Proceedings of he 8th Workshop on Challenges in the Management of Large Corpora, pages 15–23
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

French Contextualized Word-Embeddings with a sip of CaBeRnet: a New French Balanced Reference Corpus

Murielle Popa-Fabre[1,2], Pedro Javier Ortiz Suárez[1,3], Benoît Sagot[1], Eric de la Clergerie[1]
[1]ALMAnaCH - Inria, [2]LLF - Université de Paris, [3]Sorbonne Université
2 rue Simone Iff, 75012 Paris, France
{murielle.fabre, pedro.ortiz, benoit.sagot, Eric.De_La_Clergerie}@inria.fr

Abstract

This paper investigates the impact of different types and size of training corpora on language models. By asking the fundamental question of quality versus quantity, we compare four French corpora by pre-training four different ELMos and evaluating them on dependency parsing, POS-tagging and Named Entities Recognition downstream tasks. We present and asses the relevance of a new balanced French corpus, CaBeRnet, that features a representative range of language usage, including a balanced variety of genres (oral transcriptions, newspapers, popular magazines, technical reports, fiction, academic texts), in oral and written styles. We hypothesize that a linguistically representative corpus will allow the language models to be more efficient, and therefore yield better evaluation scores on different evaluation sets and tasks.

Keywords: Balanced French Corpus, Language Models, French, BERT, ELMo, Tagging, Parsing, NER

1. Introduction

The question of quality versus size of training corpora is increasingly gaining attention and interest in the context of the latest developments in neural language models' performance. The longstanding issue of corpora "representativeness" is here addressed, in order to grasp to what extent a linguistically balanced cross-genre language sample is sufficient for a language model to gain in accuracy for contextualized word-embeddings on different NLP tasks.

Several increasingly larger corpora are nowadays compiled from the web, i.e. frWAC (Baroni et al., 2009), CCNet (Wenzek et al., 2019) and OSCAR-fr (Ortiz Suárez et al., 2019). However, does large size necessarily go along with better performance for language model training? Their alleged lack of representativeness has called for inventive ways of building a French balanced corpus offering new insights into language variation and NLP.

Following Biber's definition, "representativeness refers to the extent to which a sample includes the full range of variability in a population" (Biber, 1993, 244). We adopt a balanced approach by sampling a wide spectrum of language use and its cross-genre variability, be it situational (e.g. format, author, addressee, purposes, settings or topics) or linguistic, e.g. linked to distributional parameters like frequencies of word classes and genres. In this way, we developed two newly built corpora. The French Balanced Reference Corpus - *CaBeRnet* - includes a wide-ranging and balanced coverage of cross-genre language use to be maximally representative of French language and therefore yield good generalizations from. The second corpus, the *French Children Book Test* (CBT-fr), includes both narrative material and oral language use as present in youth literature, and will be used for domain-specific language model training. Both are inspired by existing American and English corpora, respectively COCA, the balanced Corpus of Contemporary American English (Davies, 2008), and the Children Book Test (Hill et al., 2015, CBT).

The second main contribution of this paper lies in the evaluation of the quality of the word-embeddings obtained by pre-training and fine-tuning on different corpora, that are made here publicly available. Based on the underlying assumption that a linguistically representative corpus would possibly generate better word-embeddings. We provide an evaluation-based investigation of how a balanced cross-genre corpus can yield improvements in the performance of neural language models like ELMo (Peters et al., 2018) on various downstream tasks. The two corpora, CaBeRnet and CBT-fr, and the ELMos will be distributed freely under Creative Commons License.

Specifically, we want to investigate the contribution of oral language use as present in different corpora. Through a series of comparisons, we contrast a more domain-specific and written corpus like Wikipedia-fr with the newly built domain-specific CBT-fr corpus which additionally features oral style dialogues, like the ones one can find in youth literature. To test for the effect of corpus size, we further compare a wide ranging corpora characterized by a variety of linguistic phenomena crawled from internet, like OSCAR (Ortiz Suárez et al., 2019), with our newly built French Balanced Reference Corpus CaBeRnet. Our aim is assess the benefits that can be gained from a balanced, multi-domain corpus such as CaBeRnet, despite its being 34 times smaller than the web-based OSCAR.

The paper is organized as follows. Sections 2. and 3. are dedicated to a descriptive overlook of the building of our two newly brewed corpora CaBeRnet and CBT-fr, including quantitative measures like type-token ratio and morphological richness. Section 4. presents the evaluation methods for POS-tagging, NER and dependency Parsing tasks, while results are introduced in §5. Finally, we conclude in §6. on the computational relevance of word-embeddings obtained through a balanced and representative corpus, and broaden the discussion on the benefits of smaller and noiseless corpora in neural NLP.

2. Corpora Building

2.1. CaBeRnet

CaBeRnet corpus was inspired by the genre partition of the American balanced corpus COCA, which currently contains over 618 million words of text (20 million words each year 1990-2019) and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts (Davies, 2008). A second reference, guiding our approach and sampling method, is one of the earliest precursors of balanced reference corpora: the BNC (Burnard, 2007), first covered a wide variety of genres, with the intention to be a representative sample of spoken and written language.

