Extrinsic Evaluation of French Dependency Parsers on a Specialized Corpus: Comparison of Distributional Thesauri

Ludovic Tanguy, Pauline Brunet, Olivier Ferret

Extrinsic Evaluation of French Dependency Parsers on a Specialized Corpus: Comparison of Distributional Thesauri

Ludovic Tanguy[1], **Pauline Brunet**[2], **Olivier Ferret**[2]
1. CLLE-ERSS: CNRS & Université de Toulouse, France
2. CEA, LIST, F-91191 Gif-sur-Yvette, France
ludovic.tanguy@univ-tlse2.fr
{pauline.brunet,olivier.ferret}@cea.fr

Abstract

We present a study in which we compare 11 different French dependency parsers on a specialized corpus (consisting of research articles on NLP from the proceedings of the TALN conference). Due to the lack of a suitable gold standard, we use each of the parsers' output to generate distributional thesauri using a frequency-based method. We compare these 11 thesauri to assess the impact of choosing a parser over another. We show that, without any reference data, we can still identify relevant subsets among the different parsers. We also show that the similarity we identify between parsers is confirmed on a restricted distributional benchmark.
Keywords: French dependency parsing, distributional semantics, specialized corpus

1. Introduction

This article takes place in the context of a more global research about distributional semantic analysis methods for specialized domains in French. Our medium-term objective is to select the most efficient method for identifying semantic similarities between the lexical and terminological units of a small specialized corpus (a few million words at most). This task is conditioned by several parameters that must be chosen carefully along the way from the extraction of word contexts to the construction of a distributional model. Unfortunately, these choices cannot be made on the sole basis of the knowledge acquired on the large and generic corpora that are commonly used to compare distributional models. Although this is no more a commonly used configuration, we selected a method relying on syntactic contexts and based on the assumption that the small amount of data can be compensated to some extent by a rich linguistic annotation. This has been partially demonstrated in (Tanguy et al., 2015). Our focus here is how the choice of the parser impacts the construction of a distributional model using a frequency-based method. This is related to the wider issue of comparing the efficiency of different available tools and models for dependency parsing.

Many efforts have been made in the French Natural Language Processing (NLP) community to compare different parsers: the Easy (Paroubek et al., 2008), Passage (De La Clergerie et al., 2008), SPMRL (Seddah et al., 2013) and CoNLL (Zeman et al., 2018) shared tasks as well as more focused studies like (Candito et al., 2010) or (De La Clergerie, 2014). However, the benchmarks used in these tasks and studies, adopting the kind of diverse, generic corpora on which the tools have been trained, might not be the most relevant option for parsing specialized corpora. Moreover, even though some of these campaigns are recent, all the main tools currently available have not been evaluated on the same benchmarks.

In the absence of a benchmark fitting our task and domain, we compare the current main French parsers on an external task, as it was done in the EPE campaign for English (Fares et al., 2018). Given that our only benchmark for this task has limited coverage and validity, our evaluation is mainly a qualitative comparison. In this paper, we mainly examine how changing the parser affects the distributional thesauri that we produce from the parsers' outputs, with a focus on the scale and scope of these differences.

We have used the following procedure: we first applied 11 dependency parsers on the same French specialized corpus. For each of the outputs, we extracted the main syntactic contexts, which we used to build 11 distributional models with a classic frequency-based method. We then compare the distributional thesauri obtained to identify the actual impact of the parser used on the result. We also propose a mapping of the parsers based on the similarity of the models that were generated using them. In the end, we do not obtain a ranking of the parsers, but only a clustering that shows which analyzers lead to similar results, and we identify some of the syntactic and semantic phenomena that are the most concerned by the differences.

In Section 2, we present the corpus and the 11 parsers we compare. In Section 3, we present the normalization and selection steps that we apply to their outputs to obtain a common base for comparison. Section 4 deals with the construction process of the distributional models and their comparison following several different approaches.

2. Material and Studied Tools

2.1. The TALN Corpus

For this experiment, we used a small-sized specialized corpus: the TALN corpus[1], made up of the archives of the *TALN* and *RECITAL* NLP conference proceedings from 1999 to 2014, for a total size of 4.5 million words. This corpus holds several advantages for the study of distributional models: it is made of high-quality written texts, it is homogeneous in genre and topic, and we, as researchers, have a level of expertise in the domain sufficient to easily interpret the results of distributional analysis.

[1] Available at `http://redac.univ-tlse2.fr/corpus/taln_en.html`

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 5822–5830
Marseille, 11–16 May 2020

The original PDF files of the articles were converted to raw text to extract parser-compatible data. For this reason, we filtered out some text elements, like footnotes, tables, math formulas, references, captions, or section headers. All line and paragraph breaks were removed and the whole text of each article was put on a single line. Thus, the robustness of the parsers when applied to noisy data was not a parameter we considered when comparing them.

