Comparaison qualitative et extrinsèque d'analyseurs syntaxiques du français : confrontation de modèles distributionnels sur un corpus spécialisé

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

RÉSUMÉ

Nous présentons une étude visant à comparer 11 différents analyseurs en dépendances du français sur un corpus spécialisé (constitué des archives des articles de la conférence TALN). En l'absence de gold standard, nous utilisons chacune des sorties de ces analyseurs pour construire des thésaurus distributionnels en utilisant une méthode à base de fréquence. Nous comparons ces 11 thésaurus afin de proposer un premier aperçu de l'impact du choix d'un analyseur par rapport à un autre.

ABSTRACT

Extrinsic evaluation of French dependency parsers on a specialised corpus : comparison of distributional thesauri

We present a study with the goal of comparing 11 different french dependency parsers applied on a specialised corpus (consisting of articles from the archives of the TALN conference). Due to the lack of a gold standard, we use each of the parsers' output to generate distributional thesauri using a frequency-based method. We compare these 11 thesauri in order to propose a first look at the impact of choosing to use a parser over another.

MOTS-CLÉS : analyse syntaxique, analyse distributionnelle, domaine de spécialité, évaluation.

KEYWORDS: French dependency parsing, distributional semantics, specialised corpus.

1 Introduction

Cet article se place dans le cadre d'une étude des méthodes de sémantique distributionnelle en domaine spécialisé et en français. Notre objectif à moyen terme est la sélection de la méthode la plus efficace pour identifier des similarités sémantiques distributionnelles entre les unités lexicales ou terminologiques d'un petit corpus de langue spécialisée (quelques millions de mots au plus). On sait que de très nombreux paramètres interviennent et doivent faire l'objet de choix éclairés, sans que l'on puisse malheureusement se fonder sur des conclusions obtenues sur de grands corpus génériques. Nous faisons tout d'abord l'hypothèse qu'une méthode à base de contextes syntaxiques permet de compenser la faible quantité de données, ce qui a en partie été montré par (Tanguy *et al.*, 2015). La question que nous étudions plus précisément ici est l'impact du choix d'un analyseur syntaxique en amont d'une méthode fréquentielle de construction d'un modèle distributionnel. Ce faisant, nous rejoignons les questions de la comparaison de l'efficacité des différents outils et modèles disponibles

pour l'analyse syntaxique en dépendances.

Il y a eu de nombreux efforts dans la communauté du Traitement Automatique des Langues (TAL) du français pour comparer les différents analyseurs syntaxiques : campagnes Easy (Paroubek *et al.*, 2008), Passage (De La Clergerie *et al.*, 2008), SPMRL (Seddah *et al.*, 2013), CoNLL (Zeman *et al.*, 2018) ou des comparaisons plus ponctuelles comme celles de (Candito *et al.*, 2010) ou (De La Clergerie, 2014). Mais les bancs de test utilisés ne sont pas forcément pertinents pour l'analyse de corpus spécialisés, préférant les corpus génériques et variés sur lesquels les outils sont entraînés. De plus, malgré les campagnes récentes, les principaux outils disponibles ne sont pas nécessairement tous comparés sur les mêmes jeux d'évaluation.

En l'absence d'étalon adapté à nos besoins, nous avons donc procédé à une comparaison qualitative en confrontant les principaux analyseurs actuels du français sur une tâche externe, à la manière des récentes campagnes EPE pour l'anglais (Fares *et al.*, 2018). Ne disposant pour cette tâche que d'un jeu de test de couverture et de validité limitées, l'essentiel de notre évaluation reste purement comparative. Nous examinerons donc avant tout l'ampleur et la portée des modifications apportées par le changement d'analyseur syntaxique sur les thésaurus distributionnels obtenus en fin de chaîne.

Notre dispositif est donc le suivant : sur un même corpus spécialisé en français, nous avons appliqué 11 analyseurs syntaxiques en dépendances (ou versions) et extrait pour chacun d'eux les principaux contextes syntaxiques que nous avons utilisés pour construire autant de modèles distributionnels en utilisant une méthode classique à base de fréquence. Nous comparons au final les thésaurus distributionnels obtenus afin d'identifier à la fois l'impact réel de l'analyseur sur la chaîne mais nous proposons aussi un début de cartographie des analyseurs en fonction de la similarité des modèles qu'ils ont permis de générer.

Dans la section 2, nous présentons le corpus choisi et les différents analyseurs (ou configurations d'analyseurs) qui y ont été appliqués. Puis nous présentons en section 3 les différentes étapes de normalisation et de sélection des sorties de ces analyseurs que nous avons dû appliquer pour obtenir un socle commun de comparaison ainsi qu'une première comparaison fondée sur ce socle. La section 4 présente le processus de construction des modèles distributionnels et leur comparaison suivant différentes approches.

2 Matériau et outils testés

2.1 Corpus TALN

Nous avons utilisé pour cette expérimentation un corpus spécialisé de petite taille : le corpus TALN [1], constitué des archives des actes des conférences TALN et RECITAL des années 1999 à 2014 incluses. Ce corpus d'environ 4,5 millions de mots possède plusieurs avantages pour l'étude des modèles d'analyse distributionnelle : il s'agit de textes écrits de bonne qualité, homogènes en termes de genre et de thématique et relevant d'un thème pour lequel nous avons un niveau d'expertise suffisant pour pouvoir interpréter facilement les résultats d'une analyse distributionnelle.

