Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre

Jean-Baptiste Camps¹, Simon Gabay², Paul Fièvre³, Thibault Clérice¹, Florian Cafiero⁴

¹Centre Jean-Mabillon, École nationale des chartes, Université Paris, Sciences & Lettres
²Université de Neuchâtel
³Bibliothèque Nationale de France
⁴GEMASS, CNRS / Université Paris-Sorbonne

Corresponding author: Jean-Baptiste Camps, Jean-Baptiste.Camps@chartes.psl.eu

Abstract
This paper describes the process of building an annotated corpus and training models for classical French literature, with a focus on theatre, and particularly comedies in verse. It was originally developed as a preliminary step to the stylometric analyses presented in Cafiero and Camps [2019]. The use of a recent lemmatiser based on neural networks and a CRF tagger allows to achieve accuracies beyond the current state-of-the art on the in-domain test, and proves to be robust during out-of-domain tests, i.e. up to 20th c. novels.

Keywords
Lemmatisation; POS tagging; 17th c. French; Classical Theatre.

I INTRODUCTION

If many lemmatisers and POS taggers have been trained, and sometimes conceived, for French (e.g. Tellier et al. [2012], Urieli [2013]…), they usually focus on contemporary French and tools for Ancien Régime French remain scarce. One important exception is the TreeTagger [Schmid, 1995] model developed by Diwersy et al. [2017] for the Presto project [Vigier and Blumenthal, 2013-2017]. They have prepared training data with c. 60,000 tokens annotated manually, using an adapted version of MULTEX [Ide and Veronis, 1994] and GRACE [Adda et al., 1998] for the parts of speech (POS), and the Leff [Sagot, 2010], Morphalou [Romary et al., 2004] and LGeRM [Souvay and Pierrel, 2009] for the lemmas. Unfortunately, training data are not publicly available (yet?)¹, they are made of non-normalised texts from the 16th to the 18th c. which prevents any comparison task, and no detailed evaluation of the model is offered by the designers of the model.

Recent developments have renewed the field of normalisation, lemmatisation and tagging of historic language varieties, including variation rich languages. Particularly, deep learning methods have been used to establish a new state of the art for Middle Dutch [Kestemont et al., 2016], Medieval Latin [Kestemont and De Gussem, 2017], Medieval Occitan [Camps and Couffignal, 2017], Early Irish [Dereza, 2019] or Middle High German and a variety of other languages [Schmid, 2019, Manjavacas et al., 2019]. Convolutional or (bidirectional) LSTM networks have

¹The Presto data have been entirely revised after the initial publication of this paper (cf. Gabay et al. [2020b]) and are available online at the following address: https://github.com/e-ditiones/LEM17.
been put to use, in architecture involving word or character levels representations, for instance word embeddings, and usually a sequence to sequence approach using an encoder–decoder model [Bollmann, 2019], that allow for taking into account the context of a token at either word or character level.

Such architectures are now competing with character level statistical machine translation models for the state of the art of historical text normalisations [Tang et al., 2018, Bollmann, 2019, Gabay and Barrault, 2020]. The developments of such architectures allowed to obtain results in the 89 to 96% range for lemmatisation of historical languages (and even up to 97.86% for standard languages like modern French, see Manjavacas et al. [2019]), though they have been showed to be the most dependent on the quantity and quality of training data, outperforming more traditional (essentially, rule based) approaches only when substantial training data is available [Bollmann, 2019].

POS and morphological taggers, especially in the case of morphologically rich languages, have also benefited from neural architectures, with state of the art for modern languages often being attained by recurrent neural nets [Ling et al., 2015, Heigold et al., 2016, Schmid, 2019], with accuracies in the range of 97-98% for POS and 92-94% for morphology [Heigold et al., 2016], and more specifically from 90 to 98% for POS tagging of historical languages [Schmid, 2019, Manjavacas et al., 2019].

In this paper, we present in detail our corpus, the annotation choices, the training set up and finally the results obtained by two models. On the one hand, an “extended” model for lemmatisation of normalised 17th c. French theatre, i.e. tested specifically for this data, but with enough additional training material to have it perform relatively well on modern (16th c. to 18th c.) and even contemporary (19th c. to 20th c.) French. On the other hand a POS-tagging model including the morphology (gender, number, tense...) based exclusively on 17th c. century training data.

These models were initially developed to allow the stylometric analyses presented in Cafiero and Camps [2019], especially the analyses based on lemma, lemma in rhyme position, and POS 3-grams. In this context, getting results as accurate as possible was a necessary step to ensure the reliability of stylistic results. This initial motivation is reflected in some of the choices of this paper, notably the constitution of the training corpus (which is a sample from the corpus that was to be used in the analyses)².

II THE THÉÂTRE CLASSIQUE CORPUS

Because of its cultural importance for French literature and thanks to a few innovative pioneers, 17th c. French theatre (also called théâtre classique, “classical theatre”) benefits from many digital editions freely available online. Among the various projects offering texts online, the Théâtre classique database [Fièvre, 2007, Schöch, 2018] proposes one of the most comprehensive collections. The texts derive from digitisations of 17th c. prints taken from Gallica [Bibliothèque nationale de France, 1997]. The oldest print is usually used, but it is not necessarily the case (for detailed bibliographic information on 17th c. printed theatre, cf. Riffaud [2014]). The text is normalised manually and the spelling aligned with contemporary French³.

Among all the plays available, 41 comedies (cf. appendix A) have been selected to create a training corpus. They have been written by six different authors spread over two generation:

²Stylometric results are not presented in this paper. For them, see Cafiero and Camps [2019].
³About this (strange) French tradition, cf. Gabay [2014], Duval [2015].
Antoine Le Métel d'Ouville (1589-1655), Jean de Rotrou (1609-1650), Paul Scarron (1610-1660), Pierre Corneille (1606-1684), Molière (1622-1673) and Thomas Corneille (1625-1709). All the plays have been written between the 1630’s and the 1670’s, that is to say within c. 40 years.

III BUILDING AN ANNOTATED CORPUS

The annotation scheme has been conceived to cope with diachronically spread data, especially earlier periods such as middle and Renaissance French.

3.1 Choice of authority lists

Since annotation principles are rather complex, we have decided to publish guidelines separately [Gabay et al., 2020a], and we will summarise here our most important choices.

Regarding POS tags, on top of the aforementioned MULTEx, many possibilities exist: EAGLES [Leech and Wilson, 1996], UD-POS [Petrov et al., 2011] and CATtex [Prévost et al., 2013]. While EAGLES and UD-POS have been developed as international standards, CATtex has been designed specifically for French medieval texts.

We have decided to choose CATtex, because we are interested in a long diachronic perspective, and we want to maintain interoperability with several existing corpora that are already using it, such as the Base de français médiéval (“Medieval French Database”, cf. Guillot-Barbance et al. [2017]) or the Geste corpus [Camps, 2019].

The annotation manual of CATtex09 (Guillot et al. [2013b]) offers a detailed list of tagging rules that we strictly observed. Three options are offered to the annotator: morphological tagging, morpho-syntactical tagging, or both. If adding both labels is ideal (to study processes such as adjectivisation, substantivisation…) it remains far to costly in time, and we have therefore opted for a simple morpho-syntactical tagging, that appeared at the time as an interesting middle way.

Regarding lemmatisation, we have already mentioned LGERM, the Leff and Morphalou. The main interest of the last one, that we have chosen, is that not only the v3.1 includes the Leff, but it is also used by the Frantext base – the data of which is partially available online to be used as additional material for our model – and the Trésor de la Langue Française informatisé. The LGERM lexicon is irrelevant in our case, since it is an artificially archaised version of Morphalou to match 17th and 18th c. forms. Concerning proper names, we built a specific reference list, thanks to the characters and places index provided by Fièvre [2007], that we expanded when necessary.

