plus le modèle vectoriel est performant (Ferret, 2013 ; Weeds et Weir, 2005 ; van der Plas, 2008). *A contrario*, la similarité entre les mots cibles ayant un faible nombre d'occurrences est calculée à partir de très peu d'information dans les contextes. Ces mots cibles ont donc une plus grande tendance à être mal regroupés (Caraballo, 1999). Cependant, les mots avec un faible nombre d'occurrences ont un rôle essentiel sur la qualité des relations extraites, qu'ils soient en position de mot cible ou dans le contexte (Gorman et Curran, 2006), car ces mots rares peuvent correspondre à des contextes caractéristiques.

Ce premier problème a des conséquences sur le coût des traitements. Pour construire l'espace vectoriel, la méthode distributionnelle est fondée sur des éléments statistiques. Ainsi, si les données ne sont pas assez importantes, il n'est pas possible de disposer d'informations statistiques suffisamment fiables et significatives pour la construction du modèle distributionnel. De plus, la matrice de cooccurrence peut devenir extrêmement large quelle que soit la taille du corpus, et l'efficacité de l'algorithme en est alors affectée (Sahlgren, 2006). Le dilemme est donc le suivant : la plus grande quantité de données est nécessaire afin de construire un modèle suffisamment fiable, mais pour que les algorithmes puissent traiter les données à un coût raisonnable la plus petite quantité de données possible est préférable.

Pour répondre à ces problèmes et notamment pallier la dispersion des données, les solutions proposées sont de deux types : les premières visent à influer sur la définition des contextes, et les secondes interviennent au niveau de la construction ou de la réduction de la matrice des vecteurs de contexte. L'objectif est toujours de réduire l'espace, c'est-à-dire la mémoire occupée, et le temps de traitement.

Parmi les méthodes visant à influer sur les contextes, certaines s'intéressent plus particulièrement à la sélection des contextes utiles ou à l'intégration des informations sémantiques de manière à modifier la distribution des contextes. Ainsi, Broda *et al.* (2009) proposent de pondérer les contextes non pas en utilisant le nombre d'occurrences des contextes à l'état brut comme il est d'usage, mais en ordonnant les contextes en fonction de leur nombre d'occurrences. Le rang est ensuite utilisé pour pondérer puis sélectionner les contextes. D'autres approches s'appuient sur des modèles de langue pour déterminer les mots plausiblement intersubstituables, c'est-à-dire les substituts les plus probables pour représenter les contextes (Baskaya *et al.*, 2013). Ces modèles assignent des probabilités à des séquences arbitraires de mots en se fondant sur le nombre de cooccurrences dans un corpus d'entraînement (Yuret, 2012). Les mots substituts et leurs probabilités sont ensuite utilisés pour créer des paires de mots de manière à alimenter une matrice de cooccurrence, avant d'utiliser un algorithme de classification. Ces méthodes sont limitées car leur performance est proportionnelle à la taille du vocabulaire et elles nécessitent de disposer de données d'entraînement importantes. L'intégration d'informations sémantiques supplémentaires peut également être un moyen d'exercer une influence sur les contextes. En effet, Tsatsaronis et Panagiotopoulou (2009) ont démontré que la modification d'une méthode distributionnelle à l'aide de relations sémantiques calculées automatiquement ou provenant d'une ressource existante permet d'améliorer sa performance. Ainsi, avec un amorçage, Zhitomirsky-Geffet et Dagan (2009) modifient les poids des éléments au sein des contextes en s'appuyant sur les voisins sémantiques trouvés à l'aide d'une mesure de similarité distributionnelle. En s'appuyant sur ces travaux, Ferret (2013) s'intéresse au problème des mots ayant peu d'occurrences en corpus. Afin de mieux prendre en compte ces informations sémantiques, il propose d'utiliser un jeu d'exemples positifs et négatifs sélectionnés de manière non supervisée à partir d'un thésaurus distributionnel, et ainsi entraîner un classifieur supervisé. Ce classifieur est ensuite appliqué pour réordonnancer les voisins sémantiques. La méthode permet ainsi d'améliorer la qualité de la relation de similarité entre des noms ayant un nombre d'occurrences moyen ou faible.

Pour faire face aux problèmes liés à la très grande dimension des vecteurs, à la dispersion des données et au bruit statistique, une autre solution consiste à limiter le nombre de composants vectoriels avec un lissage de la matrice (Turney et Pantel, 2010). En effet, le calcul de la similarité entre toutes les paires de vecteurs est une tâche coûteuse alors que seuls les vecteurs qui partagent une dimension différente de zéro doivent être comparés[2]. La plupart des modèles connus utilisent des méthodes de réduction de dimension, généralement mises au point de manière à conserver les mots dont le nombre d'occurrences est faible. Une solution consiste à projeter les données aux dimensions élevées dans un espace ayant un nombre de dimensions plus réduit, tout en préservant approximativement les distances relatives entre les points, c'est-à-dire entre les mots cibles. La décomposition aux valeurs singulières

2. On considère que deux vecteurs ne partageant pas de dimension ne peuvent pas être similaires.

(SVD) (Deerwester *et al.*, 1990) est une méthode d'algèbre linéaire permettant la factorisation de matrice. Elle peut également être utilisée pour décomposer une matrice, afin d'obtenir une matrice finale ayant beaucoup moins de colonnes (généralement quelques centaines) mais plus dense (Turney et Pantel, 2010). Les méthodes fondées sur la SVD permettent ainsi de produire des vecteurs de contexte moins creux et moins affectés par le bruit statistique. À partir des données initiales, c'est-à-dire la matrice des contextes, cette technique divise la matrice en composants linéaires indépendants. La SVD est une méthode de factorisation de matrice coûteuse utilisée dans l'analyse sémantique latente (LSA) (Landauer et Dumais, 1997) pour réduire la matrice des contextes. La LSA est une méthode permettant de calculer des vecteurs sémantiques, ou vecteurs de contexte, à grande dimension, à partir des statistiques de cooccurrence des mots cibles.

Les méthodes décrites ci-dessus correspondent toutes à un processus d'optimisation généralement non supervisé. Face à ces méthodes, un nouvel ensemble de modèles vectoriels, fondé sur un apprentissage supervisé, a vu le jour ces dernières années et fait l'objet de nombreux travaux. Ainsi, à partir des travaux fondateurs de Bengio *et al.* (2003), des modèles prédictifs s'appuyant sur des réseaux de neurones ont été proposés (Mikolov *et al.*, 2013). Contrairement aux méthodes non supervisées qui commencent par construire les vecteurs de contexte et ensuite pondèrent ces vecteurs, les réseaux de neurones fixent directement les poids des vecteurs de manière à prédire les contextes dans lesquels les mots cibles correspondants ont tendance à apparaître. Le système apprend ainsi à assigner des vecteurs similaires à des mots cibles similaires (Baroni *et al.*, 2014).

Pour répondre au problème de la dispersion des données dans l'espace vectoriel, l'approche que nous proposons consiste à ajouter des informations sémantiques dans les contextes distributionnels, à l'instar de Tsatsaronis et Panagiotopoulou (2009) et Ferret (2013). Cependant, notre objectif diffère des travaux précédents : nous intégrons des relations sémantiques acquises automatiquement pour regrouper les contextes en faisant abstraction de la variation terminologique. Le nombre de contextes est ainsi réduit et, en revanche, la valeur associée à la cooccurrence mot cible/mot en contexte augmente. De plus, si les méthodes fondées sur la SVD réduisent la dispersion des données, leur fonctionnement n'est pas très explicite et elles s'appuient sur des méthodes mathématiques. Nous proposons au contraire de généraliser les contextes distributionnels en utilisant des connaissances linguistiques, des relations sémantiques acquises sur l'ensemble du corpus de travail. Le fonctionnement de notre méthode est explicitement décrit dans la section suivante.

3. Architecture globale de la méthode distributionnelle

Comme souligné précédemment, l'analyse distributionnelle appliquée à des corpus de spécialité ou des corpus de petite taille est limitée par une dispersion des données dans la matrice des contextes : cette matrice, représentant la distribution des mots ou des termes, contient beaucoup d'éléments ayant une valeur nulle. Pour tenter de ré-

soudre ce problème, nous proposons une approche consistant à densifier la matrice des contextes. Cette approche consiste à réaliser une abstraction des variations superficielles ou des contextes soit peu significatifs statistiquement soit liés au bruit de la méthode d'identification de ces distributions.

La méthode d'analyse distributionnelle que nous avons mise en œuvre suit le schéma présenté dans la figure 1. L'abstraction des contextes se trouve au cœur de la méthode. Elle exploite les contextes des mots cibles définis à l'étape 1 et précède le calcul de similarité sémantique (étape 3). L'abstraction des contextes, qui correspond, pour nous, à leur généralisation et à leur normalisation, est réalisée à l'aide de relations sémantiques acquises automatiquement. C'est une fois que la variation morphologique et sémantique est réduite dans les contextes que nous calculons la similarité entre les mots cibles.

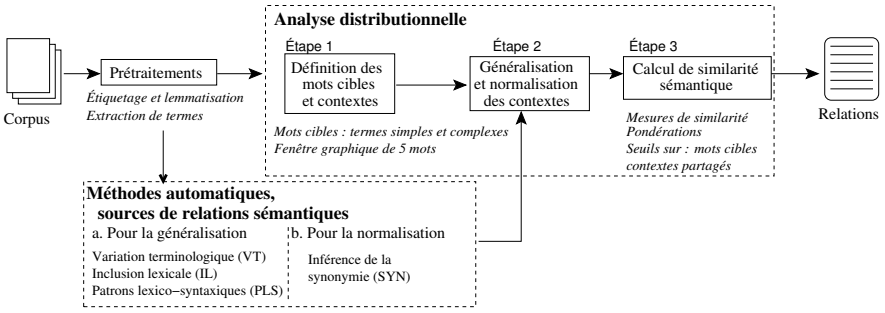

Figure 1. *Processus d'analyse distributionnelle*

3.1. *Définition des mots cibles et des contextes*

Dans le cadre d'applications en langue de spécialité, l'identification de relations sémantiques entre termes (simples et complexes) est primordiale. Les termes font référence aux notions du domaine et les relations sémantiques permettent plus d'appréhender le sens des termes. Ainsi, nous nous restreignons à l'analyse distributionnelle entre termes simples et complexes, qui constituent pour nous les mots cibles, et nous nous intéressons aux relations entre termes simples et complexes pris comme un seul ensemble (par exemple : *artère* et *souffle systolique*). Comme contextes distributionnels des mots cibles, nous avons choisi d'utiliser des fenêtres graphiques de $\pm$ 5 mots autour du mot cible. Il s'agit d'une taille reconnue comme adaptée aux textes de spécialité (Rapp, 2003 ; Généreux et Hamon, 2013). Elle permet aussi d'obtenir des contextes plus pertinents pour un mot cible, engendrant des résultats de meilleure qualité. Cependant, en limitant le nombre de contextes sélectionnés, cette taille de fenêtre accentue le problème de dispersion des données (Rapp, 2003). Les contextes sont composés de mots pleins qui cooccurent avec le mot cible au sein de la fenêtre

graphique. Nous considérons comme contexte les adjectifs, les noms, les verbes et les termes simples et complexes en écartant les mots vides (déterminants, conjonctions, adverbes, etc.).

3.2. *Calcul de similarité sémantique*

Lorsque les contextes ont été collectés, nous calculons la similarité entre deux mots cibles, en fonction de leurs contextes partagés. De nombreuses mesures de similarité et de pondération existent (Weeds *et al.*, 2004). Lors de précédentes expériences, nous avons constaté que l'indice de Jaccard obtient de meilleurs résultats avec les corpus de spécialité (Périnet et Hamon, 2014) :

$$S(w_m, w_n) = \frac{|\{\forall k, ctxt^k_{w_m}\} \cap \{\forall k, ctxt^k_{w_n}\}|}{|\{\forall k, ctxt^k_{w_m}\} \cup \{\forall k, ctxt^k_{w_n}\}|}$$

où $ctxt^k_{w_m}$ représente le contexte k du mot cible w_m, $\{\forall k, ctxt^k_{w_m}\}$ représente l'ensemble des contextes du mot cible w_m, et $|ctxt^k_{w_m}|$ correspond au nombre d'occurrences du contexte k pour le mot cible w_m.

