Part-Of-Speech Taggers Features In French Learner Texts

Nadezhda Barymova
Petrozavodsk State University
Petrozavodsk, Russia
nbarymova@gmail.com

Anna Zhestkova
Petrozavodsk State University
Petrozavodsk, Russia
zhest1806anna@gmail.com

Nadezhda Oulianova
Petrozavodsk State University
Petrozavodsk, Russia
nad.ulyanova2013@yandex.ru

Olga Nikiforova
Petrozavodsk State University
Petrozavodsk, Russia
o_nik75@rambler.ru

Abstract—**This paper presents investigation result of the comparative analysis of TreeTagger, CoreNLP and SpaCy as part-of-speech taggers for processing texts in French. Acting in the framework of a Learner Texts Corpus Creation Project a group of researchers from the Institute of Foreign Languages together with the Institute of Mathematics and Information Technologies of Petrozavodsk State University analyzed the above mentioned tagging tools mainly being focused on defining the most effective one.**

I. INTRODUCTION

Corpus technologies are actively developing and find their application in various fields, from linguistic researches to their use in the foreign languages teaching techniques. [1], [2]

The most important component of the corpus in any language is a tagset, which allows the morphological analyzer (morphologizer) automatically process linguistic objects in a formalized form [3].

The main objective of our research is to select the most suitable tagger for processing a text in French. To tackle this problem we:

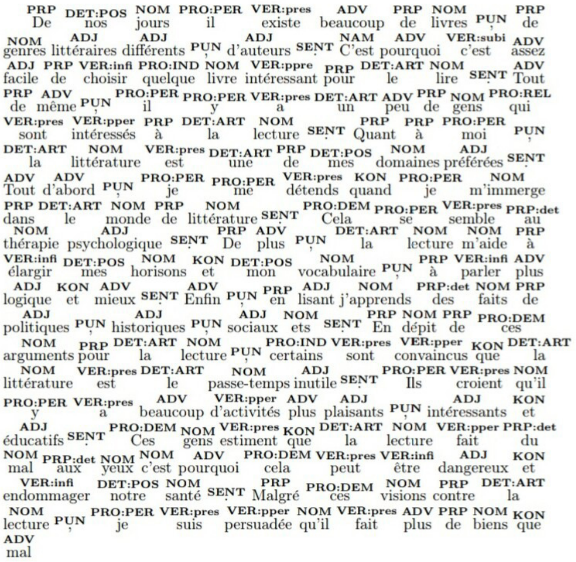
- analyzed 30 learner texts in French using the most famous and affordable taggers TreeTagger, CoreNLP and SpaCy.
- highlighted the main errors (separation of words, highlighting parts of speech, etc.) made by each of the specified taggers.
-compared the tagging results and highlight the tagger with the least number of errors
- selected the most appropriate tagger to use in work with the corpus of learner texts in French using theanalysis results

II. MAIN PART

During the research work we analyzed 30 learner texts of different types such as essays, motivation letters, topics, article evaluation using TreeTagger, CoreNLP and SpaCy taggers. All the texts under analysis were written by the students of the Institute of Foreign Languages of Petrozavodsk State University. The analysis of the texts made it possible to obtain each tagger tools comprehensive evaluation.

TreeTagger [4] is the only tool with tags in French among the three part-of-speech taggers under consideration which proves to be its main advantage comparing to other tools. Considering this part-of-speech tagger as a possible version of a tool for working with texts, it should be taken into account that it is not very efficient from the point of view of visual perception.

The reasons for that can be mentioned as the follows: first, the tagger font color is black, second, the tags are located directly above the words making the text less readable so you have to read the whole text while paying a special attention to the superscripts, and finally, the body of the text is not split into separate sentences, which also complicates the process of checking the tagger.