La conversion en texte brut depuis les fichiers PDF initiaux a été réalisée avec le seul objectif de disposer de texte compatible avec un analyseur syntaxique, au détriment d'un ensemble d'éléments comme les titres de section, notes de bas de page, tableaux, formules mathématiques, références

1. Disponible sur `http://redac.univ-tlse2.fr/corpus/taln.html`

bibliographiques et autres légendes. Les césures ont été éliminées et l'ensemble du texte ainsi filtré de chaque article est présenté sur une même ligne. La robustesse des analyseurs face à des données trop bruitées n'a donc pas été un paramètre de cette étude.

2.2 Analyseurs testés

Nous avons sélectionné 7 outils proposant une analyse syntaxique du français en dépendances, en nous concentrant sur ceux disponibles facilement et prêts à l'emploi (i.e. prenant en charge l'ensemble de la chaîne de traitement du texte brut jusqu'aux dépendances syntaxiques). Ces outils ont été appliqués avec toutes les options par défaut.

Tous ces outils se fondent sur des modèles obtenus par apprentissage sur des corpus annotés. La multiplicité des analyseurs est en fait essentiellement due à des choix d'implémentation concernant les techniques d'analyse en dépendances (par graphe ou par transitions par exemple), les modèles d'apprentissage (modèles classiques type SVM ou entropie maximale ou plus récemment réseaux de neurones récurrents) et les traitements en amont ou périphériques (segmentation, lemmatisation). Les corpus d'entraînement sont, eux, bien plus réduits en nombre, ce qui s'explique bien entendu par le coût élevé du processus d'annotation et de validation. Avant de présenter les outils que nous avons étudiés, il est donc primordial de rappeler les corpus disponibles, puisque ceux-ci ont un impact décisif sur la nature et le format des sorties.

FTB Le premier corpus annoté syntaxiquement pour le français est le *Corpus arboré du français*, plus connu sous son nom anglais de *French Treebank* ou *FTB* (Abeillé *et al.*, 2003). Constitué d'environ 600 000 mots issus du journal Le Monde, il a tout d'abord été annoté en suivant une analyse en constituants, puis converti automatiquement en dépendances par (Candito *et al.*, 2010). Une autre version a également été produite pour la campagne d'évaluation SPMRL, notamment pour le repérage des unités polylexicales (Seddah *et al.*, 2013).

UD French Afin de faciliter le développement et la comparaison des différents analyseurs ainsi que des études typologiques crosslingues à grande échelle, un schéma universel de dépendances a été proposé, fondé sur le modèle des *Stanford Dependencies*, désormais baptisé *Universal Dependencies* (ou UD)[2] (Nivre *et al.*, 2016). Le projet UD propose des lignes génériques ainsi que des jeux d'étiquettes universels (pour les catégories grammaticales et les relations syntaxiques) et a permis de regrouper sous un même format et de diffuser différents corpus annotés du français, notamment :

UD French FTB est la conversion en UD du French Treebank original et contient environ 550 000 mots.

UD French ParTUT est la partie française du corpus multilingue Parallel-TUT (Bosco *et al.*, 2012), composé d'échantillons de textes variés (textes de lois, entrées de Wikipédia, pages Facebook, etc.) pour un total d'environ 30 000 mots.

UD French GSD est le corpus initial du projet UD qui prend ses sources dans les dépendances de Stanford (McDonald *et al.*, 2013). Il contient 400 000 mots issus d'articles de journaux, de pages Wikipédia, de blogs ou de critiques de divers produits.

UD French Sequoia est un corpus développé en complément du FTB dans le but (entre autres) d'en améliorer la couverture en genre et en domaines (Candito & Seddah, 2012).

2. Voir `http://universaldependencies.org/` pour un historique détaillé et les différentes versions de son développement.

Les 70 000 mots de ce corpus sont issus de débats parlementaires, de presse régionale, de Wikipédia et de textes médicaux. Initialement annoté suivant le schéma du FTB, il a été converti au format UD.

Il est à noter que malgré la volonté de normalisation et d'universalisation du projet UD, les différents corpus cités précédemment n'utilisent pas exactement les mêmes conventions ni les mêmes jeux d'étiquettes pour les relations syntaxiques. Ces différences s'expliquent par des positions théoriques ou techniques face à certains phénomènes syntaxiques, mais aussi par les différentes étapes de conversion qu'ont connues certains corpus.

En ce qui concerne les analyseurs, nous avons sélectionnés 7 outils différents, dont certains proposent des variantes en termes de modèles pré-entraînés, autrement dit de corpus.

CoreNLP (Manning *et al.*, 2014), l'analyseur principal de l'équipe de Stanford, implémente un étiqueteur à entropie maximale et un analyseur syntaxique par transitions. Il a été entraîné sur le corpus UD GSD.

StanfordNLP (Qi *et al.*, 2018) est un outil qui, en plus de permettre d'accéder aux fonctionnalités de CoreNLP en Python, implémente une chaîne de traitement neuronale entièrement différente. Son analyseur syntaxique par graphes repose sur un réseau neuronal LSTM. StanfordNLP propose plusieurs modèles pour le français. Nous en avons utilisé deux, entraînés respectivement sur les corpus UD **GSD** et **Sequoia**.

