IMT Atlantique

Dépt. Image & traitement de l'information Technopôle de Brest-Iroise - CS 83818

29238 Brest Cedex 3

Téléphone: +33 (0)2 29 00 13 04 Télécopie: +33 (0)2 29 00 10 12 URL: www.imt-atlantique.fr



Collection des rapports de recherche d'IMT Atlantique

Etude des performances de POS-taggers appliqués à la langue française

CHARPENTIER Lucas FISE 1A

HARALAMBOUS Yannis EL MASDOURI Achraf

Date d'édition : 7 janvier 2020

Version: 1.1



Sommaire

1.	Intro	oduction	4		
2.	Anal	yse et Comparaison des systèmes d'étiquettes	4		
3.	Ana	yse et Comparaison des systèmes de traits	6		
4.	Forn	ne des résultats	9		
5.	Comparaison des précisions des taggers				
6.	Com	paraison des performances temporelles des taggers	11		
7.	Conclusion				
An	nexes	F	14		
An	nexe	1 – Informations sur les logiciels utilisés dans l'étude	14		
		Spacy			
		1.1.1. Version & Modèle			
	1.2.	Stanford POStagger			
		1.2.1. Version & Modèle			
	1.3.				
		1.3.1. Version & Modèle	14		
	1.4.	TreeTagger	14		
		1.4.1. Version & Modèle			
	1.5.	RNNTagger			
D٨	fáran	cas	15		

Liste des figures

1.	Pourcentage d'erreur en fonction du tag
	Liste des tableaux
1.	Le système Universal POS-tags
2.	Le tagset FrenchTreebank
3.	Le tagset TreeTagger
4.	Les traits associés au French TreeBank
5.	Les traits morphologiques du système Universal Dependencies utilisés par Spacy 8
6.	Les traits du système de tags Treetagger
7.	Les formes des réponses des taggers
8.	Nombre d'étiquettes dans le corpus
9.	Les traits du système de tags Treetagger
10.	Mesure de l'erreur des taggers (en %)
11.	Mesure du temps de traitement d'un fichier (en s)

Glossaire

finie Propriété d'un verbe pouvant être à la racine d'une proposition indépendante du reste du texte.. 7 **French TreeBank** corpus de référence de l'étude [2]. 2, 4, 5, 6, 7, 9, 10, 11

POS littéralement Part Of Speech, partie du discours. 4

tagset traduction anglaise de l'expression "ensemble d'étiquettes". 4

Universal Dependencies abrégé UD [4], il s'agit d'un cadre d'annotations grammaticales interlinguistiques cohérentes mais fondé sur un effort communautaire ayant conduit à la création plus de 100 banques d'arbres dans plus de 70 langues.. 2, 4, 6, 7, 9

1. Introduction

Les outils mis à disposition en ligne pour estimer les parties du discours (POS) dans les textes français se multiplient et leur dernière comparaison [5] n'est plus utilisable de nos jours. Ce rapport rend compte de l'état de l'art des outils de POS-tagging applicables à la langue française. 4 outils ont été sélectionnés pour cette étude : Spacy, le Stanford POS-Tagger, Treetagger et Talismane. Les études se baseront sur les configurations par défault de ces outils. Les versions utilisées pour l'étude sont détaillées dans l'annexe 1. Tout d'abord, je tenterai de synthétiser les différences entre les modèles de représentation des connaissances adoptés par chaque outil. Ensuite, je comparerai sur différents critères tels que la complexité en temps et la précision du résultat les performances de chacun des outils. Le corpus utilisé pour effectuer les comparaisons est le French TreeBank et le langage utilisé pour l'interprétation est Python3. Ce projet est à considérer avec l'Étude sur l'efficacité des outils d'analyse syntaxiquedes textes en français[1] rédigé par Mr. EL MASDOURI Achraf qui compare les outils de dépendances syntaxiques.

2. Analyse et Comparaison des systèmes d'étiquettes

Pour de nombreuses langues et en particulier pour la langue française, un système d'étiquettes universel Universal Dependencies a été instauré en Octobre 2014. Cet ensemble d'étiquettes particulier est partagé par plusieurs langues.

