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

Murielle Fabre, Pedro Javier Ortiz Suárez, Benoît Sagot, Éric Villemonte de La Clergerie

# 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 describes and compares the impact of different types and size of training corpora on language models like ELMo. By asking the fundamental question of quality versus quantity we evaluate four French corpora for training on parsing scores, POS-tagging and named-entities recognition downstream tasks. The paper studies the relevance of a new corpus, CaBeRnet, featuring 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 and balanced corpora will allow the language model to be more efficient and representative of a given language 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 the 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 and OSCAR-fr (Ortiz Suárez et al., 2019), but 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.

Following Biber (1993: 244), "representativeness refers to the extent to which a sample includes the full range of variability in a population", we adopt a balanced approach in sampling a wide spectrum of features 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. Thereby we contribute two newly built corpora. One purposed to be maximally representative of French language to yield good generalizations from, including a full range of language use variability, the French Balanced Reference Corpus - *CaBeRnet*. And a second that would yield a domain-specific language model training including both narrative material and oral language use, the *French Children Book Test* (CBT-fr).

Based on the underlying assumption that a linguistically representative corpus would possibly generate word-embeddings, that while being more representative of real language use, would tentatively preform better in downstream tasks. This paper provides an evaluation-based investigation of how a linguistically balanced corpus can yield improvements in the performance of neural language models like ELMo (Peters et al., 2018) in a given language. Specifically, we ask the contribution of oral language use in corpora, and therefore 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 to wide ranging corpora characterized by a variety of linguistic phenomena crawled from internet by ortizsuarez ortizsuarez, versus our newly built French Balanced Reference Corpus CaBeRnet, that features a wide and balanced coverage of cross-genre language use, including oral.

All in all, our evaluation results confirm the effectiveness of large ElMo-based language models fine-tuned or pre-trained with a balanced and linguistically representative corpus, like CaBeRnetFRanc as opposed to domain-specific ones and extra-large and noisy ones.

**Structure of the paper** The paper is organized as follows. Section 2. is dedicated to a descriptive overlook of corpus building and data collection. The construction process of our two newly brewed corpora CBT-fr and CaBeRnet is presented thoroughly in this section that summarises information details that can be found in corpus metadata. The achievement of linguistic balance in CaBeRnet is detailed in section 2.1. Statistics on the distribution of lexical, syntactic and morphological features of the different subparts of the corpus are also presented.

In section 3. the focus is give to several quantitative measures to characterize the corpora under analysis : average length of sentences, type-token ratio and morphological richness. The characteristics of CBT-fr and CaBeRnet are compared to the other corpora under analysis (OSCAR-fr, Wiki-fr) are to be found in this section.

Section 4. introduces the evaluation methods used to obtain the POS-tagging, NER and dependency Parsing results. Results are presented and discussed in Section 5.. Finally, we conclude in section 6. on the computational and linguistic relevance of fine-tuning obtained through balanced

and representative corpora. We conclude by broadening the discussion with a series of future developments to enrich CaBeRnet and further investigate the benefits of smaller and noiseless corpora in neural NLP research.
Resources associated to this paper encompass[1]: five version of FrELMo trained on the four corpora presented in this paper and two newly brewed corpora, including a French version of the balanced Corpus of Contemporary American English COCA (Davies, 2008) and one of the Children Book Test CBT (Hill et al., 2015).

## 2. Corpus Building

CaBeRnet corpus is meant to parallel COCA corpus[2], which contains more than 560 million words of text (20 million words each year 1990-2017) and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts (Davies, 2008). A second reference guiding our approach and building method is one of the precursor and a classical balanced reference: the BNC (Burnard, 2007). In that it aims at covering a wide variety of genres, with the intention to be a representative sample of spoken and written language.

### 2.1. Data Collection - CaBeRnet

CaBeRnet was obtained by compiling both existing datasets and web text from different sources (see Metadata - Lists), 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 written production).