Some of our choices diverge from those made by the authors of Morphalou. We were, for instance, more systematic in choosing the masculine singular form as a lemma for nouns (baronne is lemmatised as baron) but not only (la (det. def.) as le, sa or ses (poss.) as son). Concerning personal pronouns, the singular masculine subject (when relevant) case has been used as lemma: direct regimen forms (le, la, les, as in, il les donne) as well as indirect regimen forms (lui, elle(s)) have been lemmatised to the subject masculine singular (il) – one single pronoun can indeed be subject (il), reflexive (se), direct object (le or en), indirect object (lui or y) or disjunctive (lui). Still in line with our diachronic approach, we kept the difference between the old partitive des

It is to be noted that using Cattex does not prevent, in the future, conversions to other tag-sets. For instance, the SRCMF corpus [Prévost and Stein, 2013], originally tagged with Cattex tags, exists in a converted version using UD-POS [Prevost Sophie, Stein Achim, SRCMF, 2020].

With morpho-syntactical tagging, categories are mainly determined in context: il veut le bien (bien is NOMcom) vs il est bien grand (bien is ADVgen).
(contracted form of de les) and the new non-definite plural article des, and encoded the contracted form au(x) as a+le.

It has to be noted that, since lemmas are not numbered in Morphalou, it has not been possible to introduce a number-based disambiguation for homographs (e.g., son1 (poss. “his”) vs. son2 (noun “sound”). . . ). Such a situation is however only partially problematic, since it remains possible to distinguish forms thanks to the POS, or the morphology.

3.2 Texts preprocessing and sampling
In order to limit model biases, each play of the corpus was sampled to create training and testing data. The text of Fièvre [2007] editions have been tokenised using TXM [Heiden, 2010] XML import. During the import an XSL filter was used to retain only the character’s speeches, with exclusion of all other material (stage directions, act and scene numbering…). Out of these data, a three-tier sample was constituted with the 2,000 first tokens of our 41 plays for training data (hereafter train set), the 100 median tokens for validation data (hereafter dev set), and the last 100 for testing data (hereafter test set). Case was not normalised, in order to keep information relevant to the identification of proper names.

The complete XML and annotation workflow is presented in fig. 1⁶.

3.3 Annotation and correction process
The annotation has been done in three phases. A first Pie [Manjavacas et al., 2019] lemmatisation model has been trained only on the Frantext Open Access data [ATILF-CNRS and Université de Lorraine, 1998-2018], and has been used to annotate a first sample of c. 40,000 tokens, in combination with an available Pie model for POS tags trained on Old French⁷. After being corrected, the same corpus has been used to train new models and annotate c. 40,000 other tokens that were, once again, corrected to create the final training corpus.

Lemma and POS-tags have all been corrected manually. This work was facilitated by the use of Pyrrha [Cléré et al., 2019], a web-based correction interface able to do batch correcting as well as to handle authority lists, allowing efficient collaborative work (fig. 2). Pyrrha also keeps tracks of all changes made on the corpus (fig. 3), and makes it possible to import, correct, share, and download back corpora and authority lists.

3.4 Expanding annotation through available resources
If POS-tags have all been systematically corrected, through both linear reading and batch corrections, it is not the case of the morphology, which has only been mostly batch-corrected, because of time concerns. Thus, we can guarantee that every POS tag has been proofread at least once (and usually multiple times), which is not the case for the morphology.

Indeed, to save time, morphological information was not added manually, but was instead projected using the lexicon of inflected forms Morphalou [ATILF-CNRS and Université de

---

⁶Since the initial publication of this article, the effects of Unicode NFKD normalisation have also been tested, with an unclear effect on training accuracy Gabay et al. [2020b].

⁷A recent version of the model for Old French can be found as part of the web application Deucalion [Cléré et al., 2019]; they are also directly usable through Pyrrha’s interface [Cléré et al., 2019]. Finally, the most up-to-date version of both the Old French and Classical French models is provided, along with functionalities to ease tagging of new documents, as part of the simple pie-extended Python Package [Cléré, 2020]. The models can be procured using their linguistic code (fro for Old French, fr for Classical French) by running the commands: pie-extended download fr and texts can be annotated by pie-extended tag fr MyFile.txt. They also are available online, on the École des chartes’ Deucalion instance, at https://dh.chartes.psl.eu/deucalion/.
Lorraine, 2016]. To do so, CATTEX POS-tags were mapped to Morphalou’s categories (table 1)\(^8\). Then, a simple algorithm looked for matching forms inside Morphalou: when the form was unambiguous, the morphological information was directly retrieved, otherwise the hand-corrected POS was used to assess the correct morphological information to retrieve. If none was found, an unknown morph tag was added.

| CATTEX  | Morphalou  | CATTEX  | Morphalou  | CATTEX  | Morphalou  |
|---------|------------|---------|------------|---------|------------|
| INJ     | Interjection | DETdef  | Déterminant | PROord  | Nom commun |
| ADVgen  | Adverbe    | DETndf  | Déterminant | PROrel  | Pronom     |
| ADVneg  | Adverbe    | DETdem  | Déterminant | PROint  | Pronom     |
| ADVint  | Adverbe    | DETpos  | Déterminant | PROcom  | Déterminant |
| ADVsub  | Adverbe    | DETind  | Déterminant | ADJqua  | Adjectif qualificatif |
| CONcog  | Conjonction | DETrel  | Déterminant | ADJind  | Adjectif qualificatif |
| CONsub  | Conjonction | DETint  | Déterminant | ADJord  | Adjectif qualificatif |
| VERcjg  | Verbe      | DETcom  | Déterminant | ADJpos  | Adjectif qualificatif |
| VERinf  | Verbe      | PROper  | Pronom     | NOMcom  | Nom commun |
| VERppe  | Verbe      | PROimp  | Pronom     | ADJcar  | Nombre     |
| VERppa  | Verbe      | PROadv  | Pronom     | DETcar  | Nombre     |
| PRE     | Préposition | PROpos  | Pronom     | PROcar  | Nombre     |
|         |            | PROdem  | Pronom     |         |            |
|         |            | PROind  | Pronom     |         |            |

Table 1: Mapping of CATTEX POS-tags to Morphalou categories

Finally, Morphalou’s information was converted back to CATTEX09 format, using the mapping presented in table 2.

| Morphalou | CATTEX  | Morphalou | CATTEX  | Morphalou | CATTEX  |
|-----------|---------|-----------|---------|-----------|---------|
| Mode      |         | Temps     |         | Nomb.     |         |
| indicative| ind     | present   | pst     | singular  | s       |
| imperative| imp     | imperfect | ipf     | plural    | p       |
| conditional| con     | future    | fut     | masculine | m       |
| subjunctive| sub     | simplePast| psp     | feminine  | f       |
| infinitive|          | Pers.     |         | neuter    | n       |
| past      |          | firstPerson| 1       | invariable| x       |
| participle|          | secondPerson| 2      | 1036442   | ERROR   |
|           |          | thirdPerson | 3       |           |         |

Table 2: Mapping of Morphalou to CATTEX flexion tags.

3.5 Increasing corpus generality with Frantext data

The Frantext base offers an open access version [ATILF-CNRS and Université de Lorraine, 1998-2018], 32 texts of which have been used to increase the training data (see appendix B). Marginal interventions have been made to correct some systematic errors, but also to ensure its consistency with our annotation choices. For instance, for pronouns (labelled as CLO, CLS and PRO in Frantext data), the lemmas je, me, m’, M’, moi, Moi, were mapped to je; likewise, ils, elle, elles, le, la, les, lui, leur, eux, Ils, Lui, Elle, Elles, to il, etc. On the other hand, some forms of celui and cela were originally lemmatised to il, and we changed the lemmatisation to celui.

\(^8\)Punctuation and proper names were dealt with separately and specifically, as well as CATTEX combined label, PRE.DETdef. The correspondence between some categories reflect substantial difference in the grammatical typology used. For instance, CATTEX PROord (e.g., ‘premier’ in ‘le premier’) match Morphalou’s more generic common noun category. For more information on CATTEX categories, see Guillot et al. [2013a].
and cela. Similarly, forms of chacun (or aucun) were lemmatised to un, and we changed it to chacun (or aucun).