Classe ouverte	Signification	Exemples	
ADJ	adjectif	vert, troisième, européen	
ADV	adverbe	ensemble, très	
INTJ	interjection	bref, bon	
NOUN	nom commun	arbre, chat	
PROPN	nom propre	Pierre, ONU, Mexique	
VERB	verbe	manger, voir	
Classe fermée			
ADP	prépositions (adpositions)	pour, de	
AUX	auxilliaire	être, avoir, faire.	
CCONJ	Conjonction de coordination	mais,ou,et,donc	
SCONJ Conjonction de subordination		quand, parce que (ADV SCONJ)	
DET	déterminant	le, mon, quel, aucun	
NUM	nombre(parfois abandonné pour ADJ ou DET)	quatre, IV, 4	
PART	mot-fonction autre	ci, là, t euphonique	
PRON	pronom	je,ceux,me,que	
Autres			
PUNCT	signe de ponctuation	. ,; :	
SYM	symboles	€,:), http://universaldependencies.org/	
X	acronymes, mots étrangers, autres	etc, french, xfgh	

Table 1 – Le système Universal POS-tags

Les classes ouvertes sont par définition des classes lexicales dont le nombre de membres est indéfini, c'est à dire qu'il est simple de rajouter un élément de cette classe à la langue. Au contraire, une classe fermée est une classe contenant un nombre défini d'éléments dont le sens limite la multiplicité. Cet ensemble d'étiquettes est fréquemment complété par une étiquette "SPACE" qui permet de traiter les cas pathologiques de segmentation sans encombre. Le Universal tagset est utilisé par plusieurs outils :

- Spacy utilise le système Universal Dependencies.
- Talismane utilise le système Universal Dependencies.

Le Stanford tagger[3], créé en 2004, utilise son propre ensemble d'étiquettes, qui existait bien avant le tagset Universal Dependencies. Il est basé sur le tagset du French TreeBank, projet initié en 1997 [2] :

Classe	Signification	
N	nom	
A	adjectif	
ADV	adverbe	
P	préposition	
D	déterminant	
CL	pronom clitique	
PRO	pronom non clitique	
PREF	préfixe	
C	conjonction	
I	interjection	
V	verbe	
ET	mot étranger	
PONCT	ponctuation	

Table 2 – Le tagset FrenchTreebank

Dans la documentation du FrenchTreeBank, Anne Abeillé et Lionel Clément [2] détaillent les usages des différents tags, en particulier celui du tag ET :

On ne note ET que si rien dans le contexte ne permet de se prononcer sur la catégorie du mot, dans une citation par exemple.

De plus, dans les faits, une notation composite a été utilisée pour des mots comme "du" ou "duquel" étant les abréviations de deux mots à natures distinctes : "P+D" et "P+PRO". Ces notations sont propres au French TreeBank et pour l'analyse des performances des taggers, on les simplifiera par :

- 1 "P+D" = 1"P" + 1"D"
- 1 "P+PRO" = 1"P" + 1"PRO"

Treetagger utilise quant à lui un ensemble d'étiquettes qui lui est spécifique.

Classe	Signification
ABR	abréviation
ADJ	adjectif
ADV	adverbe
DET	déterminant
INT	interjection
KON	conjonction
NAM	nom propre
NOM	nom commun
NUM	chiffre
PRO	pronom
PRP	préposition
PUN	ponctuation
SENT	fin de phrase (signe de ponctuation)
SYM	symbole
VER	verbe

Table 3 – Le tagset TreeTagger

On constate que des trois ensembles d'étiquettes, le système Universal POS-tags est le plus précis. Cependant, la plupart des taggers renvoient des informations supplémentaires en complément du tag. Continuons notre tour d'horizon de ces trois système en analysant la nature de ces informations supplémentaires.

3. Analyse et Comparaison des systèmes de traits

Que ce soit sous le nom de trait, indication morphologique ou sous la forme de complément d'étiquette, chaque Tagger relève en plus de l'étiquette d'un mot un ensemble d'informations sur ce mot. On utilisera ici le terme "trait" pour désigner une de ces informations.