#### 2.1.1. CaBeRnet- Oral Transcriptions

The oral sub-portion gathers both oral transcriptions (ORFEO and Rhapsodie[3]) and Films subtitles (Open Subtitles, `www.opensubtitles.org/fr`), pruned from diacritics and interlocutors tagging and time stamps. To these transcriptions, the French European Parliament Proceedings (1996-2011) as presented in Koehn (2005) contribute a sample of more complex oral style with longer sentences and richer vocabulary.

#### 2.1.2. 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[4].It was selected to match popular style because it is characterized by simplified written press style and a wide range of every-day topics characterizing local regional press.

#### 2.1.3. CaBeRnet- Fiction & Literature

The whole sub-portion of Fiction & Literature was compiled from march 2019's Wikisource dump and extracted using WikiExtractor.py, a script that extracts and cleans text from a Wikipedia database dumps, by performing template expansion and preprocessing of template definitions (`https://github.com/attardi/wikiextractor`).

#### 2.1.4. 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 modified version of WikiExtractor.py. Newspaper's content gathered by the Chambers-Rostand Corpus (i.e. Le Monde 2002-2003, La Dépêche 2002-2003, L'Humanité 2002-2003) and *Le Monde diplomatique* open-source corpus were assembled to represent a high register 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.[5]

#### 2.1.5. CaBeRnet- Academic

The academic genre was also built from different sources including WikiBooks and Wikipedia dump 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[6] was included to represent more concise written style characterizing scientific abstracts and proceedings.

| CaBeRnet | nbTokens | nb Unique Forms | Mo |
|---|---|---|---|
| Oral | 122,864,888 | 291,744 | 735,4 Mo |
| Popular | 131,444,017 | 458,521 | 758,5 Mo |
| News | 132,708,943 | 462,971 | 797,2 Mo |
| Fiction | 198,343,802 | 983,195 | 1 080 Mo |
| Academic | 126,431,211 | 1433663 | 810,8 Mo |
| Tot. | 711,792,861 | 2,558,513 | 4 190 Mo |

Table 1: Comparison of number of unique forms in the different genres represent by CaBeRnet partition into Oral, Popular, News, Fiction and Academic. Mo: Mega Octet. lemmatisation and tockenisation was achieved as described in section 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

[1]The link to the corpus and FrElMos will be available upon acceptance of the paper. Following the link the reader will have access to a dedicated website *cabernet-corpus.fr* where raw text version and metadata for each sub-part of the corpus are also be available.

[2]`https://www.corpusdata.org`

[3]ORFEO `www.cocoon.huma-num.fr/exist/crdo/`; Rhapsodie `www.projet-rhapsodie.fr`

[4]`www.cnrtl.fr/corpus/estrepublicain/`

[5]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.

[6]This corpus of proceedings builds on a subset of scientific articles presented at two conferences between 2007 to 2013, namely TALN (Traitement Automatique des Langues Naturelles) and RECITAL (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues). It consists of 586 articles for a total of about 2 million words.`redac.univ-tlse2.fr/corpus/taln_en.html`

compiling a rich set of regular expressions specific of each textual source to obtain clean plain text as an outcome.

### 2.2. Data Collection - French Children Book Test (CBT-fr)

The French Children Book Test (CBT-fr) was built upon its original English version of the Children Book Test (CBT), which consists of books that are freely available thanks to Project Gutenberg (`www.gutenberg.org`). This data-set can be found from `www.fb.ai/babi/` Hill et al. (2015).

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 (see www.cabernet-corpus.fr - Metadata). Chapter heads, titles, notes and 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 in the book selection was done to obtain a comparable corpus to the English CBT.

Our effort in equating the selection method and type of the books in English and French CBT is yielding a multilingual set of comparable corpus, containing texts that are collected using the same sampling frame and similar balance and representativeness.

| CBT-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: Comparison of number of words in the corpora under study.

## 3. Corpora Descriptive Comparison - Method

Two tokanization methods were used, 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, Segmenteur-Étiqueteur Markovien standalone Dupont (2017).

All corpora were then randomly shuffled by sentence to then were shuffled sake of select samples of 3 million words, that would allow to compare then in terms of lexical composition (Type-Token Ratio).