NLPCube (Boroş *et al.*, 2018) est, comme StanfordNLP, fondé sur des réseaux de neurones récurrents LSTM. Sa particularité principale est une analyse syntaxique indépendante de l'étiquetage morphosyntaxique, l'une comme l'autre utilisant uniquement des attributs lexicalisés, sans information morphologique. Il est significativement plus lent que tous les autres outils utilisés. Nous n'avons pas trouvé d'indication précise concernant les corpus sur lesquels le modèle fourni a été entraîné et nous avons présumé que la somme des corpus UD disponibles pour le français a été utilisée.

Spacy est un outil à visée industrielle qui met en avant sa rapidité par rapport aux autres outils disponibles. L'étiqueteur est un perceptron moyenné avec des attributs liés aux clusters de Brown suivant Koo *et al.* (2008). Il implémente un analyseur syntaxique par transitions non-monotone qui peut revenir sur des décisions antérieures (Honnibal & Johnson, 2015). Le modèle fourni a été entraîné sur le corpus WikiNER (Nothman *et al.*, 2012) pour la reconnaissance d'entités nommées et sur UD Sequoia pour l'étiquetage et l'analyse syntaxique.

UDPipe (Straka & Straková, 2017) accomplit la tokenisation et la segmentation dans un même temps avec un réseau de neurones récurrent à portes (GRU). Pour l'étiquetage, il génère des étiquettes possibles à partir du suffixe du mot puis désambiguïse à l'aide d'un perceptron moyenné. L'analyse syntaxique par transitions est fondée sur un réseau neuronal simple à une couche. UDPipe propose plusieurs modèles pour le français. Nous en avons utilisé trois, entraînés respectivement sur les corpus UD **GSD**, **Sequoia** et **ParTUT**.

Talismane (Urieli & Tanguy, 2013) utilise une combinaison de modèles statistiques aux traits spécifiques à chaque langue et de règles incorporant de la connaissance linguistique. En plus de la version distribuée entraînée sur le French TreeBank converti en dépendances (**FTB**), nous avons également utilisé une version expérimentale utilisant un modèle Universal Dependencies (**UD**) entraîné sur la concaténation de tous les corpus UD décrits précédemment.

MSTParser (McDonald *et al.*, 2006) est un analyseur en dépendances par graphes. Nous l'avons

utilisé à l'aide du paquet BONSAI [3], qui le couple à l'étiqueteur MElt (Denis & Sagot, 2009) et met en œuvre le meilleur modèle MST selon le benchmark de (Candito *et al.*, 2010). Il s'appuie sur la version non-UD du FTB.

Nous sommes bien conscients que les analyseurs décrits ci-dessus ne sont comparables que sur un plan purement pratique puisqu'ils relèvent de technologies, de degrés de finalisation voire d'époques très différents et qu'ils sont fondés sur des données d'entraînement elles aussi non comparables. Néanmoins, ils forment une bonne partie du paysage actuel des solutions disponibles en termes d'analyse syntaxique robuste du français et sont à ce titre tous susceptibles d'être considérés.

3 Exploitation des sorties

La comparaison des 11 outils ou versions sélectionnés nécessite d'identifier ou de construire un terrain commun entre les différentes analyses produites. Plusieurs problèmes se posent en termes d'hétérogénéité des sorties : l'identification des unités, l'alignement des jeux d'étiquettes morphosyntaxiques, la lemmatisation et bien entendu les relations de dépendance syntaxique. Nous avons décidé de nous limiter aux mots simples relevant des classes ouvertes (noms, verbes, adjectifs et adverbes), sous leur forme lemmatisée et en leur associant leur catégorie. Pour les relations de dépendance, nous avons sélectionné les principales à la fois représentées dans tous les formalismes et jugées les plus utiles pour l'analyse distributionnelle.

3.1 Identification des mots

Notre étude, comme c'est le cas pour l'essentiel des travaux actuels en sémantique distributionnelle, porte sur des mots simples et suppose donc que la segmentation se fait de façon homogène par les différents analyseurs.

La question de la lemmatisation est cruciale pour le français et les petits corpus puisqu'elle permet de limiter la dispersion des unités et donc de lutter contre le manque de données. Elle permet également un lien (en amont ou en aval de l'analyse distributionnelle) avec des ressources lexicales ou terminologiques qui représentent généralement leurs entrées sous leur forme canonique. Mais la lemmatisation reste une opération délicate, que les analyseurs traitent de façon inégale. Dans notre liste de systèmes, notons deux cas particuliers : CoreNLP, qui ne propose pas de lemmatisation pour le français, et Spacy, qui semble ne pas prendre en compte la catégorie morphosyntaxique des mots avant de calculer leur lemme et propose donc des sorties majoritairement erronées (nous avons remarqué que toute forme fléchie pouvant correspondre à un verbe sera toujours lemmatisée en utilisant l'infinitif de ce verbe, même si cette forme a été correctement identifiée comme un nom ou un adjectif). Les autres analyseurs qui effectuent une lemmatisation standard peuvent prendre ponctuellement des décisions différentes pour certaines situations (lemmes inconnus absents, certains noms féminins lemmatisés au masculin, lemmes ambigus marqués comme tels, etc.).