Le French TreeBank possède des tags qui palient au manque de certaines catégories par rapport au système Universal Dependencies comme par exemple le tag de sous-catégorie de la classe C qui évite d'avoir à distinguer les classes CCONJ et SCONJ. Cependant, il manque toujours à cet ensemble de tags la possibilité de différencier les AUX. Bien que cette information soit marginale en français au vu du travail à faire sur les formes verbales composées, elle reste utile dans certaines circonstances. Les combinaisons "étiquette+traits" exposées par le FTB atteignent le nombre de 252.

Classe	Traits associés		
A	sous-catégorie : qualificatif, cardinal, indéfini, interrogatif, ordinal, possessif genre : f, m		
	nombre : s, p (sous-catégorie=possessif) personne : 1s, 2s, 3s, 1p, 2p, 3p		
ADV	sous-catégorie : exclamatif, négatif, interrogatif, démonstratif,∅		
CL	sous-catégorie : sujet, objet, réfléchi genre : f, m personne : 1s, 2s, 3s, 1p, 2p, 3p		
С	sous-catégorie : C (coordination), S (subordination)		
DET	sous-catégorie : démonstratif, défini, indéfini, interrogatif, exclamatif, cardinal, négatif, partitif, possessif genre : f, m nombre : s, p (sous-catégorie=possessif) personne : 1s, 2s, 3s, 1p, 2p, 3p		
PRO sous-catégorie : démonstratif, indéfini, interrogatif, négatif, cardinal, possessif, personnel, réfléchi, relatif genre, nombre et personne (selon les sous-catégories)			
N	sous-catégorie : cardinal, commun, propre genre : f, m nombre : s, p		
V	temps : C,F,G, [2] personne : 1s, 2s, 3s, 1p, 2p, 3p (temps=participe passé) genre et nombre		
PONCT	forte (S) ou faible(W)		

Table 4 – Les traits associés au French TreeBank

Pour l'analyse des traits du système Universal Dependencies, on étudie la liste définie dans la tag-map de la configuration française de Spacy. En effet, les traits du système Universal Dependencies [4] sont innombrables et pour la plupart non adaptés à la langue française. On constate que les *ADP* (prépositions) recoivent des attributs genre,nombre et personne alors qu'elles n'avaient pas de traits dans le French TreeBank. Les combinaisons "étiquette+traits" exposées par le Spacy avec Universal Dependencies atteignent le nombre de 200.

Classe	Traits associés
ADJ	type numérique : cardinal,ordinal,0 (non numérique)
	genre : Fem, Masc
	nombre : Sing, Plur
ADP	genre,nombre,personne: Fem,Plur,3; Masc,Plur,3; Masc,Sing,3; @
ADV	Polarité : Neg, Ø
	Type : interrogatif, ∅
AUX	temps : passé, présent, futur, imparfait
	voix : passive, \emptyset (active)
	mode: indicatif, conditionnel, subjonctif
	forme: finie, infinitif, adjectif verbal (Part)
	genre : Fem, Masc
	nombre : Sing, Plur
	personne: 1,2,3
DET	caractère défini : défini, indéfini
	sous-catégorie : article, interrogatif, démonstratif
	possessif: vrai, Ø (faux)
	genre : Fem, Masc
	nombre : Sing, Plur
NOUN	type numérique : cardinal, ∅
1,001,	genre : Fem, Masc
	nombre : Sing, Plur
NUM	genre, nombre : Masc, Plur
110111	type numérique : cardinal
PRON	sous catégorie : relatif, personnel, démonstratif
FRON	réflexivité : vrai, \emptyset (faux)
	genre : Fem, Masc
	nombre : Sing, Plur
	personne : 1, 2, 3
PROPN	genre: Fem, Masc
	nombre : Sing, Plur
V	temps : passé, présent, futur, imparfait,
	voix : passive, \emptyset (active)
	mode : indicatif, conditionnel, subjonctif, impératif
	forme: finie, infinitif, adjectif verbal (Part)
	genre: Fem, Masc
	nombre : Sing, Plur
	personne: 1,2,3

Table 5 – Les traits morphologiques du système Universal Dependencies utilisés par Spacy

A l'opposé des deux systèmes précédents, la configuration française de Treetagger décide de prendre le parti d'un système de traits contenant peu d'informations. Cela permet d'améliorer les performances temporelles de l'algorithme. Cependant, plus de travail va être nécessaire dans les étapes de traitement du langage en aval comme l'analyse syntaxique [1]. Treetagger ne comprend que 33 combinaisons "étiquette+traits".