The second tokenization method was run only on the 3 million words samples (see Table 5 to automatically tag them into part-of-speech and lemmatize them. For this purpose we used the TreeTagger.[7] .

### 3.1. Corpora Size and Composition

Length of sentences is a simple measure to quantify both sentence syntactic complexity and genre. Hence, the average length of a sentences reported in Table 3 shows interesting patterns of distributions across genres.

| CORPUS | WORDS | 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: Comparison of number of words in the corpora under study.

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 corpus comparison a corpus of French text extracted from Common Crawl[8]. Specifically, 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 corpora extracted from Common Crawl, from the 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., 2018a) 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 specialised 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.[9]

#### 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 - GitHub, see 2.1.3.). As shown in Table 3, this data-set is relatively large with

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

[8] `https://commoncrawl.org/about/`

[9] We did not use CCNet because of its difficult availability and download.

its around 660 million words, sentences are relatively long 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.

### 3.2. Corpora Lexical richness

Focusing on a useful measure of complexity that documents lexical richness or variety in vocabulary we present type-token ration (TTR) of the corpora under analysis. Usually 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 4 summarises the lexical variety of the five sub-portions of CaBeRnet, respectively taken as representative of Oral, Popular, Fiction, News, and Academic genres. The domain diversity of academic texts is here evident.

| CORPUS SUB SET | NB TOCKENS | NB 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 4: Comparison of proportion of Forms in 3 millions words samples from the different register represent by CaBeRnet partition into Oral, Popular, News, Fiction and Academic.

### 3.3. Corpora Morphological richness

To select a measure that would help quantifying the different corpora morphological richness, we follow (Bonami and Beniamine, 2015) and evaluated on randomly selected samples of 3 million words from each corpus under analysis the proportion of lemmas with multiple forms in a given vocabulary size, see Table 5.

| CBT-FR 3 M SAMPLE | |
|---|---|
| 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 |
| CABERNETFRANC 3 M SAMPLE | |
| number of different lemmas | 30 488 |
| total number of forms | 180.089 |
| mean number of forms per lemma | 6,19 |
| Number of lemmas having more than one form : | 15.927 |
| Percentage of lemmas with multiple forms | 52,24 |
| FRWIKI 3 M SAMPLE | |
| number of different lemmas | 31.385 |
| total number of forms | 238.121 |
| mean number of forms per lemma | 7,85 |
| Number of lemmas having more than one form : | 15.182 |
| Percentage of lemmas with multiple forms | 48,37 |
| OSCAR 3 M SAMPLE | |
| number of different lemmas | 31.204 |
| total number of forms | 190.078 |
| mean number of forms per lemma | 6,40 |
| Number of lemmas having more than one form : | 16.480 |
| Percentage of lemmas with multiple forms | 52,81 |

Table 5: Lexical Statistics comparing morphological richness of the corpora under study.

## 4. Corpora Computational Evaluation tasks

This section reports the experiments designed to better understand the computational impact of the quality and linguistic balance versus size of ELMo's (Peters et al., 2018) training corpora with the pre-training method (§4.1.) and tasks described in 4.2.

### 4.1. ELMo Pre-traing and 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 of FRrELMo-based language models. While *Method 2* is based on pre-training on a huge corpus + fine-tuning with our Reference Balanced Corpus for French CaBeRnet, $\text{ELMo}_{\text{OSCAR+CaBeRnet}}$. Hence, Method 1 implies pure pre-traing with the corpora uner compariaon end yields the following four language models : $\text{ELMo}_{\text{OSCAR}}$, $\text{ELMo}_{\text{Wikipedia}}$, $\text{ELMo}_{\text{CaBeRnet}}$ and $\text{ELMo}_{\text{CBT}}$. The fine-tuning method (i.e. Method 2) was applied only to $\text{ELMo}_{\text{OSCAR}}$ fine-tuned with CaBeRnet.

