Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1104–1113
Marseille, 20-25 June 2022

BERTrade: Using Contextual Embeddings to Parse Old French

Loïc Grobol[1,2], Mathilde Regnault[3,4], Pedro Ortiz Suárez[5,6], Benoît Sagot[5], Laurent Romary[5], Benoit Crabbé[7]

(1) Modyco, Université Paris Nanterre and CNRS, Nanterre France
(2) LIFO, Université d'Orléans and INSA Centre – Val-de-Loire, Orléans, France
(3) Lattice, CNRS, ENS, PSL and Université Sorbonne Nouvelle, Paris France
(4) Institut für Linguistik/Romanistik (ILR), Universität Stuttgart, Deutschland
(5) Inria, Paris, France
(6) Sorbonne Université, Paris, France
(7) LLF, CNRS and Université Paris Cité, France
lgrobol@parisnanterre.fr, mathilde.regnault@ling.uni-stuttgart.de,
{pedro.ortiz,benoit.sagot,laurent.romary}@inria.fr, benoit.crabbe@u-paris.fr

Abstract

The successes of contextual word embeddings learned by training large-scale language models, while remarkable, have mostly occurred for languages where significant amounts of raw texts are available and where annotated data in downstream tasks have a relatively regular spelling. Conversely, it is not yet completely clear if these models are also well suited for lesser-resourced and more irregular languages. We study the case of Old French, which is in the interesting position of having relatively limited amount of available raw text, but enough annotated resources to assess the relevance of contextual word embedding models for downstream NLP tasks. In particular, we use POS-tagging and dependency parsing to evaluate the quality of such models in a large array of configurations, including models trained from scratch from small amounts of raw text and models pre-trained on other languages but fine-tuned on Medieval French data.

Keywords: Old French, Contextual word embeddings, Dependency Parsing, Part of Speech Tagging

1. Introduction

There is a growing interest in digital humanities for automatic processing and annotation of historical texts. In this work, we study how to take advantage of current NLP models of the BERT family to advance the state of the art in processing historical languages, taking Old French (9th-13th century French) as a use case.

Old French is one of the historical languages for which we have the largest amount of syntactically annotated data, and we expect that our results on these language states may be generalised and used as a source of inspiration for researchers currently developing annotated resources for other historical languages.

Using contextual word embeddings as input representations has brought clear gains in performances for most of the NLP tasks for which they have been used. However, this has mostly been attested in languages where sufficient (raw) linguistic data is available. For less-resourced languages, the most common approach has been to leverage multilingual models such as mBERT (Devlin et al. 2019) Historical languages are typical cases where available linguistic data is limited, with no chance of acquiring new texts. They are also not normalized by spelling and institutional conventions and tend to be more heterogeneous than contemporary lesser-resourced languages.

Old French is a particularly interesting language for this kind of study, since relatively to its limited amount of available raw text, its volume of *annotated* linguistic data is quite high, due to the existence of the SRCMF dependency treebank (Prévost and Stein 2013) and its latest incarnation in the Universal Dependency project (Nivre et al. 2020), which boasts around 17.7 K sentences[1] for around 171 K words.

Another interesting property of Old French is its proximity to a well-resourced language, namely contemporary French, for which monolingual contextual embeddings models exist and have been shown to be relevant for dependency parsing (Le et al. 2020; Martin et al. 2020). Last, but certainly not least, the design of an accurate syntactic parser for Old French would be a very valuable tool for computer-assisted linguistic studies. Indeed, studying the historical variation of syntax in a language that lacks both native speakers and centralized standard variants can be very challenging, due to the prohibitive cost of manual annotation. Automatic syntactic annotations, either as a "silver-standard" truth or as a bootstrapping step towards manual annotation, can drastically reduce that cost.

In this work, exploiting this currently unique situation of Old French among lesser-resourced and historical languages, we use dependency parsing and POS-tagging of Old French as probes of the relevance of contextual embeddings in a context of high heterogeneity and relative scarcity of data. More precisely, we consider several neural language models, some of which trained or fine-tuned on a new corpus of raw Old and Middle French texts, and use their internal representations of words as inputs to train taggers and parsers on the SRCMF treebank. The resulting tagging and parsing scores then serve

[1] Putting it in the second place of all French language treebanks in number of sentences.

as an evaluation of the quality and usefulness of these representations. We claim the following contributions:

- We provide empirical evidence that contextual embeddings are relevant for historical language processing, even when no data is available beyond the treebank used to train a parser.

- We provide a comparative study of several strategies for obtaining such contextual embeddings. Specifically, we compare cases where raw data is available in the target language and cases where existing contextual embeddings are available for the contemporary counterpart of a historical language.

- We release two publicly available resources for Old French: BERTrade[2][3], a set of contextual word embedding models ; and a state-of-the-art POS-tagging and dependency parsing model[4].

The paper is organized as follows. Section 2 provides an overview of related work that aims at taking advantage of the BERT family of language models in scenarios where the amount of data is limited. In Section 3 we provide a description of the dataset we gathered to conduct our experiments, and finally we report experiments in Section 4 involving reusing BERT from other languages and training BERT models on Old French.

