ANTILLES: An Open French Linguistically Enriched Part-of-Speech Corpus

Yanis Labrak, Richard Dufour

ANTILLES: An Open French Linguistically Enriched Part-of-Speech Corpus

Yanis Labrak[1][0000−0003−1072−3862] and Richard Dufour[2][0000−0003−1203−9108]

[1] LIA - Avignon University, 84911 Avignon, France
yanis.labrak@univ-avignon.fr
[2] LS2N - Nantes University, 44300 Nantes, France
richard.dufour@univ-nantes.fr

Abstract. Part-of-speech (POS) tagging is a classical natural language processing (NLP) task. Although many tools and corpora have been proposed, especially for the most widely spoken languages, these suffer from limitations concerning their user license, the size of their tagset, or even approaches no longer in the state-of-the-art. In this article, we propose ANTILLES, an extended version of an existing French corpus (UD French-GSD) comprising an original set of labels obtained with the aid of morphological characteristics (gender, number, tense, etc.). This extended version includes a set of 65 labels, against 16 in the initial version. We also implemented several POS tools for French from this corpus, incorporating the latest advances in the state-of-the-art in this area. The corpus as well as the POS labeling tools are fully open and freely available.

Keywords: Part-of-speech corpus · POS tagging · Open tools · Word embeddings · Bi-LSTM · CRF · Transformers

1 Introduction

For a few years now, several research areas have seen significant breakthroughs in both the arrival of big data and ways of exploiting it using deep learning based approaches. Then, various natural language processing (NLP) tasks reached a level of performance and maturity allowing them to be industrialized.

Among NLP field, part-of-speech (POS) tagging is a low-level grammatical task which consists in assigning, for each word of a sentence, its corresponding morphosyntactic category, such as verb (VERB), determinant (DET), adjective (ADJ) and much more (noun, auxiliary, etc.). This labeling is usually the root for more complex linguistic tasks such as named-entity recognition, text summarization, text generation, automatic speech recognition, spell checking, etc. In other words, many applications or research issues depend on the efficiency and quality of this labeling. While POS tagging problem was initially tackled with rule-based approaches, supervised statistical learning now allows us to achieve the best performance [22, 18].

French language has been relatively well studied for POS tagging. Nonetheless, for the sake of universality, the open multilingual corpora, and their derived tools, have mainly been designed with a limited number of tags across languages. To our knowledge, no open French corpus currently makes it possible to freely and easily train such a state-of-the-art POS tagging system with a sufficient level of tags granularity, allowing us to take into account this particular inflectional language.

We therefore propose in this article ANTILLES, an extended French corpus containing a set of linguistically enriched morphosyntactic tags. We then increase the capacities of the Universal Dependencies (UD) French GSD corpus [11] by extending its morphosyntactic information in single labels, from 16 to 63, thanks to the use of additional information present in the `Conll-U` format. This choice appears to be the most relevant since French GSD has been chosen for being one of the largest French manually annotated POS corpus, with a high level of granularity and a non-restrictive license of use. We also propose to evaluate different POS tagging systems using state-of-the-art neural network architectures. The corpus as well as the POS taggers are open and freely usable.

The paper is organized as follows. Section 2 presents an overview of the current resources (corpora and tools) available for French POS tagging. Then, Section 3 describes the proposed corpus and its inherent modifications. The experiments as well as the proposed POS tagging systems are detailed in Section 4. Finally, Section 5 concludes and opens new perspectives for the task.

2 Existing resources for French POS tagging

We propose, in this section, a detailed overview of the main resources in French, including the corpora (Section 2.1), as well as the available tools and their approaches (Section 2.2).

2.1 POS corpora

Table 1 summarizes the most popular corpora annotated with morphosyntactic labels for the French language. For comparison, we provided the number of tokens (*i.e.* words, punctuation, ...), the number of POS tags (# Tags), the license of use and the nature of the documents in the corpus (Genre).