L'indice de Jaccard compare le nombre de contextes communs à deux mots cibles à l'ensemble des contextes de ces mots (Tanimoto, 1958). Nous utilisons la généralisation pondérée de l'indice de Jaccard, telle que définie par Grefenstette (1994). Cette version généralise la similarité de Jaccard à la sémantique des valeurs non binaires, de manière à représenter chaque contexte par une valeur réelle entre 0 et 1. L'intersection est remplacée par le poids minimal et l'union par le poids maximal. Afin de valoriser les contextes les plus significatifs, nous avons pondéré les contextes par le nombre d'occurrences relatif du contexte c du mot cible w_m :

$$NbOccRel(w_m, c) = \frac{|ctxt^c_{w_m}|}{|\{\forall k, ctxt^k_{w_m}\}|}$$

Cette mesure de pondération des contextes est généralement associée à l'indice de Jaccard. Elle permet de prendre en compte l'importance d'un contexte d'un mot cible, par rapport au nombre total de contextes du mot cible.

Lors du calcul de similarité entre les mots cibles un très grand nombre de relations est généré. Garder toutes ces relations n'a pas de sens : un trop grand ensemble de relations est difficile à exploiter et à analyser *a posteriori*. Nous filtrons les relations avec la combinaison de trois paramètres : deux d'entre eux sont appliqués aux contextes (le nombre des contextes partagés et leur nombre d'occurrences) et le troisième est appliqué aux mots cibles (le nombre d'occurrences des mots cibles). Pour chaque paramètre, un seuil est calculé automatiquement en fonction du corpus. Les seuils que nous utilisons pour chaque corpus (voir section 5.1) sont répertoriés dans le tableau 1.

	Textes cliniques	Menelas
Nombre de contextes partagés	1	1
Nombre d'occurrences des contextes partagés	1	2
Nombre d'occurrences des mots cibles	3	3

Tableau 1. *Paramètres : valeurs des seuils sur les contextes et mots cibles*

4. Règles d'abstraction des contextes distributionnels

Une fois que les mots cibles et les contextes ont été définis, nous réalisons une abstraction des contextes. Cette abstraction est réalisée avec des ensembles de relations sémantiques acquises par des méthodes automatiques à partir du corpus de travail (voir section 5.2). Elle comprend une généralisation et une normalisation des contextes.

Nous partons du constat que les éléments superficiels différenciant les formes d'une même unité lexicale sont parfois effacés lors du processus de lemmatisation à l'aide d'une abstraction morphologique. Il est ainsi possible de regrouper sous une même unité lexicale ces différentes variations. Par exemple, la lemmatisation des verbes conjugués *opéré*, *opérons* et *opèrent*, permet d'effacer les marques de temps, de mode et de personne, et de regrouper ces trois formes sous le lemme *opérer*.

Une telle abstraction peut être également envisagée au niveau sémantique, où les traits « effacés » ne sont plus morphologiques mais sémantiques. Cette abstraction sémantique se traduit par exemple par le passage à un niveau supérieur dans une hiérarchie de concepts. Par exemple, les termes *chaise*, *fauteuil* et *tabouret* peuvent voir certains de leurs traits sémantiques effacés, de manière à être regroupés dans la classe sémantique des *sièges*. Le terme *tabouret* perd alors ses traits sémantiques *sans dossier* et *trois pieds*, le terme *fauteuil* perd ses traits *accoudoirs* et *confort*, et enfin la *chaise* perd ses traits *dossier* et *quatre pieds*.

Ainsi, nous émettons l'hypothèse que les contextes distributionnels peuvent être regroupés dans une même classe sémantique (par exemple, la classe des *sièges*). Cette classe serait représentée par un élément de cette classe, comme par exemple son hyperonyme (*siège*), de manière similaire au lemme *opérer* par rapport à l'ensemble des formes qu'il couvre. Le représentant *siège* serait alors utilisé comme substitut pour remplacer l'ensemble des mots appartenant à cette classe dans les contextes distributionnels. Après abstraction sémantique des contextes, la diversité des contextes est alors réduite : les contextes ne comptent alors plus qu'un seul lemme, *siège*, là où il y en avait quatre avant l'abstraction (*fauteuil*, *tabouret*, *chaise* et *siège*). Nous supposons que si ce substitut est utilisé pour remplacer les contextes, il devrait permettre de faire abstraction d'éléments superficiels, tout en gardant la même base sémantique et le même sens. L'objectif est d'une part, de diminuer la diversité des contextes distributionnels (on trouve alors dans les contextes uniquement *siège*, et non plus *chaise*, *tabouret* et *fauteuil*), et, d'autre part, d'augmenter le nombre d'occurrences

des contextes. Le contexte *siège*, s'il remplace ces trois termes, a alors une occurrence de 3.

4.1. *Généralisation des contextes*

La généralisation utilise des relations d'hyperonymie acquises par les méthodes définies à la section 5.2 : les patrons lexico-syntaxiques (PLS), l'inclusion lexicale (IL) et la variation terminologique (VT). Nous disposons alors, pour chaque mot w_i dans le contexte du mot w, de plusieurs ensembles de relations d'hyperonymie, $\mathbb{H}_s(w_i) = \{H_1, \ldots, H_n\} : \mathbb{H}_{PLS}$, $\mathbb{H}_{IL}$ et $\mathbb{H}_{VT}$, l'ensemble des hyperonymes pouvant être vide. Nous avons défini deux règles de substitution permettant de généraliser les contextes.

Ainsi, pour chaque mot w_i dans le contexte d'un mot w, nous appliquons l'une des règles suivantes :

1) si $|\mathbb{H}_S(w_i)| = 1$, alors $w_i := H_1$

Si un seul hyperonyme (H_1) acquis par une ou plusieurs méthodes S correspond au mot en contexte, le mot est remplacé par cet hyperonyme. Par exemple, si l'inclusion lexicale fournit la relation *restriction/restriction du débit coronaire*, alors *restriction du débit coronaire* est remplacée par *restriction*.

2) si $|\mathbb{H}_S(w_i)| > 1$, $w_i = \text{argmax}_{|H_i|}(\mathbb{H}_S(w_i))$

Si plusieurs hyperonymes acquis par une ou plusieurs méthodes S correspondent au mot en contexte, nous prenons en compte le nombre d'occurrences des hyperonymes $|H_1|, \ldots, |H_n|$ dans le corpus, et nous choisissons l'hyperonyme dont le nombre d'occurrences est le plus élevé dans le corpus. Par exemple, si pour le terme *artère coronaire droite* dans le contexte, les patrons lexico-syntaxiques fournissent les hyperonymes suivants : *artère coronaire*, *artère*, *vaisseau*, celui qui a le plus grand nombre d'occurrences est sélectionné et utilisé pour remplacer *artère coronaire droite* dans le contexte des mots cibles.

Quand plusieurs ensembles de relations d'hyperonymie sont disponibles, la phase de généralisation des contextes est réalisée en utilisant chaque méthode individuellement (par exemple, en généralisant avec les patrons lexico-syntaxiques) ou en combinant les méthodes. Les contextes sont alors généralisés en utilisant les ensembles de relations les uns à la suite des autres (par exemple, en généralisant avec les patrons puis avec l'inclusion lexicale) ou toutes ensemble (l'union des trois méthodes). Nous n'utilisons pas la propriété de transitivité de la relation d'hyperonymie.

4.2. *Normalisation des contextes*

Quant à la normalisation des contextes, elle utilise des relations de synonymie acquises à l'aide de la méthode définie en section 5.2. Nous avons défini une règle de normalisation qui vise à réduire les variations sémantiques. Les relations de synonymie sont tout d'abord regroupées sous la forme de groupes de synonymes et le

synonyme ayant le plus grand nombre d'occurrences est choisi comme représentant de ce groupe. Ainsi, à chaque mot w_i dans le contexte du mot cible w, correspond un groupe de synonymes $\mathbb{S}(R) = \{S_1, \ldots, S_n, R\}$ avec son représentant R.

Nous définissons une règle de normalisation des contextes, appliquée à chaque mot w_i dans le contexte d'un mot w pour substituer le mot du contexte par le représentant du groupe de synonymes auquel il appartient : $si\ \exists R | w_i \in \mathbb{S}(R), alors\ w_i\ :=\ R$ (l'ensemble de synonymes peut être vide). Si un mot dans le contexte appartient à un groupe de synonymes, il est remplacé par le représentant du groupe. Par exemple, si le terme *altération métabolique* dans le contexte d'un mot cible appartient au groupe de synonymes fourni par la méthode d'acquisition de synonymes (*anomalie métabolique*, *maladie métabolique*, *troubles métaboliques* et *altération métabolique*) celui qui a le plus grand nombre d'occurrences est sélectionné comme représentant et utilisé pour remplacer *altération métabolique* dans le contexte.

Pour la généralisation et la normalisation, si deux termes ont le nombre d'occurrences le plus élevé, le choix du terme représentatif n'ayant pas d'impact sur la méthode, nous sélectionnons le premier dans l'ordre alphabétique.

5. Matériel

Nous présentons dans cette section, les corpus de travail sur lesquels nous avons mené nos expériences, ainsi que les méthodes d'acquisition de relations sémantiques utilisées lors de la généralisation des contextes.

5.1. *Corpus*

Nous avons mené nos expériences sur deux corpus médicaux de taille et de langue différentes. Ces corpus contiennent des échanges entre spécialistes et se caractérisent par un degré de spécialisation élevé. Nous avons utilisé le corpus Menelas rédigé en français (Zweigenbaum, 1994). Le second corpus est rédigé en anglais et fourni par la compétition I2B2/2012 (Sun *et al.*, 2013). Dans les deux cas, les textes sont anonymisés.

Le corpus français Menelas comporte 84 839 mots. Il est constitué de deux grandes parties : un extrait d'un manuel de référence sur la coronarographie et les maladies coronariennes (environ 15 000 mots), et un ensemble de comptes rendus d'hospitalisation et de lettres de médecins hospitaliers aux médecins traitants concernant des malades atteints d'une maladie coronarienne (environ 70 000 mots). Les phrases de ce corpus sont longues avec 17,5 mots par phrase en moyenne. Pour ce corpus, le problème de dispersion des données est lié à la petite taille du vocabulaire. Si le manuel est bien rédigé, avec des phrases ayant une syntaxe *sujet - verbe - objet*, ce n'est pas toujours le cas des lettres des médecins, parfois produites à la hâte. Ainsi, le corpus comporte des abréviations, mais également un certain nombre d'erreurs. Les phrases

ne sont pas toujours bien construites, et peuvent, par exemple, ne pas contenir de verbe ou correspondre à une prise de notes.

Le corpus en anglais est composé de 311 documents cliniques provenant d'hôpitaux américains, fournis par Partners HealthCare et the Beth Israel Deaconess Medical Center. Il comporte 178 070 mots. Ce corpus comprend des phrases plus courtes que le corpus Menelas, avec 11 mots par phrase en moyenne, mais son vocabulaire est deux fois plus important, engendrant une plus grande diversité dans les contextes et accentuant le problème de dispersion des données.

Les corpus sont analysés à travers la plate-forme de TAL Ogmios[3] (Hamon et Nazarenko, 2008). La plate-forme a été configurée pour un étiquetage morphosyntaxique et une lemmatisation du corpus, à l'aide de TreeTagger (Schmid, 1994), et une extraction de termes analysés syntaxiquement a été réalisée à l'aide de YaTeA[4] (Aubin et Hamon, 2006). Les mots cibles et les contextes distributionnels sont définis à partir de ces prétraitements. Nous identifions ainsi les termes simples et complexes (dans les mots cibles et les contextes), et les mots dans les contextes (cf. section 3.1). Les termes extraits sont également utilisés pour l'acquisition des relations sémantiques.

5.2. *Acquisition de relations sémantiques*

La généralisation des contextes distributionnels s'appuie sur des relations sémantiques existantes, acquises sur l'ensemble du corpus de travail. Pour obtenir ces relations à partir de corpus, nous avons choisi d'utiliser plusieurs approches classiques d'acquisition de relations sémantiques entre termes : des patrons lexico-syntaxiques (PLS) dédiés à l'acquisition de relations d'hyperonymie, une méthode utilisant l'inclusion lexicale (IL), et des règles de variation terminologique (VT).