Nous avons donc décidé de lemmatiser toutes les sorties avec le même outil en utilisant un lexique flexionnel de référence et en nous fondant sur la catégorie morphosyntaxique attribuée par l'analyseur et ce, pour chaque mot des classes ouvertes. Le lexique est constitué de la fusion de Morphalou (Romary *et al.*, 2004) et de Lefff (Sagot, 2010). En cas d'absence de la forme de surface du mot dans

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

le lexique, nous avons appliqué une stratégie de lemmatisation par substitution de la partie droite de la forme fondée sur le plus long suffixe trouvé dans le lexique en appliquant la méthode décrite dans (Tanguy & Hathout, 2007, p. 302). Ainsi, un mot inconnu du lexique comme *relemmatisons* catégorisé comme verbe sera lemmatisé en *relemmatiser* par analogie avec des couples (forme, lemme) dont la finale est commune comme *(schématisons, schématiser)*. Cette méthode robuste permet de traiter l'ensemble des cas de façon homogène et déterministe.

Dans toute la suite, les mots sont représentés par leur lemme ainsi calculé et leur catégorie grammaticale. Nous avons identifié 5 580 mots de classe ouverte ayant une fréquence minimale de 5 dans chacune des 11 sorties. Notre comparaison s'appuiera sur cet ensemble.

3.2 Extraction des triplets syntaxiques

L'étape suivante consiste à extraire les relations de dépendance entre deux mots qui serviront de représentation des contextes des mots pour l'analyse distributionnelle, suivant une longue tradition (Lin, 1998; Bourigault, 2002; Padó & Lapata, 2007; Baroni & Lenci, 2010; Lapesa & Evert, 2017). Dans tous ces travaux, le mode le plus classique de représentation des contextes d'apparition des mots est l'utilisation d'un triplet syntaxique du type (mot_dépendant, relation, mot_gouverneur). Par exemple de la phrase *"Nous avons utilisé un analyseur syntaxique"* (correctement analysée), on peut extraire les triplets *(analyseur, obj, utiliser)* et *(syntaxique, mod, analyseur)*. Ces triplets permettent en fait de produire chacun deux représentations contextuelles : une pour le dépendant, l'autre pour le gouverneur (avec une relation inversée), que l'on représente sous la forme de couples *(analyseur, utiliser_obj)*, *(utiliser, analyseur_obj-1)*, *(syntaxique, analyseur_mod)* et *(analyseur, syntaxique_mod-1)*. C'est sur la base de ces couples que le rapprochement des mots s'effectuera en appliquant le principe de l'analyse distributionnelle.

Il existe un grand nombre de possibilités pour générer ces triplets (et les couples correspondants) à partir des sorties d'un analyseur en dépendances. (Baroni & Lenci, 2010) ou (Lapesa & Evert, 2017) en ont proposé de nombreuses variantes, qui dépendent par exemple des relations syntaxiques considérées, du nombre de liens de dépendance suivis et de l'inclusion de certains mots (e.g. les prépositions) dans les relations syntaxiques. (Tanguy *et al.*, 2015) ont, sur le même corpus TALN, et en utilisant un jeu d'évaluation réduit, confirmé les conclusions de (Padó & Lapata, 2007) sur l'intérêt de se limiter à un jeu réduit de relations de dépendance, ce que nous avons fait ici.

Comme indiqué en section 2, les analyseurs testés sont entraînés sur des corpus différents et leurs sorties héritent donc des choix différents faits lors des campagnes d'annotation manuelle (ou d'une conversion automatique le cas échéant). Les différences principales dans notre cas sont celles existant entre les modèles issus du FTB et ceux des corpus de la famille UD, comme le montre la figure 1 pour une même phrase de notre corpus, correctement analysée par deux outils différents.

Cet exemple illustre que la normalisation des triplets nécessite de prendre en compte à la fois les différences de segmentation pour le traitement des articles contractés (*du/de le*), d'étiquettes des catégories morphosyntaxiques (N/NOUN, V/VERB), de liens de dépendance (advmod/mod) mais aussi dans la façon dont une même relation est traduite par plusieurs dépendances, comme c'est le cas pour le rattachement de *corpus* à *totalité* via la préposition *de*.

Nous avons au final choisi d'extraire les triplets correspondant aux relations suivantes :

- **`N suj V`** : sujet nominal d'un verbe ;
- **`N obj V`** : objet direct nominal d'un verbe ;

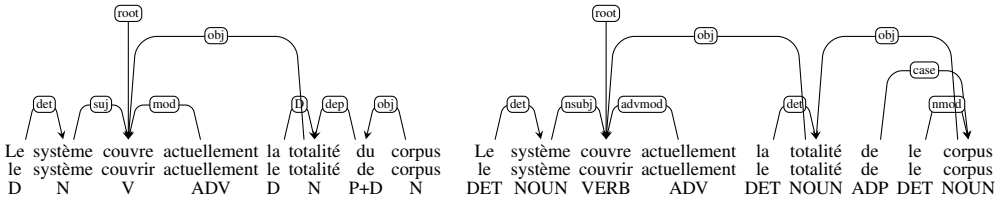

FIGURE 1 – Dépendances identifiées pour la phrase "Le système couvre actuellement la totalité du corpus" par MSTParser (à gauche) et UDPipe-Sequoia (à droite)

`ADJ mod N` : adjectif modifieur d'un nom ou attribut du sujet nominal ;

`ADV mod ADJ/V` : adverbe modifieur d'un adjectif ou d'un verbe ;

`X coord X` : coordination entre deux noms, verbes, adverbes ou adjectifs ;

`X prep_P X` : rattachement prépositionnel entre nom, verbe ou adjectif.