Classe	Traits associés	
DET article ou pronom possessif		
PRON sous-catégorie : démonstratif, indéfini, personnel, relatif, posse		
PREP "det" s'il s'agit d'une contraction : au, aux, du, de		
PUN	"cit" si il s'agit d'une ponctuation de citation	
V	temps : conditionnel, futur, impératif, imparfait, infinitif, participe passé, participe présent, passé simple, subjonctif imparfait, subjonctif présent	

Table 6 – Les traits du système de tags Treetagger

4. Forme des résultats

De manière à conclure sur la comparaison des modes de représentation des informations morphologiques, on détaille les formats des réponses renvoyées par les différents taggers : Chaque forme possède des avantages

Outil	Forme de la réponse	
Spacy	liste de Tokens, classe du module Spacy Python, comprenant à la fois les étiquettes, les traits et des méthodes diverses	
Stanford	Les tags sont directement intégrés dans le texte initial.	
Talismane	le résultat est donné sous une convention spécifique à talismane qui est cohérente avec l'étape suivante du processus d'analyse complet de talismane.	
Treetagger	retourne une liste de mots suivis de leurs étiquettes+traits et de leur lemme.	

Table 7 – Les formes des réponses des taggers

et des inconvénients, sur les critères de la complexité, de la complétude et de la compatibilité. Les deux outils partageant le même ensemble d'étiquettes (Spacy et Talismane), utilisent tous deux leurs propres formats de données en sortie.

On constate que les ensembles d'étiquettes fûrent tous concus avec une approche différente du problème en tête. Treetagger est bien plus léger que les deux autres, ce qui permet de l'étudier plus rapidement. Le système Universal Dependencies contient de nombreuses informations dans les tags mais son système de traits est peu détaillé en comparaison de celui associé au French TreeBank qui compense ses étiquettes peu nombreuses.

5. Comparaison des précisions des taggers

Les taggers ne fournissent a priori des informations comparables. Cependant, grâce à l'étude menée ci-dessus, on peut tenter de comparer leurs performances en termes de précision. De manière à éviter les problèmes sur l'interprétation de certains mots du corpus source, on décide d'utiliser les segmenteurs mis à disposition avec chaque tagger sur le texte initial plutôt que de le donner déjà segmenté aux taggers. Ce choix permet d'évaluer l'efficacité du tagger tel qu'il est utilisé dans la majorité des cas.

Cependant, ce choix a un coût : chaque segmenteur traite le texte d'une manière différente et donc les étiquettes obtenues en sortie des taggers ne sont pas comparables une à une. Par exemple, on peut observer les nombres de segments totaux trouvés par chaque segmenteur dans la table8. On constate jusqu'à une augmentation de 6,26% du nombre d'étiquettes. C'est à dire qu'en moyenne, les annotations fournies par Spacy vont prendre du retard sur celles du texte initial à hauteur de une étiquette tous les 20 mots. Pour

Annotations initiales	Spacy	Stanford	TreeTagger	Talismane
644595	684995	624990	653772	649931

Table 8 – Nombre d'étiquettes dans le corpus

comparer les précisions, on analyse l'entièreté du corpus avec chacun des taggers et on tente de simplifier tous les résultats à un système d'étiquettes simplifié obtenu à partir du système d'étiquettes Universal Dependencies. Cette simplification a été étudiée dans la perspective d'une ontologie concernant les outils. Dans le tableau 9, on a synthétisé une partie des informations contenues dans l'ontologie : chaque étiquette A présente sur la ligne de l'étiquette de première colonne S rend compte d'une relation "A is a S". De même, chaque étiquette A présente sur la colonne du système d'étiquettes E traduit l'existence de la relation : "A is part of E".

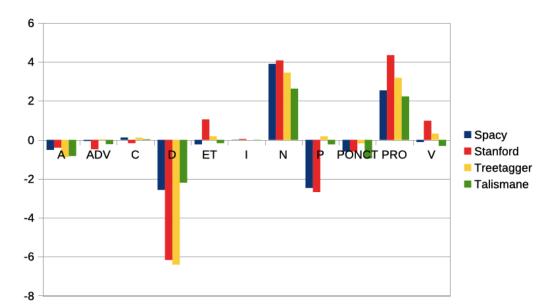
Sachant que les étiquettes ne sont pas "alignés", on ne peut vérifier une à une les relations de cette ontologie, mais seulement inférer des fréquences de chaque étiquette simplifié des résultats.