Methodologically, we seek to understand through a computational approach of fine-tuning with resources that are up to 30 times smaller than pre-training corpora has a observable impact on NLP tasks scores. It is namely for this reason, we selected ELMo which not only performs generally better on sequence tagging than other architectures, but is also better suited to pre-train on small corpora because of its inferior rage of parameter (93.6 million) compared to RoBERTa-base architecture used for CamBERT (BERT-base, 12,110 million - Transformer).

**Embeddings from Language Models** (ELMo) (Peters et al., 2018) is a neurla Language Model, that is, a model that given a sequence of $N$ input tokens, $(t_1, t_2, ..., t_N)$, computes the probability of the sequence by modeling the probability of token $t_k$ given the history $(t_1, ..., t_{k-1})$:

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1}).$$

ELMo in particular uses a biLM consisting of LSTM layers, that is, it concatenates both a forward and a backward language model generating a contextualized bi-directional representation of each token in a given sentence.
All the training experiments are performed with a fully trained model for 10 epochs. As is was done for the original English ELMo (Peters et al., 2018). Hence, all our FRrELMo-based language models build on top of the UDPipe Future parser and tagger (Straka, 2018) as implemented in Straka et al. (2019) which is open source and freely available.[10]

[10]https://github.com/CoNLL-UD-2018/UDPipe-Future

**The UDPipe Future architecture** is a multi-task model that predicts POS tags, lemmas and dependency trees jointly. It consists of an embedding step containing: character level word-embeddings that are trained along the rest of the network, pre-trained word-embeddings[11], a randomly initialized word embeddings that are trained along the rest of the network, and contextualized word-embeddings for which we plug our customly trained ELMos.
All these embeddings are then concatenated and are fed to two shared Bi-LSTMs that generate shared representations that are forwarded to two separate Bi-LSTMs; one that is followed by a softmax layer and predicts the POS tags, and another that is followed by a Deep Bi-Affine Attention Layer (Dozat and Manning, 2017) that produces dependency trees.
In other words 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[12] and it is designed to easily pre-append contextualized word embeddings to the model.

## 4.2. Evaluation Tasks - Method

We distinguish three main evaluation tasks that were performed by ELMo pre-trained on OSCAR ($ELMo_{OSCAR}$), frWIKI ($ELMo_{Wikipedia}$), CaBeRnet ($ELMo_{CaBeRnet}$) and CBT-fr ($ELMo_{CBT}$) and comparing them with ElMo pre-trained on OSCAR and fine-tuned with CaBeRnet, i.e. $ELMo_{OSCAR+CaBeRnet}$ (see Results Table 7). The focus is given here on what is evaluated of the quality of contextualized word-embeddings obtained from different pre-training corpora under comparison. Crucially, manipulating the presence of oral transcriptions and oral proceeding will be interesting to understand the impact on accuracy of our language model and their impact on several language tasks after fine-tuning. Our development experiments compare the corpora presented in Table 3.

**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.

**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 (4.2.2.). As it involves a sequence labeling task that consists in predicting which words refer to real-world objects, such as people, locations, artifacts and organisations, it directly probes the quality and specificity of semantic representations issued by the more or less balanced corpora under comparison.

### 4.2.1. Part-of-speech tagging and dependency parsing

Different types of terms of comparisons were considered on the two downstream tasks of part-of-speech (POS) tagging and dependency parsing.
For POS-tagging and Parsing we select as a baseline UDPipe Future (2.0), without any additional contextualized embeddings. Here we only report in Table xxx the publish results by xxx On the other hand, UDify, UniPipi Future + mBERT and CamemBERT represent different terms of comparison for state-of-the-art results on Parsing and Pos-tagging as detailed here under.
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.

**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 here under (cf. Table 6).
**GSD** treebank (McDonald et al., 2013) is the second-largest tree-bank available for French after the FTB (described in subsection 4.2.2.), it contains data from blogs, news articles, reviews, and Wikipedia.
The **Sequoia** treebank[13] (Candito and Seddah, 2012; 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** is a corpus converted automatically from the Rhapsodie treebank[14] (Lacheret et al., 2014; Bawden et al., 2014) with manual corrections. It consists of 57 sound samples of spoken French with orthographic transcription and 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 6 contains a summary comparing the sizes of the treebanks[15].
We evaluate the performance of our models using the standard UPOS accuracy for POS-tagging, and Unlabeled Attachment Score (UAS) and Labeled Attachment Score

[11] We use the French fastText embeddings distributed by (Grave et al., 2018b).