Dans ce dernier cas, nous avons intégré la préposition à la relation, si bien que le cas ci-dessus de "totalité du corpus" donne les triplets (*totalité, prep_de, corpus*) et (*corpus, prep_de-1, totalité*).

Nous avons également normalisé les locutions prépositionnelles et adverbiales dont certaines sont identifiées par certains des analyseurs. Talismane-FTB, par exemple, identifiera "à partir de" comme une préposition complexe directement au niveau de la segmentation (*à_partir_de*) alors que les analyseurs UD comme NLPCube utiliserons une relation de dépendance spécifique (fixed) qui rattache *de* à *partir* et *partir* à *à*. Dans les deux cas, nous avons reconstruit la préposition complexe et ses relations de dépendance externes en utilisant la première notation. Nous avons de plus exclus explicitement les cas où le verbe est un modal et ceux où l'adverbe est une négation.

Ces extractions ont nécessité le développement de règles spécifiques pour s'adapter à chaque famille de format de sortie. Ceci a pu conduire à regrouper certaines relations et assimiler des distinctions que certains formats d'annotation feraient et pas d'autres.

3.3 Comparaison des triplets syntaxiques

Le nombre de triplets (occurrences) extraits est assez stable et va de 2,13 millions pour Spacy à 2,67 millions pour Talismane-UD. Les triplets uniques (types) vont de 1,04 million pour Spacy à 1,32 million pour UDPIPE-Partut. Les triplets impliquant un mot du vocabulaire commun (cf. ci-dessus) rassemblent un total de 2,8 millions de triplets différents (dont seuls 10%, soit 261 965, ont été repérés par tous les analyseurs). La comparaison des accords entre les analyseurs sur les fréquences des triplets donne une corrélation (Spearman) moyenne de 0,49. On peut observer plus précisément certaines tendances dans les rapprochements opérés sur cette base au niveau de la figure 2.

On remarquera que ce sont les jeux d'étiquettes (UD *vs* FTB) qui semblent avoir l'impact le plus important, comme attendu, avec un isolement de MSTParser et de Talismane-FTB. Il y a en revanche de très importantes variations au sein des analyseurs entraînés sur les corpus de la famille UD, sans que l'on puisse à ce stade identifier un rôle prédominant de l'architecture ou du corpus d'entraînement.

Nous avons examiné manuellement les différences en parcourant la liste des triplets ayant une fréquence importante dans l'une des sorties et absents d'une ou plusieurs autres. Les principaux phénomènes que nous avons pu identifier à la source de ces désaccords sont les suivants :

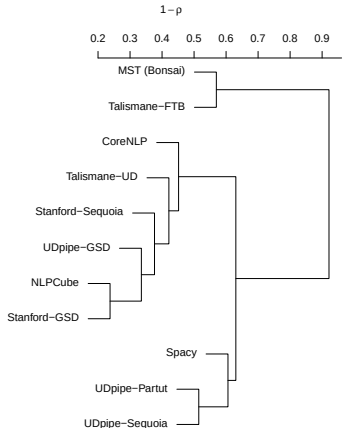

FIGURE 2 – Classification hiérarchique ascendante des analyseurs en fonction de leur corrélation sur les fréquences des triplets repérés par au moins deux analyseurs différents

— segmentation (ou non) des mots composés (*mot-cible*, *hors-contexte*, etc.) ;
— étiquetage de certains mots : *même* (ADJ, ADV, PRON), *tout* (ADJ, ADV, DET, N, PRON), *certain* (ADJ, DET, PRON), numéraux (ADJ, NUM, N) ;
— prise en compte des majuscules lors de l'étiquetage ; le cas de désaccord le plus courant est *TA* (pour traduction automatique) : N ou DET ;
— repérage de locutions (*d'abord*, *à partir de*, *par exemple* etc.), même en prenant en compte les différentes stratégies comme indiqué plus haut ;
— catégorisation (et donc lemmatisation) des participes (présents et passés : ADJ ou V) ;
— catégorisation des composés N-N (*candidat terme*, *langue cible*, *vecteur contexte* : étiquetés comme N-N, ADJ-N, N-ADJ ou autre).

Sans chercher à harmoniser ces différents points, nous avons utilisé ces jeux de triplets pour construire des modèles distributionnels.

4 Comparaison des modèles distributionnels

4.1 Construction des modèles

Pour reprendre la distinction faite dans (Baroni *et al.*, 2014), la construction des contextes distributionnels dans cette étude s'inscrit dans une tradition d'approches à base de comptes (Lin, 1998) par opposition aux approches prédictives (Mikolov *et al.*, 2013). Ce choix est doublement motivé. Tout d'abord, l'utilisation de relations de dépendance dans les approches prédictives reste rare, à l'exception de (Levy & Goldberg, 2014). Plus fondamentalement, un certain nombre de travaux récents (Pierrejean & Tanguy, 2018) ont montré que les approches prédictives se caractérisent par une certaine instabilité du point de la recherche des plus proches voisins. Afin de nous concentrer sur les différences résultant de la seule utilisation de différents analyseurs syntaxiques, nous avons donc opté pour une approche à base de comptes.

