Standardizing linguistic data: method and tools for annotating (pre-orthographic) French

Simon Gabay
forename.name@unine.ch
Universités de Neuchâtel et de Genève
Neuchâtel and Genève, Switzerland

Jean-Baptiste Tanguy
forename.name@sorbonne-universite.fr
Sorbonne Université
Paris, France

Thibault Clérice
forename.name@chartes.psl.eu
École des Chartes
Paris, France

Matthias Gille-Levenson
forename.name@ens-lyon.fr.
École normale supérieure de Lyon
Lyon, France

ABSTRACT
With the development of big corpora of various periods, it becomes crucial to standardize linguistic annotation (e.g. lemmas, POS tags, morphological annotation) to increase the interoperability of the data produced, despite diachronic variations. In the present paper, we describe both methodologically (by proposing annotation principles) and technically (by creating the required training data and the relevant models) the production of a linguistic tagger for (early) modern French (16-18th c.), taking as much as possible into account already existing standards for contemporary and, especially, medieval French.

CCS CONCEPTS
• Applied computing → Arts and humanities; • Computing methodologies → Natural language processing.

KEYWORDS
linguistic annotation, pre-orthographic language, lemmatisation, POS-tagging

ACM Reference Format:
Simon Gabay, Thibault Clérice, Jean-Baptiste Camps, Jean-Baptiste Tanguy, and Matthias Gille-Levenson. 2020. Standardizing linguistic data: method and tools for annotating (pre-orthographic) French. In Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC ’20), October 15–17, 2020, Online, Tunisia. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3423603.3423996

1 INTRODUCTION
If medievalists have been working for years on the creation of high quality corpora as well as the tools to build and analyse them (e.g. the Base de français médiéval (BFM) [Guillot et al. 2017] and TXM [Heiden et al. 2010]), the situation is different for more recent periods of French literature. As far as we know, for texts written between the 16th and the 18th c., Presto is the only project that has developed a corpus [Blumenthal et al. 2017] with some computational resources [Driwersy et al. 2017]. However, both are not maintained anymore, which leaves more than three centuries of French literature computationally understudied. It becomes therefore important to overcome this difficulty, and to create the tools that we desperately need for (early) modern French (henceforth EMF).

Such a task cannot be done without taking into account longstanding annotation practices, in order to allow (minimal) interoperability with already existing datasets. Such a statement is sadly easier said than done, because EMF is an intermediary stage between medieval (12th-15th c.) and late modern and contemporary (from c. 1750) French, two states of language that tend to have different needs regarding annotation: EMF is then caught in between two (potentially incompatible) practices, one for each extreme of the continuum. Facing such a dilemma, we have decided first to look up, in diachrony, and align (when possible) our choices with those of medievalists, rather than to look immediately down at what synchronists do – which tends to be the norm. Our project is indeed to create a tagger for EMF producing data as much as possible compatible with existing medieval corpora, and therefore build the technical frame for the computational study of pre-orthographic French, understood as the forms of this tongue prior to their standardisation based on the recommendations of the Académie française.

Textual production before the implementation of this "academic" French (c. mid-18th c., rather than the publication of its Dictionnaire in 1694) share common features [Catach 2001]. Among many others, the most important one is to undergo a significant graphematic variation, i.e. a relative instability regarding the spelling (estoit vs étoit) and the segmentation (à fin que vs afin que). Thus, to bypass this problem, the linguistic annotation has to be thought as a pivot format [Auroux 2000, p. 393] that allows users to navigate across heterogeneous data.

Several options are available, such as the conflation by phonetic identity or by lemma [Jurish 2012] – proto-forms being currently investigated as a potential solution, without being yet fully operational [Fourrier and Sagot 2020]. Among all possibilities, we have decided to opt for lemmatisation+POS tagging as a joint solution for two main reasons: on the one hand many resources are already available (cf. infra), and on the other hand it is already used for other tasks such as collation [Camps et al. 2019b] or stylometric analysis [Camps and Cafiero 2019].


2 ANNOTATION PRINCIPLES

Since all choices have been carefully discussed in our annotation manual [Gabay et al. 2020] we will here present the main problems raised by annotating EMF: tokenisation, lemmatisation, POS and morphological tagging.

2.1 Tokenisation

As any language, French has undergone phenomena of agglutination at the graphematic (ce pendant → cependant) and at the lexical (e.g. bien que) level. Such a process is not easily datable, because it varies from one region or one phenomenon to the other: in long diachrony, it is therefore impossible to decide precisely when such occurrences have to be tagged as two, or one token. For this reason, we have decided to use the blank space as the main separator for the tokenisation, with hardly any exception to this rule: all locutions, may they be conjunctive (tandis que), adverbial (à demi) or prepositional (quant à), are segmented in two parts rather than one (tandis and que, à and demi, quant and à).