[12] https://github.com/ufal/acl2019_nested_ner

[13] https://deep-sequoia.inria.fr

[14] https://www.projet-rhapsodie.fr

[15] https://universaldependencies.org

| Treebank | nb Tokens | nb Words | nb Sentences | Genre |
|---|---|---|---|---|
| GSD | 389 363 | 400 387 | 16 342 | News + Wikipedia + Blogs |
| Sequoia | 68 615 | 70 567 | 3 099 | Popular + Wikipedia + Medicine + EuroParl |
| Spoken | 34 972 | 34 972 | 2 786 | Oral transcription |
| ParTUT | 27 658 | 28 594 | 1 020 | Oral + Wikipedia + Legal |

Table 6: Sizes in Number of tokens, words and phrases of the 4 treebanks used in the evaluations of POS-tagging and dependency parsing.

(LAS) for dependency parsing. We assume gold tokenisation and gold word segmentation as provided in the UD treebanks.

**State-of-the-art** We compare our models to UDify (Kondratyuk, 2019). UDify is 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.
It is relevant to compare our results CamemBERT on those tasks because compared to UDify is the work that pushed the furthest the performance in fine-tuning end-to-end a BERT-based model on downstream POS-tagging and dependency parsing. Finally, we compare our model to UDPipe Future (Straka, 2018), a model ranked 3rd in dependency parsing and 6th in POS-tagging during the CoNLL 2018 shared task (Seker et al., 2018). UDPipe Future provides us a strong baseline that does not make use of any pre-trained contextual embedding.

#### 4.2.2. Named Entity Recognition

**Treebanks test data-set** The benchmark data set from the French Treebank[16] (FTB) (Abeillé et al., 2003) was selected in its 2008 versio, as introduced by Candito and Crabbé (2009) and complemented with NER annotations by Sagot et al. (2012).
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".
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. Specifically, (1) precision measures account for the percentage of entities found by the system that are correctly tagged, (2) recall measures sand for the percentage of named entities present in the corpus that are found, and (3) F1 score measure combines both precision and recall measures giving a global measure of a model's performance.

**NER State-of-the-art** Most of the advances in NER haven been achieved in English, particularly focusing on the CoNLL 2003 (Sang and Meulder, 2003) and the Ontonotes v5 (Pradhan et al., 2012; Pradhan et al., 2013) English corpora.

[16]This data-set has only been stored and used on Inria's servers after signing the research-only agreement.

Importantly, NER task was traditionally tackled using Conditional Random Fields (CRF) (Lafferty et al., 2001), CRFs were later used as decoding layers for Bi-LSTM architectures (Huang et al., 2015; Lample et al., 2016) showing considerable improvements over CRFs alone. Later, these Bi-LSTM-CRF architectures were enhanced with contextualised word-embeddings which yet again brought major improvements to the task (Peters et al., 2018; Akbik et al., 2018). Finally, large pre-trained architectures settled the current state of the art showing a small yet important improvement over previous NER-specific architectures (Devlin et al., 2019; Baevski et al., 2019).

In non-English NER the CoNLL 2002 shared task included NER corpora for Spanish and Dutch corpora (Sang, 2002) while the CoNLL 2003 included a German corpus (Sang and Meulder, 2003). Here the recent efforts of (Straková et al., 2019) settled the state of the art for Spanish and Dutch, while (Akbik et al., 2018) did it for German.

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, based on the RoBERTa architecture pre-trained on the French sub-corpus of the newly available multilingual corpus OSCAR ((Martin et al., 2019)).