5.2.1. *Patrons lexico-syntaxiques*

Nous avons recours à des patrons définis pour l'acquisition de relations d'hyperonymie (*artère coronaire droite*/*artère* pour le français, et *disease*/*diabetes* en anglais). Pour le français, nous utilisons les patrons définis par Morin et Jacquemin (2004), comme par exemple *{quelques | plusieurs etc.} SN : LISTE*, où SN est un syntagme nominal et LISTE une liste de syntagmes. Pour l'anglais nous reprenons les patrons définis par Hearst (1992), pour acquérir des relations entre termes simples et complexes, par exemple, *SN {, SN}*{,} or other SN*, où SN est un syntagme nominal.

5.2.2. *Inclusion lexicale*

Cette approche s'appuie sur l'hypothèse, selon laquelle, si un terme en position tête (par exemple, *infarctus*) est inclus lexicalement dans un autre terme (par exemple, *infarctus du myocarde*), il existe généralement une relation d'hyperonymie entre ces

3. `http://search.cpan.org/~thhamon/Lingua-Ogmios/`
4. `http://search.cpan.org/~thhamon/Lingua-YaTeA/`

deux termes (Grabar et Zweigenbaum, 2003). Nous utilisons ici l'analyse syntaxique des termes fournie par YaTeA.

5.2.3. *Variation terminologique*

Nous utilisons la méthode d'acquisition de variantes terminologiques proposée par Jacquemin (2001) et implémentée dans Faster. Cette méthode exploite des règles de transformation morphosyntaxique décrivant la variation terminologique. Les variantes peuvent résulter de plusieurs opérations syntaxiques, morphologiques ou lexicales : principalement la permutation (*antibiotic course/courses of antibiotics*), la dérivation (*sténose de l'aorte/sténose aortique*) et l'insertion (*abdominal pain/abdominal muscle pain*). Par ailleurs, bien que Faster offre la possibilité d'obtenir des variantes sémantiques, nous avons choisi de ne pas les acquérir de cette manière. En effet, SynoTerm propose également des relations sémantiques dont une partie pourrait être acquise par Faster. De plus, les relations acquises par SynoTerm sont typées sémantiquement alors que Faster n'offre pas cette possibilité.

Dans le corpus en français (Menelas), les règles utilisées pour identifier des relations sémantiques entre termes sont essentiellement l'insertion[5]. En revanche, pour l'anglais, les trois règles sont utilisées. L'insertion d'un modifieur, par exemple *de revascularisation*, au sein du terme complexe *chirurgie coronarienne*, permet d'identifier une relation d'hyperonymie entre les deux termes concernés, *chirurgie coronarienne* et *chirurgie de revascularisation coronarienne*. Dans le cas de la dérivation et de la permutation, nous obtenons très peu de relations avec cette règle, et les relations obtenues sont plus apparentées à des relations de synonymie que d'hyperonymie.

Comme l'approche utilisée ne propose pas de relations typées sémantiquement et comme la plupart des relations sont identifiées grâce à la règle d'insertion, nous avons considéré les relations obtenues comme des relations d'hyperonymie. Les termes hyperonyme et hyponyme sont identifiés à partir du nombre de mots présents dans chaque terme : le terme le plus court correspond alors à l'hyperonyme (*lésion significative*), et le terme le plus long à l'hyponyme (*lésion coronaire significative*).

5.2.4. *Inférence de relations de synonymie*

Pour la normalisation des contextes, nous utilisons également une méthode à base de règles visant l'acquisition de relations sémantiques (Hamon et Nazarenko, 2001). Cette méthode permet d'inférer une relation de synonymie entre des termes complexes si au moins un de leurs composants (têtes) sont synonymes. Pour cela, nous utilisons deux dictionnaires existants. Pour le français, il s'agit du dictionnaire de langue générale *Le Robert*, qui contient des relations de synonymie entre mots. Pour l'anglais, nous avons utilisé les relations de synonymie entre mots proposées par WordNet (Fellbaum, 1998).

5. Nous ne disposions pas de ressources permettant d'identifier des variantes terminologiques par dérivation.

6. Expériences et résultats

Nous présentons dans cette section les expériences que nous avons menées, la méthode d'évaluation utilisée ainsi que les résultats obtenus.

6.1. *Expériences*

L'abstraction des contextes ayant pour objectif la réduction de la dispersion des données, nous avons évalué et caractérisé l'impact de la normalisation et de la généralisation des contextes sur la qualité des regroupements et des relations sémantiques obtenus. Pour cela, nous avons utilisé les règles proposées dans la section 4. Celles-ci s'appuient sur les relations sémantiques acquises automatiquement (cf. section 5.2) pour généraliser et normaliser les contextes séparément et de manière combinée. Afin de cerner la contribution de chaque méthode d'acquisition de relations sémantiques, ainsi que leur complémentarité, nous avons réalisé deux séries d'expériences autour de l'abstraction des contextes : une première série autour de la généralisation des contextes et une seconde pour la normalisation.

Tout d'abord, les règles de généralisation des contextes distributionnels w_i sont appliquées en utilisant séparément les ensembles $\mathbb{H}_{PLS}(w_i)$, relations d'hyperonymie acquises à l'aide des patrons lexico-syntaxiques (AD/PLS), $\mathbb{H}_{IL}(w_i)$, relations d'hyperonymie issues de l'inclusion lexicale (AD/IL), et $\mathbb{H}_{VT}(w_i)$, variantes terminologiques (AD/VT). Toutes les relations d'hyperonymie ont également été prises en compte dans leur ensemble indépendamment de la méthode utilisée pour les acquérir. On considère alors l'union des trois méthodes, c'est-à-dire l'ensemble $H(w_i) = \mathbb{H}_{PLS}(w_i) \cup \mathbb{H}_{IL}(w_i) \cup \mathbb{H}_{VT}(w_i)$ – AD/ALL3, pour appliquer les règles de généralisation sur le contexte w_i.

Nous n'avons réalisé qu'une seule expérience utilisant la normalisation des contextes w_i à l'aide des groupes de synonymes $\mathbb{S}(w_i)$ acquis automatiquement.

6.2. *Évaluation*

L'évaluation des relations acquises par analyse distributionnelle reste aujourd'hui une problématique importante et il est difficile d'évaluer une méthode distributionnelle en raison de la grande variété de relations qu'elle produit (Adam *et al.*, 2013 ; Morlane-Hondère, 2013). En effet, ces ressources contiennent un large spectre de relations lexicales, aussi bien des relations dites classiques que des relations moins bien spécifiées mais qui peuvent être pertinentes dans certaines applications (Morris et Hirst, 2004). Nous présentons tout d'abord les ressources puis les mesures que nous avons utilisées pour l'évaluation de notre approche.

6.2.1. *Références*

Nous faisons le choix d'évaluer notre méthode de manière intrinsèque, c'est-à-dire d'évaluer directement les relations sémantiques produites par la méthode. À l'instar de Curran (2004) et Ferret (2013), nous considérons ici les relations obtenues comme des ensembles de voisins associés à des mots cibles, les voisins étant ordonnés suivant la similarité avec le mot cible.

Aussi, nous comparons les relations sémantiques acquises à celles fournies par des ressources existantes. Nous utilisons les relations sémantiques issues de l'UMLS[6].

Les résultats obtenus sur le corpus Menelas sont évalués par rapport aux relations présentes dans la partie française de l'UMLS, c'est-à-dire 2 434 relations entre les termes du corpus[7]. Les types des relations contenues dans la référence sont majoritairement des co-hyponymes (1 536 relations), mais on dispose également de 333 hyperonymes, de 438 synonymes, et de 128 relations spécifiques du domaine (`expanded_form_of`). Les résultats obtenus sur le corpus de textes cliniques sont comparés aux 53 203 relations de la partie anglaise de l'UMLS restreintes aux termes du corpus. Les types des relations sont majoritairement co-hyponymes (22 680 relations) mais on dispose aussi de 22 939 relations du domaine (`has_sign_or_symptom`, etc.), de 6 505 hyperonymes et de 1 079 synonymes.

6.2.2. *Mesures d'évaluation*

Nous avons utilisé plusieurs métriques habituellement utilisées pour évaluer les résultats d'une analyse distributionnelle : la macro-précision (Sebastiani, 2002), la moyenne des précisions moyennes (MAP) (Buckley et Voorhees, 2005) et la R-précision. Nous utilisons le programme standard trec_eval[8], mis au point lors des campagnes TREC.

La macro-précision est la moyenne des précisions $p(w_i)$ obtenues pour chaque mot cible (w_i) et un ensemble de voisins sémantiques I_i^j, $I_i^{j(+)}$ étant un voisin pertinent pour le mot cible w_i, et n_i le nombre de ses voisins :

$$P = \frac{\sum_{k=1}^{|w_i|} \frac{\sum_{j=1}^{n_i} I_i^{j(+)}}{\sum_{j=1}^{n_i} I_i^j}}{|w_i|}$$

Nous avons considéré quatre sous-ensembles voisins permettant d'obtenir la macro-précision après examen de 1 ($n_i = 1$, P@1), 5 ($n_i = 5$, P@5), 10 ($n_i = 10$, P@10) et 100 voisins ($n_i = 100$, P@100) :

6. http ://www.nlm.nih.gov/research/umls/

7. La partie française de l'UMLS propose 1 735 419 relations mais nous restreignons l'ensemble de référence aux seules relations entre les termes du corpus.

8. `http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz`

$$P@N = \sum_{i=1}^{|w_i|} p(w_i|n_i = N)$$

La R-précision (Buckley et Voorhees, 2005) est une alternative à la précision limitée à un rang n. Elle consiste à utiliser comme seuil n_i le nombre de voisins corrects attendu pour un mot cible w_i, n_i variant alors suivant les mots cibles. La mesure est ainsi plus équitable qu'une précision à seuil fixe, car le seuil de précision varie en fonction du nombre de voisins attendus. Nous utilisons ensuite la moyenne des R-précisions par mot cible.

Nous évaluons également les résultats à l'aide de la *Mean Average Precision* (MAP). Celle-ci est obtenue en considérant la précision non interpolée $UAP(I_i^j)$ des voisins sémantiques I_i^j au rang j, n_i étant le nombre de voisins sémantiques I_i^j du mot cible w_i. La MAP est alors la moyenne de ces précisions non interpolées :

$$MAP = \frac{1}{|w_i|} \sum_{i=1}^{|w_i|} \frac{1}{n_i} \sum_{j=1}^{n_i} UAP(I_i^j)$$

La MAP reflète la qualité du classement : elle valorise le fait que la méthode ordonne tous les voisins sémantiques corrects proches de la tête de liste. Réciproquement, le fait d'ajouter des voisins sémantiques incorrects en fin de liste (après les voisins corrects) ne pénalise pas la méthode. Ainsi, contrairement à la R-précision qui permet d'évaluer également le classement des voisins, la MAP prend en compte tous les voisins, même ceux en fin de classement, alors que la R-précision se limite aux n voisins corrects attendus.

6.3. *Résultats*

Dans cette section, nous présentons et analysons les résultats obtenus sur nos deux corpus, par l'analyse distributionnelle après généralisation ou normalisation des contextes. Nous utilisons comme point de comparaison (*baseline*) les résultats obtenus avec l'analyse distributionnelle seule, c'est-à-dire sans abstraction des contextes. Compte tenu de l'absence de typage sémantique des relations par l'analyse distributionnelle et du nombre de relations proposées par cette approche, la qualité des résultats d'une analyse distributionnelle est généralement faible et difficile à apprécier lorsque ces résultats sont confrontés à des ressources existantes. Aussi, nous nous intéressons surtout à la différence entre les résultats obtenus à l'aide des mesures d'évaluation et à la qualité des regroupements à travers une analyse manuelle.

6.3.1. *Généralisation*

Nous avons analysé les regroupements obtenus après généralisation des contextes distributionnels en utilisant individuellement chaque ensemble de relations séman-

tiques acquises (AD/IL, AD/PLS, AD/VT), en combinant l'ensemble de ces relations (AD/ALL3).