Following this maximalist approach, none of the compounds are analysed as one token: peut être, long temps... are all considered as multiple tokens (peut + être, long + temps) – and this despite the presence of the hyphen, which is considered as a token itself (peut-être → peut + + être, long-temps → long + + temps...). For the same reason (absence of blank space), words that are already welded (monsieur < mon sieur) are treated as one token.

Such a choice to respect the actual segmentation of original witnesses introduces a bias, because it follows the scripta (i.e. the writing practice) of printers in (early) modern France [Pellat 1995] and not necessarily the preference of authors. This decision is however more in line with our philological approach, because it relies less on today’s linguistic understanding of EMF and outsources complex decisions to historical speakers.

2.2 Lemmatisation

This tokenising strategy creates some lemmatisation problems. The most important one is the absence of coherent lemmas for agglutinated forms. The nouns that have undergone phenomena of agglutination have no autonomy outside a locution such as parce que, < par ce) or afin (in afin que, < à fin), but also with portmanteau words (tresoïseuant < tres+obeissant). A simple solution is the creation of compounded lemmas, with an underscore separating the original lemmas (tresoïseuant), but such a method cannot be generalised: on the one hand, in diachrony, (too) many tokens could need one (mon sieur? autre fois?), and on the other hand we take the risk to end up with (too) many compounded lemmas for a single token (audit→à_le_dit). In order to limit our logic, tokens that have subsisted into contemporary French are therefore analysed as a simple lemma (parce, afin), but those that have not subsisted are analysed as a compounded lemma (tresoïseuant): this limits the creation of lemma, and maintains interoperability with other language states.

For other words, we have decided to use, when possible, the contemporary form of each token as a canonical form. It usually is the infinitive form (for verbs: mangeons → manger) or the masculine singular (for nouns or adjectives: comtesses → comte). Such a process is diachronically problematic, since some words have lost (dominus/domina > dame but also old fr. dom), or almost lost (damosieau vs domoiseille), their masculine counterpart over time: in such cases, we only choose the feminine version if the masculine one is not attested anymore at the end of the 18th c.

In order to control the consistency of the annotation, several authority lists have been created. The main one derives from the lexicon of medieval French inflected forms LGeRM [Souvay and Pierrel 2009] in its modern version (i.e. moderne étendu) produced for the Presto project [Diverson et al. 2017]. The LGeRM lexicon is based on the Freeling version of the Leff [Sagot 2010], augmented with Morphalou [Romary et al. 2004] for contemporary lemmas, and the Dictionnaire du Moyen Français (DMF) [ATILF-CNRS and Université de Lorraine 2015] for (early) modern ones. Such an approach should help us maintain minimal interoperability with the medieval version of LGeRM, but also the Trésor de la Langue Française informatisé (TLF) [Pierrel et al. 2004]. Two other authority lists have been created: one for named entities (people, places... ) and an other one with foreign words.

2.3 POS-tagging

Regarding POS-tagging, we have decided to use CATTEX [Prévost et al. 2013] for three reasons: first it is already used for the BFM and by several other corpora [Camps et al. 2019a], second it already benefits from a detailed annotation manual [Guillot et al. 2013a], third, it has been designed to cope with the complexity of old states of the French language.

Between the options exposed by Guillot et al. [2013b], we have opted for morpho-syntactical annotation: we take into account the context when categorising tokens and adjectives can be tagged as adverbs (il parle fort) or adverbs as substantives (il fait le bien). It allows us to follow the evolution of uses over time for a single word. However, following CATTEX as well as Presto recommendations, we never categorise present and past participles as adjectives, because if such distinctions are already difficult in synchrony, they are even harder to identify in diachrony. It would be too complicated to precisely differentiate the past participle of perdre (perdu) from the adjective perdu over the very centuries during which the verb has undergone a slow process of adjectivisation, especially without clear markers such as the determiner for the substantivization (il fait bien vs il fait le bien).

Our (maximalist) tokenisation principles incite (but do not force) us to tag words with their original part of speech: bien que is an ADVgen and not a member of a potential LOCconj (that does not exist within CATTEX). This choice considerably simplifies the interoperability with Old and Middle French. It does create some problems though, and we are forced to use analogies to tag certain tokens: pourve in pourve que is therefore ADVgen, such as bien in bien que.

2.4 Morphology

We have decided to implement CATTEX-max [Prévost et al. 2013], that is to say to annotate each token with morphological information (gender, mood, tense... ). Contrary to lemmas and POS, this linguistic information has been added for technical purposes (it has helped maximise the efficiency of the final model): if it has been proofread to avoid major mistakes, its perfection is not guaranteed.
3 TRAINING A MODEL

Based on the previously introduced annotation choices, we have decided to train a lemmatiser and a POS-tagger.