2.2. Studied Parsers

We selected 7 tools able to perform dependency parsing in French, focusing on the parsers that are easily available and ready-to-use, i.e. those that take in charge the whole process, from raw text to dependencies. These tools were all used with their default options and parameters.

All these tools rely on machine learning models trained on manually annotated corpora. Their differences hinge on implementation choices like parsing techniques (graph- or transition-based, for instance), machine learning architectures (SVM, maximal entropy or more recently recurrent neural networks), and upstream processing (segmentation, lemmatization). However, there is much less choice when it comes to the corpora used for training, given the high cost of the annotation and validation processes. Before presenting the parsers we have selected, it is thus important to present the available French corpora on which these parsers were trained since they have a decisive impact on the nature and the form of the output.

FTB The first French corpus that has been syntactically annotated is the *Corpus arboré du français*, better known under its English name *French Treebank* or *FTB* (Abeillé et al., 2003). Made up of around 600,000 words from the newspaper *Le Monde*, it was first annotated for constituency parsing and later automatically converted to dependencies by (Candito et al., 2010). Another version was produced for the SPMRL evaluation campaign, mainly for identifying multi-word expressions (Seddah et al., 2013).

UD French To facilitate the development and comparison of parsers as well as large-scale crosslingual studies, a universal dependency scheme was proposed, based on the *Stanford Dependencies* model, now called *Universal Dependencies* (or UD)[2] (Nivre et al., 2016). The UD project proposes universal tagsets (for Part-Of-Speech (POS) tagging and syntactic dependencies) and was the framework for the collection and diffusion of several French annotated corpora under the same format, in particular:

- **UD French FTB** is the result of the conversion to UD of the original French Treebank and is made of around 550,000 words.
- **UD French ParTUT** is the French subset of the Parallel-TUT multilingual corpus (Bosco et al., 2012), which is composed of varied text samples (legal texts, Wikipedia articles, Facebook pages, etc.) for a total of approximately 30,000 words.
- **UD French GSD** is the original corpus for the French UD project initially annotated using the Stanford dependencies scheme (McDonald et al., 2013). It contains 400,000 words of text from newspaper articles, Wikipedia articles, blogs, or product reviews.
- **UD French Sequoia** was developed as a complement to the FTB with the main goal to improve its coverage in terms of genre and domain (Candito and Seddah, 2012). The 70,000 words in this corpus come from parliamentary debates, regional newspaper articles, Wikipedia and medical texts. Initially annotated following the FTB scheme, it has been converted to UD.

Despite the normalization efforts of the UD project, the previous corpora do not use the same tagsets and annotation conventions because of different choices made for dealing with some syntactic phenomenons, as well as conversions applied to some of them.

As mentioned previously, we selected 7 different parsers, some of which proposing several models, trained on different corpora.

CoreNLP (Manning et al., 2014), the main parser of the Stanford team, implements a maximal entropy tagger and a transition-based parser. It was trained on the UD GSD corpus.

StanfordNLP (Qi et al., 2018) is a tool that, on top of giving access to the CoreNLP chain in Python, implements an entirely different parsing algorithm. Its graph-based parser relies on a LSTM neural network. StanfordNLP proposes several French models; we used two of them trained on the UD **GSD** and **Sequoia** corpora.

NLPCube (Boroş et al., 2018) is, like StanfordNLP, based on a LSTM recurrent neural network. Its main particularity is that syntactic parsing occurs independently from the POS tagging, both of them using only lexicalized attributes and no morphological information. It must be noted that this tool is significantly slower than the others. We have not found any precise information on the corpora used to train the model: we assume that the sum of all available French UD corpora was used.

Spacy is an all-purpose NLP tool designed for industrial applications whose main characteristic is its high speed compared to most other parsers. The tagger is based on a perceptron, with attributes based on Brown clusters, following (Koo et al., 2008). It implements a non-monotonous transition-based parser that can revise previous decisions (Honnibal and Johnson, 2015). The available model was trained on the WikiNER corpus (Nothman et al., 2012) for Named Entities Recognition and on UD Sequoia for POS tagging and parsing.

UDPipe (Straka and Straková, 2017) uses a neural network with a Gated Recurrent Unit mechanism to perform

[2] See `http://universaldependencies.org/` for the detailed history and different versions.

both tokenization and sentence segmentation at once. For POS tagging, it generates the possible tags for a word according to its suffix and uses a perceptron to assign the correct one. The transition-based parsing relies on a simple one-layer neural network. UDPipe includes several French models: we used three of them, trained on the UD **GSD**, **Sequoia** and **ParTUT** corpora.

Talismane (Urieli and Tanguy, 2013) uses a mix of statistic models and hand-crafted language-specific features and rules incorporating linguistic knowledge. The main model has been trained on the French TreeBank converted to dependencies (**FTB**), but we also tested the newer Universal Dependency model (**UD**) trained on the concatenation of all UD corpora previously described.

