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

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

Inria
PR[AI]RIE
PaRis Artificial Intelligence Research InstituteE
Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL
cnrs
UNIVERSITÉ DE GENÈVE

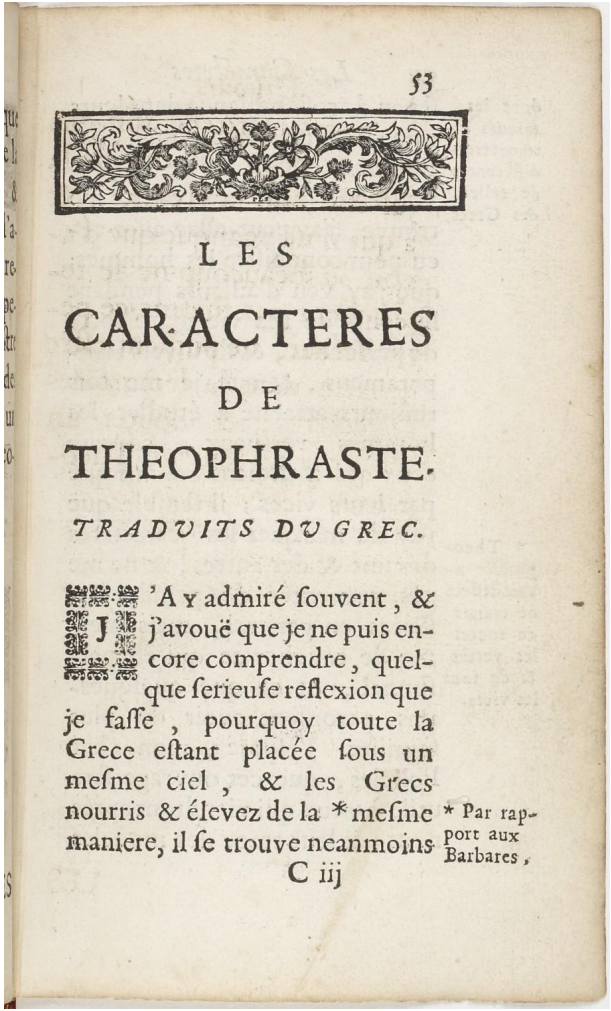
Early Modern French

Specificities:

- Unstable spelling (*était*, *estoit*, *étoit*, etc.)
- Specific letters (*ß*, *ſ*, etc.)
- Different syntax (no punctuation, longer sentences, etc.)

Paradox:

- Impossible to use the same models for Early Modern French and contemporary French
- Need from digital humanists to process Early Modern documents



53

LES

CARACTERES

DE

THEOPHRASTE.

TRADUITS DU GREC.

J'Ay admiré ſouvent, & j'avouë que je ne puis encore comprendre, quelque ſerieuſe reflexion que je faſſe, pourquoy toute la Grece eſtant placée ſous un meſme ciel, & les Grecs nourris & élevez de la *meſme maniere, il ſe trouve neanmoins

* Par rapport aux Barbares,

C iij

FreEM collection

The FreEM (for ***FRE****nch* ***E****arly* ***M****odern*) project aims at conceiving all the required tools to process optimally Early Modern French documents:

- Tasks: lemmatisation (*estoit*->être), POS tagging (*estoit*-> VERcjg), linguistic normalisation (*estoit*->*était*)), named entity recognition and disambiguation
- Creating annotated corpora for several tasks
- Training models with these corpora
- More information at https://freem-corpora.github.io

A dedicated language model should improve the accuracy of NLP tasks

➜ We have engaged in the creation of a big corpus of Early Modern French texts and a BERT language model