## 5. Results & Discussion

### 5.1. Dependency Parsing and POS-tagging

#### 5.1.1. $\text{ELMo}_{\text{CaBeRnet}}$: Spoken a test for balance

$\text{ELMo}_{\text{CaBeRnet}}$ offers representation that are not only competitive but sometimes better than Wikipedia especially considering that the majority of evaluation tree-banks are built on Wikipedia data. For Spoken $\text{ELMo}_{\text{CaBeRnet}}$ is reaching state-of-the-are results in POS-tagging on this oral specialized tree-bank (see dark gray highlight on Table 7. It performs better than CamemBERT which was the previous the state of the art on Spoken.

$\text{ELMo}_{\text{CaBeRnet}}$ shows a clear effect of balance when tested upon a purely spoken test-set like the Spoken tree-bank. Importantly, this effect is difficultly explainable by the size of oral style in CaBeRnet, because oral sub-part is only one fifth of the total. Furthermore, in this one fifth, only an even smaller amount words comes from pure oral transcripts which constitute the Spoken tree-bank. Namely, 67 444 words from the Rhapsodie corpus, and 575 894 words form ORFEO. We understand this result as a direct consequence of the fact that CaBeRnet contains a balanced amount of oral language use, which shows to pay off in POS-tagging.

These results are extremely surprising especially given the fact that our evaluation method was fundamentally aiming at comparing the quality of word-embedding representations and not beating state-of-the-art results.

| 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$_{CBT}$ | 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$_{Wikipedia}$ | **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$_{CaBeRnet}$ | 97.76 | 91.91 | 89.49 | 99.27 | 94.65 | 93.40 | ***97.32*** | 85.63 | 80.61 | 97.58 | 94.24 | 91.90 |
| +ELMo$_{OSCAR}$ | 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$_{OSCAR+CaBeRnet}$ | 97.88 | **92.67** | **90.34** | 99.26 | **94.75** | **93.54** | 97.22 | 85.77 | 80.80 | 97.50 | 94.66 | 92.43 |
| *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 7: Final POS and dependency parsing scores of CamemBERT and mBERT (fine-tuned in the exact same conditions as CamemBERT), UDify as reported in the original paper 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 | | | |
|---|---|---|---|
| Model | Precision | Recall | F1 |
| *Baselines* | | | |
| SEM (CRF) (Dupont, 2018) | 87.89 | 82.34 | 85.02 |
| LSTM-CRF (Dupont, 2018) | 87.23 | 83.96 | 85.57 |
| LSTM-CRF | 85.87 | 81.35 | 83.55 |
| +FastText | 88.53 | 84.63 | 86.53 |
| +FastText+ELMo$_{CBT}$ | 79.77 | 77.63 | 78.69 |
| +FastText+ELMo$_{Wikipedia}$ | 88.87 | 87.56 | 88.21 |
| +FastText+ELMo$_{CaBeRnet}$ | 88.82 | 87.82 | 88.32 |
| +FastText+ELMo$_{OSCAR}$ | 88.89 | ***88.43*** | ***88.66*** |
| +FastText+ELMo$_{OSCAR+CaBeRnet}$ | ***88.93*** | 88.08 | 88.50 |
| *Baselines* | | | |
| CamemBERT | 88.35 | 87.46 | 87.93 |

Table 8: Results for NER on the FTB. Best scores in bold, second to best underlined.

#### 5.1.2. ELMo$_{CaBeRnet}$: a test for coverage

From Table 7, 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. Broad coverage possibly contributes to enhancing representations about oral language, in that a balanced sample may enhance the convergence of generalization about oral style from distinct genre that still imply oral like dialogues in fiction narratives. ELMo$_{CBT}$ also features oral dialogues in youth literature but does not show the same results because of the lack of variety of genres, thus demonstrating again the advantage of a comprehensive coverage of language use.

#### 5.1.3. The effect of balance on ELMo OSCAR of CaBeRnet Fine-tuning

Comparing ELMo$_{OSCAR}$ and ELMo$_{OSCAR+CaBeRnet}$ we can observe that for GSD and Sequoia fine-tuning OSCAR pre-trained emdedding with CaBeRnet yields better representations, especially on UAS and LAS results. However, fine-tuning does not always yields better findings as one can observe in Spoken and ParTUT tree-banks, see Table 7.

