From FreEM to D'AlemBERT

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, Benoît Sagot

From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French

Simon Gabay[3], Pedro Ortiz Suarez[1,2], Alexandre Bartz[2], Alix Chagué[1]
Rachel Bawden[1], Philippe Gambette[4], Benoît Sagot[1]
Inria[1], Sorbonne Université[2], Université de Genève[3], LIGM, Université Gustage Eiffel, CNRS[4]
2 rue Simone Iff, 75012 Paris (France)[1], 21 rue de l'École de médecine, 75006 Paris (France)[2],
rue du Général-Dufour 24, 1211 Genève (Switzerland)[3],
5 boulevard Descartes, F-77454 Champs-sur-Marne (France)[4]
{pedro.ortiz, alix.chague, benoit.sagot, rachel.bawden}@inria.fr[1], alexandre.bartz@sorbonne-universite.fr[2],
simon.gabay@unige.ch[3], philippe.gambette@univ-eiffel.fr[4]

Abstract

Language models for historical states of language are becoming increasingly important to allow the optimal digitisation and analysis of old textual sources. Because these historical states are at the same time more complex to process and more scarce in the corpora available, specific efforts are necessary to train natural language processing (NLP) tools adapted to the data. In this paper, we present our efforts to develop NLP tools for Early Modern French (historical French from the 16th to the 18th centuries). We present the FreEM$_{max}$ corpus of Early Modern French and D'AlemBERT, a RoBERTa-based language model trained on FreEM$_{max}$. We evaluate the usefulness of D'AlemBERT by fine-tuning it on a part-of-speech tagging task, outperforming previous work on the test set. Importantly, we find evidence for the transfer learning capacity of the language model, since its performance on lesser-resourced time periods appears to have been boosted by the more resourced ones. We release D'AlemBERT and the open-sourced subpart of the FreEM$_{max}$ corpus.

Keywords: Digital humanities, Early Modern French, Language modelling, Neural language representation models, Less-resourced languages, Corpus creation, POS tagging

1. Introduction

With the rise of digital humanities, it is becoming increasingly important to develop high quality tools to automatically process old states of languages (*e.g.* Old or Early Modern French). Libraries or archives, among others, are digitising large numbers of historical sources, from which high quality data must be extracted for further study by specialists of human sciences following new approaches such as "distant reading" (Moretti, 2013). Many (sub)tasks such as automatic OCR post-correction (Rijhwani et al., 2021) and linguistic annotation (Camps et al., 2020) benefit from pretrained language models to improve their accuracy, and this is what motivated us to develop a BERT-like (Devlin et al., 2019) contextualised language model for Early Modern French (*i.e.* 16th-18th c. French).

Languages evolve over time on many different levels: from one century to another, we see variations in spelling, syntax, the lexicon etc. However this variation is not uniform: it tends, at least for "literate scriptors" (literature, journalism, law, etc.), to converge towards a single norm over time, and this has especially been the case for French because of the prominent role of the *Académie française* and the *remarqueurs* (Ayres-Bennett and Seijido, 2011). The result of this convergence is, for instance, that spelling and word order within sentences have become more strict, where they were less so in the past. From a computational perspective, historical states of language are therefore not only different from the contemporary state, but also more complex since they do not follow a strict and explicit norm. In French, this explicit norm appeared in the 17th c. and was slowly integrated throughout the 18th c. On top of this first linguistic problem, a second issue appears: because the production of textual sources has continued to grow exponentially, it is easier to collect a corpus for contemporary French than for 19th c. French, which is itself easier than for 18th c. French, etc. The further we go back in time, the more scarce resources are, which creates the following paradox: we have more data when the language is homogeneous and simple for the computer to process, and less when it is heterogeneous and harder to process.

The following paper will address the development of D'AlemBERT, a neural language model in a complex setting, defined here as the state of language with scarce heterogeneous resources. We will also present FreEM$_{max}$, the data used to train the model, discuss its conception, and evaluate its efficiency with a classical natural language processing (NLP) task, part-of-speech (POS) tagging, crucial for corpus linguistics and the digital humanities. We release both the D'AlemBERT model and a subset of the FreEM$_{max}$ dataset that we were allowed by the original authors to open-source .