CaBeRnet was obtained by compiling existing data-sets and web-text extracted from different sources as detailed in this section. As shown in Table 1, genres sources are evenly divided ($\sim$120 million words each) into spoken, fiction, magazine, newspaper, academic to achieve genre-balanced between oral and written modality in newspapers or popular written style, technical reports and Wikipedia entries, fiction, literature or academic production).

CaBeRnet Oral The oral sub-portion gathers both oral transcriptions (ORFEO and Rhapsodie[1]) and Films subtitles (Open Subtitles.org), pruned from diacritics, interlocutors tagging and time stamps. To these transcriptions, the French European Parliament Proceedings (1996-2011), as presented in Koehn (2005), contributed a sample of more complex oral style with longer sentences and richer vocabulary.

CaBeRnet Popular Press The whole sub-portion of Popular Press is gathered from an open data-set from the *Est Républicain* (1999, 2002 and 2003), a regional press format[2]. It was selected to match popular style as it is characterized by easy-to-read press style and a wide range of every-day topics characterizing local regional press.

CaBeRnet Fiction & Literature The Fiction & Literature sub-portion was compiled from march 2019's Wiki Source and WikiBooks dump and extracted using WikiExtractor.py, a script that extracts and cleans text from a WikiMedia database dumps, by performing template expansion and preprocessing of template definitions.[3]

CaBeRnet News The News sub-portion builds upon web crawled elements, including Wikimedia's NewsComments and WikiNews reports from may 2019 WikiMedia dump, collected with a custom version of WikiExtractor.py. Newspaper's content gathered by the Chambers-Rostand Corpus (i.e. Le Monde 2002-2003, La Dépèche 2002-2003, L'Humanité 2002-2003) and *Le Monde diplomatique* open-source corpus were assembled to represent a higher register of written news style from different political and thematic horizons. Several months of French Press Agency reports (AFP, 2007-2011-2012) competed with more simple and telegraphic style the newspaper written sample of the corpus.[4]

CaBeRnet Academic The academic genre was also built from different sources including technical and educational texts from WikiBooks and Wikipedia dump (prior to 2016) for their thematic variety of highly specialized written production. ORFEO Corpus offered a small sample of academic writings like PHD dissertations and scientific articles encompassing a wide choice of disciplinary topics, and TALN Corpus[5] was included to represent more concise written style characterizing scientific abstracts and proceedings.

CaBeRnet Sub-set	Tokens	Unique Forms	TTR
Oral	122 864 888	291 744	0.0024
Popular	131 444 017	458 521	0.0035
News	132 708 943	462 971	0.0035
Fiction	198 343 802	983 195	0.0050
Academic	126 431 211	1 433 663	0.0113
Total	711 792 861	2 558 513	0.0036

Table 1: Comparison of number of unique forms in the different genres represented by CaBeRnet partition. TTR: Type-Token Ration. Lemmatization and tokenization was performed as described in §3..

For all sub-portions of CaBeRnet, visual inspection was performed to remove section titles, redundant meta-information linked to publishing schemes of each of the six news editor includes. This was manually achieved by compiling a rich set of regular expressions specific of each textual source to obtain clean plain text as an outcome.

2.2. French Children Book Test (CBT-fr)

The French Children Book Test (CBT-fr) was built upon its original English version, the Children Book Test (CBT) Hill et al. (2015)[6], which consists of books freely available on `www.gutenberg.org`Project Gutenberg.

Using youth literature and children books guarantees a clear narrative structure, and a large amount of dialogues, which enrich with oral register the literary style of this corpus. The English version of this corpus was originally built as benchmark data-set to test how well language models capture meaning in context. It contains 108 books, and a vocabulary size of 53,628.

French version of CBT, named CBT-fr, was constructed to guarantee enough linguistic similarities between the collected books in the two languages. 104 freely available books were included. One third of the books were purposely chosen because they were classical translations of English literary classics. Chapter heads, titles, notes and

[1] ORFEO corpus available at `www.cocoon.huma-num.fr/exist/crdo/` ; Rhapsodie corpus at `www.projet-rhapsodie.fr`.

[2] Corpus available at `www.cnrtl.fr/corpus/estrepublicain/`.

[3] Script available at `https://github.com/attardi/wikiextractor`.

[4] At the time being, this part of CaBeRnet corpus is still subject to Licence restrictions. This restricted amount of AFP news reports can reasonably fall in the public domain.

[5] TALN proceedings corpus (about 2 million) builds on a subset of 586 scientific articles (from 2007 to 2013), namely TALN and RECITAL. Available at `redac.univ-tlse2.fr/corpus/taln_en.html`.

[6] This data-set can be found at `www.fb.ai/babi/`.

all types of editorial information were removed to obtain a plain narrative text. The effort of keeping proportion, genre, domain, and time as equal as possible yields a multilingual set of comparable corpora with a similar balance and representativeness.

CHILDREN BOOK TEST - FR	WORDS
number of different lemmas	25 139
total number of forms	95 058
mean number of forms per lemma	3.78
Number of lemmas having more than one form :	14 128
Percentage of lemmas with multiple forms	56.20

Table 2: Lexical statistics of French CBT, performed as described in §3.