MSTParser (McDonald et al., 2006) is a graph-based dependency parser. We used it through the BONSAI[3] package, which combines it with the MElt POS tagger (Denis and Sagot, 2009) and uses the best MST model according to the (Candito et al., 2010) benchmark. It is based on the non-UD version of the FTB.

We are fully aware that these parsers can only be compared on a practical level since the technologies they use, their goals, their training data, and even the times at which they were created are extremely different. Nevertheless, they form a large part of the current panel of available solutions for robust syntactic parsing in French, and in this respect, are suitable for consideration in this study.

3. Processing Parser Outputs

To compare the 11 selected parsers (or versions of parsers), we needed to identify or create a common ground between the produced parsing outputs. Their heterogeneous outputs raise several challenges: token identification, lemmatization, alignment of POS tags and syntactic dependency relations. We decided to limit our scope to single open-class words (nouns, verbs, adjectives, and adverbs), represented by their lemma and POS tag. As for dependency relations, we wanted to focus on units that are both present in all schemes and are the most useful for distributional semantic analysis.

3.1. Word Identification

Our study, like most current work on distributional semantics, is focused on single word units and, therefore, presumes that tokenization is homogeneous across parsers.

Lemmatization is of crucial importance for French and especially when processing small corpora since it limits the dispersion of word types because of inflection and thus, helps to compensate for the lack of data. Lemmatization also enables linking (before or after distributional analysis) with lexical or terminological resources, which always have their entries under their canonical form. But lemmatization is a sensitive process, and its quality varies greatly from one tool to another. In our list of parsers, we noted two special cases with the versions we used: CoreNLP, which does not offer lemmatization for French, and Spacy, which does not take morpho-syntactic data into account for lemmatization and outputs mostly incorrect results (for example, any word that could be a verb form is lemmatized as a verb, even if it was recognized as a noun or adjective in the POS tagging step). The other parsers may make different decisions in some situations, such as not outputting lemmas for unknown words, lemmatizing feminine nouns in the masculine form, concatenating ambiguous lemmas, etc.

Thus, we decided to re-lemmatize the outputs of all the parsers with the same tool, using an inflectional reference lexicon and relying on the POS tag assigned by the parsers. The lexicon we used is a fusion of Morphalou (Romary et al., 2004) and Lefff (Sagot, 2010). When the inflected form of the word was missing from the lexicon, we used a lemmatization strategy based on the substitution of the right-hand part of the form based on the longest suffix found in the lexicon, following the method described in (Tanguy and Hathout, 2007, p. 302). Thus, an unknown word such as *relemmatisons*, correctly tagged as a verb, will be lemmatized as *relemmatiser* by analogy with (wordform, lemma) pairs sharing the same ending like *(schématisons, schématiser)*. This robust method can process all cases in a homogeneous and deterministic way.

Thereafter, words are represented by their lemma and their POS tag. We identified 5,580 open-class words with at least 5 occurrences in each of the 11 outputs. Our comparison will use this set of words as a base.

3.2. Extraction of Syntactic Triplets

The next step is to extract the dependency relations between words, which will serve as a representation of the words' contexts for distributional analysis, following a long tradition (Lin, 1998; Bourigault, 2002; Padó and Lapata, 2007; Baroni and Lenci, 2010; Lapesa and Evert, 2017). In all these studies, the most common representation of the contexts in which a word occurs is a set of syntactic triplets (dependent_word, relation, governor_word). For instance, from the sentence *"Nous avons utilisé un analyseur syntaxique"* (*"We used a syntactic parser"*), considered as correctly parsed, we can extract the triplets *(analyseur, obj (direct object), utiliser)* and *(syntaxique, mod (modifier), analyseur)*. From each triplet, we can produce two contextual representations: one for the dependent, and the other for the governor, combining the other word with the dependency. This gives the following four pairs for the previous example: *(analyseur, utiliser_obj)*, *(utiliser, analyseur_obj-1)*, *(syntaxique, analyseur_mod)* and *(analyseur, syntaxique_mod-1)*. Following the principles of distributional analysis, the similarity between words will be computed based on these pairs.

There are different ways to generate these triplets (and the corresponding pairs) from the output of a dependency parser. Baroni and Lenci (2010) or Lapesa and Evert (2017) have proposed several variants, depending on the selected syntactic relations, on the number of dependency relations followed, and on the inclusion of certain words in the relations (prepositions and conjunctions) or not. Tanguy et

[3] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_mst.html

al. (2015), studying the same TALN corpus as we do, and using a limited evaluation set, confirmed the interest of restricting to a limited set of dependency relations as (Padó and Lapata, 2007) had identified for English. We also followed this principle.

As indicated in Section 2, the selected parsers have been trained on different corpora and the choices made during the manual annotation of these corpora (or their automated conversion) heavily influence the parsers' outputs. In our case, the main differences lie between models trained on the original FTB and models trained on corpora belonging to the UD family, as shown in Figure 1 for a given sentence in our test corpus, correctly analyzed by two different parsers.