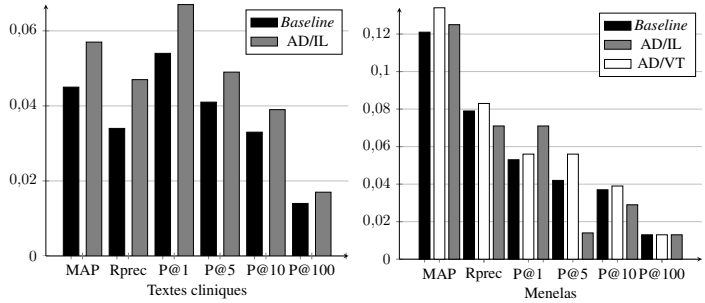

Figure 2. *Résultats obtenus pour la généralisation et comparaison avec la* baseline. *Généralisation avec l'inclusion lexicale (IL) pour le corpus de textes cliniques, et avec la variation terminologique (VT) et l'inclusion lexicale pour le corpus Menelas.*

Les résultats les plus intéressants obtenus après généralisation des contextes sont présentés à la figure 2, pour les deux corpus. Pour le corpus de textes cliniques, la généralisation avec l'inclusion lexicale (AD/IL) permet d'augmenter les résultats quelle que soit la mesure d'évaluation utilisée. L'impact de la généralisation des contextes est beaucoup plus faible lorsque les relations sont acquises à l'aide des patrons lexico-syntaxiques (AD/PLS) ou lorsqu'il s'agit de variantes terminologiques (AD/VT). Parmi les précisions à un rang donné, nous observons que les écarts les plus significatifs sont obtenus avec P@1, c'est-à-dire avec la précision prenant en compte uniquement le premier voisin (+ 0,013). Aussi, la MAP et la R-précision augmentent grâce à la généralisation des contextes avec respectivement + 0,012 et + 0,013. La généralisation des contextes semble ainsi contribuer à l'amélioration de l'ordonnancement des voisins. De plus, des termes pertinents, qui n'apparaissaient pas dans les 10 premiers éléments avec la *baseline* (analyse distributionnelle seule), remontent dans le classement des voisins grâce à la généralisation des contextes.

En ce qui concerne le corpus Menelas, seules les relations acquises à l'aide de la variation terminologique (AD/VT) ont un impact positif sur les résultats lors la généralisation. Les constats sont similaires aux précédents : la MAP et la R-précision augmentent par rapport à la *baseline*, respectivement + 0,013 et + 0,014. Les résultats de l'analyse distributionnelle en utilisant les deux autres ensembles de relations sémantiques (AD/IL et AD/PLS) sont plus mitigés : tandis que la R-précision décroît (– 0,008 et – 0,016), la MAP augmente légèrement (+ 0,004), la précision P@1 est la plus importante avec les relations d'inclusion lexicale (+ 0,018) et un peu plus faible avec les relations issues des patrons lexico-syntaxiques (+ 0,01), mais les précisions aux rangs supérieurs (P@5, P@10 et P@100) sont fortement dégradées (jusqu'à – 0,028 pour la P@5 de AD/IL) ou inchangées (P@100).

Lorsque l'on considère toutes les relations disponibles pour la généralisation des contextes, indépendamment de la méthode utilisée pour les acquérir (AD/ALL3), les résultats sont également améliorés par rapport à la *baseline*, en particulier, sur le corpus de textes cliniques (figure 3). Sur le corpus Menelas, les résultats sont similaires ou légèrement supérieurs à ceux obtenus lorsque les contextes sont généralisés avec l'inclusion lexicale.

Même si les résultats semblent varier selon le corpus, la généralisation des contextes fondée sur les relations d'inclusion lexicale améliore la qualité des résultats obtenus. Si l'on considère les mesures d'évaluation telles que la P@1, la MAP et la R-précision, la généralisation des contextes a un impact positif quelle que soit la méthode des relations employée pour acquérir ces relations. Les résultats bénéficient également de l'union des trois ensembles de relations par rapport aux résultats obtenus par la généralisation individuelle. Enfin, les observations sur le corpus Menelas indiquent que l'inclusion lexicale semble avoir une influence importante sur les relations obtenues lorsqu'elles sont comparées à une référence. Ce constat peut être dû au nombre de relations proposé par l'inclusion lexicale.

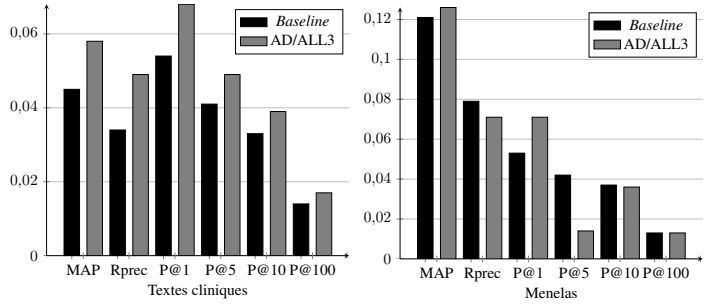

Figure 3. *Résultats obtenus sur les corpus de textes cliniques et Menelas, en généralisant les contextes à l'aide de toutes les relations à disposition (AD/ALL3) et comparaison avec la* baseline.

Étant donné la faible couverture de nos ressources et afin de mieux caractériser les voisins acquis et la qualité des relations sur les corpus, nous avons analysé les 10 premiers voisins de chaque mot cible retrouvés dans l'UMLS, pour les deux corpus.

Nous présentons dans le tableau 2, les 10 premiers voisins du mot cible *cough* (*toux*), obtenus avec la *baseline* et après généralisation des contextes avec l'inclusion lexicale (AD/IL), sur le corpus de textes cliniques. Les voisins soulignés sont ceux présents dans l'UMLS (*pain*, *fever*, *history*, *diarrhea* et *dyspnoea*). La généralisation avec l'inclusion lexicale (AD/IL) permet de mieux classer ces voisins, qui remontent alors dans le classement des 10 premiers voisins, à l'exception de *pain* déjà présent avec la *baseline*. Les 10 premiers voisins obtenus avec la généralisation sont plus pertinents et décrivent mieux le sens du mot cible *cough*. En effet, la *baseline* obtient un groupement sémantique autour de l'évanouissement et de la perte de connaissance

Mot cible : *cough*

	Baseline		**AD/IL**	
Rang	Voisin	Sim	Voisin	Sim
1.	nausea	0,00091	nausea	0,00108
2.	*pain*	0,00063	**fever**	0,00105
3.	*paroxysmal nocturnal dyspnoea*	0,00048	**vomiting**	0,00105
4.	*weakness*	0,00045	**chill**	0,00101
5.	*dizziness*	0,00044	**history**	0,0082
6.	*loss of consciousness*	0,00039	*pain*	0,00080
7.	*abd pain*	0,00036	**patient**	0,00080
8.	*numbness*	0,00036	**diarrhea**	0,00078
9.	*home*	0,00035	**dysuria**	0,00074
10.	*sweat*	0,00035	**dyspnoea**	0,00065

Tableau 2. *Corpus de textes cliniques : exemple de 10 premiers voisins obtenus pour le mot cible « cough », avec la* baseline*, et après généralisation avec l'inclusion lexicale (AD/IL). Les voisins en gras remontent dans le classement, ceux en italique descendent, et les termes soulignés sont ceux présents dans la référence.*

avec les termes *weakness*, *dizziness*, *loss of consciousness*, *numbness*. Ce groupement est sémantiquement homogène, mais sémantiquement moins proche de *cough* que les termes *fever*, *chill*, *dyspnoea*.

Nous avons réalisé une analyse similaire sur le corpus Menelas. Nous présentons ici quelques observations sur les regroupements obtenus lorsque les contextes sont généralisés avec les variantes terminologiques. Le tableau 3 illustre cette analyse avec les 10 premiers voisins du mot cible *cholestérol*. La généralisation des contextes a une influence sur le classement des voisins : 6 des 10 premiers voisins obtenus avec la *baseline* descendent dans le classement (en italique), c'est-à-dire que la généralisation augmente le score de similarité obtenu des voisins classés après les 10 premiers (en gras). Certains voisins restent inchangés du point de vue de leur similarité avec le mot cible même s'ils descendent dans le classement : *oblitération*, et *angiocoronarographie*. D'autres ont leur score de similarité qui est légèrement augmenté (*bilan lipidique*, *triglycéride*) et qui sont maintenus au même rang. Parmi ces 10 voisins, dans les deux cas de figure, aucun voisin n'est retrouvé dans l'UMLS sans que cela signifie qu'il ne s'agit pas de voisins pertinents. Ainsi, pour *cholestérol*, l'analyse distributionnelle après généralisation permet d'acquérir des relations du domaine, avec les voisins *bilan lipidique*, *bilan biologique*, *coronarographie*, *ventriculographie*, *angio-coronarographie*, mais également la relation d'hyperonymie entre *cholestérol* et *cholestérol total*. Globalement, les voisins acquis après généralisation sont un ensemble plus homogène sémantiquement, et correspondent au concept d'examen clinique.

Nous avons également observé que la combinaison AD/IL+PLS a une plus grande influence sur les mots cibles ; cette configuration réduit le nombre de mots cibles re-

trouvés dans la ressource et ceux-ci sont en partie différents des mots cibles obtenus avec la *baseline*. En revanche, en généralisant avec les variantes terminologiques, les mots cibles sont globalement les mêmes qu'avec la *baseline*, la différence se situe plus au niveau des voisins et de leur classement.

Mot cible : *cholestérol*

	Baseline		**AD/VT**	
Rang	Voisin	Sim	Voisin	Sim
1.	bilan lipidique	0,0028	bilan lipidique	0,0029
2.	triglycéride	0,0020	triglycéride	0,0021
3.	*lésion sévère*	0,0011	**cinétique ventriculaire gauche**	0,0012
4.	angio-coronarographie	0,0010	**bilan biologique**	0,0012
5.	oblitération	0,0010	**cholestérol total**	0,0012
6.	*extrasystole ventriculaire*	0,0009	**fonction ventriculaire gauche**	0,0012
7.	*examen clinique*	0,0009	**ventriculographie**	0,0011
8.	*coronaire droite*	0,0008	oblitération	0,0011
9.	*parenchyme pulmonaire*	0,0008	**coronarographie**	0,0010
10.	*pression pulmonaire*	0,0008	angio-coronarographie	0,0010

Tableau 3. *Corpus Menelas, fenêtre restreinte : exemple de 10 premiers voisins obtenus pour le mot cible* cholestérol*, avec la* baseline*, et après généralisation avec les variantes terminologiques (AD/VT). Les voisins en gras remontent dans le classement, ceux en italique descendent.*

6.3.2. *Normalisation*

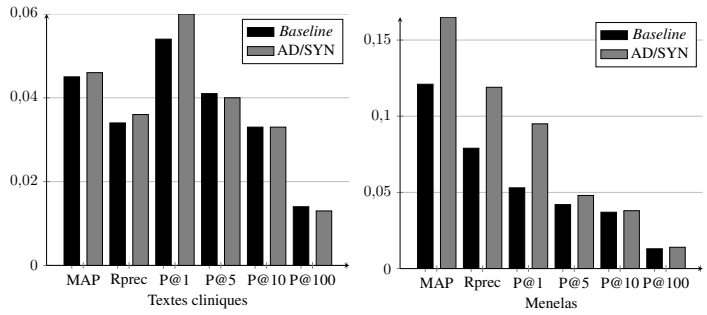

Figure 4. *Résultats obtenus pour la normalisation et comparaison avec la* baseline*. Normalisation avec les relations de synonymie pour les corpus de textes cliniques et Menelas.*

La normalisation est réalisée à l'aide des relations de synonymie acquises en corpus (SYN). Les résultats obtenus sont présentés dans la figure 4. Nous pouvons observer que l'impact de la normalisation est quasiment nul avec le corpus de textes cliniques mais plus important avec le corpus Menelas. Ainsi, pour les textes cliniques,

quelle que soit la mesure d'évaluation utilisée, les résultats sont quasiment identiques à ceux obtenus avec la *baseline*, avec un écart maximal de + 0,006 entre la *baseline* et la généralisation, obtenu en termes de P@1. En revanche, pour le corpus Menelas, l'impact de la normalisation sur la qualité de résultats est positif, quelle que soit la métrique utilisée. À l'instar de la généralisation, l'impact le plus fort de la normalisation est constaté avec la MAP, la R-précision et la P@1, pour lesquelles l'écart avec la *baseline* est respectivement de + 0,044, + 0,040 et + 0,042. Ainsi, la normalisation permet également d'améliorer le classement des termes. Et à l'inverse de la généralisation, son impact semble plus important quand le corpus est de petite taille.