3. Corpora Descriptive Comparison

We used two different tokenizers: SEM, Segmenteur-Étiqueteur Markovien standalone Dupont (2017) and TreeTagger. Both are based on cascades of regular expressions, and both perform tokenization and sentence splitting. The first was used for descriptive purposes because it technically allowed to segment and tokenize all corpora including OSCAR (23 billion words). Hence, all corpora were entirely segmented into sentences and tokenized using SEM. The second tokenization method was run only on 3 million words samples to automatically tag them with TreeTagger into part-of-speech and lemmatize them.[7] All corpora were randomly shuffled by sentence to then select samples of 3 million words, to be able to compare them in terms of lexical composition (Type-Token Ratio, see Table 4).

3.1. Corpora Size and Composition

Length of sentences is a simple measure to quantify both sentence syntactic complexity and genre. Hence, the number of sentences reported in Table 3 shows interesting patterns of distributions across genres, consider the comparison between CaBeRnet an Wiki-fr. In our effort to evaluate the impact of corpora pre-training on ELMo-based contextualized word-embedding, we introduce here our two terms of comparison, namely the crawled corpus OSCAR-fr and the Wikipedia-fr one.

3.1.1. OSCAR fr

As it has been shown that pre-trained language models can be significantly improved by using more data (Liu et al., 2019; Raffel et al., 2019), we decided to include in our comparison a corpus of French text extracted from Common Crawl[8]. We leverage on a recently published corpus, OSCAR (Ortiz Suárez et al., 2019), which offers a pre-classified and pre-filtered version of the November 2018 Common Craw snapshot.

OSCAR gathers a set of monolingual text extracted from Common Crawl - in plain text *WET* format - where all HTML tags are removed and all text encodings are converted to UTF-8. It follows a similar approach to (Grave et al., 2018) by using a language classification model based on the fastText linear classifier (Joulin et al., 2016; Grave et al., 2017) pre-trained on Wikipedia, Tatoeba and SETimes, supporting 176 different languages.

After language classification, a deduplication step is performed without introducing a specialized filtering scheme: paragraphs containing 100 or more UTF-8 encoded characters are kept. This makes OSCAR an example of unfiltered data that is nearly as noisy as to the original Crawled data.

3.1.2. FrWIKI

This corpus collects a selection of pages from Wikipedia-fr from a dump executed in April 2019, where HTML tags and tables were removed, together with template expansion using Attardi's tool (WikiExtractor, §2.1.). As reported on Table 3, in this data-set (660 million words) sentences are relatively longer compared to other corpora. It has the advantage of having a comparable size to CaBeRnet, but its homogeneity in terms of written genre is set to Wikipedia entries descriptive style.

CORPUS	WORDFORMS	TOKENS	SENTENCES
OSCAR-fr	23 212 459 287	27 439 082 933	1 003 261 066
Wiki-fr	665 599 545	802 283 130	21 775 351
CaBeRnet	697 119 013	830 894 133	54 216 010
CBT-fr	5 697 584	6 910 201	317 239

Table 3: Comparing the corpora under study.

3.2. Corpora Lexical Variety

Focusing on a useful measure of complexity that documents lexical richness or variety in vocabulary, we present the type-token ration (TTR) of the corpora under analysis. Generally used to asses language use aspects like the variety of different words used to communicate by learners or children, it represents the total number of unique words (types/forms) divided by the total number of tokens in a given sample of language production. Hence, the closer the TTR ratio is to 1, the greater the lexical richness of the corpus. Table 1 summarizes the lexical variety of the five subportions of CaBeRnet, respectively taken as representative of Oral, Popular, Fiction, News, and Academic genres. Domain diversity of texts can be observed in the lexical statistics showing a gradual increase in the number of distinct lexical forms (cf. TTR). This pattern reflects a generally acknowledged distributional pattern of vocabulary-size across genres. Oral style shows a poorer lexical variety compared to newspapers/magazines' textual typology. The lexically rich fictional/classic literature is outreached by academic writing-style with its wide-ranging specialized vocabulary. All in all, Table 1 quantitatively demonstrates that the selected textual and oral materials are indeed representative of the five types of genres of CaBeRnet.

[7] Based on the tag-set available at `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html`.

[8] More information available at `https://commoncrawl.org/about/`.

3.3. Corpora Morphological richness

To select a measure that would help quantifying the different corpora morphological richness, we follow (Bonami and Beniamine, 2015). Hence, the proportion of lemmas with multiple forms in a given vocabulary size was evaluated on randomly selected samples of 3-million-words from each corpus under analysis (see Table 4).

3 M SAMPLES	CBT-FR	CABERNET	FR-WIKI	OSCAR
nb of diff. lemmas	25 139	30 488	31 385	31 204
tot. nb forms	95 058	180 089	238 121	190 078
mean nb forms/lemma	3.78	6.19	7.85	6.40
nb lemmas > 1 form	14 128	15 927	15 182	16 480
% lemmas > 1 form	56.20	52.24	48.37	52.81

Table 4: Lexical statistics on morphological richness over randomly selected samples of 3 million words from each corpus. nb : number

Table 4 reports some more in-depth lexical and morphological statistics across corpora. Although OSCAR is 34 times bigger than CaBeRnet, their total number of forms and the proportion of lemmas having more than one form in a 3-million-word sample are comparable. FrWiki shows a radically different lexical distribution with numerous hapaxes but a lower morphological richness. Although its total number of forms is more than one third higher than in OSCAR and CaBeRnet samples, the proportion of lemmas having more than one distinct form is around four points below CaBeRnet and OSCAR. Comparatively, youth literature in CBT-fr shows the greatest morphological richness, around 56% of lemmas have more than one form.