We were also more systematic in the alignment of feminine and masculine forms to a single (masculine, singular) lemma, may it be for possessives (mienne, tienne, sienne to mien, tien, sien) or just nouns (hôtesse, amie, veuve, captive, to hôte, ami, veuf, capitif, etc).

A few minor corrections of obvious errors were also made (e.g., saurer to savoir), especially regarding homograph forms of some lemmas (e.g., between verbal forms of défaire, “undo’,’ and the noun défaite, “defeat”, or between ver, “worm” and vers, “verse”). An additional adjustment has also been made regarding the ligature æ (cœur), which has been preferred over its decomposed form (coeur).

IV TRAINING SETUP

As previously mentioned, many tools are already available. Tagger [Schmid, 1995] remains one of the most widely used, even though it is outperformed by other solutions. For the French language, TALISMAN [Urieli, 2013] or MElt Sagot [2016] are commonly used by the NLP community, but many other solutions are available such as Lemming (only for lemmas, cf. Müller et al. [2015]), Marmot (only for POS, cf. Müller et al. [2013]) or Pie (mainly for lemmas, cf. Manjavacas et al. [2019]). We have decided to use the two latter.

4.1 Lemmatisation

Concerning Pie as a lemmatiser, we tested three different configurations (table 3):

1. **base (sent-lm)**, default configuration, based on the configuration that achieved best accuracy described in Manjavacas et al. [2019], using sentence context, RNN character embeddings, as well as forward and backward language models;
2. **wembs** the same as the previous one, but with the adjunction of word embeddings, initialised using pretrained embeddings;
3. **bert** same as the previous one, but using CamemBERT embeddings [Martin et al., 2019], reduced from 768 to 150 dims;
4. **cnn+wembs** a configuration using CNN character embeddings, with word embeddings, based on skipgram on a larger unlemmatised corpus. This configuration is the one used for Cafiero and Camps [2019], with limited additional tuning.

|        | context | char embs | word embs | hidden size |
|--------|---------|-----------|-----------|-------------|
| base   | sentence| rnn 300 dims | 0          | 150         |
| wembs  | sent.+word | rnn 300 dims | 150 dims | 150         |
| bert   | sent.+word | rnn 300 dims | 768 to 150 dims | 150         |
| cnn+wembs | sent.+word | cnn 150 dims | 150 dims | 150         |

Table 3: Configurations tested with Pie

For each configuration, due to the stochastic nature of the process, five models were trained, using early stopping with threshold 0.001 and patience 6, and the best one was retained.

For the third configuration, word-embeddings were pretrained using Word2Vec Python implementation, on a large corpus of 343 theatre texts from Fièvre [2007] and those of the Frantext Open Access that we presented supra, for a total of c. 7M tokens.

---

9For this configuration, we used a config file provided to us by Enrique Manjavacas, main developer of Pie.
Figure 1: Workflow for building the training corpus and models (ellipses) and the annotated XML corpus used for stylometric analysis (rectangles); void-arrowheads specify steps particular to the initial model building phase
Figure 2: Screenshot of the *Pyrrha* interface: main correction view
Figure 3: Screenshot of the Pyrrha interface: correction history and corpus statistics
4.2 POS tagging
For POS-tagging, we trained both Marmot and Pie on the training data we produced, without further augmentation. The configurations were the following:

1. **Marmot**: base configuration provided with Marmot, using the dev set during training, and the test set for final evaluation;
2. **Pie**: same configurations that for lemmatisation (base (sent-lm).wembs and bert, see above and table 3) but with a CRF output layer to predict part-of-speech tags;
3. +aux: for each of the POS-tagging configuration, we tried to see if we could obtain a gain in accuracy by using auxiliary tasks (the tasks used can be seen in table 2 in bold, with the addition of Case, mainly used for personal pronouns). In a multi-task setting, we trained linear classifiers for each morphological feature, but sharing weights with the main task (POS prediction).

V RESULTS

5.1 Calibration and in-domain tests
Results of the Pie lemmatisation training are presented in table 4. The best Pie model (configurations 2, wembs) achieved 99.09% accuracy on the test set. The results obtained with Marmot and Pie for POS on the test set are presented in table 5, and the best results are achieved by the wembs+aux configuration. Yet, in both cases, the variation between the accuracies are relatively low, and not substantially higher than random variations between different iterations of the same configuration.

|                         | base | wembs | bert | cnn-wembs | support |
|-------------------------|------|-------|------|------------|---------|
| **all**                 | 98.80| **99.09**| 98.95| 98.80      | 4181    |
| unknown tokens          | 70.31| 71.88 | 76.56| 65.62      | 64      |
| ambiguous tokens        | 97.32| **98.02**| 97.43| 97.43      | 857     |
| unknown targets         | 50.00| 57.14 | **85.71**| 57.14      | 14      |

Table 4: Pie lemmatisation accuracies on the test set for the best model for each configuration. “Unknown tokens” are tokens never seen during training, while “ambiguous tokens” are forms that can correspond to different lemmas. “Unknown targets” are lemmas never seen in training, but that the neural network can still sometimes accurately predict, thanks to its character level modelling.

|                      | Marmot | Pie          |
|----------------------|--------|--------------|
|                      | base   | +aux | wembs | +aux | bert | +aux | support |
| All                  | 96.87  | 96.72 | 96.51 | 96.84 | **97.01** | 96.65 | 96.15 | 4181    |
| Ambiguous tokens     | NA     | 91.86 | 91.43 | **92.40** | 92.29 | 91.76 | 90.36 | 934     |
| Unknown tokens       | 82.57  | **86.24** | **86.24** | 78.44 | 81.65 | 76.61 | 73.39 | 218     |

Table 5: Results obtained for POS by the Marmot and Pie models on the test set. For the Pie models, we confronted trainings with or without morphology as auxiliary tasks.

5.2 Out-of-domain tests
To evaluate the ability of the best models to generalise for data from other periods, we conceived two out-of-domain corpora. Since we want to evaluate generality with regard to diachronic, diaphasic and diagenic (i.e. gender based) variation, we have selected samples from 16th to the 20th century texts, either from theatre plays or from a variety of genres outside theatre, literary or practical (administrative texts, correspondence, etc.), from male and female authors, in order to have:
• 20 samples of roughly 100 tokens for each century, 10 of theatre plays, 10 of a variety of other genres;
• roughly as much tokens written by men or women for each century;
• a comparable distribution of tokens by genres for each century.

In addition, for the samples concerning 17th century theatre, we excluded verse comedies in general, and the authors from which were drawn our training corpus. A complete list is given in appendix C.

We evaluate the best models (*wembs* and *wembs+aux*) for lemma and for POS on the out-of-domain data (table 6 and 7).

| Corpus | 16th | 17th | 18th | 19th | 20th | All cent. |
|--------|------|------|------|------|------|-----------|
| Theatre | 97.60 | 98.10 | 98.88 | 98.34 | 98.00 | 98.19 |
| Not theatre | 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 |

Table 6: Lemmatisation accuracies of the best models on out-of-domain data

| Corpus | 16th | 17th | 18th | 19th | 20th | All cent. |
|--------|------|------|------|------|------|-----------|
| Theatre | 95.05 | 96.59 | 95.98 | 94.81 | 93.57 | 95.18 |
| Not theatre | 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 |

Table 7: POS accuracies of the best models on out-of-domain data

The lemmatisation model proves to be relatively robust: globally, the loss of accuracy is of roughly 1 percentage point, while it is closer to 3 percentage points for the POS model. This difference can be explained by the difference between the training corpora: the use of significant additional data to improve the efficiency of the lemmatisation model seems to reflect in its greater capacity to generalise. The same reason could also explain why the lemmatisation models transpose better to non-dramatic texts than the POS model.

In both cases, though, the better accuracies are reached for the 17th and 18th centuries – and, surprisingly, more specifically for 18th century data in most cases. It progressively decreases going backward or forward in time.