Mot cible : *réseau coronarien*

	Baseline		**AD/SYN**	
Rang	Voisin	Sim	Voisin	Sim
1.	réseau circonflexe	0,00135	réseau circonflexe	0,00163
2.	*examen clinique*	0,00127	**coronaire droite**	0,00141
3.	artère coronaire	0,00121	artère coronaire	0,00140
4.	coronaire droite	0,00117	*examen clinique*	0,00140
5.	*maladie*	0,00091	**lésion**	0,00136
6.	*valve*	0,00088	**hypertension artérielle**	0,00118
7.	*index cardio-thoracique*	0,00088	**sténose**	0,00107
8.	*coeur*	0,00088	**athérome**	0,00104
9.	athérome	0,00085	**réseau coronaire**	0,00099
10.	*hypertrophie ventriculaire gauche pariétale*	0,00083	*maladie*	0,00099

Tableau 4. *Corpus Menelas : exemple de 10 premiers voisins obtenus pour le mot cible* réseau coronarien *, avec la* baseline*, et séparément après normalisation avec les synonymes (AD/SYN).*

De manière similaire à la généralisation, nous avons analysé manuellement les regroupements obtenus. Dans le tableau 4, nous présentons un exemple de regroupement avec les 10 premiers voisins du mot cible *réseau coronarien* obtenu avec la *baseline* et après normalisation (AD/SYN) des contextes. Nous pouvons observer que la normalisation permet de faire remonter dans le classement les termes *coronaire droite*, *lésion*, *hypertension artérielle*, *sténose*, *athérome* et *réseau coronaire*. Le résultat obtenu a une plus grande cohérence sémantique, avec un groupement sémantique autour des *pathologies* et *complications* qui peuvent être liées au réseau coronarien et un autre groupement autour du réseau lui-même. La normalisation permet ainsi de faire remonter la variante *réseau coronaire*.

Dans l'ensemble, nous avons constaté que l'ordre des 10 premiers voisins varie quand la normalisation est appliquée sur les contextes. En revanche, nous avons observé des similarités dans les regroupements obtenus après normalisation et après généralisation à l'aide les variantes terminologiques (AD/VT) : les voisins obtenus sont généralement les mêmes et les valeurs de similarité identiques. De même, l'impact sur le classement des voisins est moins important avec la normalisation qu'après généralisation des contextes.

6.4. *Bilan*

Les expériences présentées ci-dessus montrent que l'abstraction des contextes distributionnels permet d'obtenir des groupements sémantiques plus homogènes et cohérents. C'est essentiellement la pertinence des voisins sémantiques acquis qui est affectée par l'abstraction. L'impact de notre approche est généralement plus important lorsqu'il s'agit d'une généralisation des contextes, notamment à l'aide de l'inclusion lexicale, qu'avec une normalisation. L'importante contribution de l'inclusion lexicale peut être considérée comme un avantage car il s'agit d'informations syntaxiques qui peuvent être facilement fournies en grand nombre par un extracteur de termes. Et, contrairement aux relations acquises grâce à des patrons lexico-syntaxiques ou aux variantes terminologiques, ces relations sont stables formellement (la méthode d'acquisition reste la même et ne dépend pas de marques présentes dans les textes) et sémantiquement (à quelques exceptions près, leur interprétation est très fiable (Dupuch *et al.*, 2012)). De plus, l'analyse manuelle des relations révèle que l'abstraction des contextes distributionnels, et en particulier leur généralisation, permet d'obtenir des groupements sémantiques plus homogènes ainsi que des voisins sémantiques sémantiquement plus proches du mot cible qu'avec l'analyse distributionnelle seule. Les relations obtenues après abstraction des contextes sont majoritairement des co-hyponymes. L'abstraction permet également d'obtenir quelques relations du domaine propres au mot cible, telles que par exemple les relations *maladie*/*examen médical*, *examen médical*/*conséquence*. Cependant, notre méthode possède des limites, car même si elle permet d'identifier des regroupements sémantiques, les relations acquises ne sont pas typées, et notre évaluation manuelle des résultats reste partielle étant donné le très grand nombre de relations acquises.

7. Conclusion

Dans cet article, nous nous sommes intéressés à la réduction de la dispersion des données dans un espace vectoriel, et à la prise en compte des termes dans un modèle vectoriel, afin de pouvoir mettre en œuvre efficacement l'analyse distributionnelle sur des corpus de spécialité. Pour cela, nous avons proposé une méthode d'abstraction des contextes distributionnels s'appuyant sur des relations sémantiques acquises en corpus. Cette adaptation d'une méthode distributionnelle permet (i) de réduire le nombre de dimensions de l'espace vectoriel en diminuant la variation terminologique des contextes distributionnels, tout en conservant leur sémantique, et (ii) de faciliter le regroupement sémantique des termes simples et complexes en augmentant leur nombre de cooccurrences avec des contextes regroupés. Les relations sémantiques utilisées sont calculées en corpus grâce à trois méthodes d'acquisition de relations d'hyperonymie, et une méthode d'acquisition de relations de synonymie. L'abstraction des contextes distributionnels consiste alors à les généraliser grâce à ces relations d'hyperonymie et à les normaliser à l'aide des synonymes. Les résultats des expériences réalisées sur deux corpus du domaine médical en anglais et en français montrent que l'abstraction des contextes distributionnels améliore la qualité des résultats. D'une

part, les groupements sémantiques obtenus sont ainsi plus homogènes et cohérents, et, d'autre part, les termes complexes sont pris en compte comme des mots cibles. Dans l'ensemble, la généralisation, et en particulier les relations fournies par l'inclusion lexicale, ont un impact fort sur les regroupements obtenus. Quant à la normalisation, elle permet surtout d'améliorer le classement des voisins et la qualité des relations obtenues lorsqu'il s'agit d'un corpus de petite taille.

Ce travail ouvre plusieurs perspectives. Tout d'abord, les relations d'hyperonymie et de synonymie que nous avons utilisées ont été exploitées séparément. Or, ces relations acquises automatiquement pourraient être considérées comme une ébauche de taxonomie ou d'un réseau sémantique. L'utilisation du réseau de relations doit nous permettre de réaliser une abstraction des contextes plus précise sémantiquement notamment en prenant en compte cette taxonomie et les distances sémantiques entre les termes. Aussi, l'ensemble des relations acquises en corpus peut être bruité. En effet, les relations générées par les méthodes automatiques contiennent des erreurs ou des relations peu intéressantes en soi, qui peuvent être bénéfiques à l'abstraction des contextes, mais qui pourraient également dégrader les résultats. Pour pallier cette éventuelle dégradation des résultats, nous envisageons d'utiliser d'autres sources de relations comme les terminologies. Il nous sera alors possible d'évaluer l'impact de l'abstraction et des relations lorsque leur statut terminologique est maîtrisé, et ce, avec des relations jugées plus fiables.

8. Bibliographie

Adam C., Fabre C., Muller P., « Évaluer et améliorer une ressource distributionnelle », *Traitement Automatique des Langues*, vol. 54, n° 1, p. 71-97, 2013.

Aubin S., Hamon T., « Improving Term Extraction with Terminological Resources », *Advances in Natural Language Processing*, n° 4139 in *LNAI*, Springer, p. 380-387, 2006.

Baroni M., *Corpus linguistics : An international handbook*, vol. 2, Anke Lüdeling and Merja Kytö, Berlin, chapter Distributions in text, p. 803-821, 2009.

Baroni M., Dinu G., Kruszewski G., « Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Baltimore, Maryland, p. 238-247, June, 2014.

Baskaya O., Sert E., Cirik V., Yuret D., « AI-KU : Using Substitute Vectors and Co-Occurrence Modeling For Word Sense Induction and Disambiguation », *Proceedings of SemEval - 2013*, Association for Computational Linguistics, Atlanta, Georgia, USA, p. 300-306, 2013.

Bengio Y., Ducharme R., Vincent P., Janvin C., « A Neural Probabilistic Language Model », *J. Mach. Learn. Res.*, vol. 3, p. 1137-1155, 2003.

Bodenreider O., Rindflesch T. C., Burgun A., « Unsupervised, Corpus-based Method for Extending a Biomedical Terminology », *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3*, BioMed '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 53-60, 2002.

Broda B., Piasecki M., Szpakowicz S., « Rank-Based Transformation in Measuring Semantic Relatedness. », *in* Y. Gao, N. Japkowicz (eds), *Canadian Conference on AI*, vol. 5549, Springer, p. 187-190, 2009.

Buckley C., Voorhees E., « Retrieval System Evaluation », *in* E. Voorhees, D. Harman (eds), *TREC : Experiment and Evaluation in Information Retrieval*, MIT Press, chapter 3, 2005.

Bullinaria J., Levy J., « Extracting semantic representations from word co-occurrence statistics : A computational study », *Behavior Research Methods*, vol. 39, n° 3, p. 510-526, 2007.

Caraballo S. A., « Automatic construction of a hypernym-labeled noun hierarchy from text », *ACL*, p. 120-126, 1999.

Chatterjee N., Mohan S., « Discovering Word Senses from Text Using Random Indexing », *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'08, Springer-Verlag, Berlin, Heidelberg, p. 299-310, 2008.

Cohen K. B., Demner-Fushman D., *Biomedical Natural Language Processing*, John Benjamins publishing company, 2013.

Curran J. R., From distributional to semantic similarity, PhD thesis, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh, 2004.

Daille B., Morin E., « French-English Terminology Extraction from Comparable Corpora », *Natural Language Processing - IJCNLP 2005, Second International Joint Conference, Jeju Island, Korea, October 11-13, 2005, Proceedings*, p. 707-718, 2005.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R., « Indexing by latent semantic analysis », *Journal of the American Society for Information Science*, vol. 41, n° 6, p. 391-407, 1990.

Déjean H., Gaussier E., Sadat F., « An approach based on multilingual thesauri and model combination for bilingual lexicon extraction », *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1-7, 2002.

Dupuch M., Dupuch L., Hamon T., Grabar N., « Semantic distance and terminology structuring methods for the detection of semantically close terms », *BioNLP : Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, Association for Computational Linguistics, Montréal, Canada, p. 20-28, June, 2012.

Fellbaum C. (ed.), *WordNet*, MIT Press, Cambridge, 1998.

Ferret O., « Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel », *TALN 2013*, Les Sables d'Olonne, France, p. 48-61, 2013.

Firth J., *A synopsis of linguistic theory 1930-1955*, Oxford : Blackwell, p. 1-32, 1957.

Généreux M., Hamon T., « Experiments in synonymy : term extraction and mapping to concepts », *Terminologie et Intelligence artificielle (TIA)*, Paris, 2013.

Gorman J., Curran J. R., « Scaling distributional similarity to large corpora », *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 361-368, 2006.

Grabar N., Zweigenbaum P., « Lexically-Based Terminology Structuring », *Terminology*, vol. 10, p. 23-54, 2003.

Grefenstette G., « Corpus-Derived First, Second and Third-Order Word Affinities », *Sixth Euralex International Congress*, p. 279-290, 1994.

Hamon T., Nazarenko A., « Detection of synonymy links between terms : experiment and results », *Recent Advances in Computational Terminology*, John Benjamins, p. 185-208, 2001.

Hamon T., Nazarenko A., « Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience », *TAL*, vol. 49, nº 2, p. 127-154, 2008.

Harris Z., « Distributional structure », *Word*, vol. 10, nº 23, p. 146-162, 1954.

Hearst M. A., « Automatic acquisition of hyponyms from large text corpora », *International Conference on Computational Linguistics*, Nantes, France, p. 539-545, 1992.

Jacquemin C., *Spotting and discovering terms through natural language processing*, The MIT Press, 2001.

Landauer T., Dumais S., « A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. », *Psychological Review ; Psychological Review*, vol. 104, nº 2, p. 211, 1997.

Lund K., Burgess C., « Producing high-dimensional semantic spaces from lexical co-occurrence », *Behavior Research Methods, Instrumentation, and Computers*, vol. 28, p. 203-208, 1996.