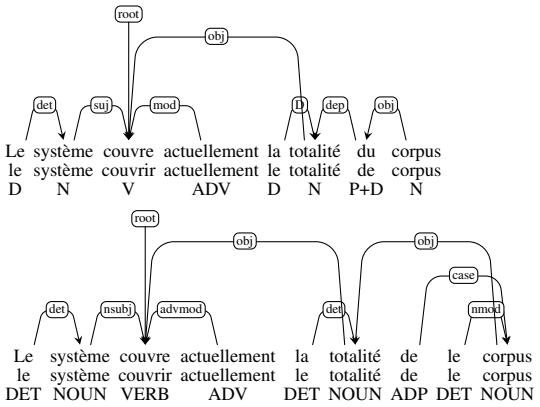

Figure 1: Syntactic dependencies for the sentence "Le système couvre actuellement la totalité du corpus" (The system currently covers the whole corpus) by MSTParser (top) and UDPipe-Sequoia (bottom).

This example illustrates how the normalization of triplets must take into account differences in tokenization for contracted articles (*du/de le*), in POS tags (N/NOUN, V/VERB), in dependency relation labels (advmod/mod), but also in the way a syntactic relation can be expressed by different dependency structures, as it is the case for the relation between *corpus* and *totalité* via the preposition *de*. We finally chose to extract triplets from the following syntactic relations:

`N suj V`: nominal subject of a verb;

`N obj V`: nominal direct object of a verb;

`ADJ mod N`: adjective modifying a noun;

`ADV mod ADJ/V`: adverb modifying an adjective or a verb;

`X coord X`: coordination between two nouns, verbs, adverbs or adjectives;

`X prep_P X`: prepositional link between noun, verb or adjective.

In the last case, we add the preposition to the relation, so that the phrase "totalité du corpus" from the example above results in the triplets (*totalité, prep_de, corpus*) and (*corpus, prep_de-1, totalité*).

We also normalized how some prepositional and adverbial locutions are identified by the parsers. For instance, Talismane-FTB identifies "*à partir de*" (from) as a compound preposition during the tokenization step (*à_partir_de*), while UD parsers like NLPCube use a specific relation (fixed) between *de* and *partir* and between *partir* and *à*. In both cases, we reconstructed the compound preposition and its external dependencies using the first notation. Finally, we also excluded triplets involving with a modal verb or a negation adverb.

3.3. Comparison of Syntactic Triplets

The number of extracted triplets (occurrences) for each parser is quite stable across parsers, with a minimum of 2.13 millions for Spacy and a maximum of 2.67 millions for Talismane-UD. Considering the types of triplets, the minimum number is 1.04 million for Spacy and the maximum is 1.32 million for UDPipe-Partut. Triplets with a word from our common vocabulary (see above) represent a total of 2.8 million different types (of which only 10%, or 261,965 exist for all parsers). We computed Spearman's ρ for each pair of parsers, measuring their agreement on the triplets' frequencies. We obtained an average value of 0.49. In Figure 2, we can more closely observe some trends in the clustering made on this basis.

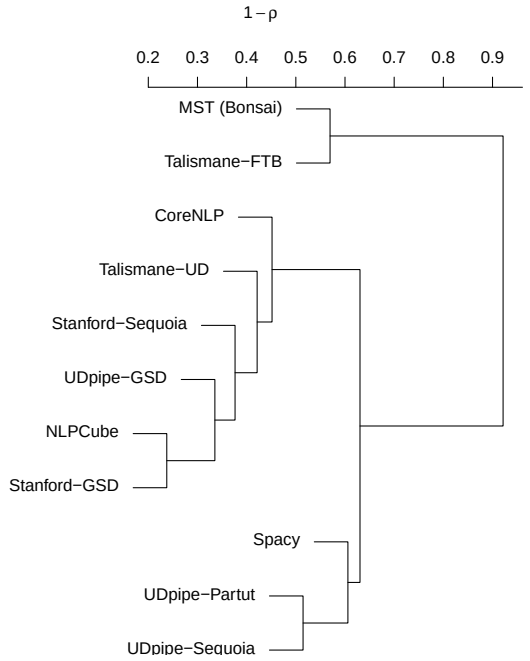

Figure 2: Hierarchical clustering of parsers according to their correlation on the frequencies of triplets found by at least two parsers.