FreEM$_{max}$

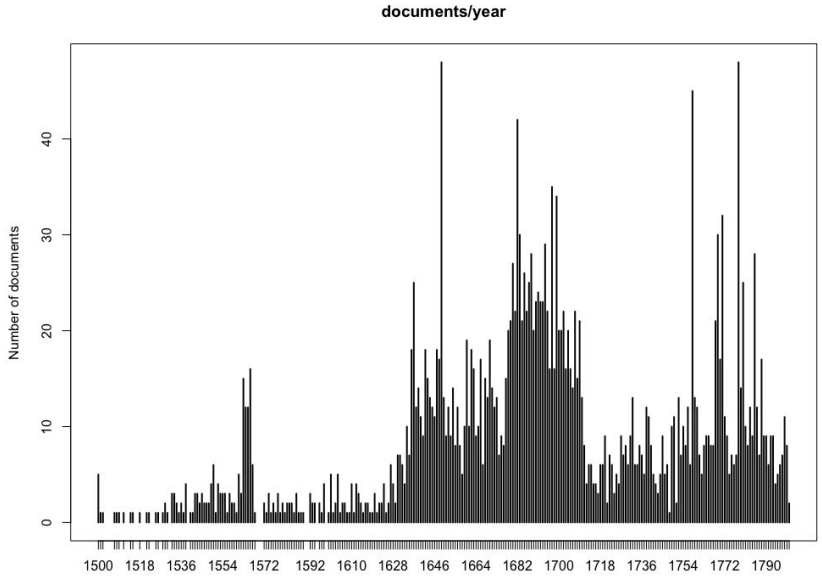
Gathering a maximum number of texts for

- French literary texts (and some non-literary)
- Written between the 16th and the 18th c.
- If possible in a machine actionable format

Sources:

- Research projects
- Existing corpora
- Personal transcriptions of researchers

Total: 185,643,482 tokens



Distribution of the documents in the FreEM$_{max}$ corpus per year. Data is more scarce for 16th French (on the left) and 18th c. French (on the right). the over-representation of 17th c. documents could be explained by its importance for French Literature.

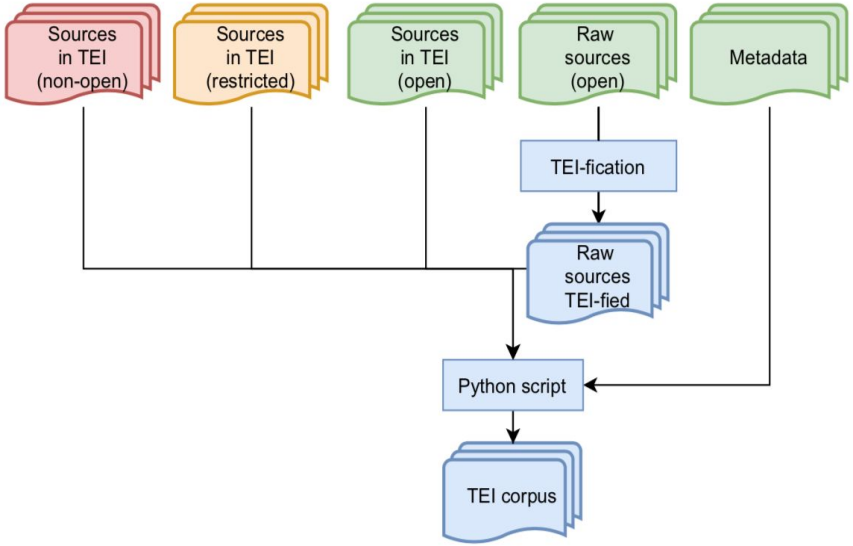
Distributing FreEM$_{max}$

Two problems:

- Modifying the data (*e.g.* CC-BY-ND)
- Distributing the data (texts under copyright, personal documents…)