2. Related Work

Large datasets for historical states of languages or extinct languages do exist. The *Corpus Middelnederlands* for Medieval Dutch (Reenen, Pieter van and Mulder, Maaike, 1998) and the *Base Geste* for Medieval French (Camps et al., 2019) are freely available online, encoded according to the guidelines of the Text Encoding Initiative (TEI). It is also the case for other corpora for later states of language, such as the *Reference corpus of historical Slovene*, covering approximately three centuries (1584–1899) of Slovene (Erjavec, 2015), and the "corpus noyau" of *Presto* (Blumenthal and Vigier, 2018). This last corpus, in its extended version, uses other French corpora such as *Espistemon* for Renaissance French (Demonet, 1998) and the University of Chicago's *American and French Research on the Treasury of the French Language* (ARTFL) (Morrissey and Olsen, 1981); or like FRANTEXT (ATILF, 1998 b), which is a generalist French corpus, covering the different states of the French language between the 11th and the 21st century. Although most of these text collections are free, the two biggest ones, FRANTEXT and ARTFL, are not freely available or open-sourced.

Concerning language modelling in French, two main models are available for contemporary French, CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020). CamemBERT was trained on a freely available, automatically web-crawled corpus called OSCAR (Ortiz Suárez et al., 2019; Ortiz Suárez et al., 2020), while FlauBERT was trained on a mix of web-crawled data and manually curated (some of which is not freely available) contemporary French corpora. Neither of these models was explicitly pre-trained for historical French.[1] However efficient language models have been trained for less-resourced or extinct Languages such as Latin (Bamman and Burns, 2020), following the approach of Martin et al. (2020) for the training of language models with less data than was previously thought. There have also been some recent projects that specifically target Early Modern French such as that of *Pie Extended* (Clérice, 2020), which uses the hierarchical encoding architecture originally proposed by Manjavacas et al. (2019), which itself is constructed by stacking multiple Bi-LSTM-CRFs. Clérice (2020) distributes pre-trained models for POS tagging and lemmatisation.

3. Corpora

Over the past few years, we have been involved in the development of linguistic resources for Early Modern French. The initiative, called FREEM (for *FREnch Early Modern*) aims to collect the corpora required for various NLP tasks such as lemmatisation, POS tagging, linguistic normalisation and named entity recognition.[2] Two of these corpora are introduced here: FREEM$_{max}$ (Section 3.2) and FREEM$_{LPM}$ (Section 3.3).

[1]Note however that texts in Old, Middle and Modern French do exist on the internet, and might have found their way into the training corpus of these two models. This is especially the case for Modern French texts, which automatic language classification tools can easily classify as contemporary French.

[2]`https://freem-corpora.github.io`.

3.1. Early Modern French

Experiments are based on data of which the core comprises Early Modern French literary texts. We loosely define Early Modern French as a state of language following Middle French in 1500—following here the *terminus ad quem* used by the *Dictionnaire de Moyen Français* (Martin, 2020)—and ending with the French Revolution in 1789. It therefore encompasses three centuries (16th, 17th and 18th c.), or two linguistic periods: the *français préclassique* or "preclassical French", 1500–1630 and the *français classique* or "classical French", 1630–1689; both periodisations are currently used in French linguistics (*e.g.* by Vachon 2010 and Amatuzzi et al. 2019).

A typical example of Early Modern French, taken from Guez de Balzac (1624), is given in Table 1. We note here the presence of several phenomena that have now disappeared in contemporary French, such as the presence of abbreviations (*dõt*→*dont*), the long *s* (ſ, see *miſeres*), the use of *v* instead of *u* (*vne* for *une*), the conservation of etymological letters (*voſtre*<Latin *vŏster* rather than *votre*) and calligraphic letters (*-y* in *Surquoy*), the absence of welding (*mal-heurs* and not *malheurs*) and the opposite (*Surquoy* and not *Sur quoi*).

For NLP tasks that process raw sequences, such differences with respect to contemporary French are not trivial, and they prevent the processing of historical texts with tools trained on recent sources.