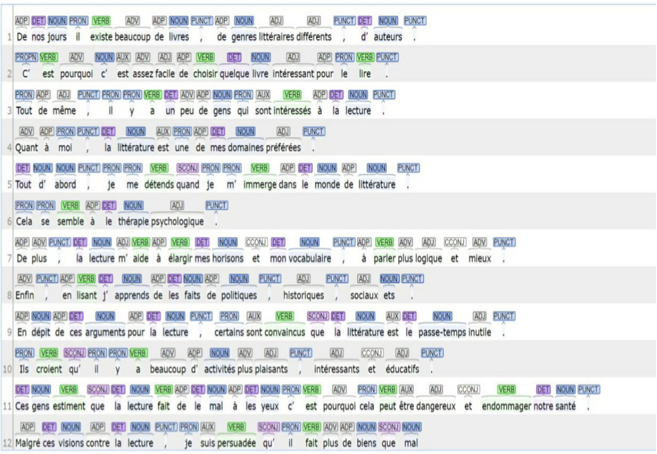


```
PRP DET:POS NOM PRO:PER VER:pres ADV PRP NOM PUN PRP
De nos jours il existe beaucoup de livres , de
NOM ADJ ADJ PUN ADJ SENT NAM ADV VER:subi ADV
genres littéraires différents , d’auteurs . C’est pourquoi c’est assez
ADJ PRP VER:infi PRO:IND NOM VER:ppre PRP DET:ART NOM SENT ADV
facile de choisir quelque livre intéressant pour le lire . Tout
PRP ADV PUN PRO:PER PRO:PER VER:pres DET:ART ADV PRP NOM PRO:REL
de même , il y a un peu de gens qui
VER:pres VER:pper PRP DET:ART NOM SENT PRP PRP PRO:PER PUN
sont intéressés à la lecture . Quant à moi ,
DET:ART NOM VER:pres DET:ART PRP DET:POS NOM ADJ SENT
la littérature est une de mes domaines préférées .
ADV ADV PUN PRO:PER PRO:PER VER:pres KON PRO:PER NOM
Tout d’abord , je me détends quand je m’immerge
PRP DET:ART NOM PRP NOM SENT PRO:DEM PRO:PER VER:pres PRP:det
dans le monde de littérature . Cela se semble au
NOM ADJ SENT PRP ADV PUN DET:ART NOM NOM PRP
thérapie psychologique . De plus , la lecture m’aide à
VER:infi DET:POS NOM KON DET:POS NOM PUN PRP VER:infi ADV
élargir mes horisons et mon vocabulaire , à parler plus
ADJ KON ADV SENT ADV PUN PRP ADJ NOM PRP:det NOM PRP
logique et mieux . Enfin , en lisant j’apprends des faits de
ADJ PUN ADJ PUN ADJ NOM SENT PRP NOM PRP PRO:DEM
politiques , historiques , sociaux ets . En dépit de ces
NOM PRP DET:ART NOM PUN PRO:IND VER:pres VER:pper KON DET:ART
arguments pour la lecture , certains sont convaincus que la
NOM VER:pres DET:ART NOM ADJ SENT PRO:PER VER:pres NOM
littérature est le passe-temps inutile . Ils croient qu’il
PRO:PER VER:pres ADV VER:pper ADV ADJ PUN ADJ KON
y a beaucoup d’activités plus plaisants , intéressants et
ADJ SENT PRO:DEM NOM VER:pres KON DET:ART NOM VER:pper PRP:det
éducatifs . Ces gens estiment que la lecture fait du
NOM PRP:det NOM NOM ADV PRO:DEM VER:pres VER:infi ADJ KON
mal aux yeux c’est pourquoi cela peut être dangereux et
VER:infi DET:POS NOM SENT PRP PRO:DEM NOM PRP DET:ART
endommager notre santé . Malgré ces visions contre la
NOM PUN PRO:PER VER:pres VER:pper NOM VER:pres ADV PRP NOM KON
lecture , je suis persuadée qu’il fait plus de biens que
ADV
mal
```

Fig.1. Screenshot of the tagged text

CoreNLP - one of the main advantages of this tagger is its visual presentation. So due to this peculiarity all the sentences in the text beginning on a new line are also numbered. This feature simplifies the word search. Moreover, this immediately allows the reader to spot/identify/ the sentence types presented in the text. At that tags for tokens are highlighted in colour, which makes it more convenient to define parts of speech.