4. Corpora Evaluation Tasks

This section reports the method of experiments designed to better understand the computational impact of the quality, size and linguistic balance of ELMo's (Peters et al., 2018) pre-training (§4.1.) and their evaluations tasks (§4.3.).

Embeddings from Language Models ELMo is an LSTM-based language model. More precisely, it uses a bidirectional language model, which combines a both forward and a backward LSTM-based language models. ELMo also computes a context-independent token representation via a CNN over characters. Methodologically, we selected ELMo which not only performs generally better on sequence tagging than other architectures, but which is also better suited to pre-train on small corpora because of its smaller number of parameters (93.6 million) compared to the RoBERTa-base architecture used for CamBERT (BERTbase, 12,110 million - Transformer) (Martin et al., 2019).

4.1. ELMo Pre-traing & Fine-tuning Method

Two protocols were carried out to evaluate the impact of corpora characteristics on the tasks under analysis. *Method 1* implies a full pre-training ELMo-based language models for each of the corpora mentioned in Table 3. While *Method 2* is based on pre-training OSCAR + fine-tuning with our French Balanced Reference Corpus CaBeRnet, yielding ELMo$_{\text{OSCAR+CaBeRnet}}$. Hence, the pure pre-traing (i.e. Method 1) yields the following four language models which were pre-trained on the four corpora under comparison : ELMo$_{\text{OSCAR}}$, ELMo$_{\text{Wikipedia}}$, ELMo$_{\text{CaBeRnet}}$ and ELMo$_{\text{CBT}}$.

4.2. Base evaluation systems

UDPipe Future (Straka, 2018) is an LSTM based model ranked 3$^{\text{rd}}$ in dependency parsing and 6$^{\text{th}}$ in POS tagging during the CoNLL 2018 shared task (Seker et al., 2018). We report the scores as they appear in Kondratyuk (2019)'s paper. We add to UDPipe Future, five differently trained ELMo language model pre-trained on the qualitatively and quantitatively different corpora under comparison. Additionally, we also test the impact of the CaBeRnet Corpus on ELMo fine-tuning.

The LSTM-CRF is a model originally concived by Lample et al. (2016) is just a Bi-LSTM pre-appended by both character level word embeddings and pre-trained word embeddings and pos-appended by a CRF decoder layer. For our experiments, we use the implementation of (Straková et al., 2019) which is readily available[9] and it is designed to easily pre-append contextualized word-embeddings to the model.

4.3. Evaluation Tasks

We distinguish three main evaluation tasks that were performed to asses the lexical and syntactic quality of contextualized word-embeddings obtained from different pre-training corpora under comparison.Crucially, comparing them with and ELMo pre-trained on OSCAR and fine-tuned with CaBeRnet, i.e. ELMo$_{\text{OSCAR+CaBeRnet}}$, will allow to control for the presence of oral transcriptions and proceeding in order to understand its impact on the accuracy of our language model and on the development experiments after fine-tuning.

Syntactic tasks The evaluation tasks were selected to probe to what extent corpus "representativeness" and balance is impacting syntactic representations, in both (1) low-level syntactic relations in POS-tagging tasks, and (2) higher level syntactic relations at constituent- and sentence-level thanks to dependency-parsing evaluation task. Namely, POS-tagging is a low-level syntactic task, which consists in assigning to each word its corresponding grammatical category. Dependency-parsing consists of higher order syntactic task like predicting the labeled syntactic tree capturing the syntactic relations between words. We evaluate the performance of our models using the standard UPOS accuracy for POS-tagging, and Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS) for dependency parsing. We assume gold tokenisation and gold word segmentation as provided in the UD treebanks.

Lexical tasks To test for word-level representation obtained through the different pre-training corpora and fine-tunings, Named Entity Recognition task (NER) was retained (4.3.2.). As it involves a sequence labeling task that

[9] Available at `https://github.com/ufal/acl2019_nested_ner`.

Treebank	Tokens	Words	Sentences	Genre
GSD	389 363	400 387	16 342	News Wiki. Blogs
Sequoia	68 615	70 567	3 099	Pop. Wiki. Med. EuroParl
Spoken	34 972	34 972	2 786	Oral transcip.
ParTUT	27 658	28 594	1 020	Oral Wiki. Legal

Table 5: Sizes of the 4 treebanks used in the evaluations of POS-tagging and dependency parsing.

consists in predicting which words refer to real-world objects, such as people, locations, artifacts and organizations, it directly probes the quality and specificity of semantic representations issued by the more or less balanced corpora under comparison.

4.3.1. POS-tagging and dependency parsing

Experiments were run using the Universal Dependencies (UD) paradigm and its corresponding UD POS-tag set (Petrov et al., 2011) and UD treebank collection version 2.2 (Nivre et al., 2018), which was used for the CoNLL 2018 shared task.