2. Related work

Since the introduction of contextualized word representations (Peters et al. 2018; Akbik et al. 2018; Devlin et al. 2019) and the many improvements proposed for them in the consumption of computational resources (Clark et al. 2020), in the amount of data required to fine-tune them (Raffel et al. 2020), and more recently in the length of the contextual window (Xiong et al. 2021); there have also been important advancements from a digital humanities point of view on *unsupervised domain adaptation* (Ramponi and Plank 2020). In this case, one specializes a language model to a particular domain with unlabeled data in order to improve performance in downstream tasks. This can be achieved by pre-training the models from scratch with specialized data (Beltagy et al. 2019) or by continuing the training of a general model with a new corpus (Lee et al. 2019; Peng et al. 2019). This last method has already been successfully implemented in the context of historical languages, in particular Han and Eisenstein (2019) showed that one can successfully adapt the original BERT (Devlin et al. 2019) to Early Modern English by continuing the pre-training on historical raw texts.

In a multilingual context, transformer-based models such as mBERT have been adapted to low-resource languages and evaluated in dependency parsing and POS-tagging, showing promising results (Chau et al. 2020; Muller et al. 2020; Gururangan et al. 2020; Z. Wang et al. 2020). However, this multilingual approach has also been criticized for favoring monolingual pre-training even when data is scarce (Virtanen et al. 2019; Ortiz Suárez et al. 2020). Indeed, even when only small pre-training corpora are available, BERT-like models have also been successfully pre-trained, resulting in well-performing models (Micheli et al. 2020). Furthermore, compact BERT-like models have also been studied (Turc et al. 2019) and might prove useful in data constrained conditions, such as monolingual pre-training of contextualized word representation for low-resource languages.

Regarding corpora for historical languages, very few of them have manually annotated syntactical resources for their medieval states. English has three such treebanks (University of Oxford 2001; Kroch et al. 2000; Traugott and Pintzuk 2008) for Old and Middle English. The TOROT treebank for Old Church Slavonic, Old East Slavonic and Middle Russian is another large resource (Berdicevskis and Eckhoff 2020). There is a treebank for Medieval Latin as well, the *Index Thomisticus Treebank* (Passarotti 2019). To our knowledge, the last large treebank containing medieval texts is IcePaHC for Icelandic (Rögnvaldsson et al. 2012). Some other corpora were annotated automatically in order to reduce the cost of annotation. For example, Rocio et al. (2003) adapted a parsing pipeline for contemporary Portuguese and Lee and Kong (2016) used a previously annotated treebank (Lee and Kong 2012) to parse a larger medieval Chinese corpus. Concerning contemporary regional Romance languages, Miletic et al. (2020) also used a smaller treebank to generate new annotations, and concluded that using similar languages to train a model does not improve parsing. Although there are many resources for Latin, and some for Ancient Greek, we do not include them here, because they do not face the same challenges as medieval states of language, in particular the high level of spelling variability. Lastly, concerning dependency parsing and POS-tagging of Old French in particular, the works of Guibon et al. (2014) and Stein (2014) and Stein (2016) are noteworthy. However, they use very different approaches to the one used in this paper and evaluate on previous versions of SRCMF, with incompatible annotation choices and slightly different texts. For the UD version of SRCMF, the most notable work is that of the winner of the *CoNLL 2018 Shared Task* (Zeman et al. 2018), UDPipe 2.0 (Straka 2018), which was later enhanced by including contextualized word embeddings (Straka et al. 2019).

3. Data

This section describes the raw corpus of Medieval French we gathered in order to train unsupervised language models for Old French. To our knowledge, it is one of the largest such dataset gathered for Medieval French, although it remains quite small (55 MiB in total) relatively to the corpora usually used for pre-training contextual embeddings models.

[2] `https://doi.org/10.5281/zenodo.6461220`

[3] *Bertrade de Laon*, also known as *Berthe au Grand Pied* was the mother of Charlemagne.