3.1 Data recycling

A substantial amount of heterogeneous data is already available (cf. tab. 1). Considering the cost of creating a gold corpus, we have decided that it would be more efficient to re-process existing data rather than create new ones from scratch.

| Name               | Gold | Norm. | Tokens | POS   | Morph |
|--------------------|------|-------|--------|-------|-------|
| CornMol            | Yes  | Yes   | 90,000 | CATTEX| Yes   |
| Frantext OA        | No   | Yes   | 2,400,000| EAGLES| No    |
| Presto gold        | Yes  | Yes   | 60,000 | MULTTEXT| No   |
| Presto core        | No   | Yes   | 6,820,000| MULTTEXT| No   |
| Presto controlled  | No   | Yes   | 11,636,000| MULTTEXT| No   |
| Presto extended    | No   | Yes   | 28,309,000| MULTTEXT| No   |

Table 1: Available training data

Two corpora come from a previous study [Camps et al. 2020] and are almost ready to use:

- The CornMol corpus has been created to carry stylometric studies [Camps and Cafiero 2019]. It is based on 41 comedies written in the 17th c., carefully sampled and proofread, which have been thoroughly described [Camps et al. 2020].
- The Frantext open access data [ATILF-CNRS and Université de Lorraine 2018] is composed of 32 texts, mainly written in the 18th (7 texts), 19th (24 texts) and 20th c. (4 texts). Because the corpus is already tagged and lemmatised (but not fully corrected) following other guidelines than ours, the lemmas have been aligned according to our standards, as much as we could.

Regarding Presto, because it is composed of four different subcorpora (for a detailed description of the different levels and variations, cf Vigier [2018]), it requires some selection, processing and correction:

- A gold corpus is made of 60,000 tokens, taken from 5 texts written in the 16th (1 text), 17th (2 texts) and 18th c. (2 texts). These texts have all been sampled, tokenised, tagged, lemmatised and proofread to create training data for a TreeTagger model.
- The final Presto corpus is a three-fold one: noyau (“core”), contrôlé (“controlled”) and étendu (“extended”). We have limited ourselves to a selection of its core version (cf. Appendix).

3.2 Preparation

Because the CornMol and Frantext Open Access corpora have already been prepared for our previous experiment on normalised-spelling French, it already follows closely our choices regarding tokenisation and POS-tagging (minor changes have been made in between), and the shift of authority list from Morphalou to LGerRM is of almost no impact since the latter derives from the former. All the effort have therefore been put on the correction and the alignment on our standards of the two Presto sub-corpora, which was problematic for three reasons. First, our choices differ strongly from Presto’s regarding tokenisation and POS-tagging – but not lemmatisation, since they also use LGerRM. Second, because their training data has been mainly prepared for POS-tagging and not lemmatisation, and also because the annotation has been done with an older and less precise tool (TreeTagger), the result is far from being perfect. Third, it is impossible to recycle some texts of their final corpus into training data, because we need (fairly) clean texts – Jean de Léry’s Voyage en terre du Brésil had therefore to be withdrawn because of the noise produced by hyphenation. Most of the correction work has been done with Pyrrha [Clérice et al. 2019].

In order to deal with more complex data than the one we use for training, we have artificially added a very limited number of glyphs typical of (early-)modern prints such as the long s (ſ), the essett (ß) or tilded letters traditionally used as abbreviations (∫, ∫, ∫, i).

Out of all these data, two different corpora have been produced:

- A primary corpus, fully annotated with lemma, POS and morphology, based on tokens with normalised and non-normalised spelling, has been created out of the two gold corpora (CornMol and the revised version of Presto gold).
- A secondary corpus, with lemmas only, has been produced out of Frantext open access data, the corrected versions of Presto gold and Presto core, and CornMol.

These two corpora are used to train two separate models: one for the POS and the morphology, and another for lemmatisation only. Therefore, two training datasets have been created using Protogénie [Clérice 2020b] (cf. tab. 2): they contain a train set, a development set (for evaluation during training) and a test set (for in-domain testing). The breakdown of tokens between the train, dev and test set is dependent on the total amount of data and has been done on an empirical basis, in order to keep as much tokens for training as possible, while still keeping a reliable test set. As such, it is dependent on the total size of the corpus (the bigger the corpus, the smaller the fraction of it which is necessary for reliable training).

| Corpus           | Train set | Dev set | Test set | Total   |
|------------------|-----------|---------|----------|---------|
| POS              | 132,905   | 9,733   | 15,303   | 157,941 |
| 84%              | 6%        | 10%     |          |         |
| Lemma            | 6,666,473 | 70,009  | 352,482  | 7,088,964|
| 94%              | 1%        | 5%      |          |         |