Nevertheless it should be noted that the texts tagging is done in English making the analysis on the one hand more simple, but imprecise at the same time.

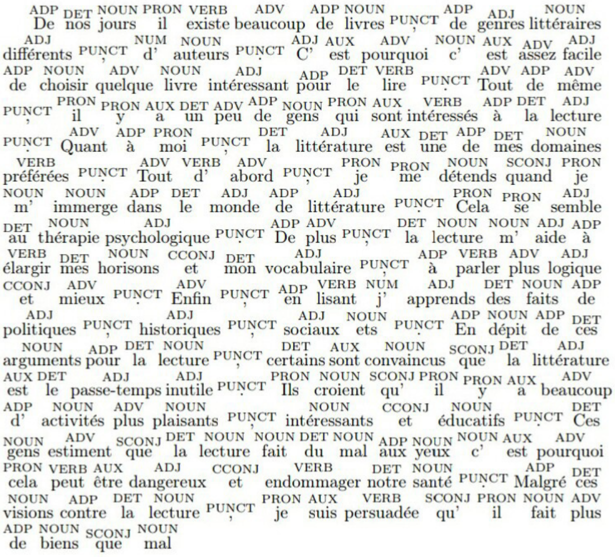


```
   ADP DET NOUN PRON VERB  ADV      ADP NOUN PUNCT ADP NOUN   ADJ        ADJ        PUNCT DET NOUN   PUNCT
1  De  nos jours il   existe beaucoup de  livres ,   de  genres littéraires différents ,   d'  auteurs .
   PROPN VERB ADV      NOUN AUX ADV   ADJ    ADP VERB    DET     NOUN  ADJ          ADP  PRON VERB PUNCT
2  C'    est  pourquoi c'   est assez facile de  choisir quelque livre intéressant pour le   lire  .
   PRON ADP ADJ   PUNCT PRON PRON VERB DET ADV ADP NOUN PRON AUX  VERB       ADP DET NOUN    PUNCT
3  Tout de  même  ,     il   y    a    un  peu de  gens qui  sont intéressés à   la  lecture .
   ADV   ADP PRON PUNCT DET NOUN       AUX PRON ADP DET NOUN     ADJ       PUNCT
4  Quant à   moi  ,     la  littérature est une  de  mes domaines préférées .
   DET  NOUN NOUN  PUNCT PRON PRON VERB    SCONJ PRON PRON VERB    ADP  DET NOUN  ADP NOUN       PUNCT
5  Tout d'   abord ,     je   me   détends quand je   m'   immerge dans le  monde de  littérature .
   PRON PRON VERB   ADP DET NOUN     ADJ           PUNCT
6  Cela se   semble à   le  thérapie psychologique .
   ADP ADV  PUNCT DET NOUN    ADJ VERB ADP VERB    DET NOUN     CCONJ DET NOUN        PUNCT ADP VERB   ADV  ADJ     CCONJ ADV   PUNCT
7  De  plus ,     la  lecture m'  aide à   élargir mes horisons et    mon vocabulaire ,     à   parler plus logique et    mieux .
   ADV   PUNCT ADP VERB   DET NOUN     ADP DET NOUN  ADP NOUN       PUNCT ADJ         PUNCT ADJ     NOUN PUNCT
8  Enfin ,     en  lisant j'  apprends de  les faits de  politiques ,     historiques ,     sociaux ets  .
   ADP NOUN  ADP DET NOUN      ADP  DET NOUN    PUNCT PRON     AUX  VERB       SCONJ DET NOUN        AUX DET NOUN        ADJ     PUNCT
9  En  dépit de  ces arguments pour la  lecture ,     certains sont convaincus que   la  littérature est le  passe-temps inutile .
   PRON VERB    SCONJ PRON PRON VERB ADV      ADP NOUN      ADV  ADJ       PUNCT ADJ          CCONJ ADJ       PUNCT
10 Ils  croient qu'   il   y    a    beaucoup d'  activités plus plaisants ,     intéressants et    éducatifs .
   DET NOUN VERB    SCONJ DET NOUN    VERB ADP DET NOUN ADP DET NOUN PRON VERB ADV      PRON VERB AUX  ADJ       CCONJ VERB        DET   NOUN  PUNCT
11 Ces gens estiment que   la  lecture fait de  le  mal  à   les yeux c'   est  pourquoi cela peut être dangereux et    endommager notre santé .
   ADP    DET NOUN    ADP    DET NOUN    PUNCT PRON AUX  VERB       SCONJ PRON VERB ADV  ADP NOUN  SCONJ NOUN
12 Malgré ces visions contre la  lecture ,     je   suis persuadée  qu'   il   fait plus de  biens que   mal
```