3.2. FREEM$_{max}$

Usable historical documents are difficult to find because, as previously mentioned, they are more rare than contemporary ones; editors also tend to normalise the language (*i.e.* use the spelling conventions of Contemporary French, see Gabay 2014) and transcriptions are not (always) distributed in a digital format. FREEM$_{max}$ (Gabay et al., 2022) is an attempt to solve this problem, and the aim of this dataset is to group together the largest number or texts possible written in Early Modern French. The texts have a variety of sources, which can be grouped into three main types:

- Two institutional datasets have been used and are not open-sourced:
  - FRANTEXT *intégral* (ATILF, 1998 b), the biggest database of French texts (only the texts between 1500 and 1800), a very small portion of which is open access: FRANTEXT *Démonstration* (ATILF, 1998 a);
  - *Electronic Enlightenment* (Bodleian Libraries, 2008), an online collection of edited correspondences of the Early Modern period;
- Several come from research projects distributing transcriptions online:

Source	Normalised	Translation
Surquoy, SIRE, s'il plaiſt à voſtre Maieſté de ſe ſouuenir des miſeres de ſon Eſtat, dõt au moins ell'a tiré cét aduantage, qu'en vne grande ieuneſse ell'a acquis vne grande experiẽce, elle verra que tous les mal-heurs de sõ bas âge ont pris leur commencement en ſemblables occaſions;	*Sur quoi, SIRE, s'il plaît à votre Majesté de se souvenir des misères de son état dont au moins elle a tiré cet avantage, qu'en une grande jeunesse elle a acquis une grande expérience, elle verra que tous les malheurs de son bas âge ont pris leur commencement en semblables occasions ;*	"Whereupon, SIR, if it pleases your Majesty to remember the miseries of her state, from which at least she has derived this advantage, that in great youth she has acquired great experience, she will see that all the misfortunes of her early life took their beginning on similar occasions;"

Table 1: Example of normalisation taken from the *Lettres* of Guez de Balzac (1624).

- The *Antonomaz project*, French *mazarinades* (`https://cahier.hypotheses.org/antonomaz`);
- The II.B section (in French) of the *Actis Pacis Westphalicae*, diplomatic letters for the Peace of Westphalia (`http://kaskade.dwds.de/dstar/apwcf/`);
- The *Bibliothèques virtuelles humanistes*, 16th c. French literature (`http://www.bvh.univ-tours.fr`);
- The *Corpus électronique de la première modernité*, 17th c. French literature (`http://www.cepm.paris-sorbonne.fr`)
- The *Condé* project, *coutumiers normands* (`https://conde.hypotheses.org`)
- The Corpus Descartes, works of René Descartes (`https://www.unicaen.fr/puc/sources/prodescartes/`);
- The *Bibliothèque dramatique* of the CELLF, 17th c. French plays (`http://bibdramatique.huma-num.fr`);
- The *Fabula numerica* project, French fables (`https://obvil.sorbonne-universite.fr/projets/fabula-numerica`);
- The *Fonds Boissy*, plays of Louis de Boissy (`https://www.licorn-research.fr/Boissy.html`);
- The *Mercure Galant* project, the famous French *gazette* and literary magazine between 1672 and 1710 (`https://obvil.sorbonne-universite.fr/corpus/mercure-galant`);
- The *Rousseau online* project, works of Jean-Jacques Rousseau (`https://www.rousseauonline.ch`);
- The *Sermo* project, sermons of the 16th and 17th c. (`http://sermo.unine.ch`);
- The *Théâtre classique* project, 17th and 18th c. French plays (`http://www.theatre-classique.fr`);

• Additional sources come from researchers who kindly accepted to offer their personal transcriptions or data scraped by our team:
  - Transcriptions of Anne-Élisabeth Spica (17th c. French novels);
  - Transcriptions found on *Wikisource* (`https://fr.wikisource.org`);
  - Transcriptions (ePub files) found on *Gallica* (`https://gallica.bnf.fr`);
  - Transcriptions found on various websites online.

Additional data for later states of the language, up to the 1920's (mainly from FRANTEXT *intégral*), are also provided for two main reasons: on the one hand, it is common to normalise Early Modern French into Contemporary French (Gabay, 2014) because of the linguistic proximity between these the two states of the language, and on the other hand, it helps to collect (precious) additional data to avoid ending up with with too small of a corpus for our needs.
The final result is far from being balanced or representative (see Figure 1). 16th c. French documents are under-represented, as well as 18th c. literature. The 17th c. is clearly over-represented, especially its second half—probably one of the most important of French literature, which could explain this situation (on top of our personal interest for this specific period).
As some texts are still (partially) protected by restrictive licences, the FREEM$_{max}$ corpus exists in both open and non-open versions, only the open one being distributed. In order to limit the impact of licences forbidding the modification of files, we have designed a pipeline to distribute the data as it was found and recreate it (see Figure 2).
Metadata is prepared manually to ensure the same categories for each document, whatever the origin. As well as the author, the title and the date (where relevant), we also provide the genre ("theatre"), sometimes a subgenre ("tragedy"), the linguistic status (normalised or not) and the licence attached to the transcription.