### 5.3 Improvement over previous corpus

In order to evaluate the importance of the addition of the newly annotated data (*Theatre*), we set-up a second training experiment where we reuse the best configuration found in 4.1 (*wembs*) over 5 iterations using the same training, dev and test set, except for the files from the *Theatre* sample in train and dev. Using the best model, we are able to measure the effect of the introduction of *Theatre* data on top of the *Frantext* data. On lemmatization, the results are clear: the overall accuracy with the same configuration goes from 97.30% to 99.09%, an increase of 1.79 points. We cannot compare unknown and ambiguous tokens as well as unknown tokens as the support from these categories moved: -88 unknown tokens, +34 ambiguous tokens, -23 unknown targets. The same impact can be seen on out-of-domain data (Table 8) with a more visible impact on the Theatre corpus of the 18th century (+0.96 points). Strangely enough, neither the 17th century or the Theatre genre benefits the most from the inclusion of 17th century Theatre data, as the *Not Theatre* corpus and the 18th century are the most improved ones.
Table 8: Score improvements between Frantext corpus and Frantext+Theatre. Reads as follow: on 17th century theatre, a model trained with Frantext+Theatre has +0.30 points of accuracy, going from 97.80 to 98.10%.

5.4 Most frequent confusions
The most frequent confusions of the best models on the out-of-domain data are presented in table 9.

| Expected | Tot. Err. | Pred. | Pred. times |
|----------|-----------|-------|-------------|
| LEA M. 9 | 9         | NOMcom 182 | ADJqua 45 |
| le 9     | il 9      | NOMpro 33 | VERinf 32 |
| cœur 6   | cuur 4    | VERcjg 18 | VERppe 11 |
| franc 6  | Franc 6   |        |            |
| Électre 5 | élECTRE 4 | ADJqua 104 | NOMcom 45 |
| noble 4  | Nobles 4  | NOMpro 81 | NOMcom 43 |
| maître 4 | maîtresse 4 | VERcjg 61 | NOMcom 18 |
| de_le 4  | un 4      | ADJqua 12 | VERppe 12 |
| un 4     | de_le 4   |        |            |
| dame 3   | Dame 2    | ADVgen 58 | NOMcom 16 |
| Phanor 3 | phanor 3  |        |            |
| voir 3   | vivre 3   | PRorel 31 | CONsub 26 |
| Vosges 3 | vosge 3   |        |            |
| Médée 3  | mèdè 3    |        |            |

Table 9: Sample from the confusion matrix for the best Pie models on the out-of-domain data.

Regarding lemmatisation, some errors related to homographs such as the token le (regimen pronoun (il) or determiner (le); or the token des (plural of the determiner un or partitive de_le). Some other errors are due to abbreviations not present in the training data (M. for monsieur).

More interestingly, for the research of potential improvements, many errors are related to rare characters classes in the training data, such as capital letters or ligatures (œ).

For POS, the most frequent confusions are in nominal rather than verbal tags. In particular, confusions between common nouns (NOMcom), proper nouns (NOMpro), adjectives (ADJqua) and nominal forms of the verbs (participle, VERppe, and infinitive, VERinf). Some errors are due to the morpho-syntactic nature of our annotation, which, for instance, labels substantive adjective as common nouns (le beau is NOMcom).

VI DISCUSSION

The scores of up to 99% for in-domain lemmatisation and 97% for POS tagging put our best models above the expected state of the art, or in it’s upper range (see I). Yet, these results were deserving of further investigation, through out-of-domain tests and more qualitative inspection.
Out-of-domain tests show that, even though they were trained for 17th century classical theatre, the model reach their best accuracies for 18th century texts. Such a fact seems counter-intuitive, but a plausible explanation might be as follow: the median date of the texts of the Frantext corpus is 1872, and mean date is 1859. By adding roughly 80k tokens of 17th century texts to the 2,3M of Frantext (with mean text date 1859), we may have somehow slightly pulled the corpus closer to the 18th century.

In any case, through the careful construction of a labelled corpus and the use of a neural lemmatiser and tagger, we were able to develop models very suitable to the annotation necessary for the stylometric analyses showed in Cafiero and Camps [2019]. But we think that the model are useful for the annotation of 17th century theatre, and even beyond that, of Early Modern French texts in normalised spelling, with encouraging results regarding generalisation beyond the original scope of the experiments.

VII FURTHER RESEARCH

Looking at our results, the main lead for improvements should be a more efficient way to deal with rare character classes, such as capital letters, diacritics or ligatures. Several methods could be used to reduce the number of classes (e.g., Unicode decomposed normalisation) or, alternatively, the training set could be sufficiently extended to provide enough cases.

More generally, three possible directions could be followed in the coming years. The first is to expand the training corpus to other dramatic genres (tragedy, tragi-comedy, . . .), and other genres in general (poetry, novels, short stories, . . .). The second would be to replace normalised tokens by non-normalised ones, and therefore offer a new model that takes full advantage of the ability of tools like Pie to deal with spelling variation in historical languages, and, doing so, strengthen the ability of model to deal with older varieties of French. The third is to expand dramatically the training with data taken from 18th or 16th c. texts. Results going in these directions and continuing the experiments described in this paper will be presented at DTUC’20 [Gabay et al., 2020b].

AUTHOR CONTRIBUTIONS

P.F. encoded the corpus and all its metadata. F.C., J.-B. C. and S.G. designed the research project. The preprocessing of the texts, the initial setup of the Pyrrha corpus and of the authority lists were performed by J.-B. C. Correction of the training data and expansion of the authority lists was shared equally between F.C., J.-B. C. and S.G. Post-processing of the trained corpus, injection of morphological data, and correction of the Frantext data was done by J.-B. C., as well as the training, testing of the models and their further use to annotate unseen data. T.C. programmed modifications to Pie code to include Bert and participated in the training and benchmarking of models, as well as additional debugging of the annotation tools. All authors contributed to the writing of this paper.

The authors have no competing interests to declare.

MATERIALS AND DATA AVAILABILITY

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 fr, and can be queried at https://tal.chartes.psl.eu/deucalion/.
The initial version created for Cafiero and Camps [2019] is available from the Science Advances website since the publication of the paper, on 27th Nov. 2019 at this address: https://advances.sciencemag.org/highwire/filestream/221312/field_highwire_adjunct_files/0/aax5489_Data_file_S1.zip. The initial version of the models is available on Zenodo (10.5281/zenodo.3353421).

The version of the best models described in this paper, as well as training, validation and test data can be found on Zenodo as well, as version 3 of the same repository (doi: 10.5281/zenodo.3828644).

Licensing is specified separately for each repository.

ACKNOWLEDGEMENTS

We thank the DIM Science du texte et connaissances nouvelles for funding the acquisition of a GPU server, as well as the École nationale des chartes for providing infrastructure and support for the server. We thank Enrique Manjavacas for his precious advice regarding lemmatisation and Pie configuration, as well as Marie Puren for her help with the annotation of 20th century texts.

References

Gilles Adda, Joseph Mariani, Josette Lecomte, Patrick Paroubek, and Martin Rajman. The grace french part-of-speech tagging evaluation task. In Proc. of LREC’98 (1st International Conference on Language Resources and Evaluation), 1998. URL https://infoscience.epfl.ch/record/98004.

ATILF-CNRS and Université de Lorraine. Base textuelle Frantext: Démonstration, 1998-2018. URL https://www.frantext.fr/repository/frantext-demo/.

ATILF-CNRS and Université de Lorraine. Morphalou v3.1, 2016. URL https://www.ortolang.fr/market/lexicons/morphalou.

Bibliothèque nationale de France. Gallica - bibliothèque numérique de la bnf, 1997. URL https://gallica.bnf.fr.

Marcel Bollmann. A Large-Scale Comparison of Historical Text Normalization Systems. Proceedings of the 2019 Conference of the North, pages 3885–3898, 2019. doi: 10.18653/v1/N19-1389. URL http://arxiv.org/abs/1904.02036.