McCray A. T., Browne A. C., Bodenreider O., « The Lexical Properties of the Gene Ontology (GO) », *Proceedings of the AMIA 2002 Annual Symposium*, p. 504-508, 2002.

Mikolov T., Yih W., Zweig G., « Linguistic Regularities in Continuous Space Word Representations », *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Association for Computational Linguistics, Atlanta, Georgia, p. 746-751, June, 2013.

Morin E., Hazem A., « Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, United States, p. 1284-1293, June, 2014.

Morin E., Jacquemin C., « Automatic Acquisition and Expansion of Hypernym Links », *Computers and the Humanities*, vol. 38, nº 4, p. 363-396, 2004.

Morlane-Hondère F., Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique, PhD thesis, Université de Toulouse, 2013.

Morris J., Hirst G., « Non-classical lexical semantic relations », *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS 04, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 46-51, 2004.

Nastase V., Nakov P., Séaghdha D. O., Szpakowicz S., *Semantic Relations Between Nominals*, Morgan and Claypool Publishers, 2013.

Périnet A., Hamon T., « Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité », *Journées d'Analyse des Données Textuelles 2014*, Paris, France, p. 507-518, 2014.

Rapp R., « Word sense discovery based on sense descriptor dissimilarity », *MT Summit'2003*, p. 315-322, 2003.

Sahlgren M., The Word-Space Model : Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces, PhD thesis, Stockholm University, Stockholm, Sweden, 2006.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *New Methods in Language Processing*, Manchester, UK, p. 44-49, 1994.

Sebastiani F., « Machine learning in automated text categorization », *ACM Computing Surveys*, vol. 34, n° 1, p. 1-47, 2002.

Sun W., Rumshisky A., Uzuner Ö., « Evaluating temporal relations in clinical text : 2012 i2b2 Challenge », *JAMIA*, vol. 20, n° 5, p. 806-813, 2013.

Tanimoto T., An element mathematical theory of classification, Technical report, I.B.M. Research, New York, NY, USA, 1958.

Tsatsaronis G., Panagiotopoulou V., « A generalized vector space model for text retrieval based on semantic relatedness », *EACL 2009*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 70-78, 2009.

Turney P. D., Pantel P., « From Frequency to Meaning : Vector Space Models of Semantics », *Journal of artificial intelligence research*, vol. 37, p. 141-188, 2010.

van der Plas L., Automatic lexico-semantic acquisition for question answering, Thèse de doctorat, University of Groningen, Groningen, 2008.

Weeds J., Weir D., « Co-occurrence Retrieval : A Flexible Framework for Lexical Distributional Similarity », *Computational Linguistics*, vol. 31, n° 4, p. 439-475, 2005.

Weeds J., Weir D., McCarthy D., « Characterising measures of lexical distributional similarity », *Proceedings of COLING'2004*, Stroudsburg, PA, USA, p. 1015-1022, 2004.

Yuret D., « FASTSUBS : An Efficient and Exact Procedure for Finding the Most Likely Lexical Substitutes Based on an N-Gram Language Model », *IEEE Signal Process. Lett.*, vol. 19, n° 11, p. 725-728, 2012.

Zhitomirsky-Geffet M., Dagan I., « Bootstrapping distributional feature vector quality », *Computational Linguistics*, vol. 35, n° 3, p. 435-461, 2009.

Zweigenbaum P., « Menelas : an access system for medical records using natural language », *Computer Methods and Programs in Biomedicine*, 1994.

Zweigenbaum P., Habert B., « Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. », *Revue Glottopol*, vol. 8, p. 22-44, 2006.

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

Ludovic Tanguy, Franck Sajous et Nabil Hathout

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

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

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

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

KEYWORDS: *distributional semantics, specialised corpus.*

1. Introduction

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

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

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

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

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

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

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

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

2. Démarche et travaux connexes

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

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

3. Données

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

3.1. *Choix des mots cibles*

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

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

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

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

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

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

3.2. *Construction du* gold standard

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

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

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

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

4. Modèles distributionnels

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

4.1. *Types de contexte*

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

4.1.1. *Contextes graphiques*

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

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

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

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

4.1.2. *Dépendances brutes*

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

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

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

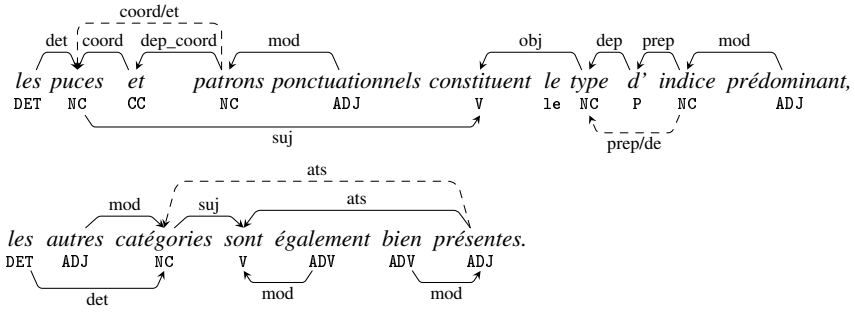

Figure 1 – Dépendances syntaxiques fournies par l'analyseur

4.1.3. *Contextes syntaxiques*

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

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

Tableau 1 – Dépendances brutes et triplets correspondants

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

Tableau 2 – Contextes syntaxiques et triplets correspondants

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

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

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

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

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

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

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

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

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

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

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

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

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

Tableau 3 – Configurations sélectionnées pour les contextes syntaxiques

4.2. *Calcul de similarité*

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

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

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

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

– dépendances brutes et contextes syntaxiques : 2 à 10

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

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

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

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

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

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

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

Tableau 4 – Répartition des 2 592 configurations étudiées

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

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

5. Analyse des résultats

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

5.1. *Méthode*

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

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

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

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

5.2. *Vue d'ensemble*

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

5.2.1. *Meilleures configurations*

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

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

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

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

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

5.2.2. *Variation suivant les catégories des mots cibles*

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

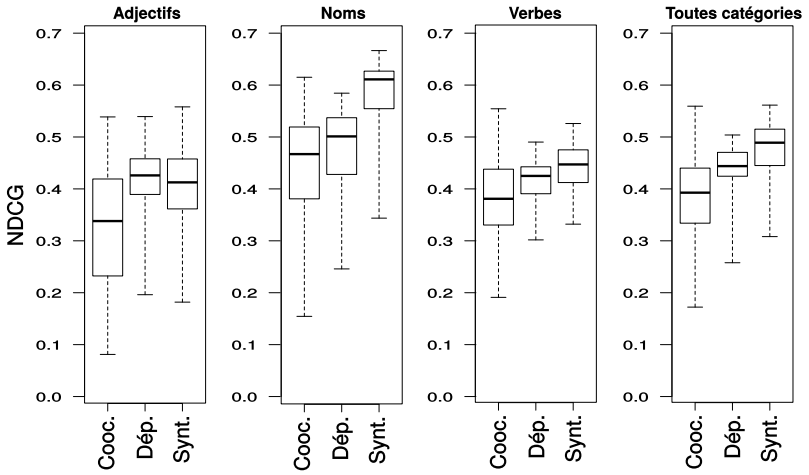

Figure 2 – Variation du NDCG en fonction des catégories et des types de contexte

5.2.3. *Variation suivant les mots cibles*

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

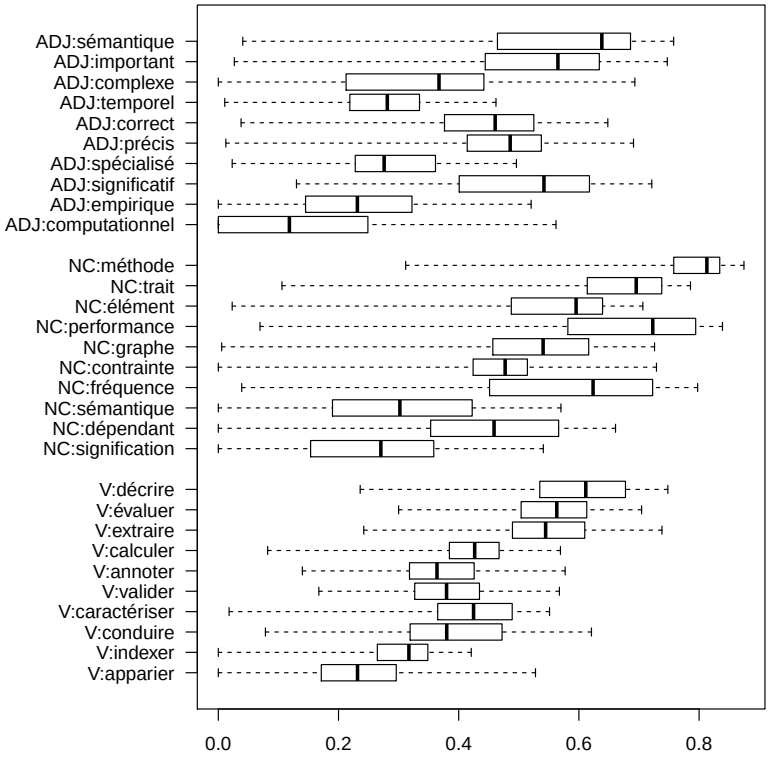

Figure 3 – Variation du NDCG moyen en fonction des mots cibles

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

5.3. *Étude des paramètres*

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

5.3.1. *Analyse globale*

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

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

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

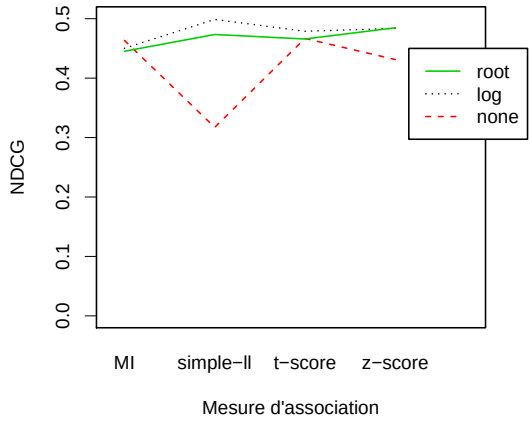

Figure 4 – Variation du NDCG moyen en fonction de la mesure d'association et de la transformation appliquée

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

5.3.2. *Impact du paramétrage des contextes graphiques*

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

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

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

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

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

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

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

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

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

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

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

5.3.3. *Impact du paramétrage des dépendances brutes*

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

5.3.4. *Impact du paramétrage des contextes syntaxiques*

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

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

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

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

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

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

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

5.3.5. *Impact des différentes relations syntaxiques*

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

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

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

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

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

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

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

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

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

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

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

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

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

– pour les noms, modifieur de nom et préposition ;

– pour les verbes, objet direct et préposition.

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

5.4. *Bilan des observations*

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

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

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

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

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

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

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

5.5. *Analyses qualitatives*

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

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

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

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

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

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

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

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

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

6. Conclusion

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

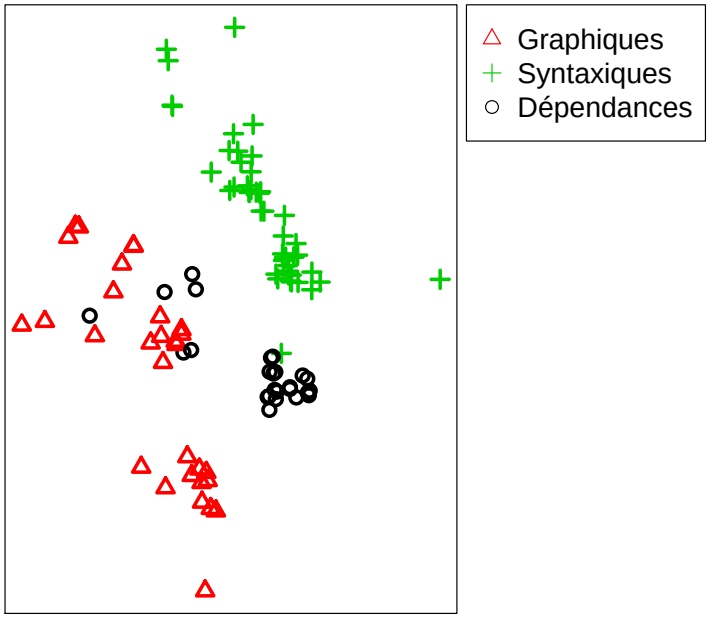

Figure 5 – Projection de Sammon des 106 meilleurs modèles sur la base de leur similarité RBO

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

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

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

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

Remerciements

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

7. Bibliographie

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Habert B., Zweigenbaum P., « Contextual Aquisition of Information Categories. What has been done and what can be done automatically? », *in* B. Nevin, S. Johnson (eds), *The legacy of Zellig Harris. Language and Information into the 21st century*, vol. 2: computability of language and computer applications, John Benjamins Publishing Company, Amsterdam / Philadelphia, chapter 8, p. 203-231, 2002.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note de lecture

Rubrique préparée par Denis Maurel

Université François Rabelais Tours, LI (Laboratoire d'informatique)

Florent PEREK. Argument Structure in Usage-Based Construction Grammar. *John Benjamins Publishing Company*. 2015. 246 pages. ISBN 978-90-272-0439-4.