3.3. FREEM$_{LPM}$

The FREEM$_{LPM}$ ("Lemma, POS tags, Morphology") has already been presented (Gabay et al., 2020b).

Origin	#Tokens
Spica corpus	691,467
Antonomaz project	119,194
Acta Pacis Westphlicae II B	2,463,047
Bibliothèque Bleue	776,838
BVH	2,434,657
CEPM	2,707,432
Condé project	3,173,845
Descartes	1,025,337
CELLF	1,873,772
Electronic enlightenment	6,568,047
Fabula project	145,978
FRANTEXT *intégral* (>1500, <1800)	60,018,390
FRANTEXT *intégral* (>1800)	71,504,440
FRANTEXT *Démonstration*	1,255,454
Gallica	5,212,333
Boissy project	438,215
Mercure galant	5,427,469
Rousseau Online project	2,428,587
Scrapping	1,936,835
Sermo project	529,647
Théâtre classique project	13,916,169
Wikisource	996,329
TOTAL	185,643,482

Table 2: Breakdown of the FREEM$_{max}$ corpus by text origin.

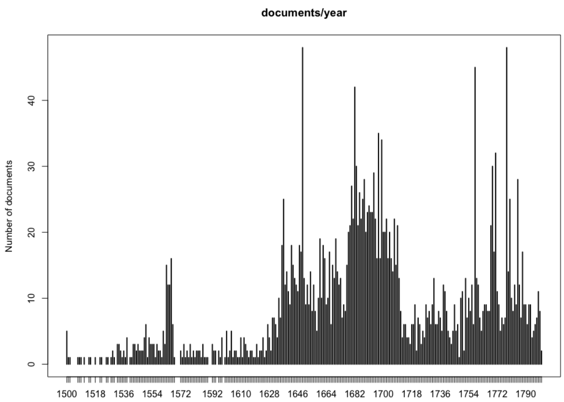

Figure 1: Distribution of the documents in the FREEM$_{max}$ corpus per year

The POS-annotated data is a mixture of two different sources. On the one hand, there is the *CornMol* corpus (Camps et al., 2020), made up of normalised 17th c. French comedies. On the other hand, there is a gold subset of the *Presto* corpus (Blumenthal et al., 2017), made up of texts of different genres written during the 16th, 17th and 18th c., which have previously been used to train annotation tools (Diwersy et al., 2017), and was heavily corrected by us to match our annotation principles (Gabay et al., 2020a).
On top of traditional in-domain tests, an out-of-domain testing dataset was prepared to control the capacity of the model to generalise to other genres and periods. Centuries covered are the 16th, 17th, 18th, 19th and 20th. There are two test sets for each century: one made up only of theatre, the other of everything but theatre. Each test set comprises 10 short samples (c. 100 tokens), as representative as possible of the linguistic production of the century (female and male authors, decade of publication, genre, etc.).
All the data from FREEM$_{LPM}$ (but almost none of the out-of-domain) can be found in FREEM$_{max}$.

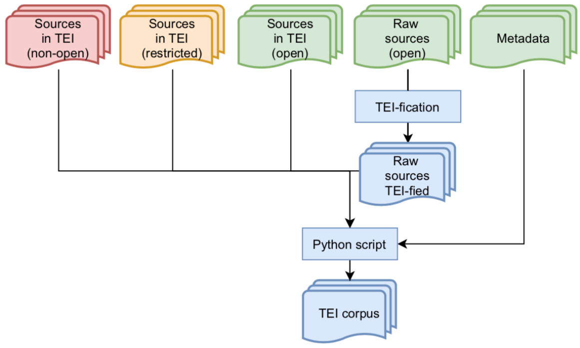

Figure 2: FREEM$_{max}$ compilation pipeline. All files are kept in their original format. Metadata is manually prepared in separate files in order to automatically transform and clean (in blue) all the available documents into XML TEI files following the same encoding. It allows us to distribute open data (in green) but also data distributed with restrictions regarding the modification of the original format (in orange). Non-open texts (in red) are not distributed.