Florian Cafiero and Jean-Baptiste Camps. Why Molière most likely did write his plays. Science Advances, 5(11), 2019. doi: 10.1126/sciadv.aax5489. URL https://advances.sciencemag.org/content/5/11/eaax5489.

Jean-Baptiste Camps, editor. Geste: un corpus de chansons de geste, 2016-… (Version 02). École nationale des chartes, Paris, April 2019. URL http://doi.org/10.5281/zenodo.2630574. textes du domaine public, développements CC-BY-SA.

Jean-Baptiste Camps and Gilles Guilhem Couffignal. La production de corpus d’occitan médiéval et prémorderné: problèmes et perspectives de travail. In Actes du XIIème Congrès International de l’Association Internationale d’Études Occitanes, Albi 2017, 2017. URL https://halshs.archives-ouvertes.fr/halshs-02050089/.

Thibault Clérice. pie-extended 0.0.13, 2020. URL https://pypi.org/project/pie-extended/.

Thibault Clérice, Julien Pilla, and Jean-Baptiste Camps. hipster-philology/pyrrha: 2.0.0, 2019. https://doi.org/10.5281/zenodo.2541730.

Thibault Clérice, Jean-Baptiste Camps, and Ariane Pinche. Deucalion, modèle ancien francais, 2019. URL http://dx.doi.org/10.5281/zenodo.2539134.

Oksana Dereza. Lemmatisation for under-resourced languages with sequence-to-sequence learning: A case of early irish. In Proceedings of Third Workshop Computational linguistics and language science, volume 4, pages 113–124, 2019.

Sascha Diversy, Achille Falaise, Marie-Hélène Lay, and Gilles Souvay. Ressources et méthodes pour l’analyse diachronique. Langages, N 206(2):21–44, August 2017. ISSN 0458-726X. URL https://www.cairn.info/revue-langages-2017-2-page-21.htm.
Frédéric Duval. Les éditions de textes du XVIIe siècle. In David Trotter, editor, *Manuel de la philologie de l’édition*, pages 369–394. De Gruyter, 2015. ISBN 978-3-11-030260-8. doi: 10.1515/9783110302608-017. URL https://www.degruyter.com/view/books/9783110302608/9783110302608-017.xml.

Paul Fiévre. Théâtre classique, 2007. URL http://www.theatre-classique.fr.

Simon Gabay. Pourquoi moderniser l’orthographe? principes d’écrit et littérature du XVIIe siècle. *Vox Romanica*, 73(1):27–42, 2014. ISSN 0042-899X. URL https://elibrary.narr.digital/article/99.125005/vox201410027.

Simon Gabay and Loïc Barrault. Traduction automatique pour la normalisation du français du XVIIe siècle. In *TALN 2020*, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, Nancy, France, June 2020. ATALA. URL https://hal.archives-ouvertes.fr/hal-02596669.

Simon Gabay, Jean-Baptiste Camps, and Thibault Clériche. Manuel d’annotation linguistique pour le français moderne (xvie-xviiiie siècles), 2020a. URL https://hal.archives-ouvertes.fr/hal-02571190.

Simon Gabay, Thibault Clériche, Jean-Baptiste Camps, Jean-Baptiste Tanguy, and Matthias Gille-Levenson. Standardizing linguistic data: method and tools for annotating (pre-orthographic) French. In *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, pages 1–7, 2020b.

Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. *Manuel de référence du jeu Cattex09*. École normale supérieure de Lyon, Lyon, 2013a. URL http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf. Version 2.0 – 8 avril 2013.

Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. Principes d’annotation cattex09. Technical report, École normale supérieure de Lyon, Lyon, 2013b. version 2.0. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf.

Céline Guillot-Barbance, Serge Heiden, and Alexei Lavrentiev. Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques*, 7:168–184, 2017. URL https://halshs.archives-ouvertes.fr/halshs-01809581.

Serge Heiden. The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme. In *24th Pacific Asia Conference on Language, Information and Computation*, pages 389–398. DECODE, Waseda University, 2010.

Georg Heigold, Guenter Neumann, and Josef van Genabith. Neural morphological tagging from characters for morphologically rich languages. *arXiv preprint arXiv:1606.06640*, 2016.

Nancy Ide and Jean Veronis. Multext: Multilingual text tools and corpora. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, 1994. URL https://www.aclweb.org/anthology/C94-1097.

Mike Kestemont and Jeroen De Gussem. Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning. *Journal of Data Mining & Digital Humanities*, 2017. URL https://jmdh.episciences.org/3835.

Mike Kestemont, Guy de Pauw, Renske van Nie, and Walter Daelemans. Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities*, page fqw034, 2016. ISSN 2055-7671, 2055-768X. doi: 10.1093/llc/fqw034. URL http://dsh.oxfordjournals.org/content/early/2016/08/26/llc.fqw034.

Geoffrey Leech and Andrew Wilson. Eagles: Recommendations for the morphosyntactic annotation of corpora. Technical report, Expert Advisory Group on Language Engineering Standards, 1996. URL https://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/annotate.pdf.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W. Black, and Isabel Trancoso. Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*, 2015.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. Improving Lemmatization of Non-Standard Languages with Joint Learning. *arXiv e-prints*, art. arXiv:1903.06939, Mar 2019. URL https://www.aclweb.org/anthology/N19-1153/.

Enrique Manjavacas, Akos Kádár, and Mike Kestemont. Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*, 2019.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seilléh, and Benoît Sagot. CamemBERT: a Tasty French Language Model. *arXiv e-prints*, art. arXiv:1911.03894, Nov 2019.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. Efficient higher-order crfs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, 2013.
A PLAYS SAMPLED TO CREATE THE INITIAL TRAINING CORPUS

The following plays selected from Fièvre [2007] were sampled.

| id   | auteur       | titre                                      | date | genre  | inspiration | structure | type | periode        |
|------|--------------|--------------------------------------------|------|--------|-------------|-----------|------|----------------|
| CP_DONSANCH | CORNEILLE, Pierre | DON SANCHE D'ARAGON, COMÉDIE HERIQUÉ | 1649 | Comédie | moeurs espagnoles | 5 act. | vers | 1641-1650     |
| CP_GALERIEDEPALAIS | CORNEILLE, Pierre | LA GALERIE DU PALAIS ou L'AMEE RINALE | 1637 | Comédie | moeurs françaises | 5 act. | vers | 1631-1640     |
| CP_ILLUSIONCOMIQUE | CORNEILLE, Pierre | L'ILLUSION COMIQUE, COMÉDIE | 1639 | Comédie | moeurs françaises | 5 act. | vers | 1631-1640     |
| CP_MELITE33 | CORNEILLE, Pierre | MELITE OU LES FAUSSES LETRES, COMÉDIE | 1633 | Comédie | moeurs françaises | 5 act. | vers | 1621-1630     |
| CP_MENTEUR | CORNEILLE, Pierre | LE MENTEUR, COMÉDIE | 1644 | Comédie | moeurs françaises | 5 act. | vers | 1641-1650     |
| PP | DATE | AUTEUR | TITRE | GENRE | ACTES | ANNEE | PUBLICATIONS |
|---|---|---|---|---|---|---|---|
| 1673 | Comédie | histoire chrétienne | 5 | vers | 1671-1680 |
| 1645 |  Comédie moeurs françaises | 5 | act. | vers | 1641-1650 |
| 1637 | Comédie moeurs françaises | 5 | act. | vers | 1631-1640 |
| 1671 | Comédie | histoire romaine | 5 | act. | 1661-1670 |
| 1634 | Comédie moeurs françaises | 5 | act. | vers | 1631-1640 |
| 1651 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |
| 1658 |  Comédie moeurs italiennes | 5 | act. | vers | 1651-1660 |
| 1671 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |
| 1660 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |
| 1651 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |
| 1677 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |
| 1659 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |
| 1660 | Comédie moeurs italiennes | 5 | act. | vers | 1651-1660 |
| 1675 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |
| 1680 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |
| 1655 | Comédie moeurs italiennes | 5 | act. | vers | 1651-1660 |
| 1656 | Comédie moeurs italiennes | 5 | act. | vers | 1651-1660 |
| 1666 | Comédie moeurs italiennes | 5 | act. | vers | 1651-1660 |
| 1665 | Comédie moeurs italiennes | 5 | act. | vers | 1651-1660 |
| 1672 | Comédie moeurs françaises | 5 | act. | vers | 1651-1660 |
| 1667 | Comédie moeurs italiennes | 5 | act. | vers | 1651-1660 |
| 1669 | Comédie moeurs italiennes | 5 | act. | vers | 1651-1660 |
| 1643 | Comédie | moeurs grec | 5 | act. | vers | 1641-1650 |
| 1643 | Comédie | moeurs grec | 5 | act. | vers | 1641-1650 |
| 1643 | Comédie moeurs espagnoles | 5 | act. | vers | 1641-1650 |
| 1643 | Comédie moeurs espagnoles | 5 | act. | vers | 1641-1650 |
| 1643 | Comédie moeurs espagnoles | 5 | act. | vers | 1641-1650 |
| 1650 | Comédie moeurs espagnoles | 5 | act. | vers | 1641-1650 |
| 1639 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |
| 1640 | Comédie moeurs espagnoles | 5 | act. | vers | 1631-1640 |
| 1638 | Comédie moeurs italiennes | 5 | act. | vers | 1631-1640 |
| 1655 | Comédie moeurs espagnoles | 5 | act. | vers | 1641-1650 |
| 1654 | Comédie moeurs espagnoles | 5 | act. | vers | 1641-1650 |
| 1650 | Comédie moeurs espagnoles | 5 | act. | vers | 1641-1650 |
| 1648 | Comédie moeurs francs | 5 | act. | vers | 1641-1650 |
| 1646 | Comédie moeurs espagnoles | 5 | act. | vers | 1641-1650 |
| 1656 | Comédie moeurs espagnoles | 5 | act. | vers | 1651-1660 |