Il y a perte d'information lors de la simplification, mais cela permet à la comparaison d'être sensée. Il faut en particulier abandonner les étiquettes suivantes

- Le traitement de la classe *PREF* préfixe (Cf. table 2) est abandonné car dépendant grandement de la segmentation du texte et ayant peu d'influence sur le reste des valeurs. En effet, il n'y a que 466 préfixes dans le corpus sur plus de 600 000 annotations et voici la définition donnée par Anne Abeillé et Lionel Clément[2], qui justifie le fait que certains outils ne puissent détecter d'équivalent à un préfixe uniquement à cause de l'étape de segmentation :
 - L'étiquette PREF est employée pour catégoriser des mots du type : méta-, anti-, franco-, auto-, ex-, super- puis d'une liste ouverte de termes. Le préfixe ne doit pas être composant d'un terme lexical. Le tiret fait partie du préfixe.
- Le tag "SPACE" (Cf. table 1) utilisé par Spacy est lui aussi négligé car il est uniquement dépendant de la logique de segmentation. De fait, aucune comparaison avec les autres taggeurs n'est possible car Spacy est le seul à considérer ces segments du texte comme parties du discours.

Système UD simplifié	Système French TreeBank	Système UD	Système Treetagger
A	A	ADJ	ADJ
ADV	ADV	ADV	ADV
C	C	CCONJ, SCONJ	KON
D	D (P+D)	DET	DET
ET	ET	X	ABR
I	I	INTJ	INT
N	N	NOUN, PROPN, NUM	NAM, NOM, NUM
P	P(P+D, P+PRO)	ADP	PRP
PONCT	PONCT	PUNCT, SYM	SYM, PUN, SENT
PRO	PRO, CL (P+PRO)	PRON	PRO
V	V	VERB, AUX	VER

Table 9 – Les traits du système de tags Treetagger



Après avoir traité l'intégralité du French TreeBank, on obtient les résultats suivants :

FIGURE 1 – Pourcentage d'erreur en fonction du tag

Lire : au moins 6% des mots sont annotés dans le French TreeBank comme déterminants (D) et ni Stanford ni Treetagger ne les détectent comme déterminants.

On constate ici que le plus grand nombre d'erreurs d'étiquetage est lié à la détection des déterminants, noms et pronoms. En sommant tous les résultats précédents, on obtient des mesures du pourcentage d'erreur de chaque tagger. Sachant que les segmentations sont différentes, cette mesure de la précision est l'une des meilleures que l'on puisse effectuer. Avec une segmentation commune, on aurait pu comparer les étiquettes associés à chacun des segment et obtenir des erreurs plus précises. Cependant, certains taggers auraient été désavantagé du fait de la forme des segments qui ne correspond pas exactement à leur fonctionnement optimal.

Spacy Stanford		Treetagger	Talismane
6.57	10.50	7.47	4.92

Table 10 – Mesure de l'erreur des taggers (en %)

Selon ce critère, le Stanford tagger est le moins précis des quatre outils, tandis que Talismane est le plus précis.

Cependant, la précision n'est pas le seul critère à apprécier dans un tel outil. Sa performance est également importante.

6. Comparaison des performances temporelles des taggers

Pour cette comparaison, on a sélectionné de manière empirique un texte du corpus et on a comparé les temps de traitement des différents algorithmes. Voici les résultats de l'usage des modules Python cProfile et pstats :

Taille de l'échantillon	Spacy	Stanford	Treetagger	Talismane
1 document du corpus	2.631	4.166	2.449	31.343

Table 11 – Mesure du temps de traitement d'un fichier (en s)

Ces mesures ne sont pas exactes du fait des commandes annexes utilisées de manière à appeler les différents outils dans le script python. Cependant, on constate que (sur le même fichier que précédemment) le temps d'exécution de tels outils est de l'ordre du dixième de seconde. On peut donc faire le constat quantitatif que l'algorithme Talismane est bien plus lent que ses pairs, avec une vitesse de l'ordre de 10 fois inférieure. De plus, Spacy et Treetagger sont les deux POS taggers les plus rapides à analyser un fichier, avec un temps d'exécution voisin.