Different terms of comparisons were considered on the two downstream tasks of part-of-speech (POS) tagging and dependency parsing.

Treebanks test data-set We perform our work on the four freely available French UD treebanks in UD v2.2: GSD, Sequoia, Spoken, and ParTUT, presented in Table 5.

GSD treebank (McDonald et al., 2013) is the second-largest tree-bank available for French after the FTB (described in subsection 4.3.2.), it contains data from blogs, news, reviews, and Wikipedia.

Sequoia tree-bank (Candito et al., 2014) comprises more than 3000 sentences, from the French Europarl, the regional newspaper *L'Est Républicain*, the French Wikipedia and documents from the European Medicines Agency.

Spoken was automatically converted from the Rhapsodie tree-bank (Lacheret et al., 2014) with manual corrections. It consists of 57 sound samples of spoken French with phonetic transcription aligned with sound (word boundaries, syllables, and phonemes), syntactic and prosodic annotations.

Finally, **ParTUT** is a conversion of a multilingual parallel treebank developed at the University of Turin, and consisting of a variety of text genres, including talks, legal texts, and Wikipedia articles, among others; ParTUT data is derived from the already-existing parallel treebank, Par(allel)TUT (Sanguinetti and Bosco, 2015). Table 5 contains a summary comparing the sizes of the treebanks.

State-of-the-art For POS-tagging and Parsing we select as a baseline UDPipe Future (2.0), without any additional contextualized embeddings (Straka, 2018). This model was ranked 3rd in dependency parsing and 6th in POS-tagging during the CoNLL 2018 shared task (Seker et al., 2018). Notably, UDPipe Future provides us a strong baseline that does not make use of any pre-trained contextual embedding.

We report on Table 6 the published results on UDify by (Kondratyuk, 2019), a multitask and multilingual model based on mBERT that is near state-of-the-art on all UD languages including French for both POS-tagging and dependency parsing.

Finally, it is also relevant to compare our results with CamemBERT on the selected tasks, because compared to UDify it is the work that pushed the furthest the performance in fine-tuning end-to-end a BERT-based model.

4.3.2. Named Entity Recognition

Treebanks test data-set The benchmark data set from the French Treebank (FTB) (Abeillé et al., 2003) was selected in its 2008 version, as introduced by Candito and Crabbé (2009) and complemented with NER annotations by Sagot et al. (2012)[10]. The tree-bank, shows a large proportion of the entity mentions that are multi-word entities. We therefore report the three metrics that are commonly used to evaluate models: precision, recall, and F1 score.

NER State-of-the-art English has received the most attention in NER in the past, with some recent developments in German, Dutch and Spanish by Straková et al. (2019). In French, no extensive work has been done due to the limited availability of NER corpora. We compare our model with the stable baselines settled by (Dupont, 2018), who trained both CRF and BiLSTM-CRF architectures on the FTB and enhanced them using heuristics and pre-trained word-embeddings.

And additional term of comparison was identified in a recently released state-of-the-art language model for French, CamemBERT (Martin et al., 2019), based on the RoBERTa architecture pre-trained on the French sub-corpus of the newly available multilingual corpus OSCAR (Ortiz Suárez et al., 2019).

5. Results & Discussion

5.1. Dependency Parsing and POS-tagging

ELMo$_{\textbf{CaBeRnet}}$: a test for balance The word-embeddings representations offered by ELMo$_{\text{CaBeRnet}}$ are not only competitive but sometimes better than Wikipedia ones. One should keep in mind that almost all of the four treebanks we use in this section include Wikipedia data. ELMo$_{\text{CaBeRnet}}$ is reaching state-of-the-are results in POS-tagging on Spoken. Notably, it performs better than CamemBERT, the previous state of the art on this oral specialized tree-bank (cf. dark gray highlight on Table 6). We understand this results as a clear effect of balance when testing upon a purely spoken test-set. Importantly, this effect is difficultly explainable by the size of oral-style data in CaBeRnet. The oral sub-part is only one fifth of the total, and in this one fifth, only an even smaller amount of data comes from purely oral transcripts comparable the ones in the Spoken tree-bank, namely 67,444 words from Rhapsodie corpus, and 575,894 words form ORFEO. Hence, CaBeRnet's balanced oral language use shows to pay off in POS-tagging. These results are extremely surprising especially given the fact that

[10] The NER-annotated FTB contains approximately than 12k sentences, and more than 350k tokens were extracted from articles of *Le Monde* newspaper (1989 - 1995). As a whole, it encompasses 11,636 entity mentions distributed among 7 different types : 2025 mentions of "Person", 3761 of "Location", 2382 of "Organisation", 3357 of "Company", 67 of "Product", 15 of "POI" (Point of Interest) and 29 of "Fictional Character".