Table 2: Breakdown of the tokens per set (numbers are rounded)
3.3 Set up

We use Pie [Manjavacas et al. 2019] as our main training and tagging software. For training, Pie is used for lemma- and POS-tagging. At inference time, we use Pie-extended [Clérisce 2020] which wraps the former with more user-friendly functionalities (model sharing, advanced post- and preprocessing options . . . ). Several architectures have been tested for our previous study on normalised texts (for more detailed scores, cf. Camps et al. [2020])², the best one being:

- RNN with 300 dimensions for characters
- Sentence and word embeddings (150 dimensions) using 343 drama texts from [Fièvre 2007] and those of the FranText Open Access that we presented supra, for a total of c. 7M tokens.
- Hidden size of 150
- Forward and backward language models

However, additional tests have proven since the possibility of increasing even more the accuracy with a bigger number of dimensions for the RNN (300 → 400) and the hidden size (150 → 256), and NFKD unicode normalisation of the data, without word embeddings.

| LEMA          | M2   | M3   | M3+NFKD | support |
|---------------|------|------|---------|---------|
| In-domain     | 99.09| 99.12| 99.28   | 4,181   |
| Out-of-domain | 97.92| 98.59| 98.67   | 13,497  |

| POS           | M1   | M1+aux | M2 | M2+aux | support |
|---------------|------|--------|----|--------|---------|
| All           | 96.72| 96.51  | 96.84 | 97.01   | 4,181   |
| Ambiguous token | 91.86| 91.43  | 92.40 | 92.29   | 934     |
| Unknown tokens | 86.24| 86.24  | 78.44 | 81.65   | 218     |

Table 3: Scores (normalised spelling data only). M1 is the basic Pie configuration, M2 adds word embeddings to M1 architecture, and +aux indicates the use of auxiliary tasks – these architectures being fully described in Camps et al. [2020] as base (sent-lm) and wembs. M3 uses a bigger number of dimensions (400) for the RNN than for the M1 configuration (150), and +NFKD indicates the additional use of unicode normalisation.

Based on these new results, we have decided to test this new successful M3 configuration rather than the M1 and M2 previously presented as the most efficient. Two versions have been tested: the first with one linear layer, the second with two linear layers. For each configuration, due to the stochastic nature of the process, three models were trained, using early stopping with threshold 0.001 and patience 7. The best one was retained. For time and ecological purposes, NFKD normalisation has been tested on the most relevant architecture only.

3.4 Results

First, the existence of spelling variation seems to have a limited impact, or to be efficiently counter-balanced by our data augmentation: the accuracy remains similar to the one of models trained exclusively on normalised spelling data for lemmatisation (98.62% vs 99.28%) and even higher for POS-tagging (97.18% vs 97.01%). Such results are very promising. When looking into the details, we observe that two linear layers have an almost null (+0.001 pt of % for POS) or slightly negative impact (-0.0002 pt of % for lemmas). Considering the (much) larger training time and energy use, it does not seem to be a satisfactory solution.

| POS          | 1 layer | +NFKD | 2 layer | support |
|--------------|---------|-------|---------|---------|
| all          | 98.62   | 98.48 | 98.46   | 352,483 |
| unknown tokens | 68.96  | 68.07 | 68.31   | 3,843   |
| ambiguous tokens | 98.3   | 98.16 | 98.12   | 182,472 |
| unknown targets | 54.27  | 54.13 | 54.32   | 2,047   |

Table 4: Lemmas (normalised+original spelling)

Because NFKD lowers the amount of characters (235 → 164) we have decided to lower the amount of dimensions for the RNN character embeddings (400 → 300). The final results show a slightly negative impact on the accuracy, which proves that the efficiency of NFKD on the CornMol corpus is linked to the smaller size of the corpus and that unicode decomposition and normalisation is probably less relevant on big corpora – similar scores could be achieved with fine tuning, especially by lowering the number of dimensions.

| POS          | 1 layer | 2 layers | support |
|--------------|---------|----------|---------|
| all          | 97.09   | 97.18    | 14,303  |
| unknown tokens | 87.57  | 88.4     | 724     |
| ambiguous tokens | 93.81  | 94.11    | 4,281   |

Table 5: POS (normalised+original spelling)

The significant increase of training data (90,000 to 150,000 tokens) have a significant effect on the efficiency of the POS-tagging model, and we improve the results of the previous model based on normalised-spelling data only.

3.5 Impact of spelling variation

The score for “ambiguous tokens” given by Pie evaluates its efficiency regarding homographs, i.e. its capacity to disentangle tokens like entre which can be both a form of the verb entrer (il entre dans la pièce) or the preposition (entre les murs). However, lemmatising historical variation-rich languages is also about handling graphematic polymorphism, because a same word can frequently have different graphematic realisations (e.g. besoin vs besoing). We have therefore decided to evaluate the robustness of our best model regarding the graphematic volatility of (early-modern) texts, as Millour and Fort [2019] did for their study on dialectal variation.