[4] `https://github.com/hopsparser/hopsparser/blob/main/docs/models.md#srcmf-ud`

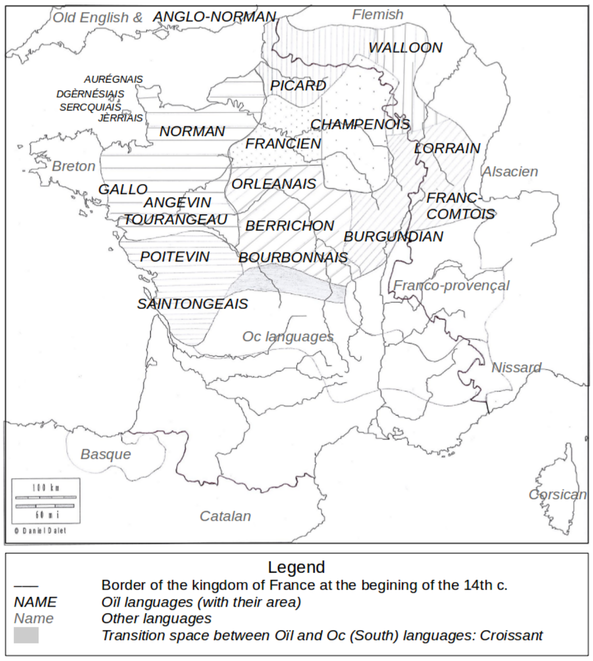


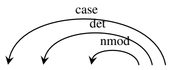
Figure 1: Oïl languages

Medieval French covers both Old French (9th-13th c.) and Middle French (14th-15th c.). These stages are linguistically close, and both precede the adoption of spelling norms. Middle French is more regular than Old French in some respects such as word order (Marchello-Nizia et al. 2020) and less in others such as NP structure and pronouns system (Marchello-Nizia 1979). Medieval French covers a set of *Oïl* Romance languages spoken in the kingdom of France between the 9th and the 15th century (fig. 1). There are around twenty such languages. Older texts are close to Late Latin, and verse is prevalent until the end of the 13th century. Old French has a relatively free word order. Until the mid-11th century, the prevalent order is *Subject-Object-Verb* (SOV), which is then gradually supplanted by SVO, which is the default order in contemporary French. Unlike most languages with free word order, the functions of verbal arguments are not always given away by morphological clues, the already simplistic case system of Old French disappears progressively through the covered period.
There are also many cases of syntactic ambiguity. For example, in the following quote from *Lancelot*,[5] (verse 5436), both “la dame” and “Lancelot” could be the subject or the object of “Vit” and only the context enables the reader to understand that “la dame” is the subject.

Dolant	*et*	*pansif*	*Lancelot*	*Vit*	*la*	*dame*
Mournful	and	meditative	Lancelot	saw	the	lady

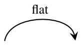
‘The lady saw that Lancelot was mournful and meditative.’

Word order is also relatively free within constituents. For example, a noun modifier can be on the left or on the right of its governor, and it is not necessarily preceded by a preposition. In contemporary French, it can only appear on the right, and it is found without a preposition only in some cases like named entities. Because of the general free word order and the absence of punctuation in our treebank, this adds up to the ambiguity of the analysis.
In each of the following examples from the SRCMF corpus, the noun following *roi* (“king”) has a different analysis: head of *roi*, modifier, argument of the same verb or a different one, with no explicit marking:



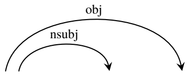
Fus	*tu*	*donc*	*pus*	*a*	*la*	***roi***	*cort*
Were	you	then	no more	at	the	king	court

“Then were you not at the king’s court anymore?” (*Beroul Tristan*)



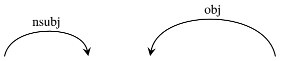
la	*fille*	*au*	*riche*	***roi***	*pescheor*
the	daughter	of the	rich	king	fisher

“the daughter of the rich Fisher King” (*Queste del Saint Graal*)



De	*Guenelun*	*atent*	*li*	***reis***	*nuveles*
From	Ganelon	waits	the	king	news

“The king waits for news from Ganelon.” (*Chanson de Roland*)



Biax	*sire*	*fet*	*li*	***rois***	*escu*	*vos*	*envoiera*	*Diex*
Dear	Sir	says	the	king	shield	you	send-FUT	God

“Dear Sir, says the king, God will send you a shield.” (*Queste del Saint Graal*)

Furthermore, overt subjects are not mandatory, and are often dropped in texts written in verse until the 12th century, after which the presence of subjects increases through time. These phenomena are particularly prevalent in verse, where metric and rhyming constraints often lead to more contrived syntactic forms than in prose.
Another source of ambiguity is the variety of spellings, due to the lack of spelling standard. For example, the word *moult* (transl. *a lot (of), very*), emblematic of this period, is initially an adjective, and it is progressively grammaticalized, becoming an adverb. Several forms appear at the same time, some with a declension, some without, and the radical does not have a fixed spelling: *molt(e)(s), molz, mult(e)(s), mul(t)z, mou(l)t…*
We chose to include a few texts from the early Middle French period (14th-15th c.) in this raw corpus, which brings a valuable complement of the prose documents

[5] In the edition from Pierre Kunstmann, from the online *Base de français médiéval*: `http://catalog.bfm-corpus.org/CharretteKu`.

Corpus	Size (MiB)	Size (Mwords)
BFM (Guillot-Barbance et al. 2017)	20.7	3.91
AND (Rothwell and Trotter 2005)	17.2	3.25
NCA (Stein et al. 2006)	9.7	2.05
Chartes Douai (Gleßgen 2003)	3.1	0.56
OpenMedFr (Wrisley 2018)	1.7	0.33
Geste (Camps et al. 2016)	1.5	0.32
MCVF (Martineau 2008)	1.4	0.26
Chartes Aube (van Reenen et al. 2006)	0.2	0.04
Total	55.3	10.53

Table 1: Data sources for the raw corpus used for model pretraining, with sizes in bytes and number of words. Due to the nature of the documents, mixing prose, verse, titles, annotations… estimating a number of sentence would be error-prone and can be abstracted over, given that the models trained here do not depend on strict sentence boundaries.

that are lacking for Old French, while staying close enough to late Old French, the boundary between the two epochs being somewhat fuzzy. These texts precede the adoption of norms established by editors after the invention of Gutenberg's printing press. Middle French is more regular than Old French in some respects such as word order (Marchello-Nizia et al. 2020) and less in others such as NP structure and pronouns system (Marchello-Nizia 1979), but they share most of their lexicon and for these relatively early texts, the syntax is not too different from that of late Old French texts.