4. D'AlemBERT: a neural language model for Early Modern French

In this section, we describe the pretraining data, architecture, training objective and optimisation setup we use for D'AlemBERT, our new neural language model for Early Modern French.

4.1. Pre-processing

Similar to RoBERTa (Liu et al., 2019) we segment the input text data into subword units using Byte-Pair encoding (BPE) (Sennrich et al., 2016) in the implementation proposed by (Radford et al., 2019) that uses bytes instead of unicode characters as the base subword units. The BPE encoding does not require pre-tokenisation (at the word or token level), thus removing the need to develop a specific tokeniser for Early Modern French. We use a vocabulary size of 32,768 subword tokens. These subwords are learned on the entire FREEM$_{max}$ dataset.

4.2. Language Modelling

Transformer D'AlemBERT uses the exact same architecture as RoBERTa, which is a multi-layer bidirectional Transformer (Vaswani et al., 2017). D'AlemBERT uses the original *base* architecture of RoBERTa (12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters).

ORIGINAL							NORMALISED OR CONTEMPORARY						
Model	16	17	18	19	20	Avg	Model	16	17	18	19	20	Avg
Drama							*Drama*						
Pie Extended	*90.34*	*94.47*	*94.64*	-	-	*93.15*	Pie Extended	*93.69*	*95.75*	*95.61*	*95.03*	*93.71*	*94.76*
CamemBERT	87.06	89.01	90.92	-	-	89.00	CamemBERT	90.18	91.51	91.37	91.13	91.42	91.12
D'AlemBERT	**94.17**	**96.59**	**96.28**	-	-	**95.68**	D'AlemBERT	**96.25**	**96.97**	**96.80**	**96.25**	**95.00**	**96.25**
Varia							*Varia*						
Pie Extended	*89.85*	*93.44*	*95.98*	-	-	*93.09*	Pie Extended	*92.52*	*94.81*	*95.98*	*92.24*	*94.03*	*93.94*
CamemBERT	86.90	88.85	92.85	-	-	89.53	CamemBERT	89.79	90.69	93.06	90.54	89.78	93.94
D'AlemBERT	**93.86**	**95.73**	**96.95**	-	-	**95.51**	D'AlemBERT	**94.52**	**96.64**	**96.88**	**94.90**	**95.30**	**95.65**
Both							*Both*						
Pie Extended	*90.08*	*93.95*	*95.33*	-	-	*93.12*	Pie Extended	*93.08*	*95.28*	*95.80*	*93.65*	*93.87*	*94.35*
CamemBERT	86.98	88.93	91.89	-	-	89.27	CamemBERT	89.99	91.10	92.22	90.84	90.60	92.53
D'AlemBERT	**94.02**	**96.16**	**96.62**	-	-	**95.60**	D'AlemBERT	**95.39**	**96.81**	**96.84**	**95.58**	**95.15**	**95.95**

Table 3: Comparison between D'AlemBERT, CamemBERT and Pie Extended performance on FREEM$_{LPM}$.

Pretraining Objective We train our model on the Masked Language Modelling (MLM) task as proposed by RoBERTa's authors (Liu et al., 2019): given an input text sequence composed of N tokens $x_1, ..., x_N$, we select 15% of tokens for possible replacement. Among those selected tokens, 80% are replaced with the special `<MASK>` token, 10% are left unchanged and 10% are replaced by a random token. The model is then trained to predict the masked tokens using cross-entropy loss.

Again, following the RoBERTa approach, we dynamically mask tokens instead of fixing them statically for the whole dataset during preprocessing. We also choose not to use the next sentence prediction (NSP) task originally used in BERT (Devlin et al., 2019), as it has been shown that it does not improve downstream task performance (Conneau and Lample, 2019; Liu et al., 2019).

Optimisation We optimise our model in the exact same way as (Liu et al., 2019) using Adam (Kingma and Ba, 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.98$) for 31k steps with large batch sizes of 8,192 sequences, each sequence containing at most 512 tokens.

