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

Jean-Baptiste Camps, Simon Gabay, Paul Fièvre, Thibault Clérice, Florian Cafiero

Journal of Data Mining and Digital Humanities

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

**Jean-Baptiste Camps**[1], **Simon Gabay**[2], **Paul Fièvre**[3], **Thibault Clérice**[1], **Florian Cafiero**[4]

[1]Centre Jean-Mabillon, École nationale des chartes, Université Paris, Sciences & Lettres
[2]Université de Neuchâtel
[3]Bibliothèque Nationale de France
[4]GEMASS, 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 Le*fff* [Sagot, 2010], *Morphalou* [Romary et al., 2004] and *LGeRM* [Souvay and Pierrel, 2009] for the lemmas. Unfortunately, training data are not publicly available (yet?), and rely mainly on non-normalised texts from the 16th to the 18th c.

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.

## 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 collection. 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 French.

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: 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 diachronicaly spread data, especially earlier period 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., 2020], 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. For the interoperability reason, we have therefore decided to choose CATTEX, currently used for the *Base de français médiéval* ("Medieval French Database", cf. Guillot-Barbance et al. [2017]).

The annotation manual of CATTEX09 (Guillot et al. [2013]) 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 Le*fff* and *Morphalou*. The main interest of the last one, that we have chosen, is that not only the v3.1 includes the Le*fff*, 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 as 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* (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.) vs. *son2* (*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).

The complete XML and annotation workflow is presented in fig. 1.

### 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 French[1]. 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* [Clérice 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,

[1] A recent version of the model for Old French can be found as part of the web application *Deucalion* [Clérice et al., 2019]; they are also directly usable through *Pyrrha*'s interface [Clérice 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 [Clérice, 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`.

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 Lorraine, 2016]. To do so, CATTEX POS-tags were mapped to *Morphalou*'s categories (table 1)[2]. 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 |
| CONcoo | 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 | **Genre** | |
| subjunctive | sub | simplePast | psp | masculine | m |
| infinitive | - | **Pers.** | | feminine | f |
| past | - | firstPerson | 1 | neuter | n |
| participle | - | secondPerson | 2 | *Varia* | |
| | | thirdPerson | 3 | - | - |
| | | | | invariable | x |
| | | | | 1036442 | ERROR |

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

[2]Punctuation and proper names were dealt with separately and specifically, as well as CATTEX combined label, `PRE.DETdef`.

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* 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*, *captif*, 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 *œ* (*cœur*), which has been preferred over its decomposed form (*coeur*).

## IV TRAINING SETUP

As previously mentioned, many tools are already available. *TreeTagger* [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* (only for POS, cf. 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[3];
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.

[3]For this configuration, we used a config file provided to us by Enrique Manjavacas, main developer of *Pie*.

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.

### 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 configuration 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. 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* | 98.8 | **99.09** | 98.95 | 98.8 | *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** | **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 in diachrony, in diaphasy and in diageny, 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 token 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).

LEMMA

| *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

POS

| *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 Most frequent confusions

The most frequent confusions of the best models on the out-of-domain data are presented in table 8.