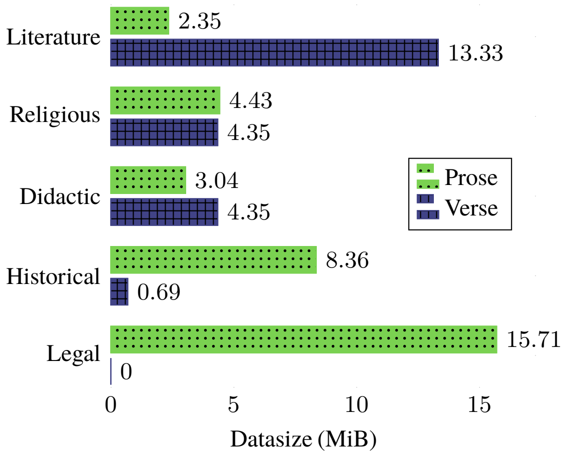


Figure 2: Distribution of form and domain, gathered from documents metadata and manual annotation.

Medieval French has many factors of variation: language evolution, dialects, domains, forms of text (verse or prose) and lack of standard. Our dataset gives us a representation of Medieval French that is as accurate and diversified as possible, given the limited amount of material that survived to these days. The detailed instructions to replicate this dataset are described in the Appendix. No particular processing is done on the original documents. In order to get a sound evaluation of the contextual embeddings trained with this dataset, we filter out the documents that are also present in the SRCMF treebank used for evaluation purposes in section 4[6]. The resulting corpus is quite heterogeneous: legal texts and verse literature are in the majority, whereas other domains, such as historical and didactic texts, are under-represented, as can be seen in fig. 2.

4. Experiments

We evaluate a set of alternative word representations on Old French, using their usefulness for POS-tagging and dependency parsing as a downstream evaluation. To that end, we train and evaluate a parser/tagger using the annotated treebank of Old French (SRCMF, Prévost and Stein (2013)) as provided by the 2.7 version of the UD dataset (Zeman et al. 2020) as a reference treebank.

Our parser/tagger probe uses Dozat and Manning (2018)'s neural graph parser made as reimplemented by Le et al. (2020) and Grobol and Crabbé (2021), using the same hyperparameters. Word representations are obtained by concatenating subword embeddings, averaged over transformer layers together with character embeddings and non contextualized word embeddings. This representation is similar to those used by Straka et al. (2019) and Ling et al. (2015). In all of our experiments, the contextual embeddings are fine-tuned while training the parser. Unlike the recent CoNLL challenges settings, we assume gold tokenization, since the syntactic annotations we target provide a reference word-based segmentation. Using a predicted one could only add noise to our experiments. Furthermore, for most European languages using a Latin script—including Old and Middle French—, word segmentation is acceptably approximated by simple typographic tokenization.

The remaining of this section presents our experimental results, sorted by nature of required data. We report UPOS POS-tagging scores as well as unlabeled and labeled attachment scores for dependency parsing (respectively UAS and LAS), as given by the CoNLL-2018

[6] As noted by Gururangan et al. (2020), pre-training on task specific data provides an additional boost, that would muddle our results, since our objective here is not so much task optimization as embeddings benchmarking.

scorer, computed on the development set of SRCMF to avoid overfitting the architecture and transfer learning procedure to the test set. Results on the test set are provided only for the dev-best models to allow us to compare our results to the state of the art.

Due to the number of costly experiments,[7] the results are reported on single runs. The results should therefore be interpreted only with respects to the broad trends: small score differences between competing settings should be taken with care.

4.1. Baselines

Embeddings	UPOS	UAS	LAS
Vanilla	93.51	87.60	81.54
Random-base	93.17	86.97	80.71
finBERT	94.44	88.44	82.47

Table 2: Results on SRCMF dev — no additional data.

We first compare a baseline where contextual embeddings are not used at all (Vanilla) with two settings using models with no preexisting knowledge of Old French: Random-base, a randomly initialized model using the same architecture and model size as RoBERTa-base (Liu et al. 2019) and finBERT (Virtanen et al. 2019), a contextual embedding model from Finnish, a Uralic language that is unrelated to Old French. These baselines are meant to check that the gain in performances observed when using models with some (possibly indirect) knowledge of Old French are linked to this knowledge and not simply due to an increase in the number of trainable parameters (for the random baseline) or to a weight distribution induced by training on a language modeling task that would be universally good for all languages (for the finBERT baseline, which can thus be seen as a different kind of weight initialization).