Pre-training We use the RoBERTa implementation in the Zelda Rose library,[3] and again, in the same way as Liu et al. (2019) our learning rate is warmed up for 10k steps up to a peak value of 0.0003 instead of the original 0.0001 used by the original implementation of RoBERTa (Liu et al., 2019), as our model diverged with the 0.0001 value. We hypothesise that this is either due to the smaller size of FREEM$_{max}$ (compared to the corpora used for RoBERTa or CamemBERT) or to our large batch size. We train our model for 31k steps, which amounts to 41 epochs. The total pre-training times, the details of the infrastructure we used and even the carbon emissions of our model are reported in Appendix A.

5. Evaluation and Discussion

In order to evaluate our D'AlemBERT model, we fine-tune it for POS tagging on the FREEM$_{LPM}$ corpus. We use the `flair` framework[4] for sequence tagging (Akbik et al., 2019). To fine-tune D'AlemBERT for POS tagging, we follow the same approach as Schweter and Akbik (2020) with some modifications: we append a linear layer of size 256 that takes as input the last hidden representation of the `<s>` special token and the mean of the last hidden representation of the subword units of each token (token as defined for FREEM$_{LPM}$), that is, we use a *"mean"* subword pooling strategy. We fine-tune D'AlemBERT with a learning rate of 0.000005 for a total of 10 epochs. We also fine-tune CamemBERT using the exact same hyperparameters as the ones we use for D'AlemBERT.

FREEM$_{LPM}$ provides a standard split (train, dev, test). However it also proposes an evaluation on a *out-of-domain* subcorpus that is not contained in the standard split and that is separated by century (from the 16$^{\text{th}}$ to the 20$^{\text{th}}$ century) and that also contains both the *Normalised* and *Original* versions of the texts for the 16$^{\text{th}}$, 17$^{\text{th}}$ and 18$^{\text{th}}$ centuries. The idea of this out-of-domain evaluation corpus is to have a fine-grained evaluation of the models to better assess their performance in all the different types of text that one might encounter when working with Early Modern French data.

Following the approach of Clérice (2020), we report the scores obtained on the out-of-domain testing dataset of FREEM$_{LPM}$ in Table 3. We use the scores previously reported by Clérice (2020) using *Pie Extended* as our baseline as well as the fine-tuned CamemBERT that serves as a second baseline as well as a rough estimation of how much knowledge can D'AlemBERT transfer from the FREEM$_{max}$ into this task.

We can see that D'AlemBERT consistently outperforms Pie Extended and CamemBERT in both the normalised and original versions of our out-of-domain testing data and for all different periods by a considerable margin. We can also see that on average the differ-

[3] `https://github.com/LoicGrobol/zeldarose`

[4] `https://github.com/flairNLP/flair`

ence in score between D'AlemBERT and Pie Extended is greater for the original split than the normalised one. This suggests that D'AlemBERT can generalise more effectively to non-normalised data than the more traditional architecture used by Pie Extended. Moreover we can also see that the difference in scores is also greater for the 16^{th} c. and 17^{th} c. data. This is interesting, especially for the 16^{th} c, because, as we can see in Figure 1, this is the least represented period in the FREEM$_{max}$ corpus. This result actually suggests that D'AlemBERT might be able to do effective transfer learning from the 18^{th} c., 19^{th} c. and 20^{th} c. data to the 16^{th} c. and 17^{th} c. data.

As for CamemBERT, we can see that it consistently scores lower than both D'AlemBERT and Pie Extended. Moreover, we can see that it struggles particularly with the non-normalised data of the 16^{th} c., 17^{th} c. and 18^{th} c.. This results clearly shows that CamemBERT cannot easily generalise to these earlier states of languages, or at least not with the quantity of data found in the training set of FREEM$_{LPM}$. These results also show the impressive capacity of D'AlemBERT to quickly generalise to diverse set of states of language, as well as its capacity to transfer knowledge from the FREEM$_{max}$ corpus into this task. The obtained results are also a testament to the importance of the pretraining data, specially taking in account that the pretraining set of CamemBERT is more than 100 times bigger than that of D'AlemBERT.