214 pairs of two different tokens sharing the same lemma, POS-tag and morphology have been extracted from the whole prestogold...
corpus, so that we have a Form A and a Form B in each pair: afin vs affin, changements vs changemens. . . . It is to be noted that, by including only pairs that are actually present in the final training data, we have limited the significance of our results for out-of-domain and/or rare graphematic realisations.

In order to evaluate the effect of spelling variation, we designed a simple algorithm: in the test set, for each pair, all sentences containing a token matching the lemma and one of the variant of the pairs are extracted. This group of sentences is then duplicated: the first set is left untouched, while, in the second set, all occurrences of variant i are replaced by the alternative variant j. Both groups are then lemmatised and the accuracy is compared to evaluate the robustness of the model despite spelling variation in strictly equivalent contexts.

By doing so, we are able to evaluate the recognising of the spelling variants while neutralising variations due to the left and right contexts (each form is artificially given the same number of occurrences and the same contexts). We then test our model on both set, and compare accuracy for i (Acci) and for j (Accj), with a simple difference:

$$\Delta_{Acc} = |Acci - Accj|$$

For all these pairs, the median $\Delta_{Acc}$, the geometric mean $\Lambda_{Acc}$, and the weighted geometric mean, using the frequency of variant j in the training data as weight, are all equal to 0 (arithm. mean is 0.025; weight. arithm. mean is 0.011). We can therefore conclude that spelling variation has a very limited impact on the accuracy, as long as both forms are present in the training data. In the future, it would be necessary to extend this kind of analysis to pairs for which one variant is unseen by the model.

### 3.6 Out-of-domain tests

We have tested our best lemma and our best POS models on the same out-of-domain testing data than the one used for our previous study. There are two test sets for each century: one made only of the theatre, the other one of everything but theatre. Each test set is composed of 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, . . .).

Because the model built for the previous study was trained on normalised-spelling data, out-of-domain data had been normalised too. For this new study, the out-of-domain test set has been duplicated and now exists in both original and normalised transcriptions for the 16th, 17th and 18th c. texts only, and not for 19th and 20th c. texts, which do not require any spelling normalisation.

We observe that accuracies on normalised-spelling out-of-domain testing data are relatively similar for both the model trained on normalised data and the one trained on original data. The fact that the main difference are observed for theatrical plays could imply that the model trained on normalised data has been tailored for this genre. Without surprise, regarding non-normalised out-of-domain testing data, the accuracy of the model diminishes with centuries as we go back in time.

Results regarding POS annotation are relatively similar to the one of lemmatisation, and we observe again a significant impact of the theatrical genre on the accuracy.

| Corpus | 16th | 17th | 18th | 19th | 20th | All cent. |
|--------|------|------|------|------|------|-----------|
| Drama  | 97.6 | 98.10| 98.88| 98.34| 98.00| 98.19     |
| Not drama | 97.78| 98.02| 98.06| 96.97| 97.39| 97.65     |
| Both   | 97.69| 98.06| 98.46| 97.66| 97.70| 97.92     |

| Test 2 |
|--------|
| Drama  | 96.65| 97.42| 97.69| 98.2 | 97.5 | 97.51     |
| Not drama | 97.48| 98.24| 98.27| 97.12| 96.79| 97.59     |
| Both   | 97.08| 97.83| 97.99| 97.66| 97.15| 97.55     |

| Test 3 |
|--------|
| Drama  | 93.93| 96.36| 96.95| 98.2 | 97.5 | 96.64     |
| Not drama | 96.23| 97.25| 98.2 | 97.12| 96.79| 97.13     |
| Both   | 95.12| 96.8 | 97.59| 97.66| 97.15| 96.89     |

### 3.6 Out-of-domain tests

We have tested our best lemma and our best POS models on the same out-of-domain testing data than the one used for our previous study. There are two test sets for each century: one made only of theatre, the other one of everything but theatre. Each test set is composed of 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, . . .).

Because the model built for the previous study was trained on normalised-spelling data, out-of-domain data had been normalised too. For this new study, the out-of-domain test set has been duplicated and now exists in both original and normalised transcriptions for the 16th, 17th and 18th c. texts only, and not for 19th and 20th c. texts, which do not require any spelling normalisation.

We observe that accuracies on normalised-spelling out-of-domain testing data are relatively similar for both the model trained on normalised data and the one trained on original data. The fact that the main difference are observed for theatrical plays could imply that the model trained on normalised data has been tailored for this genre. Without surprise, regarding non-normalised out-of-domain testing data, the accuracy of the model diminishes with centuries as we go back in time.