† The edition used as base is from 1690, but the play dates back to 1670.
The edition used as base is from 1709, but the play dates back to 1652.
‡ Created in 1661, but published 1680.
B ANNOTATED TEXTS FROM THE FRANTEXT BASE

The following texts, from ATILF-CNRS and Université de Lorraine [1998-2018], were used as complementary data for lemmatisation training.

| id   | auteur     | titre                              | date | place | publisher   | N tok |
|------|------------|------------------------------------|------|-------|-------------|------|
| K699 | 1790:52916 | LETO Pierre                        | 1891| Paris | Gallimard   | 61,437|
| K934 | 1798:53797 | ROSTAND Edmund                     | 1908| Paris | Faussquelle | 47,213|
| K999 | 1809:45606 | GOURMONT René de                   | 1899| NA    | Soc. Mercure de Fr. | 53,640|
| L233 | 1895:13727 | JARRY Alfred                       | 1922| s.p.  | Faussquelle | 38,813|
| L266 | 1905:90455 | POINÇON Henri                      | 1905| Paris | Flammannier | 66,125|
| L435 | 1852:37455 | BICHOû Henry                       | 1907| Paris | Faussquelle | 79,058|
| L466 | 1858:68155 | HUYSMANS Josi-Jean                 | 1902| Paris | Offenlof   | 71,159|
| L299 | 1850:53336 | JAURES Jean                        | 1904| Paris | E. Bartlet  | 24,085|
| L784 | 1898:100999| LEROUX Gaston                      | 1904| Paris | Flammannier | 50,232|
| L448 | 1857:68366 | ABOUT Edmund                       | 1867| Paris | M. Levy    | 27,381|
| L854 | 1846:20871 | MEILHAC Henri et Ludovic HALÉVY   | 1867| Paris | M. Levy    | 27,381|
| M232 | 1873:70999 | VERNE Jules                        | 1877| Paris | Hetzel     | 84,925|
| M289 | 1874:49844 | FROMENTIN Eugène                   | 1877| Paris | Plon       | 80,612|
| M362 | 1825:11446 | BRILLAS-SAUVARIN Jean-Athelmine    | 1847| Paris | Charpentier | 127,968|
| M374 | 1830:28471 | FOURIER Charles                    | 1830| Paris | Bossange   | 32,566|
| M425 | 1794:4621  | CHENIER André                      | 1910| Paris | Delagrange | 5,473 |
| M433 | 1807:57291 | STAEL Germain de                   | 1807| s.p.  | Peltier     | 65,012|
| M464 | 1801:74280 | DESTUTT DE TRACY                   | 1804| Paris | Cuvier     | 80,433|
| M466 | 1803:62746 | KREUZNER Barbare Juliane von       | 1840| Paris | Charpentier | 73,463|
| M473 | 1809:82487 | LAMARCK Jean-Baptiste              | 1809| s.p.  | Denti      | 95,111|
| M492 | 1805:12259 | CUVIER Georges                     | 1805| Paris | Baudouin   | 138,674|
| M528 | 1798:12661 | GUILHERM DE PIÈREXCOURT            | 1798| Paris | Barbe      | 66,878|
| M548 | 1832:199497| SAY Jean-Baptiste                  | 1841| Paris | Gaullmann  | 223,038|
| M622 | 1798:24358 | SIÈYES Emmanuel                   | 1888| Paris | Soc. Hist. R. Fr. | 27,002|
| M629 | 1792:31251 | FLORIAN Jean-Pierre                | 1783| Paris | Didot      | 37,172|
| M993 | 1869:61374 | GONCOURT Édouard et Jules de       | 1876| s.p.  | Charpentier | 71,174|
| M939 | 1852:84555 | COMTE Auguste                      | 1909| Paris | Garnier    | 95,569|
| N245 | 1838:32486 | HUGO Victor                        | 1905| Paris | Offenlof   | 41,487|
| N268 | 1802:65535 | BONAVENTUR Louis de               | 1802| Paris | Le Clerc   | 72,539|
| N429 | 1778:64388 | BUFFON Georges-Louis de            | 1778| Paris | Imperial    | 70,866|
| P556 | 1872:177870| VIOLETTE-DUC Eugène                | 1872| Paris | A. Morel    | 199,695|
| Q543 | 1784:54328 | RETIF DE LA BRONTE önicus          | 1784| s.p.  | Ver Duchesne | 64,424|

C OUT-OF-DOMAIN TEXTS

Here follows the list of the texts that were sampled for building the out-of-domain test set. Spelling was modernised when necessary.

Theatre texts were sampled from Fièvre [2007], with the exception of:

AUBIGNAC_PUCELLE sample transcribed and modernised from the edition Paris, 1642, available on Gallica.

GRINGORE_SAINT-LOUIS sample transcribed and modernised from the Œuvres complètes, éd. Ch. d'Héricault et A. Montaiglon, Paris, 1858-1877, available on Gallica.

ANOUILH_MEDEE sample transcribed from the Nouvelles pièces noires, Paris, La Table Ronde, 1976.

GIRAUDOUX_ELECTRE sample transcribed from the Théâtre complet, éd. J. Body, Paris, Gallimard, 1982.

CESAIRE_CHRISTOPHE sample transcribed from Poésie, Théâtre, Essais et Discours, éd. Albert James Arnold, Paris, CNRS Editions, 2013.

For the non-dramatic texts, the main sources, as shown in the table, are:

ELEC Éditions en ligne de l'École des chartes, http://elec.enc.sorbonne.fr/.

GALL Gallica, Bibliothèque nationale de France, https://gallica.bnf.fr/.

WS Wikisource: la bibliothèque libre, https://fr.wikisource.org/.

Other online editions were occasionnaly used:

Correspondance d'Isabelle de Charrière, éd. Suzan van Dijk and Madeleine van Strien-Chardonneau, https://charriere.huygens.knaw.nl/.

Les Nouvelles Nouvelles (1663) by Jean Donneau de Visé, éd. Claude Bourqui and Christophe Schuwey, http://www.unifr.ch/nouvellesnouvelles/.