Lu par **Antonio BALVET**

Université de Lille, CNRS, UMR 8163 – STL – Savoirs Textes Langage, F-59000 Lille, France.

La structure argumentale des verbes, c'est-à-dire la spécification des participants à une situation médiée par une forme grammaticale, se trouve de fait à l'interface entre syntaxe et sémantique. L'ouvrage de F. Perek aborde la structure argumentale en anglais du point de vue des approches guidées par l'usage (« usage-based *») et des approches constructionnelles (*« construction grammar *»). Il passe en revue et évalue la pertinence de plusieurs conceptions de la structure argumentale, des approches lexicalistes projectionnistes classiques aux approches plus cognitivistes. Il examine, en production comme en compréhension, les effets de la fréquence d'usage des unités lexicales prédicatives et de leurs éventuels arguments ou adjoints. Il prend position en faveur d'une approche fondée sur l'usage de la structure argumentale, tout en mettant en évidence un réseau de relations entre constructions.*

L'ouvrage de Florent Perek, *Argument Structure in Usage-Based Construction Grammar*, présenté dans la collection « Constructional Approaches to Language », prend sa source dans la thèse que l'auteur a défendue en 2012 à l'université de Freiburg, intitulée *Verbs, Constructions, Alternations: Usage-based perspectives on argument realization*. Il s'agit donc d'une version révisée, augmentée et en partie restructurée de la thèse, centrée sur une problématique cognitiviste de la question de la prédication et de la réalisation de la structure argumentale. La description et la formalisation de la structure argumentale intéressent au premier chef aussi bien la linguistique fondamentale et théorique que les domaines les plus appliqués (didactique des langues, lexicologie et lexicographie, TAL), ou encore ceux à l'intersection avec les sciences cognitives. Avec cet ouvrage, l'auteur propose des pistes méthodologiques et théoriques allant à rebours des pratiques descriptives usuelles, notamment dans la constitution de ressources lexicales électroniques.

L'ouvrage est structuré en huit chapitres et trois parties qui traitent de façon méthodique de la problématique choisie : les différentes approches descriptives et formelles de la réalisation de la structure argumentale, en l'occurrence des verbes anglais. L'auteur passe en revue et évalue la pertinence du traitement proposé par les

approches « classiques », à savoir les approches lexicalistes et projectionnistes. Par approches lexicalistes et projectionnistes, l'auteur désigne toutes les approches dans lesquelles une dichotomie est posée entre le niveau lexical et le niveau grammatical, suivant en cela l'orientation et la répartition des tâches que l'on retrouve en particulier dans les approches générativistes des faits linguistiques, mais également dans la plupart des approches en syntaxe formelle et en TAL. Les approches lexicalistes et projectionnistes posent que l'essentiel des propriétés syntaxiques et sémantiques doit être spécifié dans le lexique. Les contraintes de réalisation de la structure argumentale sont ainsi vues comme projetées du niveau lexical vers le niveau syntaxique, puis vers le niveau sémantique, *via* des règles de liage (« *linking rules* ») notamment. F. Perek montre que la prise en compte de phénomènes de Performance (créativité dans l'emploi des verbes), et plus particulièrement des données acquisitionnelles[1] et expérimentales met à mal les propositions des approches lexicalistes et projectionnistes.

L'auteur oppose à ces approches les propositions faites par les tenants d'une linguistique d'inspiration cognitiviste, et plus particulièrement par les approches guidées par l'usage (« *usage based* »). Pour lui, ces approches proposent une conception plus harmonieuse de la structure argumentale, de la souplesse observée dans les réalisations et de la créativité syntaxique, d'une part. D'autre part, il montre de façon convaincante comment la structure syntaxique qui sert de support à la réalisation des actants sémantiques, autrement dit les « constructions », a des effets tant en production qu'en compréhension. Enfin, l'auteur affirme qu'une approche cognitiviste, ancrée dans la notion de construction, doit également intégrer les paramètres liés à l'usage des unités prédicatives, du point de vue des locuteurs. Il montre ainsi comment les représentations mentales de la valence des verbes chez les locuteurs intègrent des éléments de fréquence, sans lesquels il devient difficile de rendre compte de certaines réalisations observées en corpus ou provoquées dans un cadre expérimental.

L'ouvrage est structuré de façon très « académique », progressant de façon méthodique de la problématique de départ pour arriver à un ensemble de conclusions et de perspectives méthodologiques. Soulignons ici la qualité d'écriture de l'ouvrage, au service d'une argumentation qui cherche systématiquement à étayer ses propositions grâce à des données extraites de corpus et des expérimentations, notamment concernant les biais observés dans la valence de certains verbes, l'influence des constructions (en fait des « collostructions ») sur les réalisations syntaxiques, ainsi que le potentiel d'alternances (« *alternations* » entre constructions, qui l'amène à poser l'existence chez les locuteurs d'un réseau de relations entre constructions : des « allostructions ». La maîtrise des alternances

1 Surgénéralisations produites par des apprenants et des enfants à partir des exemples auxquels ils ont été exposés, emplois « fautifs » de verbes intransitifs dans une structure transitive...

entre constructions est posée comme une compétence fondamentale dans la maîtrise de la langue.

Bien que, de par sa formation, l'auteur soit informé des enjeux pour le TAL de la problématique traitée dans l'ouvrage, le lecteur trouvera peu de propositions concrètes pour l'implémentation, par exemple, d'une analyse syntaxique ou sémantique automatique guidée par l'usage dans un cadre cognitiviste. D'autre part, l'ouvrage dans son ensemble adopte un point de vue centré sur celui du locuteur et non sur celui du linguiste ou du lexicographe. L'auteur cible en effet en priorité les aspects méthodologiques et théoriques liés à la prise en compte de phénomènes d'usage, de l'influence des constructions et de leur potentiel d'alternances dans l'abord de la structure argumentale. Il prend ainsi clairement position en faveur d'un continuum entre arguments et adjoints et sur la nécessité d'intégrer les effets de fréquence d'usage dans les descriptions et modélisations de la structure argumentale. Il prend donc position contre la méthode introspective sur laquelle reposent pourtant nombre de propositions théoriques en syntaxe ou en sémantique, mais également contre des notions aussi apparemment établies que celle de la polysémie, en les déplaçant du niveau des entrées lexicales individuelles à celui, plus général, des constructions avec lesquelles elles sont susceptibles d'être associées. L'auteur porte un regard critique sur le traitement du sens dans les approches constructionnelles : il montre que la charge sémantique des constructions n'est pas réductible aux unités lexicales particulières qu'elles convoquent. Il montre également que la recherche d'un sens le plus générique possible pour une construction ne rend pas fidèlement compte des observations. Plutôt que des classes unifiées de constructions (par exemple, les « constructions conatives »), l'ouvrage propose une conception du lexique mental faite d'associations forme syntaxique-charge sémantique cohérentes à un niveau local mais pas nécessairement au niveau global. L'auteur va ainsi à contre-courant de la position dominante en grammaire constructionnelle, en faveur d'une orientation plus « radicale ».

Les propositions et résultats présentés incitent donc le lecteur à regarder d'un œil nouveau nombre de modèles établis, mais également la plupart des réalisations pratiques utilisées en TAL : classement des verbes de Beth Levin (et donc nomenclature des verbes dans le Princeton Wordnet), ressources lexicales électroniques (de type Verbnet) et grammaires formelles, en particulier. Bien que l'ouvrage ne cherche pas à créer la polémique sur une notion aussi centrale en linguistique, gageons qu'il suscitera nécessairement le débat, et à tout le moins invitera tant les linguistes « théoriques », les psycholinguistes, que les praticiens du TAL à une nouvelle réflexion méthodologique sur la question du rapport entre Compétence et Performance, norme et usage dans la description et la formalisation des faits en syntaxe et en sémantique.

Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

INRIA, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
sylvain.pogodalla@inria.fr

Anaïs LEFEUVRE-HALFTERMEYER : anais.lefeuvre@live.fr

Titre : Sémantique des temps du français : une formalisation compositionnelle

Mots-clés : Sémantique formelle, temporalité des éventualités, lexique compositionnel, récits de voyage, lambda-calcul simplement typé.

Title: *French Tenses Semantics: a Compositional Formalization*

Keywords: *Formal semantics, eventualities temporality, compositional lexicon, travel novels, simply typed lambda-calculus.*

Thèse de doctorat en Informatique, LaBRI/INRIA, unité de formation d'informatique, Université de Bordeaux, sous la direction de Christian Retoré (Pr, Université de Bordeaux) et Mauro Gaio (Pr, Université de Pau et des Pays de l'Adour). Thèse soutenue le 23/06/2014.

Jury : M. Christian Retoré (Pr, Université de Bordeaux, codirecteur), M. Mauro Gaio (Pr, Université de Pau et des Pays de l'Adour, codirecteur), M. Henk Verkuyl (Pr émérite, Université d'Utrecht, Pays-Bas, rapporteur), M. Patrice Enjalbert (Pr émérite, Université Paris Ouest Nanterre La Défense – Paris 10, rapporteur), Mme Delphine Battistelli (Pr, Université Paris Ouest Nanterre La Défense – Paris 10, examinatrice), M. Richard Moot (CR, CNRS, LaBRI, Bordeaux, examinateur), M. Jean-Yves Antoine (Pr, Université François Rabelais, Tours, examinateur), M. Nicolas Hannusse (DR, CNRS, LaBRI, Bordeaux, président).

Résumé : *Cette thèse s'inscrit dans le cadre du projet Région Aquitaine–INRIA : ITIPY dont le but est à terme l'extraction automatique d'itinéraires à partir de récits de voyage du XIXème et du début du XXème siècle.*

Notre premier travail fut de caractériser le corpus comme échantillon du français, par une étude contrastive d'une part de données quantitatives, et d'autre part de la structure des récits de voyage. Cette étude montrant que les segments textuels portant la narration de l'itinéraire sont difficiles à isoler du reste du récit, nous avons adopté une approche centrée sur le verbe comme support privilégié à l'expression du déplacement ou de la localisation.

Nous nous sommes consacrée à l'étude de la composante temporelle de la sémantique de ces verbes, et plus particulièrement à l'analyse automatique des temps verbaux du français. Disposant d'un analyseur syntaxique et sémantique à large échelle du français, Grail, basé sur les grammaires catégorielles et la sémantique compositionnelle en λ-DRT, notre tâche a été de prendre en compte les temps des verbes pour reconstituer la temporalité des événements et des états, notions regroupées sous le terme d'éventualité.

Cette thèse se concentre sur la construction d'un lexique sémantique traitant des temps verbaux du français. Nous proposons une extension et une adaptation d'un système d'opérateurs compositionnels conçu pour les temps du verbe néerlandais et anglais, aux temps et à l'aspect du verbe français du XIXème siècle à nos jours. Pour cela, nous nous appuyons sur une étude sémantique des temps et aspect du français d'un point de vue diachronique. Nous proposons une modélisation en terme d'intervalles de cette nouvelle version du système et nous proposons les entrées du lexique sémantique pour quelques adverbiaux de temps, eux aussi adaptés dans le cadre de ce système.

Cette formalisation est de facto opérationnelle, car elle est définie en terme d'opérateurs du λ-calcul dont la composition et la réduction, déjà programmées, calculent automatiquement les représentations sémantiques souhaitées : des formules multisortes de la logique d'ordre supérieur.