For POS-tagging in GSD and ParTUT the results of ELMo$_{OSCAR}$ and ELMo$_{OSCAR+CaBeRnet}$ are in second place position compared to ELMo Wikipedia, but are still extremely close.

As for parsing results, we can observe in Table 7 a interesting pattern of results across treebanks highlighted in light gray. We see that for GSD and Sequoia the CaBeRnet fine-tuned version ELMo$_{OSCAR+CaBeRnet}$ compared to the pure Oscar pre-trained ELMo$_{OSCAR}$ is achieving higher scores. while the revers pattern is observable for the other two treebacks, namely Spoken and ParTUT. This configuration can be explained if we understand this pattern as due to the reinforcement and unlearning of ELMo$_{OSCAR}$ of some of its representations during the process of fine-tuning. Specifically, we can observe that parsing scores are better on treebaks 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.

#### 5.1.4. ELMo$_{CBT}$: small but relevant contribution

ELMo$_{CBT}$ shows a very interesting pattern of results. Even if its results are under the baseline in GSD and Sequoia, it yields better results than the baseline for Spoken and ParTUT. Given its reduced size, we were expecting it to overfit, which would explain an under baseline performance. However, this was not the case on Spoken and ParTUT treebanks. These results demonstrate ELMo$_{CBT}$ contribution in generating representations that are useful to UDPipe model to achieve better results in POS-tagging and dependency parsing tasks on the ParTUT treebank. The presence of oral dialogues is certainly playing a role in this pattern of results. This astonishing result calls for further investigation on the impact of pre-training wit reduced-size, noiseless and domain-specific corpora of the kind of CBT-fr.

### 5.2. NER

For named entity recognition, our experiments show that LSTM-CRF+FastText+ELMo$_{OSCAR+CaBeRnet}$ achieves a better precision than the traditional CRF-based SEM architectures described above in Section 4.2.2. (CRF and Bi-LSTM+CRF) and CamemBERT, which is currently state-of-the-art.

Importantly, LSTM-CRF + FastText + ELMo$_{CaBeRnet}$ 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 forms reported in Table 5. We can conclude that fine-tuning with CaBeRnet on ELMo OSCAR generates better word-embedding representations than Wikipedia in this task. Overall, NER scores shows improvements compared to CamemBERT.

Fine-tuning with CaBeRnet has a second effect on recall, we understand this slight drop as possibly due to unlearning of the wide spectrum of vocabulary that is in OSCAR and not in CaBeRnet. For instance the whole french Wikipedia is included in OSCAR and not in CaBeRnet. Nonetheless, it has to be noted that these scores are still better than previous state-of-the-art.

CBT-fr is under the baseline LSTM-CRF. This can possibly be explained because the corpus is very distant from FTB tree-bank (i.e. newspaper articles) in terms of topics and domain, or that the size of the corpus is too little to yield good-enough representation to perform entity mentions recognition.

All in all, we showed that CaBeRnet corpus can reliably be used as a basis for training neural language models that perform in down-stream tasks, as well as suited for the creation of balanced lexical frequency-based dictionary entries, grammar studies, other language reference materials.

## 6. Perspectives & Conclusion

The paper investigates the relevance of different types of corpora on ELMo pre-training and fine-tuning, and confirms the effectiveness of pre-trained language models with a balanced and linguistically representative corpus, like CaBeRnetFRanc, on several downstream tasks.

By adding to UDPipe Future 5 differently trained ELMo language model that were pre-trained on qualitatively and quantitatively different corpora, our French Balanced Reference Corpus CaBeRnet shows on three different downstream tasks for French (POS-tagging, dependency parsing, named-entity recognition), achieves to improve the state-of-the-art for POS-tagging over previous monolingual and multilingual approaches.