Results regarding POS annotation are relatively similar to the one of lemmatisation, and we observe again a significant impact of the theatrical genre on the accuracy.

| Corpus | 16th | 17th | 18th | 19th | 20th | All cent. |
|--------|------|------|------|------|------|-----------|
| Drama  | 95.05| 96.59| 95.98| 94.81| 93.57| 95.18     |
| Not drama | 92.89| 94.27| 96.53| 91.87| 91.35| 93.42     |
| Both   | 93.93| 95.44| 96.27| 93.36| 92.48| 94.30     |

| Test 2 |
|--------|
| Drama  | 93.53| 95.68| 95.01| 93.37| 93.42| 94.19     |
| Not drama | 92.75| 94.5 | 95.21| 92.09| 93.59| 93.64     |
| Both   | 93.12| 95.09| 95.12| 92.74| 93.5 | 93.92     |

| Test 3 |
|--------|
| Drama  | 89.3 | 92.72| 94.49| 93.37| 93.49| 92.72     |
| Not drama | 89.42| 92.6 | 95.49| 92.09| 93.59| 92.67     |
| Both   | 89.36| 92.66| 95.01| 92.74| 93.5 | 92.69     |

### 4 INTERPRETATION

Interestingly, none of the most important errors are linked to graphematic polymorphism. Most of the problems are linked to homographs: let pronoun (il le veut) vs determiner (le château), the verb a (il a) vs the preposition (à Genève) when it does not have the grave accent (a Genève), the determiner des (il fait des ronds) vs the enclise de+le (le père des enfants). If such results confirm those obtained on normalised data, the presence of noise in the Presto core corpus could explain some errors.
5 FURTHER WORK
The next step, already in preparation, is the extension of the annotation to named entities to train a NER for EMF. With texts published on three centuries, and especially with a long excerpt of the Encyclopédie (c. 1,500,000 tokens), we do believe that our corpus is rich enough to carry preliminary tests. The presence of already corrected lemmas and POS tags for proper nouns (to which a special attention was given during the correction of the corpus) should dramatically ease the annotation process.

Regarding lemmatisation and POS-tagging, the most important task will be the extension of the training data for 16th c., and even 15th c. French, which is clearly underrepresented in our corpora: the complexity of middle French would certainly require more data, and new strategies to maintain the interoperability between medieval and (early) modern datasets.

6 CONTRIBUTIONS
This project is a follow-up to another one designed by Florian Cañiero, J.-B. C. and S.G. It has been led by S.G., who prepared the data and the final article, with the help of many: J.-B. T. for the conversion of Presto gold from MULTEX to CATTEX, Lucence Ing for the annotation of Presto gold, and Matthias Gille Levenson for the correction of Presto core. Trainings have been configured and executed by T.C. J.-B. C. has accompanied the entire process, providing technical guidance and philological feedback on the compatibility with medieval data. All authors discussed the results and contributed to the final manuscript.

7 DATA
The most up-to-date version of the models can be easily obtained and used thanks to the pie-extended Python package, available on PyPi (https://pypi.org/project/pie-extended/), with the command pie-extended download frem. All the data are available online at https://github.com/e-ditiones/LEM17.

ACKNOWLEDGMENTS
We would like to thank Lucence Ing and Frédéric Duval for their help regarding complicated annotation choices.