MODEL	GSD			SEQUOIA			SPOKEN			PARTUT		
	UPOS	UAS	LAS	UPOS	UAS	LAS	UPOS	UAS	LAS	UPOS	UAS	LAS
Baseline UDPipe Future	97.63	90.65	88.06	98.79	92.37	90.73	95.91	82.90	77.53	96.93	92.17	89.63
+ELMo$_{\text{CBT}}$	97.49	90.21	87.37	98.40	92.18	90.56	96.60	85.05	79.82	97.27	92.55	90.44
+ELMo$_{\text{Wikipedia}}$	97.92	92.13	89.77	99.22	94.28	92.97	97.28	85.61	80.79	**97.62**	94.01	91.78
+ELMo$_{\text{CaBeRnet}}$	97.87	92.02	89.62	99.33	94.42	93.14	***97.30***	85.39	80.63	97.43	94.02	91.86
+ELMo$_{\text{OSCAR}}$	97.85	92.41	90.05	99.30	94.43	93.25	97.10	85.83	**80.94**	97.47	**94.74**	**92.55**
+ELMo$_{\text{OSCAR+CaBeRnet}}$	**97.98**	**92.57**	**90.22**	**99.34**	**94.51**	**93.38**	97.24	**85.91**	80.93	97.58	94.47	92.05
State-of-the-art												
UDify	97.83	93.60	91.45	97.89	92.53	90.05	96.23	85.24	80.01	96.12	90.55	88.06
UDPipe Future + mBERT	97.98	92.55	90.31	*99.32*	94.88	93.81	97.23	*86.27*	*81.40*	*97.64*	94.51	92.47
CamemBERT	*98.19*	*94.82*	*92.47*	99.21	*95.56*	*94.39*	96.68	86.05	80.07	97.63	95.21	*92.90*

Table 6: Final POS and dependency parsing scores on 4 French treebanks (French GSD, Spoken, Sequoia and ParTUT), reported on test sets (4 averaged runs) assuming gold tokenisation. Best scores in bold, second to best underlined, state-of-the-art results in italics.

NER - RESULTS on FTB	Precision	Recall	F1
Baselines Models			
SEM (CRF) (Dupont, 2018)	87.89	82.34	85.02
LSTM-CRF (Dupont, 2018)	87.23	83.96	85.57
LSTM-CRF test models	85.87	81.35	83.55
+FastText	88.53	84.63	86.53
+FastText+ELMo$_{\text{CBT}}$	79.77	77.63	78.69
+FastText+ELMo$_{\text{Wikipedia}}$	88.87	87.56	88.21
+FastText+ELMo$_{\text{CaBeRnet}}$	88.91	87.22	88.06
+FastText+ELMo$_{\text{OSCAR}}$	88.89	88.43	88.66
+FastText+ELMo$_{\text{OSCAR+CaBeRnet}}$	**90.70**	**89.12**	**89.93**
State-of-the-art Models			
CamemBERT (Martin et al., 2019)	88.35	87.46	87.93

Table 7: NER Results on French Treebank (FTB): **best scores**, second to best.

our evaluation method was aiming at comparing the quality of word-embedding representations and not beating the state-of-the-art.

ELMo$_{\text{CaBeRnet}}$: a test for coverage From Table 6, we discover that not only balance, but also the broad and diverse genre converge of CaBeRnet may play a role in its POS-tagging success is we compare its results with ELMo$_{\text{CBT}}$ that also features oral dialogues in youth literature. The fact that ELMo$_{\text{CBT}}$ does not show a comparable performance in POS-tagging, can be interpreted as linked to its size, but possibly also to its lack of variety in genres, thus, suggesting the advantage of a comprehensive coverage of language use. This suggests that a balanced sample may enhance the convergence of generalization about oral-style from distinct genre that still imply oral-like dialogues like in fiction. In sum, broad coverage may contribute to enhancing representations about oral language.

The effect of balance on Fine-tuning For POS-tagging in GSD the results of ELMo$_{\text{OSCAR}}$ are in second place position compared to ELMo$_{\text{OSCAR+CaBeRnet}}$ that is extremely close to ELMo$_{\text{Wikipedia}}$. While in POS-tagging in ParTUT, ELMo$_{\text{Wikipedia}}$ exhibits better results than ELMo$_{\text{OSCAR}}$, and ELMo$_{\text{OSCAR+CaBeRnet}}$ is in second position.

Further comparing GSD and Sequoia scores from ELMo$_{\text{OSCAR}}$ and ELMo$_{\text{OSCAR+CaBeRnet}}$, we observe that fine-tuning with CaBeRnet the emdeddings that were pre-trained on OSCAR, yields better representations for the three tasks compared to both the original ELMo$_{\text{OSCAR}}$ and ELMo$_{\text{CaBeRnet}}$. However, fine-tuning does not always yield better findings than ELMo$_{\text{OSCAR}}$ on Spoken and ParTUT, where ELMo$_{\text{OSCAR+CaBeRnet}}$ places in second after ELMo$_{\text{OSCAR}}$ for parsing scores UAS/LAS (cf. Table 6).

A closer look on Parsing results reveals an interesting pattern of results across treebanks (see light gray highlights on Table 6). We see that for GSD and Sequoia the CaBeRnet fine-tuned version ELMo$_{\text{OSCAR+CaBeRnet}}$ compared to the pure OSCAR pre-trained ELMo$_{\text{OSCAR}}$ is achieving higher scores. While a reverse and less clear-cut pattern is observable for the other two treebanks, namely Spoken and ParTUT. This configuration can be explained if we understand this pattern as due to the reinforcement and unlearning of ELMo$_{\text{OSCAR}}$ representations during the process of fine-tuning. Specifically, we can observe that parsing scores are better on treebanks that share the kind of language use represented in CaBeRnet, while they are worst on corpora that are closer in language sample to OSCAR corpus, like Spoken and ParTuT. This calls for further developments of CaBeRnet (§6.).