As we can see, the most important factor seems to be the difference between the annotation schemes (UD *vs* FTB), as predicted, with MSTParser and Talismane-FTB being isolated from the others, despite the normalization of triplets. However, there are also very large variations between the parsers based on UD corpora, even though we cannot say whether the architecture or the training corpus has the biggest role. We manually examined the differences, searching for triplets who reach a high frequency in

one parser's output and are missing in one or several others. The main phenomena at the source of these disagreements were:

- tokenization (or not) of compound words (*mot-cible*, *hors-contexte*, etc.);
- ambiguity in the POS-tagging of some specific words: *même* (ADJ, ADV, PRON), *tout* (ADJ, ADV, DET, N, PRON), *certain* (ADJ, DET, PRON), numerals (ADJ, NUM, N);
- consideration of uppercase letters in POS tagging: the most common disagreement was on *TA* (standing for "Traduction Automatique" [machine translation]), misinterpreted as an uppercase *ta* [your]: N or DET;
- identification of prepositional or adverbial phrases (*d'abord* [firstly], *à partir de* [from], *par exemple* [for example] etc.), even with the strategies mentioned earlier;
- POS-tagging and lemmatization of participles (present and past: ADJ or V);
- POS-tagging of N-N compounds (which are rare in French but can be common in technical texts): *candidat terme* [candidate term], *langue cible* [target language], *vecteur contexte* [context vector]): tagged as N-N, ADJ-N, N-ADJ or other.

Without further effort to harmonize the outputs for these cases, we used these triplets to build distributional models.

4. Comparison of Distributional Models

4.1. Building the Models

Following the distinction made in Baroni et al. (2014), we built our distributional models according to a count-based approach, such as in (Lin, 1998), rather than according to a predictive model such as in (Mikolov et al., 2013). The first justification of this choice is that, except for (Levy and Goldberg, 2014), the number of studies relying on dependency relations is very limited among predictive approaches. More importantly, some recent studies (Pierrejean and Tanguy, 2018) have shown that predictive approaches are unstable to some extent concerning the nearest distributional neighbors of a word. Since we specifically want to concentrate on the effects resulting from the use of different syntactic parsers, we adopted a count-based approach.

We based our method on the findings of recent studies in the field (Kiela and Clark, 2014; Baroni et al., 2014; Levy et al., 2015) and more particularly took up two main options from (Ferret, 2010): the use of Positive Pointwise Mutual Information (PPMI) for weighting the (co-occurrent, relation) pairs and the application of a very loose filter that removes the pairs with only one occurrence in these contexts. The second choice is justified by both the small size of our target corpus and the experiments of (Ferret, 2010) with linear co-occurrents.

We classically computed the similarity of two words by measuring the Cosine similarity score between their contexts vectors. For a given model, this computation was done for each pair of words with contexts sharing at least one element. The results of this process can also be viewed as a distributional thesaurus in which each entry corresponds to a word of the considered vocabulary and is associated with the list of all other words of this vocabulary, sorted in descending order of their similarity value with the entry. In practice, only the first 100 distributional neighbors are kept.

4.2. Global Comparison of Models

First, we compared each model to all others by measuring the Spearman correlation coefficient on the similarity scores (Cosine) of all pairs of words common to all models. Hierarchical clustering was applied to the resulting similarity matrix for giving a global view of this comparison, as illustrated in Figure 3. As for syntactic triplets, three models appear atypical (MST, Talismane-FTB, and Spacy) while the others are very close to each other.

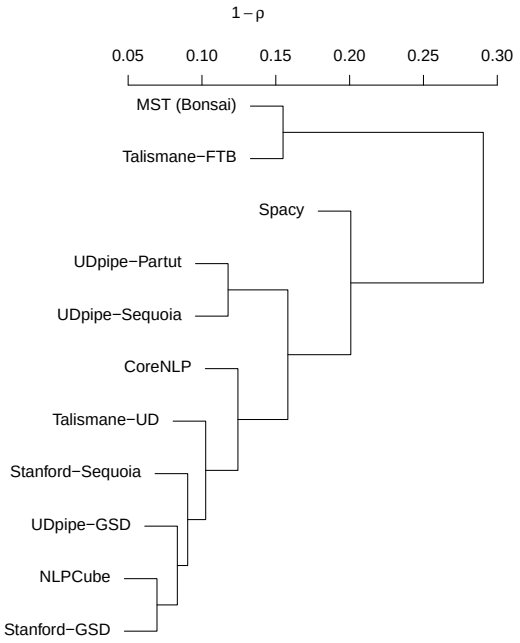

Figure 3: Hierarchical clustering of models according to their agreement on the Cosine similarity values of the common pairs of words.

Then, we computed the agreement of our models on the nearest neighbors retrieved for each word. Among the vocabulary shared by all models, only 4,469 words had at least one distributional neighbor in this vocabulary. For each pair of models, the agreement on the nearest neighbor retrieved for each word was computed[4] and used for building a similarity matrix. As previously, hierarchical clustering was performed from this matrix, which leads to Figure 4 (left side). Compared to Figure 2, the global picture of the similarities between models is a little bit different, even if the models built with Spacy, MST, and Talismane-FTB are still the most different from the others.

[4]Ratio of the number of words sharing the same nearest neighbor to the size of the considered vocabulary.