Table 2 shows the results obtained in these configurations, which show that using a model with random weights, even fine-tuned for these tasks, does not bring any improvement, and is in fact even worse than using no contextual embeddings at all. In contrast, using a model that has been pretrained for language modeling—even for an unrelated language—brings some modest improvements. This suggests that pretraining gives a structure to this kind of model that makes it suitable for fine-tuning on the downstream task, but the impact of this gain is clearly—and predictably—very limited compared to what can be expected for representations that have been trained on relevant linguistic data.

4.2. With related contextual embeddings

When a low-resource language is close to a well-resourced one, it is possible to leverage models designed for the latter. For Old French, contemporary French is an obvious candidate and two contextual embeddings models are available: FlauBERT (Le et al. 2020) and CamemBERT (Martin et al. 2020). Furthermore, mBERT (Devlin et al. 2019), a model trained on a multilingual corpus which does not include Old French (possibly apart from some fragments in its contemporary French training data), has been shown to be suitable for many languages, and in particular for Indo-European and Romance languages (Straka et al. 2019; Muller et al. 2020). We report in table 3 the results obtained when using these language models directly, without additional fine-tuning involving Old French data. As expected, these results show significant improvements over the baselines, confirming that using contextual embeddings for a related language works better than both randomly initialized embeddings and embeddings pretrained for an unrelated language—even after fine-tuning. More surprisingly, the best results here are obtained with mBERT. This could mean that mBERT benefits from having been pretrained for a wider range of languages, including in particular other Romance languages that share with Old French some features, lost in contemporary French: for instance null subjects.

Base model	UPOS	UAS	LAS
FlauBERT	95.70	90.43	85.45
CamemBERT	95.86	91.15	86.31
mBERT	96.06	91.52	86.83

Table 3: Results on SRCMF dev — monolingual models.

4.3. With raw linguistic data

We now try to take advantage of the raw Medieval French data described in section 3. To that end, we explore two strategies: training a model from scratch and refining existing models by "post-training" them—running a few more training epochs on the Medieval French raw data.

In the "from scratch" strategy, we first train a BBPE subword tokenizer (C. Wang et al. 2020) on our raw corpus, then train a RoBERTa (Liu et al. 2019) masked language model. Taking inspiration from Micheli et al. (2020), who worked in a setting close to ours: a small and noisy pretraining corpus used to create a model from scratch, we used a RoBERTa architecture. As reported in table 4, we tested several parametrizations of the architecture also inspired by Turc et al. (2019). Out of these alternatives, the "BERTrade-petit" configuration was the most successful and this is the one we keep for the following experiments.

For the "post-training" strategy, we continue the training of the pre-trained models used in sections 4.1 and 4.2, for 12 epochs on our raw corpus. We used the same RoBERTa masked language modeling task, using the same parameters as Z. Wang et al. (2020) (but without vocabulary modifications), resulting in the BERTrade-X models, where X is the name of the base model.

The results of these experiments are reported in Table 5. Comparing these to our results of section 4.2 shows that training a model from scratch, even on such limited amounts of data, yields a better model than a

[7] See the Appendix for elements on the carbon footprint of our experiments.

Name	Layers	Embeddings	Heads	UPOS	UAS	LAS
BERTrade-tiny	2	128	2	94.03	88.66	82.79
BERTrade-small	4	512	8	96.53	86.30	87.49
BERTrade-petit	12	256	4	97.14	91.90	89.18
BERTrade-medium	8	512	8	96.62	91.92	87.60
BERTrade-base	12	768	12	96.74	92.37	88.42

Table 4: Results on SRCMF dev — Performances of different model sizes when training from scratch

Base model	UPOS	UAS	LAS
BERTrade-petit	97.14	92.95	89.18
BERTrade-finBERT	96.28	92.12	87.92
BERTrade-mBERT	96.95	93.33	89.60
BERTrade-CamemBERT	97.16	93.75	90.06
BERTrade-FlauBERT	96.94	93.75	90.07

Table 5: Results on SRCMF dev — using raw data.

simple task-specific fine-tuning of mBERT. However, post-training mBERT yields even better results, and the best ones are obtained by post-training the models for contemporary French.

Model	UPOS	UAS	LAS
Straka et al. (2019)	96.26	91.83	86.75
mBERT	96.19	92.03	87.52
BERTrade-petit	96.60	92.20	87.95
BERTrade-mBERT	97.11	93.86	90.37
BERTrade-FlauBERT	97.15	93.96	90.57
BERTrade-CamemBERT	97.29	94.36	90.90

Table 6: Results on SRCMF test

4.4. Putting it all together

Finally, in table 6, we compare the performances of our models on the test set of SRCMF with those obtained by Straka et al. (2019), with similar methods. The difference between the models is that we fine-tune the word embeddings, while Straka et al. (2019) keep them frozen. Our mBERT baseline, which is the closest to their configuration, shows that even without any additional data, task-specific fine-tuning already brings significant improvements, while our models refined using our raw corpus of Medieval French bring further improvements, leading to state-of-the-art results that are consistent with their results on the development set.