Fig.2. Screenshot of the tagged text

SpaCy

When working with this tagger, you need to take into account the fact that its list of tags in English is the same as that one of the CoreNLP tagger on the one hand, at the same time the total error amount is bigger than the number in the above described part- of-speech taggers.



```
ADP DET NOUN  PRON VERB   ADV      ADP NOUN   PUNCT ADP ADJ    NOUN
De  nos jours il   existe beaucoup de  livres ,     de  genres littéraires
ADJ        PUNCT NUM NOUN    PUNCT ADJ AUX ADV      NOUN AUX ADV   ADJ
différents ,     d'  auteurs .     C'  est pourquoi c'   est assez facile
ADP NOUN    ADV     NOUN  ADJ         ADP  DET VERB PUNCT ADV  ADP ADV
de  choisir quelque livre intéressant pour le  lire .     Tout de  même
PUNCT PRON PRON AUX DET ADV ADP NOUN PRON AUX  VERB       ADP DET ADJ
,     il   y    a   un  peu de  gens qui  sont intéressés à   la  lecture
PUNCT ADV   ADP PRON PUNCT DET ADJ         AUX DET ADP DET NOUN
.     Quant à   moi  ,     la  littérature est une de  mes domaines
VERB      PUNCT ADV  VERB ADV   PUNCT PRON PRON NOUN    SCONJ PRON
préférées .     Tout d'   abord ,     je   me   détends quand je
NOUN NOUN    ADP  DET ADJ   ADP ADJ         PUNCT PRON PRON ADJ
m'   immerge dans le  monde de  littérature .     Cela se   semble
DET NOUN     ADJ           PUNCT ADP ADV  PUNCT DET NOUN    NOUN ADJ  ADP
au  thérapie psychologique .     De  plus ,     la  lecture m'   aide à
VERB    DET NOUN     CCONJ DET ADJ         PUNCT ADP VERB   ADV  ADJ
élargir mes horisons et    mon vocabulaire ,     à   parler plus logique
CCONJ ADV   PUNCT ADV   PUNCT ADP VERB   NUM ADJ      DET NOUN  ADP
et    mieux .     Enfin ,     en  lisant j'  apprends des faits de
ADJ        PUNCT ADJ         PUNCT ADJ     NOUN PUNCT ADP NOUN  ADP DET
politiques ,     historiques ,     sociaux ets  .     En  dépit de  ces
NOUN      ADP  DET NOUN    PUNCT DET      AUX  NOUN       SCONJ DET ADJ
arguments pour la  lecture ,     certains sont convaincus que   la  littérature
AUX DET ADJ         ADJ     PUNCT PRON NOUN    SCONJ PRON PRON AUX ADV
est le  passe-temps inutile .     Ils  croient qu'   il   y    a   beaucoup
ADP NOUN      ADV  NOUN      PUNCT NOUN         CCONJ NOUN      PUNCT DET
d'  activités plus plaisants ,     intéressants et    éducatifs .     Ces
NOUN ADV      SCONJ DET NOUN    NOUN DET NOUN ADP NOUN NOUN AUX ADV
gens estiment que   la  lecture fait du  mal  aux yeux c'   est pourquoi
PRON VERB AUX  ADJ       CCONJ VERB       DET   NOUN  PUNCT ADP    DET
cela peut être dangereux et    endommager notre santé .     Malgré ces
NOUN    ADP    DET NOUN    PUNCT PRON AUX  VERB      SCONJ PRON NOUN ADV
visions contre la  lecture ,     je   suis persuadée qu'   il   fait plus
ADP NOUN  SCONJ NOUN
de  biens que   mal
```

Fig.3. Screenshot of the tagged text

It should be also noted that the number of tokens in the same text varied depending on the used part-of-speech tagger. The data is presented in diagram 1.