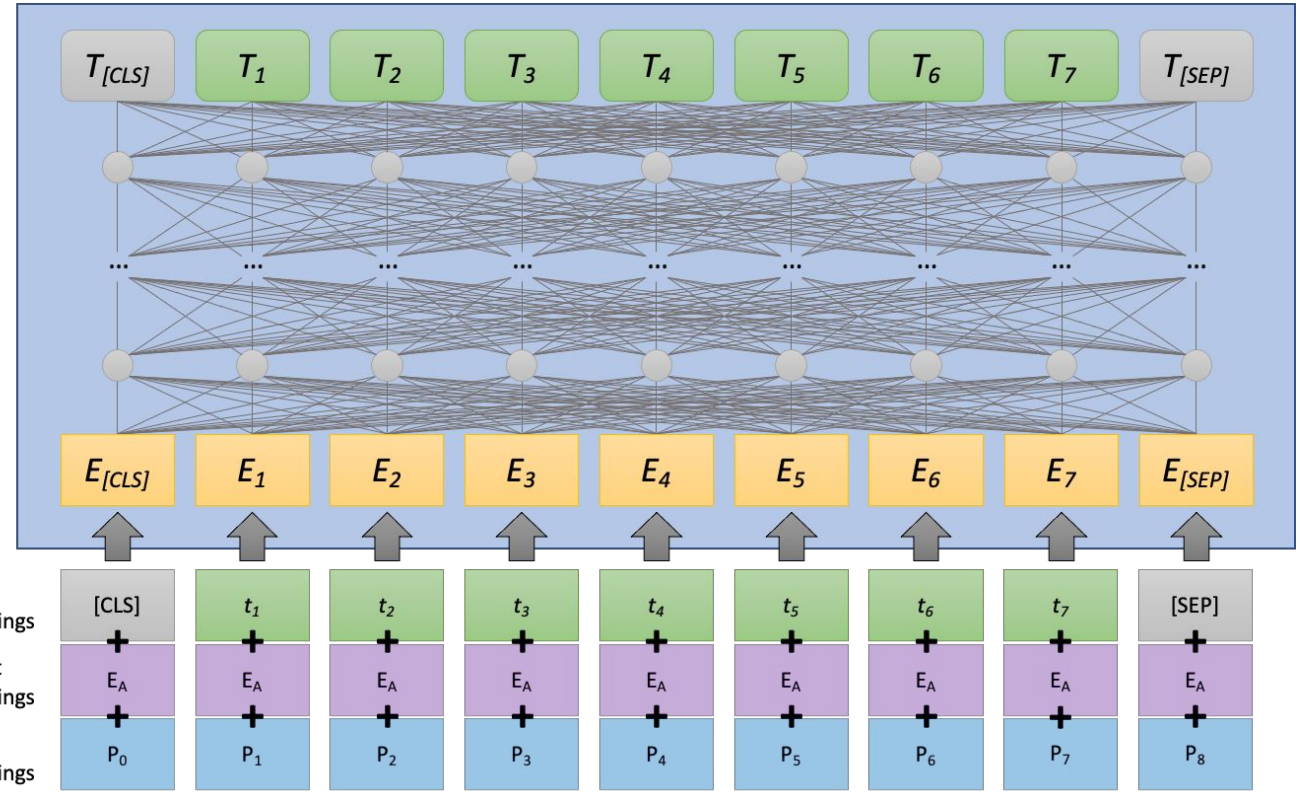
Two consequences:

- Creation of a compilation pipeline
- Two versions:
  - **Full version (not distributed)**
  - **Open Access version (distributed without all the documents)**