5. Conclusion

In this work, we have shown that building a monolingual contextual word embeddings model for Medieval French is possible even with limited and heterogeneous linguistic data and that it can bring significant performance gains in parsing and POS-tagging. To that end, the best strategy seems to be post-training a contextual word embedding model for contemporary French on raw Medieval French documents. We have not directly addressed the internal heterogeneity issue in both our pretraining and fine-tuning data, relying instead on the versatility of the representation models we considered to bypass it, but it seems a promising perspective for future work—for instance by using finer-grained post-training, concentrating on specific linguistic sub-periods or genres.

For historical languages in general, this suggests that language-specific fine-tuning is more efficient when applied to a model pre-trained for their contemporary counterpart than when applied to a multilingual model. While this study is not currently easy to replicate for other languages due to the lack of annotated data for a suitable downstream task, it suggests that the considerable amount of work required to gather even a small amount of raw texts in the target language is a sound investment, given the significant improvements it can bring to contextual word representations. Beyond historical languages, these findings could also help for processing minority dialectal variants and contact languages of well-resourced languages, and we leave for future work the exploration of these generalizations.

Acknowledgments

This work was funded by the ANR projet ANR-16-CE38-0010 “Profiterole” (PRocessing Old French Instrumented TExts for the Representation Of Language Evolution) and used the HPC resources of IDRIS under the allocation 2020-AD011011637 made by GENCI.

6. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (Aug. 2018). Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Beltagy, I., Lo, K., and Cohan, A. (Nov. 2019). SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620. Hong Kong, China. Association for Computational Linguistics.

Bender, E. M. et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. FAccT '21. Virtual Event, Canada. Association for Computing Machinery.

Berdicevskis, A. and Eckhoff, H. (May 2020). A Diachronic Treebank of Russian Spanning More Than a Thousand Years. English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5251–5256. Marseille, France. European Language Resources Association.

Camps, J.-B. et al. (2016). Geste: un corpus de chansons de geste:

Chau, E. C., Lin, L. H., and Smith, N. A. (Nov. 2020). Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334. Online. Association for Computational Linguistics.

Clark, K. et al. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Desrochers, S., Paradis, C., and Weaver, V. M. (2016). A Validation of DRAM RAPL Power Measurements. In *Proceedings of the Second International Symposium on Memory Systems*, pages 455–470. MEMSYS '16. Alexandria, VA, USA. Association for Computing Machinery.

Devlin, J. et al. (June 2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Minneapolis, Minnesota. Association for Computational Linguistics.

Ding, S., Renduchintala, A., and Duh, K. (Aug. 2019). A Call for Prudent Choice of Subword Merge Operations in Neural Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213. Dublin, Ireland. European Association for Machine Translation.

Dozat, T. and Manning, C. D. (July 2018). Simpler but More Accurate Semantic Dependency Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490. Melbourne, Australia. Association for Computational Linguistics.

Gleßgen, M.-D. (2003). L'élaboration philologique et l'étude lexicologique des Plus anciens documents linguistiques de la France à l'aide de l'informatique. *Mémoires et documents de l'École des chartes*, 71: 371–386.

Grobol, L. and Crabbé, B. (June 2021). Analyse en dépendances du français avec des plongements contextualisés. In *28e Conférence sur le Traitement Automatique des Langues Naturelles*. Lille (virtuel), France.

Guibon, G. et al. (Dec. 2014). Parsing Poorly Standardized Language Dependency on Old French. In V. Henrich et al., editors, *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories*, pages 51–61 (Tübingen, Deutschland).

Guillot-Barbance, C., Heiden, S., and Lavrentiev, A. (July 2017). Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques*, 7: 168–184. eprint: `halshs-01809581`.

Gururangan, S. et al. (July 2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Online. Association for Computational Linguistics.

Han, X. and Eisenstein, J. (Nov. 2019). Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248. Hong Kong, China. Association for Computational Linguistics.

Kroch, A., Taylor, A., and Santorini, B. (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). *CD-ROM, second edition, release 4*: U. o. P. Department of Linguistics, editor.

Le, H. et al. (May 2020). FlauBERT: Unsupervised Language Model Pre-training for French. English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490. Marseille, France. European Language Resources Association.

Lee, J. et al. (Sept. 2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36.4: 1234–1240. eprint: `https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf`.

Lee, J. and Kong, Y. H. (2012). A dependency treebank of classical Chinese poems. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 191–199.

– (2016). A dependency treebank of Chinese Buddhist texts. *Digital Scholarship in the Humanities*, 31.1: 140–151.

Ling, W. et al. (Sept. 2015). Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530. Lisbon, Portugal. Association for Computational Linguistics.

Liu, Y. et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692: arXiv: `1907.11692`.

Marchello-Nizia, C. (1979). *Histoire de la langue française aux XIVe et XVe siècles*. Bordas.

Marchello-Nizia, C. et al. (2020). *Grande Grammaire historique du français*. de Gruyter.