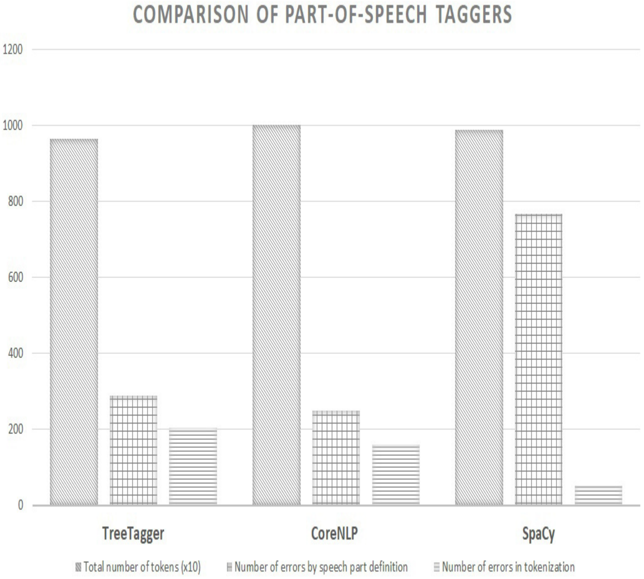


Fig.3. Diagram I

So we found out the indicator of the total number of tokens in the TreeTagger was the lowest for the errors in tokenization are often encountered due to abbreviated articles and pronouns which are written with an apostrophe followed by a word ending in a vowel and which are not separated into certain tokens. As you can see in Table II, TreeTagger, compared to CoreNLP and SpaCy, makes the most of word separating errors (2.1% versus 1.6% and 0.5%).

CoreNLP, on the contrary, splits parts of merged articles into different tokens, therefore, the total number of tokens is much higher than in other taggers. In SpaCy, the percentage of errors in the division of words with the total number of tokens is the smallest, based on it, we can conclude that this indicator is more accurate in this tagger. Thus, if tokenization is the prevailing goal of the text analysis, then SpaCy can be effective in the process of choosing a part-of-speech tagger.

TABLE II .POS TAGGERS COMPARISON

Percent /indicators (of the total number of tokens)	TreeTagger	CoreNLP	SpaCy
Percent of errors by speech part definition	3%	2,5%	7,8%
Percent of errors in tokenization	2,1%	1,6%	0,5%
Total percent of errors	5,1%	4,1%	8,3%

Therefore, with this error percent criterion (determining parts of speech in 3 taggers) TreeTagger and CoreNLP have similar indicators: 3% and 2.5%. Whereas in SpaCy error percent is 7.8%.

The lowest percent of the total number of errors and the number of tokens, 4.1% is characteristic for CoreNLP tagger, as for the highest one it is for SpaCy namely 8,3% ; and as to TreeTagger this percent is 5.1%. Consequently, the TreeTagger and CoreNLP part-of-speech taggers are more accurate in defining part of speech and allocating tokens.

We have compiled the following diagram III and TABLE III with marked erorrs using one of the texts (essays) analyzed through all three taggers as an example.