On notera cependant que Spacy est un outil développé pour le langage Python, et ne subit donc pas la surcouche algorithmique liée au changement de langage.

7. Conclusion

Talismane est le tagger le plus précis, mais cela se paye par un temps d'éxecution bien plus long. Au contraire, Treetagger est le plus rapide des quatre algorithmes d'étiquettage, mais les informations qu'il fournit ne sont pas aussi approfondies que tous les autres du fait des lacunes de son système d'étiquettes. On constate que les tags qu'il faut vérifier en priorité après le traitement d'un texte par de tels outils (en configuration standard), sont les déterminants et les pronoms en une moindre mesure pour Spacy et le Stanford POS-tagger.

L'usage du Stanford POStagger n'est selon les résultats de cette étude plus intéressant au vu des autres outils à disposition aujourd'hui.

Une comparaison plus pousséede la précision des taggers pourrait être obtenue à partir d'une analyse excluant l'étape de segmentation.

Annexes

Annexe 1 – Informations sur les logiciels utilisés dans l'étude

L'étude a été faite sous un macOS Mojave Version 10.14.4 (18E226). Les logiciels ont été légèrement adaptés pour leur utilisation dans cet environnement.

1.1. Spacy

1.1.1. Version & Modèle

Version: 2.0.18

Modèle: fr_core_news_sm

1.2. Stanford POStagger

1.2.1. Version & Modèle

Version: Stanford POS Tagger, v3.9.2 - 2018-10-16

Modèle : french-ud.tagger

1.3. Talismane

1.3.1. Version & Modèle

Fichier de configuration : talismane-fr-ud-output-5.2.0.conf

1.4. TreeTagger

1.4.1. Version & Modèle

Version: tree-tagger-MacOSX-3.2

Configuration: française.

Remarque: Le wrapper de Treetagger en Python disponible en ligne a été utilisé, en version 2.3.

1.5. RNNTagger

L'étude avait pour objectif d'analyser RNNTagger, variante de Treetagger basée sur un algorithme de deep-learning. Cependant, l'installation ayant été infructueuse, cette étude a été abandonnée.

Références

- [1] Achraf El Masdouri, Étude sur l'efficacité des outils d'analyse syntaxique des textes en français, Juin 2019
- [2] Anne Abeillé et Lionel Clément, *ANNOTATION MORPHO-SYNTAXIQUE*; Les mots simples Les mots composés, LLF Université Paris 7, 10 janvier 2003
- [3] Kristina Toutanova et Christopher D. Manning, *Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger*, Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 2000
- [4] Slav Petrov, Dipanjan Das et Ryan McDonald, *A Universal Part-of-Speech Tagset*, ArXiv:1104.2086., 11 Avril 2011
- [5] A. Allauzen et H. Bonneau-Maynard, Training and evaluation of POS taggers on the French MULTITAG corpus, Proceedings of the Sixth Language Resources and Evaluation (LREC'08), Language Resource and Evaluation Conference, Marrakech, 28-30 May 2008

OUR WORLDWIDE PARTNERS UNIVERSITIES - DOUBLE DEGREE AGREEMENTS

3 CAMPUS, 1 SITE





IMT Atlantique Bretagne-Pays de la Loire - http://www.imt-atlantique.fr/

Campus de Brest

Technopôle Brest-Iroise CS 83818 29238 Brest Cedex 3 France T +33 (0)2 29 00 11 11

F +33 (0)2 29 00 10 00

Campus de Nantes

4, rue Alfred Kastler CS 20722 44307 Nantes Cedex 3 France

T +33 (0)2 51 85 81 00 F +33 (0)2 99 12 70 08

Campus de Rennes

2, rue de la Châtaigneraie CS 17607 35576 Cesson Sévigné Cedex France T +33 (0)2 99 12 70 00 F +33 (0)2 51 85 81 99

Site de Toulouse

10, avenue Édouard Belin BP 44004 31028 Toulouse Cedex 04 France T +33 (0)5 61 33 83 65