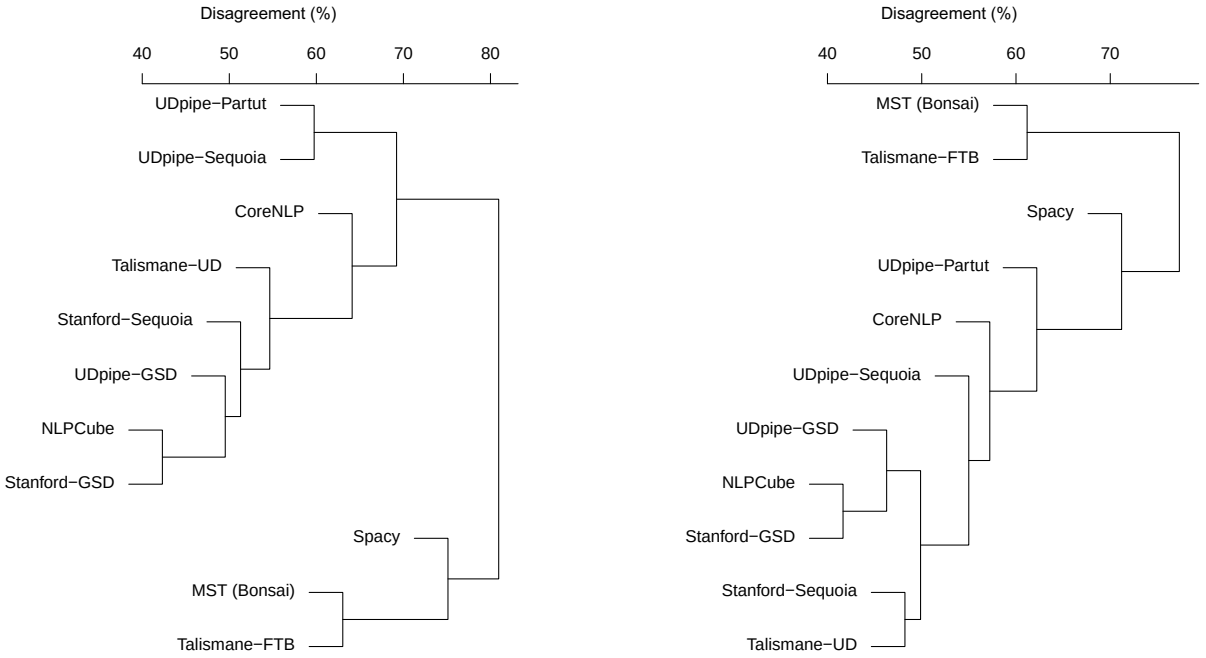

Figure 4: Hierarchical clustering of models according to their agreement on the nearest neighbor (on the left side) and the 25 nearest neighbors (on the right side).

Following (Pierrejean and Tanguy, 2018) for their study of the distributional stability of neural language models, we extended the comparison to the nearest 25 neighbors (ignoring their rank) and obtained an average ratio of 0.58, meaning that only 42% of the 25 nearest neighbors are shared by two models on average. Applying hierarchical clustering as before leads to the right part of Figure 4. The overall clustering, in that case, is closer to the one we obtained with the correlation about the global similarity of words in Figure 3 than to the agreement on the nearest neighbor in the left part of Figure 4, which is not surprising since considering the first 25 neighbors of each word also implies considering a larger number of similarities between words. However, once again, the most atypical models are those built with Spacy, MST, and Talismane-FTB while a core of UD-based models built with the Stanford, NLPCube, and UDPipe parsers can be identified.

These global trends are confirmed by comparing the neighbors of words by the means of the *Rank-Biased Overlap* measure (Webber et al., 2010), as illustrated by Figure 5. This measure is applied to all neighbors of our thesaurus' entries (100 neighbors in practice) and extends the notion of average overlap – the average of the overlap between two lists at different ranks – for decreasing the importance of overlap as the rank of the considered neighbors increases. As a consequence, nearest neighbors are given greater importance. This importance is defined by the p parameter, which can be interpreted as the probability, starting from the beginning of the list of neighbors, to continue to consider the next neighbors of the list. The value $p = 0.98$ used in our case means that the first 50 nearest neighbors of an entry account for around 85% of the evaluation. Figure 5 is based on the distance $1 - RBO$, which can be considered as a metric.

From a qualitative point of view, we found 322 entries for which all 11 models agree concerning their nearest neighbor. They correspond to different patterns:

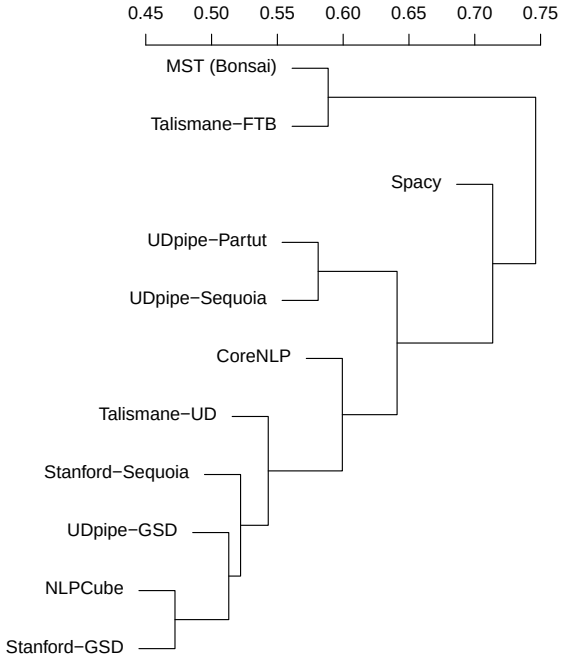

Figure 5: Hierarchical clustering of models according to the RBO measure.