REFERENCES
ATILF-CNRS and Université de Lorraine. 1998-2018. Base textuelle Frantext: Démonstration. https://www.frantext.ens.fr/repository/frantext-demo/
ATILF-CNRS and Université de Lorraine. 2015. Dictionnaire du Moyen Français (1330-1500). http://www.atilf.fr/dmf
Sylvain Auroux. 2000. Les langues universelles. In Histoire des idées linguistiques: L’hégémonie du comparatisme. Philosophie et langage, Vol. 3. Mardaga, Liège / Bruxelles.
Peter Blumenthal, Sascha Diwersy, Achille Falaise, Marie-Hélène Lay, Gilles Sourvay, and Denis Vigier. 2017. Presto, un corpus diachronique pour le français des XVIe-XVIIIe s. École nationale des chartes.
Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. 2020a. Pie-Extended: an extension for Pie with pre-processing and post-processing. École nationale des chartes. https://doi.org/10.5281/zenodo.3883589
Thibault Clérice. 2020b. Protoingen, post-processing for NLP dataset. École nationale des chartes. https://doi.org/10.5281/zenodo.3883585
Thibault Clérice, Julien Pilla, Jean-Baptiste Camps, Vincent Jolivet, and Ariane Pinche. 2019. Pyrrha, A language independent post correction app for POS and lemmatisation. École nationale des chartes. https://doi.org/10.5281/zenodo.2325427
Sascha Diwersy, Achille Falaise, Marie-Hélène Lay, and Gilles Sourvay. 2017. Ressources et méthodes pour l’analyse diachronique. Langages N° 206, 2 (Aug. 2017), 21–44. https://www.cairn.info/revue-langages-2017-2-page-21.htm
Paul Füreve. 2007. Théâtre classique. http://www.theatre-classique.fr
Clémence Fourier and Benoît Sagot. 2020. Comparing Statistical and Neural Models for Learning Sound Correspondences. In LTHAL 2020: First Workshop on Language Technologies for Historical and Ancient Languages. Marseille, France. https://hal.inria.fr/hal-02529929
Simon Gabay, Jean-Baptiste Camps, and Thibault Clérice. 2020. Guidelines for linguistic annotation of modern French (16th-18th c.). https://hal.archives-ouvertes.fr/hal-02571190 Manuel d’annotation en vue de la création de modèle de lemmatisation et d’annotation morpho-syntaxique et morphologique du français des XVIe-XVIIIe s. Céline Guillot, Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. 2013a. Manuel de référence du jeu CatteX09. http://fbmüns-laon.fr/spip.php?article323
Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. 2017. Base de français médiéval: une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. Diachroniques. Revue de Linguistique française diachronique 7. (Dec. 2017), 168–184. https://halshs.archives-ouvertes.fr/halshs-01809581
Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. 2013b. Principes d’annotation CatteX09. http://fbm.ens-lyon.fr/spip.php?article173
Serge Heiden, Jean-Philippe Magué, and Bénédicte Pincemin. 2010. TXM : Une plateforme logicielle open-source pour la taxéméttrie - conception et développement, Vol. 2. Éditions Universitariè de Lettere Economia Diritto, Rome, Italy, 1021–1032. https://halshs.archives-ouvertes.fr/halshs-00549779
Bryan Jurish. 2012. Finite-state canonicalization techniques for histori
german. Doctoral thesis. Universität Potsdam, Potsdam, Germany. https://publisup.uni-potsdam.de/frontdoor/index/index/docId/5562
Enrique Manjavacas, Ako Kádár, and Mike Kestemont. 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). Association for Computational Linguistics, Minneapolis, Minnesota, 1493–1503. https://doi.org/10.18653/v1/N19-1153
Alice Millour and Karé Font. 2019. Unsupervised Data Augmentation for Less-
Resourced Languages with no Standardized Spelling. In RANLP, Varna, Bulgaria, 579–584. https://halshs.archives-ouvertes.fr/halshs-02280002/document
Jean-Christophe Pellat. 1995. L’évolution de l’orthographe des imprimés au XVIIe s. (libraires français et hollandais). In « Ces mots qui sont nos mots ». Mélanges d'Histoire de la Langue Française, de Dialectologie et d'Onomastique offerts au professeur Charles Jacqur. Institut Charles Buneau, Charleville-Mézières, 83–96.
Jean-Marie Pierrel, Jacques Dessendin, and Pascale Bernard. 2004. Le TLPI ou Trésor de la Langue Française informatisé. In Proceedings of the 11th EURALEX Interna
tional Congress (4-10). Geoffrey Williams and Sandra Vessier (Eds.). Université de Bretagne-Sud. Faculté des lettres et des sciences humaines, Lorient, 165–170.
Sophie Prévost, Céline Guillot, Alexei Lavrentiev, and Serge Heiden. 2013. Jeu d’étiquettes morphosyntaxiques CATTEX2009 (v2.0). http://fbm.ens-lyon.fr/spip.php?article176
Laurent Romary, Susanne Salmon-Alm, and Gil Franchoupo. 2004. Standards go
ing concrete: from LMF to Morphalou. In The 20th International Conference on Computational Linguistics (COLING 2004) - TextDict ’04 Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries. Geneva, Switzerland. 22–28. https://hal.inria.fr/inria-00121489
Benoît Sagot. 2010. Le Lefff, a freely available and large-cov
erage morphological dictionary, with focus on pre-processing and post-processing. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Nellecita Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (Eds.). European Language Resources Association (ELRA), Valletta, Malta, 2744–2751. https://hal.inria.fr/inria-00521242
Gilles Sourvay and Jean-Marie Pierrel. 2009. LGERM, a corpus of the most
data in Moyen Français. Traitement Automatique des Langues 50, 2 (2009), 149–172. https://halshs.archives-ouvertes.fr/halshs-00396452
### APPENDIX

The full breakdown of the *CornMol* and *FranText Open Access* corpora can be found in Camps et al. [2020]. We therefore describe only the *Presto* data we have used.