ELMo$_{\text{CBT}}$: small but relevant ELMo$_{\text{CBT}}$ shows an intriguing pattern of results. Even if its scores are under the baseline on GSD and Sequoia, it yields over the baseline results for Spoken and ParTUT. Given its reduced size, one would expect it to overfit, this would explain the under baseline performance. However, this was not the case on Spoken and ParTUT treebanks, thus showing ELMo$_{\text{CBT}}$ contribution in generating representations that are useful to UDPipe model to achieve better results in POS-tagging and parsing tasks on the ParTUT and Spoken tree-banks. The presence of oral dialogues is certainly playing a role in this results' pattern. This unexpected result calls for further investigation on the impact of pre-training with reduced-size, noiseless, domain-specific corpora.

5.2. NER

For named entity recognition, LSTM-CRF +FastText +ELMo$_{\text{OSCAR+CaBeRnet}}$ achieves a better precision, recall and F1 than the traditional CRF-based SEM architectures (§ 4.3.2.) and CamemBERT, which is currently state-of-the-art.Importantly, LSTM-CRF +FastText +ELMo$_{\text{CaBeRnet}}$ reaches better results in finding entity mentions, than Wikipedia which is a highly specialized corpus in terms of vocabulary variety and size, as can be seen in the overwhelming total number of unique forms it contains (see Table 4). We can conclude that both pre-training and fine-tuning with CaBeRnet on ELMo OSCAR generates better word-embedding representations than Wikipedia in this downstream task.

CBT-fr NER results are under the LSTM-CRF baseline. This can possibly be explained by the distance in terms of topics and domain from FTB tree-bank (i.e. newspaper articles), or by the reduced-size of the corpus to yield good-enough representation to perform entity mentions recognition.

All in all, our evaluations confirm the effectiveness of large ELMo-based language models fine-tuned or pre-trained with a balanced and linguistically representative corpus, like CaBeRnet as opposed to domain-specific ones, or to an extra-large and noisy one like OSCAR.

6. Perspectives & Conclusion

The paper investigates the relevance of different types of corpora on ELMo's pre-training and fine-tuning. It confirms the effectiveness and quality of word-embeddings obtained through balanced and linguistically representative corpora.

By adding to UDPipe Future 5 differently trained ELMo language models that were pre-trained on qualitatively and quantitatively different corpora, our French Balanced Reference Corpus CaBeRnet unexpectedly establishes a new state-of-the-art for POS-tagging over previous monolingual (Straka, 2018) and multilingual approaches (Straka et al., 2019; Kondratyuk, 2019).

The proposed evaluation methods are showing that the two newly built corpora that are published here are not only relevant for neural NLP and language modeling in French, but that corpus balance shows to be a significant predictor of ELMo's accuracy on Spoken test data-set and for NER tasks.

Other perspective uses of CaBeRnet involve it use as a corpus offering a reference point for lexical frequency measures, like association measures. Its comparability with English COCA further grants the cross-linguistic validity of measures like Point-wise Mutual Information or DICE's Coefficient. The representativeness probed through our experimental approach are key aspects that allow such measures to be tested against psycho-linguistic and neuro-linguistic data as shown in previous neuro-imaging studies (Fabre et al., 2018).

The results obtained for the parsing tasks on ParTUT open a new perspective for the development of the French Balanced Reference Corpus, involving the enhancement of the terminological coverage of CaBeRnet. A sixth sub-part could be included to cover technical domains like legal and medical ones, and thereby enlarge the specialized lexical coverage of CaBeRnet. Further developments of this resource would involve an extension to cover user-generated content, ranging from well written blogs, tweets to more variable written productions like newspaper's comment or forums, as present in the CoMeRe corpus (Chanier et al., 2014).The computational experiments conducted here also show that pre-training language models like ELMo on a very small sample like the French Children Book Test corpus or CaBeRnet yields unexpected results. This opens a perspective for languages that have smaller training corpora. ELMo could be a better suited language model for those languages than it is for others having larger size resources.

Results on the NER task show that size - usually presented as the more important factor to enhance the precision of representation of word-embeddings - matters less than linguistic representativeness, as achieved through corpus linguistic balance. ELMo$_{\text{OSCAR+CaBeRnet}}$ sets state-of-the art results in NER (i.e. Precision, Recall and F1) that are superior than those obtained with a 30 times larger corpus, like OSCAR.

To conclude, our current evaluations show that linguistic quality in terms of *representativeness* and balance is yielding better performing contextualized word-embeddings.

Acknowledgments

We acknowledge Benoit Crabbé for his helpful suggestions at the beginning of reflection on balanced corpora. We are indebted to Yoann Dupont for his help in collecting data from Wikimedia dumps and for his critical comments. Olivier Bonami and Kim Gerdes conversations were instrumental. This work was supported by the French National Research Agency (ANR) under grant ANR-14-CERA-0001 and BASNUM (ANR-18-CE38-0003). The authors are grateful to Inria Sophia Antipolis - Méditerranée "Nef" computation cluster for providing resources and support.