The corpora are detailed in the table under four parts: 1) those completely free to use (most UD corpora); 2) those free except for commercial use (TCOF-POS, UD French-ParTUT and French Treebank); 3) those that are not directly downloadable and with limited usage constraints (Valibel and Orféo); 4) those with a paid license, especially for commercial use (Crater). Among these corpora, we observe a great disparity in the number of tokens, with around 20k/30k tokens for the smallest and up to 10 millions for the largest. As we can see, most of the large corpora are distributed under a restricted license, requiring either to pay to be able to access the data, or to go through a registration form with acceptance (*Limited*).

Regarding the nature of the documents used, we can highlight two main types: those that are purely textual (news, wiki, blogs, etc.) and those that are speech-oriented, relying on speech transcriptions. The second is clearly less present in terms of number of corpus, and limited either in terms of annotated data or license of use.

Corpus	# Tokens	# Tags	License	Genre
UD French-FQB [21]	23,349	16	LGPL-LR	nonfiction, news
UD French-PUD [24]	24,131	15	CC BY-SA 3.0	wiki, news
UD French-ParisStories [12]	29,438	15	CC BY-SA 4.0	spoken
UD French-Rhapsodie [13]	43,700	15	CC BY-SA 4.0	spoken
UD French-Sequoia [6]	68,596	16	LGPL-LR	medical, nonfiction, wiki, news
UD French-GSD [11]	400,399	16	CC BY-SA 4.0	blog, reviews, wiki, news
UD French-FTB [11]	556,064	16	LGPL-LR	news
TCOF-POS [4]	22,240	62	BY-NC-SA 2.0	spoken
UD French-ParTUT [20]	27,658	17	CC BY-NC-SA 4.0	legal, wiki, news
French Treebank [1]	664,500	15	CC-BY-NC-ND	news
Valibel [9]	6 millions	107	Limited	spoken, thesis
Orféo [3]	10 millions	20	Limited	interview, meeting, spoken
Crater 1 [15]	1 million	105	Paying	telecommunication manuals
Crater 2 [17]	1,5 millions	105	Paying	telecommunication manuals

Table 1. List of major French POS corpora, including the number of tag occurrences (# Tokens), the number of different tags (# Tags), their availability nature (License), and the genre of annotated documents (Genre).

2.2 POS taggers

As for the corpora, different morphosyntactic tagging tools specific to the French language have been proposed, each with its own specificities in terms of approaches and tags granularity. The list of the most popular POS taggers for the French language is detailed in Table 2.

	# Tags	License	Approach	Corpus
spaCy	18	MIT	Convolutional Neural Network	UD French Sequoia
Talismane	27	AGPL-3.0	Support Vector Machine	French Treebank
MElt	29	LGPL-3.0	Maximum-Entropy Markov models	French TreeBank
LGTagger	29	LGPL-LR	Conditional Random Fields	French Treebank
SoMeWeTa	29	GPL-3.0	Brown clusters	French Treebank
MarMoT	29	GPL-3.0	Conditional Random Field	French Treebank
gilf/french-postag-model	29	Unknown	bert-base-multilingual-cased	free-french-treebank
SEM	30	GNU	Hidden Markov Model	French Treebank
Stanford	33	GPL-3.0	Cyclic Dependency Network	French TreeBank
TreeTagger	33	GPL-3.0	Hidden Markov Model	French Treebank
Morfette	33	BSD-3-Clause	Logistic Regression	French TreeBank
NLTK	33	Apache-2.0	Cyclic Dependency Network	French TreeBank
DisMo	64	GPL-3.0	Conditional Random Field	PFC
LIA-Tagg	103	GPL-2.0	Second-Order HMM	Private corpus

Table 2. List of the most popular POS taggers for the French language.

We can first note that all POS taggers are based on statistical approaches and are, for the most part, trained on open corpora. Only LIA-Tagg relies on a private corpus, on which we found no description. Nevertheless, if we look at the open corpora, we see that French TreeBank is mostly used: its license being non-commercial (i.e. CC-BY-NC-ND, as seen in Table 1), only the spaCy tool, under a non-restrictive MIT license, is completely license free right of use.

spaCy however suffers from a limited number of tags (only 18). In general, the number of tags is very limited (between 18 and 33 tags). This is because most tools rely on corpora following the UD annotation guideline or [7] which sought to produce a set of labels that could apply to multiple languages.