Œuvres de Frédéric le Grand - Werke Friedrichs des Großen Digitale Ausgabe der Universitätsbibliothek Trier, dir. Hans-Ulrich Seifert, Trier, http://friedrich.uni-trier.de/.

Testaments de Poilus (1914-1918): transcription collaborative, dir. Florence Clavaud, 2018-...., https://testaments-de-poilus.huma-num.fr/.

The sample from MONTEGUT_ISMENE was transcribed and modernised from the edition Paris, 1768, available on Google Books.
### 3.1 Theatre

| id        | author                      | title                                      | date       | genre                  | inspiration                      | structure     | type    | period     |
|-----------|-----------------------------|--------------------------------------------|------------|------------------------|----------------------------------|---------------|---------|------------|
| 3.1.1     | ANONYME_PARDONNEUR          | anonyme                                    | 1500 c.    | Farce                  | Sermon joyeux                    | 1 act.        | vers    | 1501-1600 |
| 3.1.2     | ANONYME_PONTAUXANES         | anonyme                                    | 1500 c.    | Farce                  | Sermon joyeux                    | 1 act.        | vers    | 1501-1600 |
| 3.1.3     | ANONYME_RESURRECTION-       | anonyme                                    | 1500 c.    | Farce                  | JEUVEE AU PONT AUX ANES           | 1 act.        | vers    | 1501-1600 |
| 3.1.4     | ANONYME_SERMONJOYNEUX       | anonyme                                    | 1500 c.    | Farce                  | SERMON JOYEUX DE BIEN BOIRE       | 1 act.        | vers    | 1501-1600 |
| 3.1.5     | BEZE_ABRAHAM                | BEZE, Théodore                             | 1550       | Tragédie               | bible                             | 1 act.        | vers    | 1541-1550 |
| 3.1.6     | GRINGORE-SAINT-LOUIS        | GRINGORE, Pierre                           | 1514       | mystère                | histoire médiévale               | 8 livres      | vers    | 1511-1520 |
| 3.1.7     | JODELLE_CLOEPATRE           | JODELLE, Étienne                           | 1574       | Tragédie               | histoire romaine                 | 5 act.        | vers    | 1571-1580 |
| 3.1.8     | JODELLE_DIDON               | JODELLE, Étienne                           | 1574       | Tragédie               | mythe grec                        | 5 act.        | vers    | 1571-1580 |
| 3.1.9     | JODELLE_EUGENE              | JODELLE, Étienne                           | 1574       | Comédie                | moeurs françaises                 | 5 act.        | vers    | 1571-1580 |
| 3.1.10    | TURNÈBE.CONTENTS            | TURNÈBE, Odet de                          | 1584       | Comédie                | moeurs françaises                 | 5 prose      | 1581-1590|            |
| 3.1.11    | AUBIGNAC_FUCELLE            | AUBIGNAC, François                        | 1642       | Tragédie               | histoire médiévale                | prose         | vers    | 1641-1650 |
| 3.1.12    | CHAMPEPLUS_ULYSSE           | CHAMPEPLUS, Jacques de                     | 1663       | Tragédie               | mythe grec                        | 5 act.        | vers    | 1601-1610 |
| 3.1.13    | CHAPPEZEAU_ARMETZAR         | CHAPPEZEAU, Samuel                        | 1656       | Tragécomédie           | histoire turque                   | 5 act.        | vers    | 1651-1660 |
| 3.1.14    | DESHOULIERES_GERESERIC      | DESHOULIERES, An-                        | 1680       | Tragédie               | histoire romaine                 | 5 act.        | vers    | 1671-1680 |
| 3.1.15    | DESJARDINS_FAVORI           | DESJARDINS, Manti-                        | 1665       | Tragédie               | moeurs espagnoles                 | 5 act.        | vers    | 1661-1670 |
| 3.1.16    | DURANT_OISVETE              | DURANT, Catherine                         | 1699       | Povretrie              | moeurs françaises                 | 1 act.        | vers    | 1691-1700 |
| 3.1.17    | MATHIEU_MAGICCIENNE-        | MATHIEU, Pierre                           | 1617       | Tragédie               | histoire française                | 4 act.        | vers    | 1611-1620 |
| 3.1.18    | SCUDERY_LIGDAMONELIDAS      | SCUDERY, Georges de                       | 1631       | Tragédie               | histoire médiévale                | 5 act.        | vers    | 1631-1640 |
| 3.1.19    | URBES_Sylvain              | URBES, Honoré d                           | 1627       | Pastoralement          | pastoralement                     | 5 act.        | vers    | 1621-1630 |
| 3.1.20    | VONDREBECK-ALARD_FORSDELAMOUR | VONDREBECK, Maurice et ALARD, Charles    | 1678       | Diversissement         | moeurs françaises                 | 3 act.        | vers    | 1671-1680 |
| 3.1.21    | BERGASSE_JOURNEE-DESJUPES-1790 | BERGASSE, Isaak                           | 1790       | Tragédie               | histoire française                | 4 act.        | prose   | 1781-1790 |
| 3.1.22    | BIEVRE_VINCENCIORIXE        | BIEVRE, François                          | 1770       | Tragédie               | histoire française                | 1 act.        | prose   | 1761-1770 |
| 3.1.23    | BOISSY_VIEUSTUNSONGE        | BOISSY, Louis de                          | 1732       | Comédie                | moeurs polonaises                 | 5 act.        | vers    | 1731-1740 |
| 3.1.24    | DANCOURT_SANCHO            | DANCOURT, Florent                         | 1712       | Comédie                | moeurs françaises                 | 5 act.        | vers    | 1711-1720 |
| 3.1.25    | DIDIEROT_FILNATUREL         | DIDIEROT, Denis                           | 1757       | Comédie                | moeurs françaises                 | 5 prose      | vers    | 1751-1760 |
| 3.1.26    | DUFRESNY_MARIEGEFAIT-ETROMPU| DUFRESNY, Charles                         | 1721       | Comédie                | moeurs françaises                 | 3 act.        | vers    | 1721-1730 |
| 3.1.27    | GUIDIN_LOTHAIRE            | GUIDIN de la BRENNELERIE, Paul-Philippe   | 1759       | Tragédie               | histoire française                | 5 act.        | vers    | 1751-1760 |
| 3.1.28    | MARNAUX_PEREPRUDENT         | MARNAUX, Pierre de                        | 1712       | Comédie                | moeurs françaises                 | 1 act.        | vers    | 1711-1720 |
| 3.1.29    | REGNARD_MENHECHES          | REGNARD, Jean-Français                    | 1705       | Comédie                | moeurs françaises                 | 5 act.        | vers    | 1701-1710 |
| 3.1.30    | VOLTAIRE_MORTDECESAR        | VOLTAIRE                                  | 1736       | Tragédie               | histoire romaine                 | 3 act.        | vers    | 1731-1740 |
| 3.1.31    | ALLAS_BONNOURGE            | ALLAS, Alphonse                           | 1889       | Féeerie                | moeurs françaises                 | 1 act.        | prose   | 1881-1890 |
| 3.1.32    | AUDE_ECOLETRAGIQUE         | AUDE, Joseph                               | 1802       | Comédie                | moeurs françaises                 | 1 act.        | mixte   | 1801-1810 |
| 3.1.33    | BANVILLE_ANCENPIERROET     | BANVILLE, Théodore                         | 1877       | Monologue              | moeurs français                   | 1 act.        | prose   | 1871-1880 |
| 3.1.34    | BONNETAIN_APRESLE-DIVORCE  | BONNETAIN, Paul                           | 1890       | Comédie                | moeurs français                   | 1 act.        | prose   | 1881-1890 |
| 3.1.35    |CONSTANT_WALLSTEIN          | CONSTANT, Ben-                            | 1809       | Tragédie               | histoire allemande                | 5 act.        | vers    | 1801-1810 |
| 3.1.36    | DUMAS_DONUAN               | DUMAS, Alexandre                          | 1836       | Drame                  | moeurs espagnoles                 | 5 act.        | mixte   | 1831-1840 |
| 3.1.37    | GENLIS_BELLETLAETE         | GENLIS, Stéphanie-Félicité Du Crest de    | 1829       | Comédie                | moeurs françaises                 | 2 prose      |        | 1821-1830 |
| 3.1.38    | HUGO_HERNAIn              | HUGO, Victor                              | 1830       | Drame                  | histoire espagnole                | 5 act.        | vers    | 1821-1830 |
| 3.1.39    | JARRY_UBURIOI             | JARRY, Alfred                             | 1896       | Drame                  | Fantaisie historique              | 5 mixte       | prose   | 1881-1890 |
| 3.1.40    | SAND_MOLIERE                | SAND, George                              | 1851       | Comédie                | histoire littéraire               | 2 act.        | prose   | 1851-1860 |
| 3.1.41    | ANOUILLI_MEDIE             | ANOUILLI, Lae                             | 1946       | Tragédie               | mythe grec                        | NA            | prose   | 1941-1958 |
# 3.2 Varia