- high frequency synonyms or antonyms for the NLP domain:
  sortie/entrée [input/output], *qualité/performance* [quality/performance], *ordonner/trier* [rank/sort], *valeur/score* [value/score], *texte/document* [text/document] etc.
- low frequency synonyms/antonyms:

empiler/dépiler [push/pop], *intimement/étroitement* [intimately/closely], *expérimentalement/empirique -ment* [experimentally/empirically], *mélodique/ intonatif* [melodic/intonational], *itérer/réitérer* [iterate/reiterate]

- pairs of words having the same role in a collocation with another word:
  parti/leçon (*tirer_obj*) [benefit_from/learn], *routier/ hydraulique* (*barrage_mod-1*) [roadblock/hydraulic_dam], *adjacence/covariance* (*matrice_prep_de*) [adjacency_matrix/covariance_matrix], *metteur/ mettre* (*scène_prep_en-1*) [director/stage]

Conversely, we found 10 cases of complete disagreement (words with a different first neighbor for each of the 11 models), all with a low frequency. For some of them, several nearest neighbors can be considered as relevant, as for the word *auxiliaire* [auxiliary], for which these neighbors mainly correspond to various grammatical concepts: *adverbe, déterminant, copule, croisé, gérondif, numéral, transitif, fraction, économiser, laps, subordination* [adverb, determiner, copula, cross, gerund, numeral, transitive, fraction, save, lapse, subordination]. In other cases, most of the neighbors do not have any meaningful relation with the entry, as for the word *post* [*ibid.*]: *subtilité, multi, bilan, jeudi, billet, chargement, délocaliser, syllabation, SRI, mercredi, pool* [subtlety, multi, summary, thursday, post, loading, relocate, syllabification, SRI, wednesday, pool], where only *billet* [post] can be considered as relevant.

4.3. Evaluation on a Small Benchmark

Lastly, we wanted to have some kind of estimation, even limited in scope, of the relative quality of these distributional thesauri. More precisely, we wanted to assess their ability to identify relevant word similarities in the NLP domain. In a previous study on the impact of different parameters of distributional semantic models, Tanguy et al. (2015) developed a small test set on the TALN corpus. They asked four NLP expert judges to assess the relevance of the nearest neighbors of 15 selected target words, according to a number of distributional models. This data set[5] is built around 5 verbs (*annoter* [annotate], *calculer* [compute], *décrire* [describe], *extraire* [extract], *évaluer* [evaluate]), 5 nouns (*fréquence* [frequency], *graphe* [graph], *méthode* [method], *sémantique* [semantics], *trait* [feature]) and 5 adjectives (*complexe* [complex], *correct* [correct], *important* [important], *précis* [precise], *spécialisé* [specialized]). For each of these target words the data set provides several words deemed similar by the annotators, with a score for each neighbor corresponding to the number of annotators who considered it as semantically related. For example, for *trait* [feature] the words considered as similar are *attribut* [attribute] (4), *caractéristique* [characteristic] (4), *propriété* [property] (4), *étiquette* [label/tag] (4), *catégorie* [category] (3), *descripteur* [descriptor] (2), *feature* [*ibid.*] (2), *indice* [clue] (2), *information* [information] (2) ... *marque* [mark] (1), *représentation* [representation] (1), *structure* [structure] (1).

[5]Available here: http://redac.univ-tlse2.fr/datasets/semdis-gold/TAL56-2/index_en.html

Model	Average rank
Talismane-UD	2.38
MST (Bonsai)	2.88
Talismane-FTB	3.00
Stanford-Seq	4.25
UDPipe-Seq	5.75
NLPCube	6.13
UDPipe-Partut	6.25
UDPipe-GSD	6.38
CoreNLP	7.50
Stanford-GSD	9.38
Spacy	11.00

Table 1: Average ranking of the parsers on the test set developed by (Tanguy et al., 2015), based on their cumulative score at ranks (1, 5, 10, 15, 20, 25, 50, 100)

This data set has only a partial coverage as, for each of the target words, the only candidates presented to the annotators were the 3 nearest neighbors according to at least one of the distributional models initially considered (720 different configurations with varying parameter values). This means that it is possible that relevant distributional neighbors identified by one of our own 11 models had not been assessed by the judges and thus evaluated as noise. However, after a first examination, we consider such cases to be marginal.