Martin, L. et al. (July 2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. Online. Association for Computational Linguistics.

Martineau, F. (July 2008). Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus*, 7: eprint: `http://journals.openedition.org/corpus/1508`.

Micheli, V., d'Hoffschmidt, M., and Fleuret, F. (Nov. 2020). On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858. Online. Association for Computational Linguistics.

Miletic, A. et al. (May 2020). Building a Universal Dependencies Treebank for Occitan. English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2932–2939. Marseille, France. European Language Resources Association.

Muller, B. et al. (Oct. 2020). When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. *arXiv:2010.12858 [cs]*: arXiv: `2010.12858 [cs]`.

Nivre, J. et al. (May 2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043. Marseille, France. European Language Resources Association.

Ortiz Suárez, P. J., Romary, L., and Sagot, B. (July 2020). A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714. Online. Association for Computational Linguistics.

Passarotti, M. (2019). The Project of the Index Thomisticus Treebank. In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*. M. Berti, editor. De Gruyter Saur, pages 299–320.

Peng, Y., Yan, S., and Lu, Z. (Aug. 2019). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65. Florence, Italy. Association for Computational Linguistics.

Peters, M. et al. (June 2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. New Orleans, Louisiana. Association for Computational Linguistics.

Prévost, S. and Stein, A., editors (2013). *Syntactic Reference Corpus of Medieval French (SRCMF)*. UPDATE VERSION NUMBER. Lyon/Stuttgart. ENS de Lyon; Lattice, Paris; ILR University of Stuttgart.

Raffel, C. et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21.140: 1–67.

Ramponi, A. and Plank, B. (Dec. 2020). Neural Unsupervised Domain Adaptation in NLP—A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855. Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rocio, V. et al. (2003). Automated creation of a medieval portuguese partial treebank. In *Treebanks: Building and Using Parsed Corpora*. Anne Abeillé, Kluwer Academic Publishers.

Rögnvaldsson, E. et al. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *LREC*, pages 1977–1984. Citeseer.

Rothwell, W. and Trotter, D. (2005). Anglo-Normand Dictionary 2:

Schwartz, R. et al. (Nov. 2020). Green AI. *Commun. ACM*, 63.12: 54–63.

Stein, A. (May 2014). Parsing Heterogeneous Corpora with a Rich Dependency Grammar. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2879–2886 (Reykjavík, Ísland). European Language Resources Association.

– (2016). Old French Dependency Parsing: Results of Two Parsers Analysed from a Linguistic Point of View. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 707–713.

Stein, A., Kunstmann, P., and Gleßgen, M.-D. (2006). Nouveau Corpus d'Amsterdam:

Straka, M. (Oct. 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. Association for Computational Linguistics.

Straka, M., Straková, J., and Hajič, J. (Aug. 2019). Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing. *arXiv:1908.07448 [cs]*: arXiv: `1908.07448 [cs]`.

Strubell, E., Ganesh, A., and McCallum, A. (July 2019). Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. Florence, Italy. Association for Computational Linguistics.

Traugott, E. C. and Pintzuk, S. (2008). Coding the York-Toronto-Helsinki Parsed Corpus of Old English Prose to investigate the syntax-pragmatics interface. *Studies in the History of the English Language IV. Empirical and Analytical Advances in the Study of English Language Change. Berlin/New York: Mouton de Gruyter*: 61–80.

Turc, I. et al. (Sept. 2019). Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. *arXiv:1908.08962 [cs]*: arXiv: `1908.08962`

[cs].

University of Oxford (2001). *The York-Helsinki parsed corpus of Old English poetry (YCOEP)*. Oxford Text Archive.

Van Reenen, P., Wattel, E., and van Mulken, M. (2006). Champagne 1270-1300, Chartes en langue française conservées aux Archives de l'Aube:

Virtanen, A. et al. (2019). Multilingual is not enough: BERT for Finnish. *CoRR*, abs/1912.07076: arXiv: 1912.07076.

Wang, C., Cho, K., and Gu, J. (Apr. 2020). Neural Machine Translation with Byte-Level Subwords. en. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34.05: 9154–9160.

Wang, Z. et al. (Nov. 2020). Extending Multilingual BERT to Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656. Online. Association for Computational Linguistics.

Wolf, T. et al. (Oct. 2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Online. Association for Computational Linguistics.

Wrisley, D. (2018). The Open Medieval French Initiative (OpenMedFr):

Xiong, Y. et al. (2021). Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention. *CoRR*, abs/2102.03902: arXiv: 2102.03902.

Zeman, D. et al. (Nov. 2020). *Universal Dependencies 2.7*.

Zeman, D. et al. (Oct. 2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21 (Brussels, Belgium). Association for Computational Linguistics.

A. Collecting the Data

The following data can be downloaded directly from their website:

- Chartes de l'Aube:
  `https://sites.google.com/site/achimstein/research/resources`
  Extract raw text from XML files: <body>, then <s>, then <word>.

- Geste:
  `https://github.com/Jean-Baptiste-Camps/Geste`
  Raw text is available under /txt/norm/.