#### Presto Gold

| Author                  | Title                              | 1st ed | 2nd ed | Tokens |
|-------------------------|------------------------------------|--------|--------|--------|
| SCÈVE M.                | Saulsaye                           | 1547   | 1547   | 6,186  |
| DU RYER P.              | Lisandre et Caliste                | 1632   | 1632   | 11,984 |
| BUSSY-RABUTIN R. de     | Les Lettres                        | 1666   | 1720   | 11,807 |
| VOLTAIRE                | Essay sur l’histoire générale et sur les moeurs | 1756   | 1756   | 16,224 |
| RÉTIF DE LA BRETONNE N.-E. | Le Paysan perverti ou les Dangers de la ville | 1776   | 1776   | 16,144 |

#### Presto core

| Author                  | Title                              | 1st ed | 2nd ed | Tokens |
|-------------------------|------------------------------------|--------|--------|--------|
| RABELAIS Fr.            | Pantagruel                         | 1532   | 1542   | 46,571 |
| RABELAIS Fr.            | Gargantua                          | 1534   | 1542   | 50,499 |
| FLORES J. de            | La Disputeable fin du Florentin    | 1535   | 1536   | 30,500 |
| SCÈVE M.                | Saulsaye                           | 1547   | 1547   | 6,182  |
| DU BELLAY J.            | La défense, et illustration        | 1549   | 1549   | 20,974 |
| ANONYME                 | Sortie pour le croy de la bourgeoisie | 1549   | 1549   | 5,053  |
| DES PÉRIERS B.          | Nouvelles recreations et joyeux de vis | 1558   | 1561   | 81,471 |
| RONSARD P. de           | Discours des Muses de ce temps     | 1562   | 1563   | 6,451  |
| BOCHANAY G.             | Apht, ou le veau                   | 1567   | 1567   | 17,147 |
| MONTAIGNE M. de         | Essai                             | 1580   | 1580   | 224,183|
| URIÉ R. d.              | L’Artis, première partie          | 1607   | 1607   | 241,180|
| BÉRIOALDE DE VERVILLE Fr. | Le Moyen de parueresser             | 1616   | 1616   | 146,480|
| COLLETET G.             | Le Travaillement de l’Yvresgue    | 1627   | 1627   | 2,713  |
| BALZAC J.-L. GUEZ de    | Le Prince                          | 1631   | 1631   | 61,629 |
| DU RYER P.              | Lisandre et Caline                | 1632   | 1632   | 21,907 |
| PERESE N. Cl.           | Lettres : t. 7 : lettres à divers : 1602-1637 | 1657   | 1696   | 87,031 |
| ASSOUCY Ch. d’          | Poësie et lettres...              | 1653   | 1653   | 32,495 |
| BUSSY-RABUTIN R. de     | Les Lettres                        | 1666   | 1720   | 192,622|
| SCUDERY M. de           | Malachié                          | 1667   | 1667   | 77,031 |
| QUINault P.             | Thésée                            | 1675   | 1675   | 13,328 |
| ESPIRIT J.              | La Faunasté des vertus humaines    | 1678   | 1710   | 171,131|
| CHARLEVAL J.-L.         | Poësie                            | 1693   | 1759   | 12,111 |
| RIGNARD J.-Fr.          | Les Folies amoureuses              | 1704   | 1820   | 18,161 |
| Dacier A.               | Des censures de la corruption de nos royaumes | 1714   | 1714   | 63,273 |
| RACINE L.               | La Grece                           | 1720   | 1742   | 13,308 |
| BOSSY L. de             | Les Divers trompeurs, ou l’Homme du jour | 1740   | 1815   | 24,796 |
| VOLTAIRE                | Essay sur l’histoire générale et sur les moeurs | 1756   | 1756   | 416,624|
| COLLECTIF               | Encyclopédie (tome 7)              | 1757   | 1757   | 1,600,020|
| ALEMBERT d’             | Lettre à M. Boisseau               | 1759   | 1759   | 14,182 |
| SAINT-LAMBERT J.-Fr.    | Les Saisons                        | 1769   | 1769   | 29,575 |
| RÉTIF DE LA BRETONNE N.-E. | Le Paysan perverti ou les Dangers de la ville | 1776   | 1776   | 239,051|
| MIRABEAU H.-G. Ruperti  | Lettres originales écrites du duc.... | 1780   | 1792   | 346,662|
| COLLIN D’HARLEVILLE J.-Fr. | L’inconstante                   | 1786   | 1805   | 16,579 |
| LA PEROUSE J.-Fr.       | Wringage de La Pérouse autour du monde | 1797   | 1797   | 338,344|
This figure "deltaAcc_hist.png" is available in "png" format from:

http://arxiv.org/ps/2011.11074v1
This figure "deltaL_assocplot.png" is available in "png" format from:

http://arxiv.org/ps/2011.11074v1