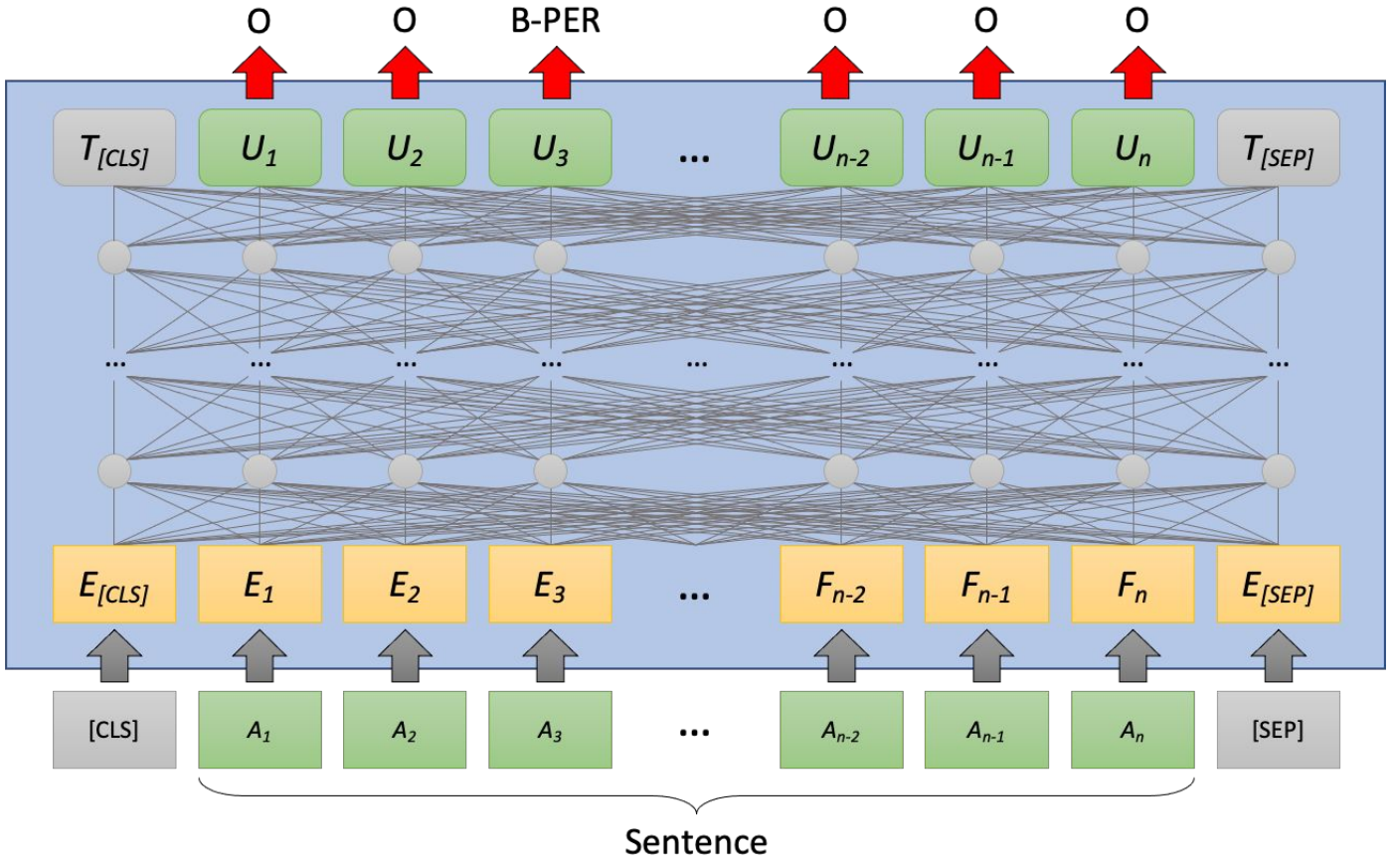


FreEM$_{max}$ compilation pipeline. All files are kept in their original format. Metadata is manually prepared in separate files in order to automatically transform and clean (in blue) all the available documents into XML-TEI files following the same encoding. It allows us to distribute open data (in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

BERT



BERT



D'AlemBERT

- Transformer
  - RoBERTa-BASE architecture: 12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters.
- 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.
- We optimise our model using Adam for 31k steps (**41 epochs**) with **large batch sizes** of 8,192 sequences, each sequence containing at most 512 tokens.
- We use a learning rate of 0.0003 instead of the original 0.0001 as our model diverged otherwise.

FreEM$_{LPM}$

FreEM$_{LPM}$ is a corpus annotated for ***L***emmata, ***P***OS tags and ***M***orphology. Regarding POS tags, annotated data comes from two different sources:

- 17th c. French theatre (normalised French, c. 90,000 tokens)
- The *Presto* gold corpus (original French, c. 60,000 tokens)

Models have already been trained and are available with *Pie extended* (https://pypi.org/project/pie-extended)

Out-of-domain data has been annotated to test the models:

- Sets for 16th, 17th, 18th, 19th and 20th c. French
- Excerpts of c. 100 tokens of 10 texts/per century
- Same distribution per century (one per decade, minimum number of female writers, etc.)

D'AlemBERT in PoS

- We use the flair framework for sequence tagging.
- We append a linear layer of size 256.
- We use a "mean" subword pooling strategy, i.e., we take the mean of the last hidden representation of the subword units of each token.
- 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.
- We evaluate the models on the out-of-domain set of FreEM$_{LPM}$.

D'AlemBERT

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**

D'AlemBERT's 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

Thank you for your attention!

https://freem-corpora.github.io - https://doi.org/10.5281/zenodo.6481135

This work was 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). It also used HPC resources from GENCI-IDRIS (Grant 2021-AD011011330R1).