TABLE III

Word	Errors by speech part definition	Correction	Total number	Errors in tokenization	Total number
Treetagger:					
d'auteurs	adj	prp+nom	12	d'auteurs	8
c'est	nam	pro:per+ver:pres		c'est(x2)	
c'est	ver:subi	pro:per+ver:pres		d'abord	
interresant	ver:ppre	adj		m'immerge	
lire	nom	ver:infini		m'aide	
m'immerge	nom	pro:per+ver:pres		j'apprends	
m'aide	nom	pro:per+ver:pres		qu'il	
(en) lisant	adj	ver:ppre		d'activités	
j'apprends	nom	pro:per+ver:pres			
qu'il	nom	kon+pro:per			
d'activités	ver:pper	prp+nom			
c'est	nom	pro:per+ver:pres			
CoreNLP:					
nos	det	pron	13	au therapie	4
d'	det	adp		des faits	
C'	propn	pron		du mal	
est X4	aux	verb		aux yeux	
c'	noun	pron			
quelque	det	adj			
intéressé	verb	adj			
j'	det	pron			
apprends	noun	verb			
convaincus	verb	adj			
sont X2	aux	verb			
notre	det	pron			
persuadé	verb	adj			
SpaCy					
genres	adj	noun	32	-	0
littéraires	noun	adj			
d'	num	adp			
C'	adj	pron			
C'(x2)	noun	pron			
choisir	noun	verb			
quelque	adv	pron			
(pour) le	det				
(lire)		pron			
Tout	adv	pron			
lecture	adj	noun			
Quant	adv	adp			
littérature(x3)	adj	noun			
d'	verb	adp			
abord	adv	noun			
détends	noun	verb			
m'(x2)	noun	pron			
immerge	noun	verb			
monde	adj	noun			
semble	adj	verb			

III. Conclusion

Due to our research analysis of TreeTagger, CoreNLP and SpaCy as part-of-speech taggers for processing texts in French we chose the CoreNLP tagger as the most effective tool in use for our project. CoreNLP proved to have almost the same number of errors in determining parts of speech, moreover, it made fewer errors in separating words and it turned out to be more illustrative at work tan other tools.

Finally, it can be summarized as follows: using TreeTagger, CoreNLP and SpaCy universal taggers it is necessary to take into account that every one of them has its merits and shortcomings in terms of various kinds of errors. So when choosing a tagger the latter must always be carefully considered.

In 2020 the Department of German and French Languages of the Institute of Foreign Languages and the Center of Artificial Intelligence of Petrozavodsk State University initiative was the starting point for learner texts corpus creation project in German and French languages with the subsequent development of various virtual applications based on it. This project on the one hand, provided teachers of a foreign language with efficient tools of checking a large number of written works, on the other hand, it gave the students the opportunity to improve their writing skills while completing student assignments [5].

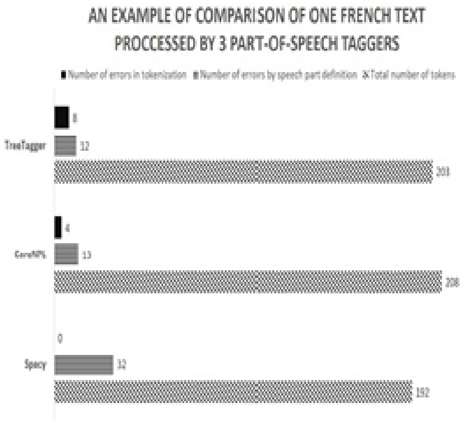


Fig.4. Diagram III

IV Acknowledgment

This work is implemented within collaboration with the Artificial Intelligence Center of Petrozavodsk State University (PetrSU).

References

[1] Nagel O.V. "Crpus Linguistics and its use in computer-based language teaching"//Language and culture – 2008 - № 4. Web : HTTPS://WWW.ELIBRARY.RU/ITEM.ASP?ID=11990991

[2] Maltseva M.S. Corpus technologies in foreign languages teaching methodology// Socio-economic phenomena and processes- 2011- №8. Web: HTTPS://CYBERLENINKA.RU/ARTICLE/N/KORPUSNYE-TEHNOLOGII-V-METODIKE-PREPODAVANIYA-INOSTRANNYH-YAZYKOV

[3] Julien Fago. Conception et réalisation d'une chaîne de traitement automatique des langues adaptée à des projets littéraires. Web : HTTPS://DUMAS.CCSD.CNRS.FR/DUMAS-02987314/DOCUMENT

[4] French TreeTagger part-of-speech tagset Web https://www.sketchengine.eu/french-treetagger-part-of-speech-tagset/

[5] I.A. Kotyurova, "Sozdanie korpusov uchebnykh tekstov kak razvivayushheesya napravlenie korpusnoj lingvistiki", *International Scientific Journal*, №5, 2020, pp. 100-109.