Sa mise en œuvre est très classique et reprend les acquis d'un certain nombre d'études récentes (Kiela

& Clark, 2014; Baroni *et al.*, 2014; Levy *et al.*, 2015), et plus particulièrement de (Ferret, 2010) au travers de deux principaux éléments : l'adoption de l'information mutuelle ponctuelle positive pour pondérer les couples (cooccurrent, relation) au sein des contextes distributionnels et l'application d'un filtrage très limité se contentant de supprimer les couples n'ayant qu'une seule occurrence dans les contextes. Ce dernier choix est motivé à la fois par la taille restreinte du corpus d'étude et les expérimentations réalisées dans (Ferret, 2010) dans le cas des cooccurrents linéaires.

Nous avons calculé la similarité entre deux mots d'un modèle en utilisant la mesure classique du cosinus sur chaque paire de mots pour laquelle cela était possible (i.e. chaque paire de mots ayant au moins un contexte commun).

4.2 Comparaison globale des modèles

Nous avons tout d'abord comparé chaque modèle aux autres en calculant le coefficient de corrélation de Spearman sur les scores de similarité (cosinus) sur l'ensemble des paires de mots du vocabulaire commun. Comme précédemment, nous avons synthétisé cette comparaison en faisant une analyse hiérarchique ascendante à la figure 3. On y retrouve, comme précédemment, que trois modèles sont identifiés comme plus atypiques (MST, Talismane-FTB et Spacy), les autres étant très regroupés.

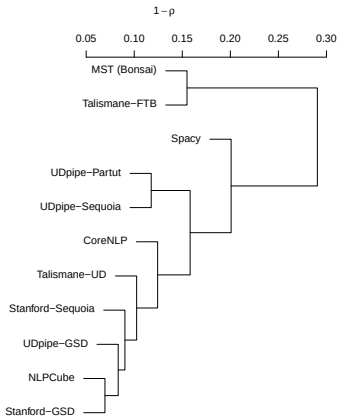

FIGURE 3 – Classification hiérarchique ascendante des modèles en fonction de leur accord sur les cosinus des paires de mots communs

Nous avons ensuite calculé l'accord entre les modèles concernant les premiers voisins proposés pour chaque mot par les différents modèles. Parmi les mots du vocabulaire commun décrit précédemment, seuls 4 469 avaient au moins un voisin distributionnel dans ce même ensemble. Nous avons calculé pour chaque paire de modèles le taux d'accord relatif sur le premier voisin et utilisé cette similarité pour faire une analyse hiérarchique ascendante présentée à la figure 4. On peut y voir une organisation différente de celle de la figure 2, bien que Spacy, MST et Talismane-FTB y restent les plus excentriques.

En étendant la comparaison des plus proches voisins aux 25 premiers voisins, nous reprenons la méthode utilisée par (Pierrejean & Tanguy, 2018) pour mesurer la stabilité des méthodes neuronales d'analyse distributionnelle. Le taux moyen est de 0,58, ce qui signifie que seuls 42% des 25 premiers voisins se retrouvent (en moyenne) d'un modèle à un autre. Le score de variation moyen peut

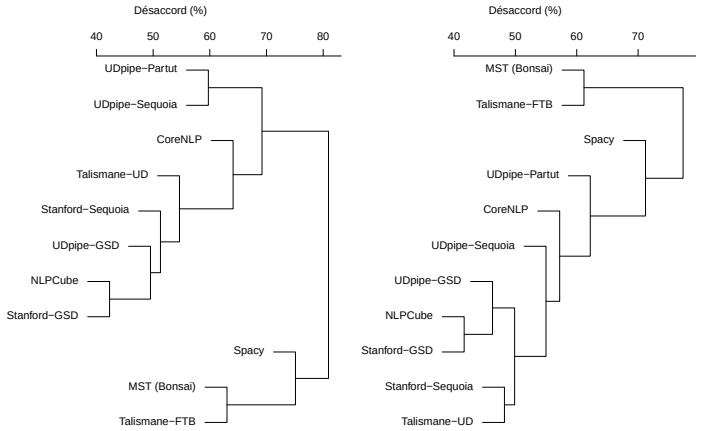

FIGURE 4 – Classification hiérarchique ascendante des modèles en fonction de leur accord sur le plus proche voisin (à gauche) et sur les 25 plus proches voisins (à droite)

être utilisé pour comparer deux à deux chacun des modèles, ce que nous avons synthétisé par une classification hiérarchique ascendante à la figure 4 (à droite). On y voit une modification par rapport à la prise en compte du seul premier voisin et au final, un rapprochement des résultats obtenus avec la corrélation sur les cosinus en figure 3. Là encore, les trois analyseurs les plus atypiques sont Spacy, MST et Talismane-FTB, tandis qu'un noyau dur des analyseurs fondés sur UD est formé par Stanford, NLPCube et UDPipe-GSD.

Ces grandes tendances se retrouvent au niveau de la comparaison des voisins des différents thésaurus considérés réalisée grâce à la mesure *Rank-Biased Overlap* (Webber *et al.*, 2010) et illustrée par la figure 5. Cette mesure est appliquée à tous les voisins (100) des entrées extraits pour chaque thésaurus et étend la notion de recouvrement moyen – la moyenne du recouvrement entre deux listes pour les différents rangs de ces listes – afin de donner une importance décroissante aux recouvrements à mesure de l'augmentation des rangs, donnant ainsi un poids plus important aux premiers voisins. Cette importance est fixée au travers du paramètre p, vu comme la probabilité de poursuivre le parcours des listes comparées après chaque item depuis leur début. Ainsi, la valeur $p = 0,98$ utilisée ici revient à dire que les 50 premiers voisins concentrent de l'ordre de 85% du poids de la mesure. La figure 5 est obtenue grâce à la distance 1 - RBO, qui a les propriétés d'une métrique.