Le passage de l'énoncé comportant une éventualité seule au discours, dont le maillage référentiel est complexe, est discuté, et nous concluons par les perspectives qu'ouvrent nos travaux pour l'analyse du discours.

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01136420

Elisa OMODEI : elisa.omodei@ens.fr

Titre : Modélisation des dynamiques socio-sémantiques dans les communautés scientifiques

Mots-clés : Linguistique computationelle, modélisation statistique, réseaux sémantiques, extraction lexicale, dynamiques socio-sémantiques, réseaux de collaboration.

Title: *Modeling the Socio-semantic Dynamics of Scientific Communities*

Keywords: *Computational linguistics, statistical modeling, semantic networks, automatic term extraction, socio-semantic dynamics, co-authorship networks.*

Thèse de doctorat en Mathématiques appliquées aux sciences sociales, école doctoral transdisciplinaire lettres/sciences, LaTTiCe (UMR 8094), Institut des Systèmes Complexes de Paris Île-de-France (ISC-PIF), École Normale Supérieure, Paris, sous la direction de Thierry Poibeau (DR, CNRS, LaTTiCe) et Jean-Philippe Cointet (IR, INRA). Thèse soutenue le 19/12/2014.

Jury : M. Thierry Poibeau (DR, CNRS, LaTTiCe, codirecteur), M. Jean-Philippe Cointet (IR, INRA, codirecteur), M. Jean-Pierre Nadal (DR, CNRS, Laboratoire de Physique Statistique — UMR 8550 — et Centre d'Analyse et de Mathématique Sociales — UMR 8557, président), Mme Clémence Magnien (DR, CNRS, LIP6, rapporteur), M. Roger Guimera (Adjunct Professor, Rovira i Virgili University, Tarragona, Espagne, rapporteur), M. Emmanuel Lazega (Pr, Sciences Po, Paris, examinateur).

Résumé : *Comment les structures sociales et sémantiques d'une communauté scientifique guident-elles les dynamiques de collaboration à venir ? Dans cette thèse, nous combinons des techniques de traitement automatique des langues et des méthodes provenant de l'analyse de réseaux complexes pour analyser une base de données de publications scientifiques dans le domaine de la linguistique computationnelle : l'*ACL Anthology. *Notre objectif est de comprendre le rôle des collaborations entre les chercheurs dans la construction du paysage sémantique du domaine, et, symétriquement, de saisir combien ce même paysage influence les trajectoires individuelles des chercheurs et leurs interactions. Nous employons des outils d'analyse du contenu textuel pour extraire des textes des publications les termes correspondant à des concepts scientifiques. Ces termes sont ensuite connectés aux chercheurs pour former un réseau socio-sémantique, dont nous modélisons la dynamique à différentes échelles. Nous construisons d'abord un modèle statistique, à base de régressions logistiques multivariées, qui permet de quantifier le rôle respectif des propriétés sociales et sémantiques de la communauté sur la dynamique microscopique du réseau socio-sémantique. Nous reconstruisons par la suite l'évolution du champ de la linguistique computationelle en créant différentes cartographies du réseau sémantique, représentant les connaissances produites dans le domaine, mais aussi le flux d'auteurs entre les différents champs de recherche du domaine. En résumé, nos travaux ont montré que la combinaison des méthodes issues du traitement automatique des langues et de l'analyse des réseaux complexes permet d'étudier d'une manière nouvelle l'évolution des domaines scientifiques.*

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01097702

Amandine PÉRINET : amandine.perinet@yahoo.fr

Titre : Analyse distributionnelle appliquée aux textes de spécialité : réduction de la dispersion des données par abstraction des contextes

Mots-clés : Traitement automatique des langues, textes de spécialité, terminologie, analyse distributionnelle, modèle vectoriel, groupements sémantiques, termes complexes, relations sémantiques, abstraction de contextes.

Title: *Distributional Analysis Applied to Specialized Corpora: Reduction of Data Sparsity through Context Abstraction*

Keywords: *Natural language processing, specialised corpora, terminology, distributional analysis, vector space model, semantic cluster, complex terms, semantic relations, context abstraction.*

Thèse de doctorat en Informatique, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS-INSERM), Université Paris Nord – Paris 13, sous la direction de Sylvie Després (Pr, Université Paris Nord – Paris 13, LIMICS-INSERM) et Thierry Hamon (MC, Université Paris Nord – Paris 13, LIMSI). Thèse soutenue le 17/03/2015.

Jury : Mme Sylvie Després (Pr, Université Paris Nord – Paris 13, LIMICS-INSERM, codirectrice), M. Thierry Hamon (MC, Université Paris Nord – Paris 13, LIMSI, codirecteur), Mme Cécile Fabre (Pr, Université de Toulouse II, rapporteur), M. Emmanuel Morin (Pr, Université de Nantes, rapporteur), M. Thierry Charnois (Pr, Université Paris Nord – Paris 13, président), M. Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay, examinateur), M. Olivier Ferret (CR, CEA LIST, examinateur).

Résumé : *Dans les domaines de spécialité, les applications telles que la recherche d'information ou la traduction automatique s'appuient sur des ressources terminologiques pour prendre en compte les termes, les relations sémantiques ou les regroupements de termes. Pour faire face au coût de la constitution de ces ressources, des méthodes automatiques ont été proposées. Parmi celles-ci, l'analyse distributionnelle s'appuie sur la redondance d'informations se trouvant dans le contexte des termes pour établir une relation. Alors que cette hypothèse est habituellement mise en œuvre grâce à des modèles vectoriels, ceux-ci souffrent du nombre de dimensions considérable et de la dispersion des données dans la matrice des vecteurs de contexte.*

En corpus de spécialité, ces informations contextuelles redondantes sont d'autant plus dispersées et plus rares que les corpus ont des tailles beaucoup plus petites.

De même, les termes complexes sont généralement ignorés étant donné leur faible nombre d'occurrences. Dans cette thèse, nous nous intéressons au problème de la limitation de la dispersion des données sur des corpus de spécialité et nous proposons une méthode permettant de densifier la matrice des contextes en réalisant une abstraction des contextes distributionnels. Des relations sémantiques acquises en corpus sont utilisées pour généraliser et normaliser ces contextes. Nous avons évalué la robus-

tesse de notre méthode sur quatre corpus de tailles, de langues et de domaines différents. L'analyse des résultats montre que, tout en permettant de prendre en compte les termes complexes dans l'analyse distributionnelle, l'abstraction des contextes distributionnels permet d'obtenir des groupements sémantiques de meilleure qualité mais aussi plus cohérents et homogènes.

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01202371

Simon PETITJEAN : simon.petitjean@hhu.de

Titre : Génération modulaire de grammaires formelles

Mots-clés : Langages dédiés, modularité.

Title: *Modular Generation of Formal Grammars*

Keywords: *Domain specific languages, modularity.*

Thèse de doctorat en Informatique, Laboratoire d'Informatique Fondamentale d'Orléans (LIFO), UFR Sciences, Université d'Orléans, sous la direction de Denys Duchier (Pr, Université d'Orléans, LIFO). Thèse soutenue le 11/12/2014.

Jury : M. Denys Duchier (Pr, Université d'Orléans, LIFO, directeur), Mme Claire Gardent (DR, CNRS, LORIA, Nancy, présidente), Mme Laura Kallmeyer (Pr, Université de Düsseldorf, Allemagne, rapporteur), M. Kim Mens (Pr, Université catholique de Louvain, Belgique, rapporteur), M. Olivier Bonami (MC, Université Paris-Sorbonne, examinateur), M. Éric Villemonte de la Clergerie (CR, INRIA Paris – Rocquencourt, examinateur), Mme Ann Copestake (Pr, Université de Cambridge, Royaume-Uni, examinatrice), M. Yannick Parmentier (MC, Université d'Orléans, LIFO, examinateur).

Résumé : *Les travaux présentés dans cette thèse visent à faciliter le développement de ressources pour le traitement automatique des langues. Les ressources de ce type prennent des formes très diverses, en raison de l'existence de différents niveaux d'étude de la langue (syntaxe, morphologie, sémantique, ...) et de différents formalismes proposés pour la description des langues à chacun de ces niveaux. Les formalismes faisant intervenir différents types de structures, un unique langage de description n'est pas suffisant : il est nécessaire pour chaque formalisme de créer un langage dédié (ou DSL), et d'implémenter un nouvel outil utilisant ce langage, ce qui est une tâche longue et complexe.*

Pour cette raison, nous proposons dans cette thèse une méthode pour assembler modulairement et adapter des cadres de développement spécifiques à des tâches de génération de ressources langagières. Les cadres de développement créés sont construits autour des concepts fondamentaux de l'approche XMG (eXtensible MetaGrammar), à savoir disposer d'un langage de description permettant la définition modulaire d'abs-

tractions sur des structures linguistiques, ainsi que leur combinaison non-déterministe (c'est-à-dire au moyen des opérateurs logiques de conjonction et disjonction). La méthode se base sur l'assemblage d'un langage de description à partir de briques réutilisables, et d'après un fichier unique de spécification. L'intégralité de la chaîne de traitement pour le DSL ainsi défini est assemblée automatiquement d'après cette même spécification.

Nous avons dans un premier temps validé cette approche en recréant l'outil XMG à partir de briques élémentaires. Des collaborations avec des linguistes nous ont également amené à assembler des compilateurs permettant la description de la morphologie de l'Ikota (langue bantoue) et de la sémantique (au moyen de la théorie des frames*).*

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01202647

Marc SPANIOL : marc.spaniol@unicaen.fr

Titre : Un cadre pour l'analyse temporelle d'Internet

Mots-clés : Analyse temporelle d'Internet, entités nommées, l'analyse de l'Internet sur le niveau entité.

Title: *A Framework for Temporal Web Analytics*

Keywords: *Temporal Web analytics, named entities, entity-level Web analytics.*

Habilitation à diriger des recherches en Informatique, GREYC (UMR 6072), UFR Sciences, Université de Caen Basse-Normandie, sous la direction de Gaël Dias (Pr, Université de Caen Basse-Normandie, GREYC). Habilitation soutenue le 09/12/2014.

Jury : M. Gaël Dias (Pr, Université de Caen Basse-Normandie, GREYC, directeur), M. Patrice Bellot (Pr, Aix Marseille Université, LSIS, rapporteur, président), M. Éric Gaussier (Pr, Université Joseph Fourier, LIG/AMA, Grenoble, rapporteur), M. Mathieu Roche (Chercheur HDR, Cirad, TETIS, Montpellier, rapporteur), M. Mohand Boughanem (Pr, Université Paul Sabatier, IRIT, Toulouse, examinateur), M. Aldo Gangemi (Pr, Université Paris Nord – Paris 13, LIPN, examinateur).

Résumé : *Web-preservation organizations like the Internet Archive not only capture the history of born-digital content but also reflect the* zeitgeist *of different time periods over more than a decade. This longitudinal data is a potential gold mine for researchers like sociologists, politologists, media and market analysts, or experts on intellectual property.*

Longitudinal data analytics—the Web of the Past—poses research challenges, but has not received due attention. The sheer size and content of Web archives render them relevant to analysts within a range of domains. The Internet Archive holds more than 350 billion versions of Web pages, captured since 1996. This coverage can no longer

be maintained, as Web content is growing at enormous rates. A high-coverage archive would have to be an order of magnitude larger.

A Web archive of timestamped versions of Web sites over a long-term time horizon opens up great opportunities for analysts. However, difficulties arise from name ambiguities, requiring a disambiguation mapping of mentions (noun phrases in the text) onto entities. For example, "Bill Clinton" might be the former US president William Jefferson Clinton, or any other William Clinton contained in Wikipedia. Ambiguity further increases if the text only contains "Clinton" or a phrase like "the US president". The temporal dimension introduces additional complexity, for example when names of entities have changed over time (e.g. people getting married or divorced, or organizations that undergo restructuring in their identities). By mapping names and phrases onto canonicalized entities, we raise the entire analytics to a semantic rather than keyword-level in order to make sense of the raw and often noisy Web contents.

URL où le mémoire pourra être téléchargé :
https://spaniol.users.greyc.fr/HDR(Spaniol).pdf

Liste des auteurs