The proposed evaluation methods are showing that the two newly built corpora that we publish here are relevant for neural NLP and language modelling in French. Corpus balance shows to be a significant predictor of ELMo's accuracy on Spoken test data-set and for NER tasks. It goes without saying that a balanced corpus like CaBeRnet will be useful to calculate stable lexical frequency measures, like association measures and grant their comparability cross-linguistic comparability with English. The stability and representativeness probed through our experimental approach are key aspects that allow measures like Pointwise Mutual Information or DICE's Coefficient to be tested against psycho-linguistic and neuro-linguistic data as show in previous neuro-imaging studies (Fabre et al., 2018; Fabre et al., 2019; Fabre et al., 2020)

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 development of this resource would additionally consider a further 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.[17]

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 thorough balanced corpus building. $\text{ELMo}_{\text{CaBeRnet}}$ and $\text{ELMo}_{\text{OSCAR+CaBeRnet}}$ set new state-of-the art results that are superior than those obtained with a 30 times larger corpus, respectively on POS-tagging and NER. The computational experiments conducted here, namely, show that pre-training language models like ELMo on a very small sample like the French Children Book Test corpus (6 million words), or on a relatively small corpus like CaBeRnet yields unexpected results. This opens a perspective for languages that have a smaller training thesaurus : ELMo could a better suited language model for those languages than it is for others having larger size resources. In the same line, an additional perspective to this work is to better understand why we observe better NER scores with ELMo architecture than we do with BERT-base language model.

In sum, this paper offers three main contributions: (1) two newly built corpora one French Balanced Reference Corpus and a second domain-specific corpus having both oral and written style, (2) five versions of FrELMo, and (3) a whole array of computational results that deepen our understanding on the effects of balance and register in NLP evaluation. To conclude, our current evaluations results 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.

## Bibliographical References

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

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In Emily M. Bender, et al., editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1638–1649. Association for Computational Linguistics.

[17] https://repository.ortolang.fr/api/content/comere/v2/comere.html

Georges Antoniadis, et al., editors. (2012). *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN, Grenoble, France, June 4-8, 2012*. ATALA/AFCP.

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., and Auli, M. (2019). Cloze-driven pretraining of self-attention networks. *CoRR*, abs/1903.07785.

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.

Bawden, R., Botalla, M.-A., Gerdes, K., and Kahane, S. (2014). Correcting and validating syntactic dependency in the spoken French treebank rhapsodie. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2320–2325, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Bonami, O. and Beniamine, S. (2015). Implicative structure and joint predictiveness.

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. and Seddah, D. (2012). 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 Antoniadis et al. (Antoniadis et al., 2012), pages 321–334.

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).

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

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dozat, T. and Manning, C. D. (2017). Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

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.*

Fabre, M., Bhattasali, S., Luh, W.-M., Saied, H. A., Constant, M., Pallier, C., Brennan, J. R., Spreng, R. N., and Hale, J. (2019). Localising memory retrieval and syntactic composition: an fmri study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, 34(4):491–510.

Fabre, M., Bhattasali, S., Pallier, C., and Hale, J. (2020). Modeling conventionalization and predictability in multiword expressions at the brain level. In *Proceedings of the Society for Computation in Linguistics (SCiL 2020) New Orleans, LA.*, pages 2331–2336.

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. (2018a). 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).

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018b). Learning word vectors for 157 languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan, May. European Language Resource Association.

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

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

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).

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley et al., editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

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 H`ông, 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., Nguy˜ên Thị, L., Nguy˜ên 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 Com-*

*putational 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*.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In Sameer Pradhan, et al., editors, *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, EMNLP-CoNLL 2012, July 13, 2012, Jeju Island, Korea*, pages 1–40, Jeju Island, Korea.

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y., and Zhong, Z. (2013). Towards robust linguistic analysis using ontonotes. In Julia Hockenmaier et al., editors, *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 143–152. ACL.

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 Antoniadis et al. (Antoniadis et al., 2012), pages 535–542.

Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Walter Daelemans et al., editors, *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Sang, E. F. T. K. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In Dan Roth et al., editors, *Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002, Held in cooperation with COLING 2002, Taipei, Taiwan, 2002*. ACL.

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.