| ID | auteur | titre | genre | forme | date | source | ed. |
|----|--------|-------|-------|-------|------|--------|-----|
| ANONYME_PAIX_BERGERAC | Chancelière royale | Paix de Bergenc. Edit. de Poitiers. | légal | prose | 1577 | ELEC | éd. Bernand Barbeque, École nationale des chartes. |
| ANONYME_RETRIEVES | anonyme | Le vray discours sur la route et admirable descripture des Restires. | discours | prose | 1587 | WS | éd. Éd. Foumer, Var. hist. et litt., t. 9 |
| ARBEAU_BELLE_FRANCOIS-PREMIER_CORREISP | Thénot Arbeau | Belle qui tiens ma vie | chanson | vers | 1588 | WS | maq. |
| LENONCOURT_MARTIN_HYPNEROTOMACHINE | Robert de Lenoncort, Jean Martin (traits) | La correspondance du chancelier Antoine Du Bourg (1555-1558) | chanson | vers | 1590 | WS | maq. |
| LIEBAUT_MISIERS | Liebaut, Nicolas Estienne | Les Misères de la femme mariée. . . | poésie morale | vers | 1593 | WS | maq. |
| LOUSSEL&ELEGIE_MARGUERITIONAVAR_REMIRE | Louise Labé | Elégie | poésie lyrique | vers | 1555 | WS | maq. |
| RABELAIS_PANTAGRUEL | Ronsard, Pierre de | Pantagruel | roman | vers | 1532 | WS | maq. |
| ANNERHOAN_PORTRAIT | Anne de Rohan-Soubise | Portrait de feue la duchesse de Nevres . . . | chanson | vers | 1620 | GALL | Ed. de Barthélemy, Paris, 1862 |
| ANONYME_ARAIBOIRE | anonyme | Air à boire | chanson | vers | 16 . . | WS | maq. |
| ANONYME_ARREST | anonyme | Arrest notable donné au profit des femmes contre l'impuissance des maris | légal | prose | 1626 | WS | maq. |
| BOILEAU_SATIRE | Nicolas Boileau | Satires (satire I) | satire | vers | 1660 | WS | maq. |
| ELISabethBOHLINE_CORRESPONDANCE | Élisabeth de Bohlin | Correspondance avec René Descartes | correspondance | vers | 1643 | WS | maq. |
| CHAPERON | Charles Perrault | Le petit Chaperon rouge | conte | vers | 1697 | WS | maq. |
| POULLAIN_EGALITE | François Poullain de La Barre | De l'Égalité des deux sexes. Discours physique et moral | discours | vers | 1679 | GALL | maq. |
| SCUDERY_ARTAMENE | Madeleine de Scudéry | Artamène ou le Grand Cyrus | roman | vers | 1654 | WS | maq. |
| VALDARE_VIEIL-JOUR | Bréval de Verville | Un jour reconnaissant que je suis incapable | poésie | vers | 1626 | WS | maq. |
| ANONYME_DECLARATION | anonyme | Déclaration des Décès de l'Homme et du Cloyos de 1793 | légal | prose | 1795 | WS | maq. |
| CHARriere_LETTRE | Isabelle de Charrière | lettre | correspondance | prose | 1755 | other | maq. |
| CHATELET_DISCOURS | Émilie du Châtelet | Discours sur le bonheur | poésie | vers | 1734 | WS | maq. |
| CHERIEN_ODE | André Chénier | À Die (à Charlotte Corday) | poésie | vers | 1793 | WS | maq. |
| FREDEDIQUE_DIVINE_EMLIE | Frédéric de Puissie | À la Divine Émilie | poésie | vers | 1737 | GALL | maq. |
| MONTAGET_LSEME | Jeanne de Montgallet-Seigné | lisse, élégie | poésie | vers | 1739 | other | maq. |
| ROUGET_MARSEILLES | Rouget de Lisle | La Marseillaise | chanson | vers | 1792 | WS | maq. |
| TENCIN_SIEGE | Madame de Tencin | Le Siège de Calais | roman | vers | 1739 | WS | maq. |
| VOLTAIRE_ZADIG | Voltaire | Le Rêve de Corneille, Comédie dramatique | comédie | disparue | 1747 | WS | maq. |

**CMR PV**

Commission des monuments historiques: PV 29 février 1884 administré

**DEOLEEUZE_LETTRE**

Henri Deleusche: Lettre à Charles Deleusche (Carrets de prison)

**DEROULEDE_CHANTS**

Paul Desro đèe: Nouveaux chants du soldat

**DESBOISSELVALMORE_INSIMIE**

Marcelle Desboisset-Valmore: L'Inso

**DERIED_FEPEE**

Hérédia, José Maria de: Légifère

**MERIMEE_ACADEMIE**

Prosper Mérimée: Discours de réception à l'Académie française

**POTTIER_INTERNATIONALE**

Eugénie Pottier: L'Internationale

**SAND_MARIANNE**

George Sand: Mariène

**SEGUR_BROSSOU**

Flavien de Ségur: François le Bossu

**STAE_ALLEMANNE**

Mme de Stael: De l'Emigré

**ANONYME_CONSORTS**

Blom: Lettres à un homme

**BRASILLACH_JEANNE**

Robert Brasillach: Le Proce de Jeanne d'Arc

**BRIMONT_MIRAGE**

Renée de Brimont: Mirages

**COLETTE_MAISON**

Colette: La Maison de Claudine

**LAHIRE_HOUE**

Jean de La Hire: La Rose fulgurante

**ANONYME_ARREST**

Promotion des morts

**BOILEAU_SATIRE**

Poésie morale et admirable desconfiture des morts

**CHATELIER_DISCOURS**

Discours sur la route et admirable descripture des Restires.

**CHERRIERE_LETTRE**

Isabelle de Charrière: lettre

**CHIENNE_ODE**

André Chénier: À Die (à Charlotte Corday)

**FREDERIC_DVINE_EMILIE**

Frédéric de Puissie: À la Divine Émilie

**MONTAGET_LSEME**

Jeanne de Montgallet-Seigné: lisse, élégie

**ROUGET_MARSEILLES**

Rouget de Lisle: La Marseillaise

**TENCIN_SIEGE**

Madame de Tencin: Le Siège de Calais

**VOLTAIRE_ZADIG**

Voltaire: Le Rêve de Corneille, Comédie dramatique
| LASTEYRIE TESTAMENT | Testament | légal | prose | 1915 | other | Testaments de Poilus |
|---------------------|-----------|-------|-------|------|-------|---------------------|
| Gaspard Louis Guy de Lasteyrie marquis du Saillant | Testament légal | prose | 1915 | other | Testaments de Poilus |
| LONDON_BAGNE | Albert Londres | Au bagne | reportage | prose | 1924 | WS | Paris, 1924 |
| NOAILLES_COEUR | Anna de Noailles | Le Coeur innombrable | poésie | vers | 1901 | WS | Paris, 1901 |
| RACHILDE_GRENOUTILLES | Rachilde | Le Tueur de Grenouilles | nouvelle | prose | 1900 | WS | Mercure de France |