The semi-free of use TCOF-POS corpus is nevertheless distinguished by its high number of tags (62). In reality, these additional tags are already an extension of the UD tags with morphosyntactic information (*e.g.* the ADJ tag is derived in 7 tags: demonstrative adjective, indefinite, etc.). Finally, note that although the UD corpora have a limited number of tags, morphosyntactic information exists, but is most often not used in the form of its own label.

In general, the sets of tags offered by the POS tools take little - or even no - account of the specificities of the French language, seeking to maintain their universality as much as possible. Only LIA-Tagg integrates a very complete set of tags, but suffers from an unavailable corpus and the use of an approach that is no longer in the state-of-the-art. Concretely, only spaCy and french-postag-model are maintained and implement state-of-the-art methods but on a restricted set of tags.

3 Extended corpus proposal

Each existing corpus for French language has interesting specificities, whether in relation to the size of the annotated data, their license free right of use, or the large number of labels offered, but no corpus combines all these advantages at the same time.

However, we found that although the associated tagset is often small, a lot of data related to linguistic information is available. This is particularly the case for UD corpora.

We have chosen to focus on the annotated French corpus UD French-GSD [11] because it includes all the necessary features to implement a linguistic enrichment of POS labels. Moreover, it is one of the few corpora completely free of use (see Section 2.1), allowing a complete redistribution of its improvements. It contains 16 POS tags and is composed of 16,341 documents, for approximately 400k manually annotated word occurrences. It also offers data integrating the morphological characteristics of words (gender, number, tense, verbal form, person, etc.) which has been automatically annotated and then partially manually corrected.

The UD French-GSD corpus follows the standard UD [19] annotation scheme. The new annotation scheme that we propose follows the morphosyntactic tags

proposed by the LIA-Tagg tagger, as they allow a complete and deep representation of the spelling as well as the grammar of French language. We achieve this enrichment by transforming the `CoNLL-U` tags (`UPOS`) and the features (`FEATS`) to our 65 new tags in order to give information on morphological characteristics such as the gender (feminine and masculine), number / person (first person, etc.), tense (past participle, etc.), types of pronouns (relative, indefinite, demonstrative, etc.) in a single label. The initial tags of the UD French-GSD corpus as well as the new tags of the ANTILLES extended corpus are detailed in Table 3. The ANTILLES corpus is freely accessible online[3] under CC-BY-SA 4.0 License.

UD French-GSD		**ANTILLES**	
Abbreviation	Description	Abbreviation	Description
ADP	Adposition	PREP	Preposition
		PART	Demonstrative particle
SCONJ	Subordinating conjunction	COSUB	Subordinating conjunction
CCONJ	Coordinating Conjunction	COCO	Coordinating Conjunction
ADV	Adverb	ADV	Adverb
PROPN	Proper noun	PROPN	Proper noun
		XFAMIL	Family name
NUM	Numerical Adjective	NUM	Numerical Adjective
		CHIF	Number
AUX	Auxiliary Verb	AUX	Auxiliary Verb
VERB	Verb	VERB	Verb
		VPP**XX** (x4)	FS/FP/MS/MP Past participle verb
		VPPRE	Present participle verb
DET	Determinant	DET	Determinant
		DET**XX** (x2)	FS/MS Determinant
		ADJ	Adjective
ADJ	Adjective	ADJ**XX** (x4)	FS/FP/MS/MP Adjective
		DINT**XX** (x2)	FS/MS Numerical adjectives
NOUN	Noun	NOUN	Noun
		N**XX** (x4)	FS/FP/MS/MP Noun
		PRON	Pronoun
		PINT	FS Interrogative pronoun
		PDEM**XX** (x4)	FS/FP/MS/MP Demonstrative pronoun
		PIND**XX** (x4)	FS/FP/MS/MP Indefinite pronoun
		PPOBJ**XX** (x4)	FS/FP/MS/MP Pronoun complements of objects
		PPER1S	Personal pronoun - First person singular
		PPER2S	Personal pronoun - Second person singular
PRON	Pronoun	PPER3**XX** (x4)	Personal Pronoun - Third Person FS/FP/MS/MP
		PREFS	Reflexive pronoun - First person of singular
		PREF	Reflexive pronoun - Third person of singular
		PREFP	Reflexive pronoun - First / Second Person of plurial
		PREL	Relative pronoun
		PREL**XX** (x4)	FS/FP/MS/MP Relative pronoun
INTJ	Interjection	INTJ	Interjection
SYM	Symbol	SYM	Symbol
PUNCT	Punctuation	YPFOR	Final point
		PUNCT	Punctuation
X	Other	MOTINC	Unknown word
		X	Typos & Other