D'un point de vue qualitatif, nous avons observé 322 mots pour lesquels les 11 modèles sont unanimes concernant leur plus proche voisin. Nous avons pu identifier différents cas de figure parmi ceux-ci :

— des antonymes ou synonymes (dans le domaine TAL) de haute fréquence : *sortie/entrée*, *qualité/performance*, *ordonner/trier*, *valeur/score*, *texte/document* etc.
— des antonymes/synonymes de basse fréquence : *empiler/dépiler*, *intimement/étroitement*, *expérimentalement/empiriquement*, *mélodique/intonatif*, *itérer/réitérer*
— des paires accidentelles que l'on peut expliquer par des contextes exclusifs et systématiques : *parti/leçon* (*tirer_obj*) *routier/hydraulique* (*barrage_mod-1*), *adjacence/covariance* (*matrice_prep_de*), *metteur/mettre* (*scène_prep_en-1*)

À l'inverse, on trouve 10 cas de désaccord total (un plus proche voisin différent pour chaque modèle),

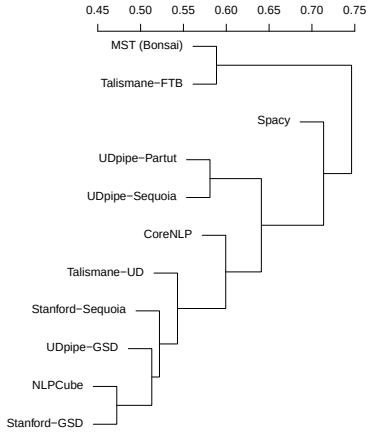

FIGURE 5 – Classification hiérarchique ascendante des modèles en fonction de la mesure RBO

tous de basse fréquence. Dans certains de ces cas, plusieurs des plus proches voisins peuvent être considérés comme pertinents, comme pour *auxiliaire*, où l'on retrouve majoritairement différentes notions grammaticales : *adverbe, déterminant, copule, croisé, gérondif, numéral, transitif, fraction, économiser, laps, subordination*. Dans d'autres cas, le bruit domine nettement, comme pour *post* : *subtilité, multi, bilan, jeudi, billet, chargement, délocaliser, syllabation, SRI, mercredi, pool*, où seul *billet* peut être considéré comme pertinent.

4.3 Évaluation sur un jeu de test *ad hoc*

Nous avons enfin voulu avoir un aperçu, même limité et partiel, des performances relatives de ces différents analyseurs, en évaluant leur capacité à identifier des similarités pertinentes pour le domaine du TAL à partir du corpus. Dans une étude visant à comparer l'impact des différents paramètres gouvernant la construction d'un modèle distributionnel, Tanguy *et al.* (2015) avaient développé un petit jeu d'évaluation sur le corpus TALN en faisant évaluer par quatre juges experts du domaine la pertinence des voisins proposés par un ensemble de systèmes sur 15 mots pivots sélectionnés.

Le jeu de données [4] contient ainsi, pour 5 verbes (*annoter, calculer, décrire, extraire, évaluer*), 5 noms (*fréquence, graphe, méthode, sémantique, trait*) et 5 adjectifs (*complexe, correct, important, précis, spécialisé*) une série de mots déclarés similaires par les juges, avec comme score synthétique pour chaque voisin d'un mot-cible le nombre de juges l'ayant retenu. Par exemple, pour le nom *trait*, on y trouve les mots suivants avec leur score : *attribut* (4), *caractéristique* (4), *propriété* (4), *étiquette* (4), *catégorie* (3), *descripteur* (2), *feature* (2), *indice* (2), *information* (2) ... *marque* (1), *représentation* (1), *structure* (1).

Le jeu est partiel puisque, pour chacun des pivots, seuls les mots ayant été ramenés comme un des trois plus proches voisins par un des systèmes initialement considérés ont été évalués (720 configurations différentes au total). Il est donc possible que des voisins pertinents proposés par un de nos 11 systèmes n'aient pas été considérés ; mais nous supposerons ici que le cas est marginal.

Sur la base de ce jeu de test, nous avons donc calculé pour chaque pivot et pour chaque modèle la

4. Disponible ici : `http://redac.univ-tlse2.fr/datasets/semdis-gold/TAL56-2/`

Modèle	Rang moyen
Talismane-UD	2,38
MST (Bonsai)	2,88
Talismane-FTB	3,00
Stanford-Seq	4,25
UDPipe-Seq	5,75
CubeNLP	6,13
UDpipe-Partut	6,25
UDPipe-GSD	6,38
CoreNLP	7,50
Stanford-GSD	9,38
Spacy	11,00

TABLE 1 – Rang moyen des analyseurs sur le jeu de test de (Tanguy *et al.*, 2015) fondé sur leur score cumulé aux rangs (1, 5, 10, 15, 20, 25, 50, 100)

somme des scores de pertinence des plus proches voisins à différents rangs (1, 5, 10, 15, 20, 25, 50 et 100). L'ordre entre les modèles variant peu au final, nous reportons en table 1 leur rang moyen sur ces différentes positions.