- OpenMedFr:
  `https://github.com/OpenMedFr/texts`
  Remove the header of each file (until *** *START*), its last line (*** *END*), paragraph breaks (#|) and folios or pages numbers.

Special permissions are required to access and use these sources:

- AND:
  `https://anglo-norman.net/project-members`

- BFM:
  `http://bfm.ens-lyon.fr/spip.php?article19`
  Raw text is available.

- Chartes Douai:
  `https://www.rose.uzh.ch/docling`

- MCVF: `http://www.voies.uottawa.ca`

- NCA:
  `https://sites.google.com/site/achimstein/research/resources`
  Extract raw text from the XML files: <body> then <txm:form>.

B. Details on the Models

B.1. Models Trained From Scratch

These are trained for 32 epochs in a masked language modeling task using the same parameters as RoBERTa (Liu et al. 2019) but a smaller batch size of 256 samples[8], which amounts to a magnitude of 1×10^5 steps. We also use a smaller vocabulary size (8192) than other works, in line with the observations of Ding et al. (2019) that learning large vocabularies on small corpora defeats the purpose of sub-word tokenization. Using a larger vocabulary size of 5×10^4 (like FlauBERT) also did not seem to bring any improvements in our preliminary experiments and made pre-training more expensive.

B.2. Post-training

The pretrained models we used in the post-training settings are those available in the 4.2.0 version of Huggingface Transformers (Wolf et al. 2020) and the exact handles are:

mBERT bert-base-multilingual-cased

flauBERT flaubert/flaubert_base_cased

camemBERT camembert-base

finBERT TurkuNLP/bert-base-finnish-cased-v1

The post-trained models are those with MLM heads, which we did not reset before post-training, so the post-training phase can be seen as a language transfer task for masked language modeling out of which we extract a contextual word embeddings model.

C. Carbon Footprint

In light of recent concerns about the power consumption and carbon footprint of deep learning models (Schwartz

[8] Preliminary experiments with larger batch sizes showed no significant improvement to compensate for the heavier computational load.

Model	Power (W)	# Models	Duration (h)	Consumption (kWh)	CO_2e (kg)
Pre-train	10 756	11	6	11 216.36	358.92
Post-train	1520	4	20	192.13	6.15
Total emissions					365.07

Table 7: Average power draw, number of models trained, training times in hours, mean power consumption including power usage effectiveness (PUE), and CO_2 emissions; for each setting.

et al. 2020; Bender et al. 2021) we report the power consumption and carbon footprint of our main experiments following the approach of Strubell et al. (2019). Two different configurations were used in our experiments, one for pre-training models from scratch (Pre-train) and another one for continuing the training of existing models (Post-train).

Pre-train: We use a cluster of 4 machines each one having 8 GPU Nvidia Tesla V100 SXM2 32 GiB, 384 GiB of RAM, and two Intel Xeon Gold 6226 processors. One Nvidia Tesla V100 card is rated at around 300 W,[9] while the Xeon Gold 6226 processor is rated at 125 W,[10]. For the DRAM we can use the work of Desrochers et al. (2016) to estimate the total power draw of 384 GiB of RAM at around 39 W. The total power draw of this setting adds up to around 10 756 W. We train 11 different models in this configuration.

Post-train: We use a single machine having 4 GPU Nvidia Tesla V100 SXM2 32 GiB, 192 GiB of RAM and two Intel Xeon Gold 6248 processors. The Xeon Gold 6248 processor is rated at 150 W,[11], and the DRAM total power draw can be estimated at around 20 W. The total power draw of this setting adds up to around 1520 W. We train 4 different models in this configuration.

Having this information, we can now use the formula proposed by Strubell et al. (2019) in order to compute the total power required for each setting:

$$p_t = \frac{1.58t(cp_c + p_r + gp_g)}{1000}$$

Where c and g are the number of CPUs and GPUs respectively, p_c is the average power draw (in W) from all CPU sockets, p_r the average power draw from all DRAM sockets, and p_g the average power draw of a single GPU. We estimate the total power consumption by adding GPU, CPU and DRAM consumption, and then multiplying by the *Power Usage Effectiveness* (PUE), which accounts for the additional energy required to support the compute infrastructure. We use a PUE coefficient of 1.58, the 2018 global average for data centers (Strubell et al. 2019). In table 7 we report the training times in hours, as well as the total power draw (in Watts) of the system used to train the models. We use this information to compute the total power consumption of each setting, also reported in table 7.

We can further estimate the CO_2 emissions in kilograms of each single model by multiplying the total power consumption by the average CO_2 emissions per kWh in our region which were around 32 g kW^{-1} h in January 2021,[12] when the models were trained. Thus the total CO_2 emissions in kg for one single model can be computed as:

$$\mathrm{CO_2e} = 0.032 p_t$$

All emissions are also reported in table 7.

[9] Nvidia Tesla V100 specification

[10] Intel Xeon Gold 6226 specification

[11] Intel Xeon Gold 6248 specification

[12] Rte - éCO_2mix.