Table 3. Labels of the initial corpus UD FRENCH-GSD and of the proposed extended corpus ANTILLES. The suffix **XX** at the end of a label corresponds to a declension among feminine singular (FS), feminine plural (FP), masculine singular (MS), masculine plural (MP).

[3] https://github.com/qanastek/ANTILLES

4 Experiments

In addition to the extended corpus, we provide a comparison of several taggers using different approaches based on neural networks. In Section 4.1, we describe the implemented approaches. We then detail the results in Section 4.2.

4.1 Proposed approaches

We implement three different state-of-the-art architectures to evaluate current performance on the POS tagging task by means of our extended French corpus:

1. Word embedding + Bi-LSTM-CRF. The first proposed system consists of a Bidirectional Long Short-Term Memory (Bi-LSTM) [10] with Conditional Random Field (CRF) [14] using, as inputs, different kinds of word embeddings. Our core system incorporates FastText embeddings [5] pre-trained specifically for French. Once this reference was obtained, we independently evaluated other state-of-the-art representations: Flair [2] and BERT [8] (here, CamemBERT [16] for the French).

2. Concatenation of word embeddings + Bi-LSTM-CRF. We propose to keep the same neural network, but train it here on the combination of several word embeddings concatenated at the input of the system. We explore all the possible concatenation combinations starting from the same word embeddings as before: FastText, Flair and CamemBERT.

3. CamemBERT Fine-Tuning. For the last system, rather than using the CamemBERT word embeddings as input to a Bi-LSTM-CRF as described in the previous architectures, we propose to directly perform a fine-tuning of the CamemBERT model by adding a linear layer dedicated to the POS labeling task after model outputs.

The complete training procedure was performed using the `Transformers` [23] library maintained by HuggingFace.

4.2 Results

Table 4 summarizes the results obtained by the three approaches proposed on the POS labeling task of the ANTILLES corpus test set. Overall, except for a few simple word embeddings (Bi-LSTM-CRF + FastText and Bi-LSTM-CRF + Flair), the performance obtained is quite similar regardless of the approach considered, from 95.24% to 97.97% for the success rate (Accuracy) and 95.19% to 97.98% for the F-measure (F1).

Our best performing model combines a Bi-LSTM-CRF architecture with a concatenation of two word embeddings Flair and Camembert as inputs (f-measure of 97.98%). The two word embeddings integrate quite different information, one coming from word sub-units (CamemBERT) and the other from

characters (Flair), which could explain their complementary performances. It outperforms our benchmark based on a Bi-LSTM-CRF combined with FastText word embeddings by 2.73%.

Model	Acc.	Prec.	Rec.	F1	# Params	Inf.
Simple Embeddings (Baseline)						
Bi-LSTM-CRF						
+ FastText	95.24%	95.26%	95.24%	95.19%	**1.27 M**	**34.91 s**
+ Flair	96.96%	96.97%	96.96%	96.94%	18.80 M	320.42 s
+ CamemBERT$_{oscar-4gb-base}$	97.77%	97.80%	97.77%	97.75%	113.35 M	151.44 s
+ CamemBERT$_{oscar-138gb-base}$	97.76%	97.80%	97.76%	97.74%	113.35 M	147.37 s
Multi-Embeddings						
Bi-LSTM-CRF						
+ FastText + Flair	97.29%	97.33%	97.29%	97.28%	20.73 M	337.46 s
+ FastText + CamemBERT$_{oscar-138gb-base}$	97.88%	97.90%	97.88%	97.85%	114.52 M	152.14 s
+ Flair + CamemBERT$_{oscar-4gb-base}$	97.89%	97.90%	97.89%	97.87%	134.73 M	411.77 s
+ Flair + CamemBERT$_{oscar-138gb-base}$	**97.97%**	**98.02%**	**97.97%**	**97.98%**	134.73 M	418.57 s
+ Flair + CamemBERT$_{ccnet-135gb-large}$	97.87%	97.92%	97.87%	97.87%	362.80 M	476.07 s
+ Flair + FastText + CamemBERT$_{oscar-138gb-base}$	97.91%	97.93%	97.91%	97.91%	137.13 M	439.95 s
Fine-tuning						
CamemBERT$_{oscar-138gb-base}$	97.78%	97.85%	97.78%	97.80%	110.08 M	53.94 s