6. Conclusion

In this paper we presented the manually curated FREEM$_{max}$ corpus of Early Modern French as well as D'AlemBERT, a RoBERTa-based language model trained on FREEM$_{max}$. With D'AlemBERT, we showed that it is possible to successfully train a transformer-based language model for historical French with even less data than originally shown in previous works (Martin et al., 2020). Moreover with our POS tagging evaluation we were able to observe some form of transfer learning and generalisation across multiple states of the language corresponding to different periods of time. Both our corpus and our model will be of use to digital humanists and linguists interested in Early Modern French. For our future work, we hope that will be able to study the application of our D'AlemBERT model to other NLP tasks such as text normalisation, named entity recognition and even document structuring, where we hope to more extensively study the transfer learning capabilities of our approach.

7. Acknowledgements

We would like to warmly thank Karine Abiven, Bertrand Gaiffe, Annette Gerstenberg, Pierre Larrivée, Gaël Lejeune, Laurent Romary, Anne-Élisabeth Spica and Martin Wynne for their help in gathering the data. This work was also performed using HPC resources from GENCI-IDRIS (Grant 2021-AD011011330R1). This work was also partly funded by Rachel Bawden's and Benoît Sagot's chairs in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001, as well as the BASNUM ANR project (ANR-18-CE38-0003).

8. Bibliographical References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Amatuzzi, A., Skupien Dekens, C., Ayres-Bennett, W., Gerstenberg, A., and Schoesler, L. (2019). Améliorer et appliquer les outils numériques. ressources et approches pour l'étude du changement linguistique en français préclassique et classique. In *Le Français en diachronie*, Travaux de Linguistique Romane, pages 337–364. Editions de linguistique et de philologie.

Wendy Ayres-Bennett et al., editors. (2011). *Remarques et observations sur la langue française. Histoire et évolution d'un genre*. Number 1 in Histoire et évolution du français. Classiques Garnier.

Bamman, D. and Burns, P. J. (2020). Latin BERT: A contextual language model for classical philology. *CoRR*, abs/2009.10053.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (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*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Blumenthal, P., Diwersy, S., Falaise, A., Lay, M.-H., Souvay, G., and Vigier, D. (2017). Presto, un corpus diachronique pour le français des XVI^{e}-XX^{e} siècles. In *Actes de la 24ème conférence sur le Traitement Automatique des Langues Naturelles - TALN'17*. Association pour le traitement automatique des langues.

Camps, J.-B., Gabay, S., Fièvre, P., Clérice, T., and Cafiero, F. (2020). Corpus and models for lemmatisation and POS-tagging of Classical French theatre. *Journal of Data Mining & Digital Humanities*.

Clérice, T. (2020). *Pie Extended, an Extension for Pie with Pre-Processing and Post-Processing*. Zenodo. https://doi.org/10.5281/zenodo.3883589.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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, MEMSYS '16, page 455–470, New York, NY, USA. Association for Computing Machinery.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Diwersy, S., Falaise, A., Lay, M.-H., and Souvay, G. (2017). Ressources et méthodes pour l'analyse diachronique. *Langages*, 206(2):21–44.

Gabay, S., Camps, J.-B., and Clérice, T. (2020a). Manuel d'annotation linguistique pour le français moderne (XVI[e] -XVIII[e] siècles).

Gabay, S., Clérice, T., Camps, J.-B., Tanguy, J.-B., and Gille-Levenson, M. (2020b). Standardizing linguistic data: Method and Ttols for annotating (pre-orthographic) french. In *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC '20)*, Hammamet, Tunisia.

Gabay, S. (2014). Pourquoi moderniser l'orthographe? principes d'ecdotique et littérature du XVII[e] siècle. *Vox Romanica*, 73(1):27–42.

Guez de Balzac, J.-L. (1624). *Lettres du sieur de Balzac*. T. Du Bray.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français (FlauBERT : Unsupervised language model pre-training for French). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 268–278, Nancy, France, 6. ATALA et AFCP.

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. *arXiv e-prints*, July. arXiv:1907.11692.

Manjavacas, E., Kádár, Á., and Kestemont, M. (2019). Improving lemmatization of non-standard languages with joint learning. 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 1493–1503, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July. Association for Computational Linguistics.

Martin, R. d. (2020). *Dictionnaire du Moyen Français*. ATILF - CNRS & Université de Lorraine.