L'ordonnancement des outils laisse voir des extrémités très marquées, dans lesquelles on retrouve en fait les analyseurs les plus excentriques des comparaisons précédentes. Spacy semble celui qui produit les résultats les moins en accord avec l'annotation initiale alors que le trio constitué de MST/Bonsai et des deux versions de Talismane semble se détacher. On remarquera également que si pour Stanford le choix du corpus d'entraînement est critique, ce n'est pas le cas pour UDPipe.

5 Conclusion et perspectives

Les variations que nous avons mesurées entre des modèles sémantiques ne différant que par les analyseurs syntaxiques utilisés sont très importantes : le choix d'un analyseur syntaxique n'est certainement pas anodin pour l'analyse distributionnelle de petits corpus spécialisés. Les différences observées en fin de chaîne, lorsque l'on compare les voisins distributionnels, ne sont pas nécessairement corrélées à celles que l'on trouve en comparant directement les sorties. On a toutefois pu observer que tous les mots, contextes et paires de mots similaires ne sont pas tous affectés de la même façon par le changement d'analyseur. Cette première étape aura néanmoins permis d'identifier, parmi les différents outils testés, ceux dont les comportements sont les plus différents. Des sondages plus pointus et des examens manuels des cas de désaccord devraient nous permettre d'identifier les contextes syntaxiques les plus critiques, et par là même nous aider à choisir l'analyseur le plus adapté dans ce dispositif.

Remerciements

Le travail présenté dans cet article a été réalisé dans le cadre du projet ADDICTE[5] (Analyse distributionnelle en domaine spécialisé), financé par l'Agence Nationale de la Recherche (ANR-17-CE23-0001). Les auteurs tiennent en outre à remercier tous les membres partenaires du projet ADDICTE, et plus particulièrement Nabil Hathout pour son aide pour la relemmatisation des sorties.

5. https://anr-addicte.ls2n.fr/

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In *Treebanks*, p. 165–187. Springer.

BARONI M., DINU G. & 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)*, p. 238–247, Baltimore, Maryland.

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

BOROŞ T., DUMITRESCU S. D. & 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*, p. 171–179, Brussels, Belgium : Association for Computational Linguistics.

BOSCO C., SANGUINETTI M. & LESMO L. (2012). The parallel-TUT : a multilingual and multiformat treebank. In *Proceedings of LREC*, p. 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 J.-M. PIERREL, Ed., *Actes de TALN 2002 (Traitement automatique des langues naturelles)*, p. 75–84, Nancy : ATALA ATILF.

CANDITO M., NIVRE J., DENIS P. & ANGUIANO E. H. (2010). Benchmarking of statistical dependency parsers for French. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, p. 108–116 : Association for Computational Linguistics.

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

DE LA CLERGERIE É. V. (2014). Jouer avec des analyseurs syntaxiques. In *Actes de TALN*.

DE LA CLERGERIE E. V., HAMON O., MOSTEFA D., AYACHE C., PAROUBEK P. & VILNAT A. (2008). Passage : from French parser evaluation to large sized treebank. In *Proceedings of LREC*.

DENIS P. & 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*, p. 22–33.

FERRET O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. In *17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Canada.

HONNIBAL M. & 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*, p. 1373–1378, Lisbon, Portugal : Association for Computational Linguistics.

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

KOO T., CARRERAS X. & COLLINS M. (2008). Simple semi-supervised dependency parsing. *Proceedings of ACL-08 : HLT*, p. 595–603.

LAPESA G. & 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 : Volume 2, Short Papers*, volume 2, p. 394–400.

LEVY O. & GOLDBERG Y. (2014). Dependency-Based Word Embeddings. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, p. 302–308, Baltimore, Maryland.

LEVY O., GOLDBERG Y. & DAGAN I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, **3**, 211–225.

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

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & 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*, p. 55–60.

MCDONALD R., LERMAN K. & PEREIRA F. (2006). Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, p. 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 (Volume 2 : Short Papers)*, p. 92–97.

MIKOLOV T., CHEN K., CORRADO G. & 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. & CURRAN J. R. (2012). Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, **194**, 151–175.

PADÓ S. & LAPATA M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, **33**(2), 161–199.

PAROUBEK P., ROBBA I., VILNAT A. & AYACHE C. (2008). EASY, Evaluation of Parsers of French : what are the results ? In *Proceedings of LREC*.

PIERREJEAN B. & 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*, p. 32–39.

QI P., DOZAT T., ZHANG Y. & 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*, p. 160–170, Brussels, Belgium : Association for Computational Linguistics.

ROMARY L., SALMON-ALT S. & FRANCOPOULO G. (2004). Standards going concrete : from LMF to Morphalou. In *COLING 2004 Enhancing and using electronic dictionaries*, p. 22–28, Geneva, Switzerland : COLING.

SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of LREC*, Valletta, Malta : 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. & 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*, p. 146–182, Seattle, Washington, USA : Association for Computational Linguistics.

STRAKA M. & 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*, p. 88–99, Vancouver, Canada : Association for Computational Linguistics.

TANGUY L. & HATHOUT N. (2007). *Perl pour les linguistes*. Hermès.

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

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

WEBBER W., MOFFAT A. & ZOBEL J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, **28**(4), 1–38.

ZEMAN D., HAJI J., POPEL M., POTTHAST M., STRAKA M., GINTER F., NIVRE J. & 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*, p. 1–21.