Table 4. Results on the POS labeling task of the ANTILLES test set.

Table 4 also integrates the size of the models (# Params) and their inference times (Inf.) to sequentially process 5,000 sentences with an RTX 2080 Ti graphics card. The performance gap between our least efficient system (Bi-LSTM-CRF + FastText) and the most efficient (Bi-LSTM-CRF + Flair + Camembert$_{OSCAR-138gb-base}$) appears small (difference in F1-score of 2.79%) considering the number of parameters as well as the inference time (12 times slower). Note that the large CamemBERT model trained on the CCNET corpus (*ccnet 135gb large*) is provided for information only in the table: we have not seen any improvement by using it, while its number of parameters is at least 2.5 times higher than any other model.

Finally, the fine-tuning approach of CamemBERT seems to be one of the best choices for this task compared to the Bi-LSTM-CRF, since it obtains results close to those obtained with the best system, but with an inference time at least 8 times faster (53.94s against $\approx$ 420s).

For information, we also compared our systems to one of the most widely used tool: spaCy. We used the POS tag and the morphological information given by spaCy to map their outputs to our tags and make the systems comparable with each other. Likewise, we also skipped the entities without annotation and represented in the test file as underscores to remove some noise in the metric.

This evaluation raised one big issue, which is the dissonance between the annotation guidelines of UD French GSD and UD French Sequoia v2.8, the first being used for training our systems and the second for training spaCy. For example, in UD French Sequoia corpus, and by extension spaCy:

– The symbols like €, $ and % are for most of the time unhandled, but they are sometimes described as NOUN, which is worse.

– Last names are not always tagged as proper nouns (PROPN) which make the mapping even more complicated.
– And last but not least, the biggest issue comes from the lack of information about the gender and number in the original annotation for the adjectives.

Finally, to have a fair comparison, we removed from this evaluation the tags involved in the previously raised annotation issues. We then obtained an F1-score of 85.81% for the spaCy system and 91.29% for the proposed Flair one. To conclude, we can expect a performance increase using our systems compared to the existing annotation tools. Note that the choice of using Sequoia to train spaCy makes it less optimized for in-depth analysis of languages such as French. This difference in performance between the systems would surely be much lower if spaCy was trained on data with more consistent annotations like UD French GSD.

All developed taggers presented in this article are available and freely usable[4].

5 Conclusion and Perspectives

In this article, we proposed an extended corpus for POS tagging in the French language. This corpus fills the observed limitations of existing corpora, whether in terms of labels, user license or existing state-of-the-art tools.

This corpus, named ANTILLES, is an extended version of the free-to-use UD French-GSD corpus, integrating additional POS tags based on a set of associated morphological data. We have also implemented numerous POS tagging tools, then evaluated the performance of various state-of-the-art neural network architectures to give an idea of the current performance level in French POS tagging. ANTILLES as well as the associated POS labeling tools are freely distributed and can be used by academics or industrialists.

The corpus is intended to be enriched over time, by extending, in the same way, the other freely accessible corpora offered by the Universal Dependencies (UD) such as PUD, ParisStories, or Rhapsodie, using the same strategy and the same set of proposed labels. All the scripts necessary to perform this transformation are available on the GitHub repository[5], the models are also available on HuggingFace[6]. The extension to other languages can also be a possibility.

6 Acknowledgements

This work was financially supported by Zenidoc and the DIETS project financed by the Agence Nationale de la Recherche (ANR) under contract ANR-20-CE23-0005. It was granted access to the HPC resources of IDRIS under the allocation 2021-A0111012991 made by GENCI.