Moretti, F. (2013). *Distant reading*. Verso.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Piotr Bański, et al., editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Ortiz Suárez, P. J., Romary, L., and Sagot, B. (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, July. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.

Rijhwani, S., Rosenblum, D., Anastasopoulos, A., and Neubig, G. (2021). Lexically aware semi-supervised learning for OCR post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green ai. *Commun. ACM*, 63(12):54–63.

Schweter, S. and Akbik, A. (2020). FLERT: Document-Level Features for Named Entity Recognition. *arXiv e-prints*, page arXiv:2011.06993, November.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Strubell, E., Ganesh, A., and McCallum, A. (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, July. Association for Computational Linguistics.

Vachon, C. H. (2010). *Le Changement linguistique au XVI[e] siècle: une étude basée sur des textes* *littéraires français*. ELiPhi, Éditions de linguistique et de philologie.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

9. Language Resource References

ATILF. (1998–a). *Frantext Démonstration*. ATILF - CNRS & Université de Lorraine.

ATILF. (1998–b). *Frantext intégral*. ATILF - CNRS & Université de Lorraine.

Blumenthal, Peter and Vigier, Denis (dir.). (2018). *Presto: corpus noyau*.

Bodleian Libraries. (2008–). *Electronic Enlightenment*. Oxford University Press.

Camps, Jean-Baptiste and LAKME-ENC and Cochet, Alice and Ing, Lucence and Paulinelvq. (2019). *Jean-Baptiste-Camps/Geste: Geste: un corpus de chansons de geste, 2016-....* Zenodo.

Demonet, Marie-Luce (dir.). (1998–). *Epistemon*. Centre d'Etudes Supérieures de la Renaissance.

Erjavec, Tomaž. (2015). *Reference corpus of historical Slovene goo300k 1.2*.

Gabay, Simon and Bartz, Alexandre and Gambette, Philippe and Chagué, Alix. (2022). *FreEM max OA: A Large Corpus for Early modern French - Open access version*. Université de Genève, 1.0.

Morrissey, Robert and Olsen, Mark. (1981–). *American and French Research on the Treasury of the French Language (ARTFL)*. University of Chicago.

Reenen, Pieter van and Mulder, Maaike. (1998). *Corpus Middelnederlands*. Instituut voor de Nederlandse Taal, 1.0.

A. Carbon Footprint

Model	Power (W)	Time (h)	(PUE·kWh)	CO^2e (kg)
Pre-train	48640	20	1537.02	46.11
Evaluation	589	1	0.93	0.03
Total CO^2e				46.14

Table 4: 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.

In light of recent interest concerning the energy consumption and carbon emission of machine learning models and specifically of those of language models (Schwartz et al., 2020; Bender et al., 2021), we have decided to report the power consumption and carbon footprint of all our experiments following the approach of Strubell et al. (2019). We report the energy consumption and carbon emissions of both the pre-training of D'AlemBERT and its evaluation.

Pre-training: We use a cluster of 32 machines, each one having 4 GPU Nvidia Tesla V100 SXM2 32GiB, 192GiB of RAM, and two Intel Xeon Gold 6248 processors. One Nvidia Tesla V100 card is rated at around 300W,[5] while the Xeon Gold 6248 processor is rated at 150W.[6] For the DRAM we can use the work of Desrochers et al. (2016) to estimate the total power draw of 192GiB of RAM at around 20W. Thus, the total power draw of the pre-training adds up to around 48640W.

Evaluation: We use a single machine with a single GPU Nvidia Tesla V100 SXM2 32GiB, 384GiB of RAM and two Intel Xeon Gold 6226 processors. The Xeon Gold 6226 processor is rated at 125 W,[7] and the DRAM total power draw can be estimated at around 39W. Therefore, the total power draw of the evaluation adds up to around 589W.

With this information, we use the formula proposed by Strubell et al. (2019) 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 centres (Strubell et al., 2019). In Table 4 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 4.

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 30g/kWh between the 30th and the 31st of December,[8] when the models were trained. Thus the total CO^2 emissions in kg for one single model can be computed as:

$$\text{CO}_2\text{e} = 0.030p_t$$

All emissions are also reported in Table 4.

[5] Nvidia Tesla V100 specification

[6] Intel Xeon Gold 6248 specification

[7] Intel Xeon Gold 6226 specification

[8] Rte - éCO^2mix.