Bibliographical References

Abeillé, A., Clément, L., and Toussenel, F., (2003). *Building a Treebank for French*, pages 165–187. Kluwer, Dordrecht.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226, 09.

Douglas Biber, editor. (1993). *Representativeness in Corpus Design.* In: Literary and Linguistic Computing 8.4.

Bonami, O. and Beniamine, S. (2015). Implicative structure and joint predictiveness. In Vito Pirelli, et al., editors, *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*, Pisa, Italy.

Burnard, L. (2007). 520 million words, 1990-present. In *The British National Corpus, version 3 - BNC XML Edition.*

Candito, M. and Crabbé, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proc. of IWPT'09*, Paris, France.

Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., and de la Clergerie, É. V. (2014). Deep syntax annotation of the sequoia french treebank. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources* *and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 2298–2305. European Language Resources Association (ELRA).

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J., and Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *JLCL - Journal for Language Technology and Computational Linguistics*, 29(2):1–30. Final version to Special Issue of JLCL (Journal of Language Technology and Computational Linguistics (JLCL, http://jlcl.org/): BUILDING AND ANNOTATING CORPORA OF COMPUTER-MEDIATED DISCOURSE: Issues and Challenges at the Interface of Corpus and Computational Linguistics (ed. by Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel).

Davies, M. (2008). 520 million words, 1990-present. In *The Corpus of Contemporary American English (COCA)*.

Dupont, Y. (2017). Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, page 42.

Dupont, Y. (2018). Exploration de traits pour la reconnaissance d'entit'es nomm'ees du français par apprentissage automatique. In *24e Conf'erence sur le Traitement Automatique des Langues Naturelles (TALN)*, page 42.

Fabre, M., Bhattasali, S., and Hale, J. (2018). Processing mwes: Neurocognitive bases of verbal mwes and lexical cohesiveness within mwes. In *Proceedings of the 14th Workshop on Multiword Expressions (COLING 2018), Santa Fe, NM.*

Grave, E., Mikolov, T., Joulin, A., and Bojanowski, P. (2017). Bag of tricks for efficient text classification. In Mirella Lapata, et al., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 427–431. Association for Computational Linguistics.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.* European Language Resources Association (ELRA).

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2015). The goldilocks principle: Reading children's books with explicit memory representations.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Kondratyuk, D. (2019). 75 languages, 1 model: Parsing universal dependencies universally. *CoRR*, abs/1904.02099.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., and Tchobanov, A. (2014). Rhapsodie: a prosodic-syntactic treebank for spoken French. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 295–301, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In Kevin Knight, et al., editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, É., Seddah, D., and Sagot, B. (2019). CamemBERT: a Tasty French Language Model. *arXiv e-prints*, page arXiv:1911.03894, Nov.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Bauer, J., Bellato, S., Bengoetxea, K., Bhat, R. A., Biagetti, E., Bick, E., Blokland, R., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Burchardt, A., Candito, M., Caron, B., Caron, G., Cebiroğlu Eryiğit, G., Celano, G. G. A., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dickerson, C., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Ephrem, B., Erjavec, T., Etienne, A., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ion, R., Irimia, E., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kahane, S., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Peng, S., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Poibeau, T., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rießler, M., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Sadde, S., Saleh, S., Samardžić, T., Samson, S., Sanguinetti, M., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shohibussirri, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Takahashi, Y., Tanaka, T., Tellier, I., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Vincze, V., Wallin, L., Washington, J. N., Williams, S., Wirén, M., Woldemariam, T., Wong, T.-s., Yan, C., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhang, M., and Zhu, H. (2018). Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In Piotr Bański, et al., editors, *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom, July. Leibniz-Institut für Deutsche Sprache.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Marilyn A. Walker, et al., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Sagot, B., Richard, M., and Stern, R. (2012). Annotation référentielle du corpus arboré de Paris 7 en entités nommées (referential named entity annotation of the paris 7 french treebank) [in french]. In Georges Antoniadis, et al., editors, *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN, Grenoble, France, June 4-8, 2012*, pages 535–542. ATALA/AFCP.

Sanguinetti, M. and Bosco, C. (2015). PartTUT: The Turin University Parallel Treebank. In Roberto Basili, et al., editors, *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, volume 589 of *Studies in Computational Intelligence*, pages 51–69. Springer.

Seker, A., More, A., and Tsarfaty, R. (2018). Universal morpho-syntactic parsing and the contribution of lexica: Analyzing the onlp lab submission to the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 208–215.

Straka, M., Straková, J., and Hajic, J. (2019). Evaluating contextualized embeddings on 54 languages in POS tagging, lemmatization and dependency parsing. *CoRR*, abs/1908.07448.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Straková, J., Straka, M., and Hajic, J. (2019). Neural architectures for nested NER through linearization. In Anna Korhonen, et al., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5326–5331. Association for Computational Linguistics.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2019). CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. *arXiv e-prints*, page arXiv:1911.00359, Nov.