[4] https://huggingface.co/qanastek/pos-french-camembert
[5] https://github.com/qanastek/ANTILLES
[6] https://huggingface.co/qanastek

7 Bibliographical References

References

1. Abeillé, A., Clément, L., Toussenel, F.: Building a treebank for french. In: Treebanks, pp. 165–187. Springer (2003)
2. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: COLING 2018, 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018)
3. Benzitoun, C., Debaisieux, J.M., Deulofeu, H.J.: Le projet orféo: un corpus d'étude pour le français contemporain. Corpus (15) (2016)
4. Benzitoun, C., Fort, K., Sagot, B.: Tcof-pos: un corpus libre de français parlé annoté en morphosyntaxe. In: JEP-TALN 2012-Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles. pp. 99–112 (2012)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
6. Candito, M., Seddah, D.: Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In: Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN. pp. 321–334. ATALA/AFCP, Grenoble, France (Jun 2012), https://aclanthology.org/F12-2024
7. Crabbé, B., Candito, M.: Expériences d'analyse syntaxique statistique du français. In: 15ème conférence sur le Traitement Automatique des Langues Naturelles - TALN'08. pp. pp. 44–54. Avignon, France (Jun 2008), https://hal.archives-ouvertes.fr/hal-00341093
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
9. Eshkol, I., Tellier, I., Samer, T., Billot, S.: Etiqueter un corpus oral par apprentissage automatique a l'aide de connaissances linguistiques. arXiv preprint arXiv:1003.5749 (2010)
10. Graves, A., rahman Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks (2013)
11. Guillaume, B., de Marneffe, M.C., Perrier, G.: Conversion et améliorations de corpus du français annotés en Universal Dependencies. Traitement Automatique des Langues **60**(2), 71–95 (2019), https://hal.inria.fr/hal-02267418
12. Kahane, S., Guillaume, B., Nakhle, M., Gaudray-Bouju, V., Mahamdi, M., Gerdes, K., Courtin, M., Guibon, G., Gendrot, C.: Ud_french-parisstories (2021)
13. Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.P., Obin, N., Pietrandrea, P., Tchobanov, A.: Rhapsodie: a prosodic-syntactic treebank for spoken french. In: LREC (2014)
14. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. p. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
15. Leech, G., McEnery, A., Oakes, M.: Multilingual corpus resources and tools developed in crater. In: Proceedings of SNLR: International Workshop on Sharable Natural Language Resources, Nara, Japan. Citeseer (1994)

16. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, E., Seddah, D., Sagot, B.: Camembert: a tasty french language model. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.645, http://dx.doi.org/10.18653/v1/2020.acl-main.645
17. McEnery, T., Wilson, A., SáNchez-LeóN, F., Nieto-Serrano, A.: Multilingual resources for european languages: Contributions of the crater project. Literary and Linguistic Computing **12**(4), 219–226 (Nov 1997). https://doi.org/10.1093/llc/12.4.219
18. Nguyen, D.Q., Verspoor, K.: An improved neural network model for joint pos tagging and dependency parsing. arXiv preprint arXiv:1807.03955 (2018)
19. Nivre, J., de Marneffe, M.C., Ginter, F., Hajič, J., Manning, C.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D.: Universal dependencies v2: An evergrowing multilingual treebank collection (2020)
20. Sanguinetti, M., Bosco, C.: PartTUT: The Turin University Parallel Treebank, pp. 51–69. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-14206-7_3, https://doi.org/10.1007/978-3-319-14206-7_3
21. Seddah, D., Candito, M.: Hard Time Parsing Questions: Building a QuestionBank for French. In: Tenth International Conference on Language Resources and Evaluation (LREC 2016). Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016), Portorož, Slovenia (May 2016), https://hal.archives-ouvertes.fr/hal-01457184
22. Wang, P., Qian, Y., Soong, F.K., He, L., Zhao, H.: Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. arXiv preprint arXiv:1510.06168 (2015)
23. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Huggingface's transformers: State-of-the-art natural language processing (2020)
24. Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C.D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H.F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., Li, J.: CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 1–19. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/K17-3001, https://aclanthology.org/K17-3001