According to this test set, we computed for each model the sum of the relevance scores for the N nearest neighbors for several values of N (1, 5, 10, 15, 20, 25, 50 and 100). The ranking between the models being quite stable, we report in Table 1 their average rank based on these values.

The ranking of the parsers shows deep differences between both ends, and we can see the same outliers we found in the previous experiments. Spacy appears to be the parser whose output and performance is the most different from the MST/Bonsai - Talismane trio. We can also note that if, for the Stanford Parser, the training corpus is critical, this is not the case for UDPipe.

5. Conclusion and Perspectives

The main objective of the work presented in this article was to study the impact of syntactic parsers on the distributional models built from a small corpus in a specialized domain. The results of this study show that this impact is very significant, which means that the choice of a syntactic parser in such a context is far from being neutral. Moreover, we observed that the differences between parsers concerning their outputs, i.e. syntactic triplets, are not strictly correlated with the differences in the resulting distributional models concerning their distributional neighbors but without being too far from them. We also noticed that all words, contexts and similar pairs of words are not impacted similarly when the parser is changed. Globally, this first study made it possible to split the set of tested parsers in such a way that the most different parsers are clearly identified. Deeper evaluations and manual analyses of disagreements between parsers should lead to pinpoint the most important syntactic contexts and as a consequence, to help in choos-

ing the most appropriate parser in the kind of context we consider in this article.

6. Acknowledgments

This work has been funded by French National Research Agency (ANR) under project ADDICTE (ANR-17-CE23-0001).

7. Bibliographical References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French. In *Treebanks*, pages 165–187. Springer.

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

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 238–247, Baltimore, Maryland.

Boroş, T., Dumitrescu, S. D., and Burtica, R. (2018). NLP-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium, October. Association for Computational Linguistics.

Bosco, C., Sanguinetti, M., and Lesmo, L. (2012). The parallel-TUT: a multilingual and multiformat treebank. In *Proceedings of LREC*, pages 1932–1938. European Language Resources Association (ELRA).

Bourigault, D. (2002). UPERY: un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In Jean-Marie Pierrel, editor, *Actes de TALN 2002 (Traitement automatique des langues naturelles)*, pages 75–84, Nancy, June. ATALA, ATILF.

Candito, M. and Seddah, D. (2012). Le corpus Sequoia: annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN*, pages 321–334.

Candito, M., Nivre, J., Denis, P., and Anguiano, E. H. (2010). Benchmarking of statistical dependency parsers for French. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 108–116. Association for Computational Linguistics.

De La Clergerie, E. V., Hamon, O., Mostefa, D., Ayache, C., Paroubek, P., and Vilnat, A. (2008). Passage: from French parser evaluation to large sized treebank. In *Proceedings of LREC*.

De La Clergerie, É. V. (2014). Jouer avec des analyseurs syntaxiques. In *Actes de TALN*.

Denis, P. and Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*.

Fares, M., Oepen, S., Øvrelid, L., Bj, J., Johansson, R., et al. (2018). The 2018 shared task on extrinsic parser evaluation: On the downstream utility of english universal dependency parsers. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 22–33.

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

Honnibal, M. and Johnson, M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September. Association for Computational Linguistics.

Kiela, D. and Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. In *2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 21–30, Gothenburg, Sweden.

Koo, T., Carreras, X., and Collins, M. (2008). Simple semi-supervised dependency parsing. *Proceedings of ACL-08: HLT*, pages 595–603.

Lapesa, G. and Evert, S. (2017). Large-scale evaluation of dependency-based DSMs: Are they worth the effort? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 394–400.

Levy, O. and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 302–308, Baltimore, Maryland, June.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

McDonald, R., Lerman, K., and Pereira, F. (2006). Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 216–220. Association for Computational Linguistics.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., et al. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR 2013, workshop track*.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*.

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2012). Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Paroubek, P., Robba, I., Vilnat, A., and Ayache, C. (2008). EASY, Evaluation of Parsers of French: what are the results? In *Proceedings of LREC*.

Pierrejean, B. and Tanguy, L. (2018). Towards qualitative word embeddings evaluation: Measuring neighbors variation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 32–39.

Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2018). Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October. Association for Computational Linguistics.

Romary, L., Salmon-Alt, S., and Francopoulo, G. (2004). Standards going concrete : from LMF to Morphalou. In *COLING 2004 Enhancing and using electronic dictionaries*, pages 22–28, Geneva, Switzerland, August 29th. COLING.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of LREC*, Valletta, Malta, May. European Languages Resources Association (ELRA).

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A., and Villemonte de la Clergerie, E. (2013). Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA, October. Association for Computational Linguistics.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Tanguy, L. and Hathout, N. (2007). *Perl pour les linguistes*. Hermès.

Tanguy, L., Sajous, F., and Hathout, N. (2015). Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé: comparaison des approches par contextes syntaxiques et par fenêtres graphiques. *Traitement automatique des langues*, 56(2).

Urieli, A. and Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de TALN*, pages 188–201, Les Sables d'Olonne, France.

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, November.

Zeman, D., Haji, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.