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

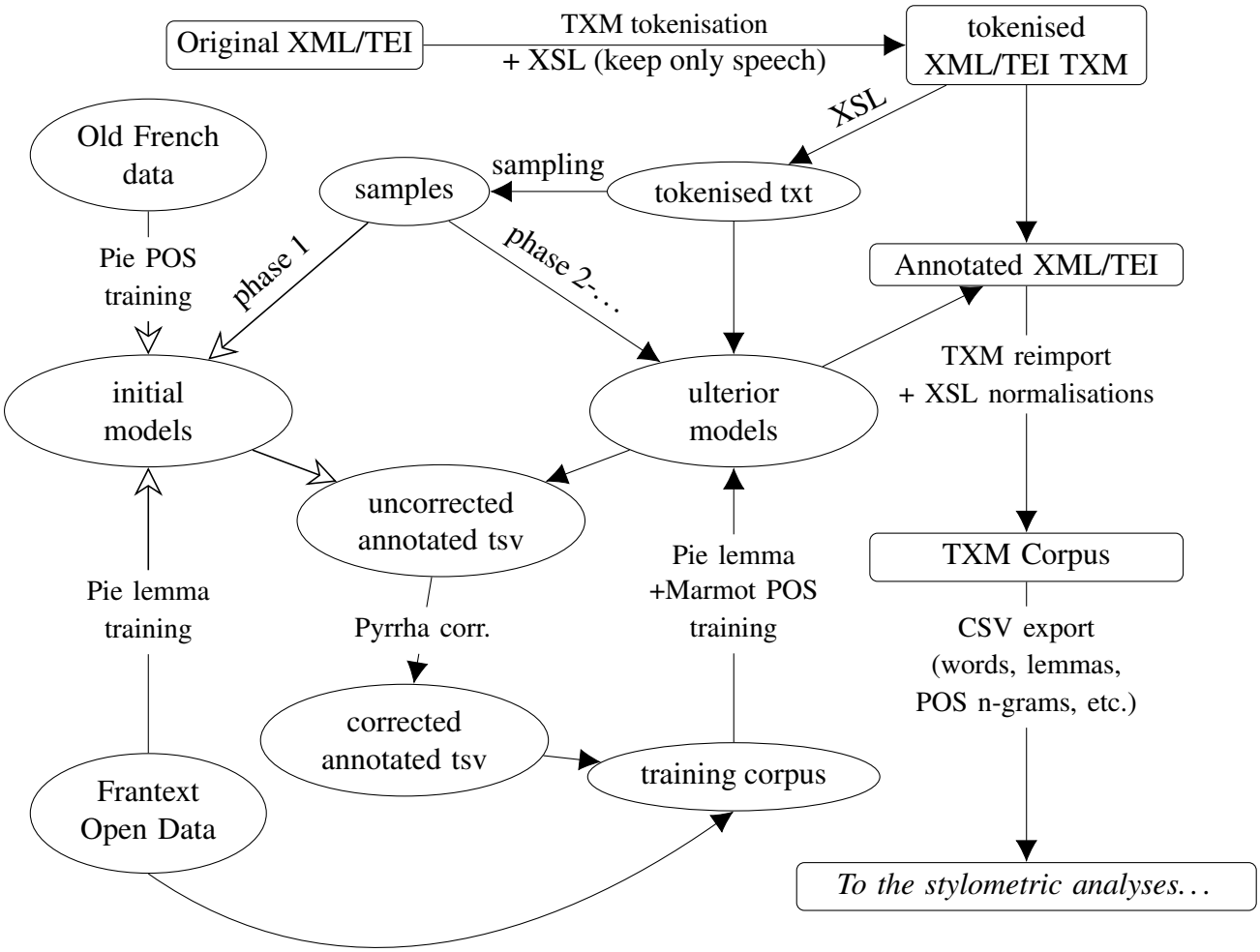


Figure 1 – Workflow for building the training corpus and models (ellipses) and the annotated XML corpus used for stylometric analysis (rectangles); void-arrowheads specify steps particular to the initial model building phase

| Expected | Tot. Err. | Pred. | Pred. times | Expected | Tot. Err. | Pred. | Pred. times |
|---|---|---|---|---|---|---|---|
| LEMMA | | | | POS | | | |
| monsieur | 9 | M. | 9 | NOMcom | 182 | ADJqua | 45 |
| le | 9 | il | 9 | | | NOMpro | 33 |
| cœur | 6 | cuur | 4 | | | VERinf | 32 |
| | | crur | 2 | | | VERcjg | 18 |
| franc | 6 | Franc | 6 | | | VERppe | 11 |
| Électre | 5 | éLECTRE | 4 | ADJqua | 104 | NOMcom | 45 |
| | | électre | 1 | | | VERppe | 17 |
| noble | 4 | Nobles | 4 | | | VERcjg | 12 |
| maître | 4 | maîtresse | 4 | NOMpro | 81 | NOMcom | 43 |
| de_le | 4 | un | 4 | VERcjg | 61 | NOMcom | 18 |
| un | 4 | de_le | 4 | | | ADJqua | 12 |
| dame | 3 | Dame | 2 | | | VERppe | 12 |
| | | damer | 1 | ADVgen | 58 | NOMcom | 16 |
| Phanor | 3 | phanor | 3 | | | VERcjg | 14 |
| voir | 3 | vivre | 3 | PROrel | 31 | CONsub | 26 |
| Vosges | 3 | vosge | 3 | | | | |
| Médée | 3 | médé | 3 | | | | |

Table 8 – Sample from the confusion matrix for the best *Pie* models on the out-of-domain data.

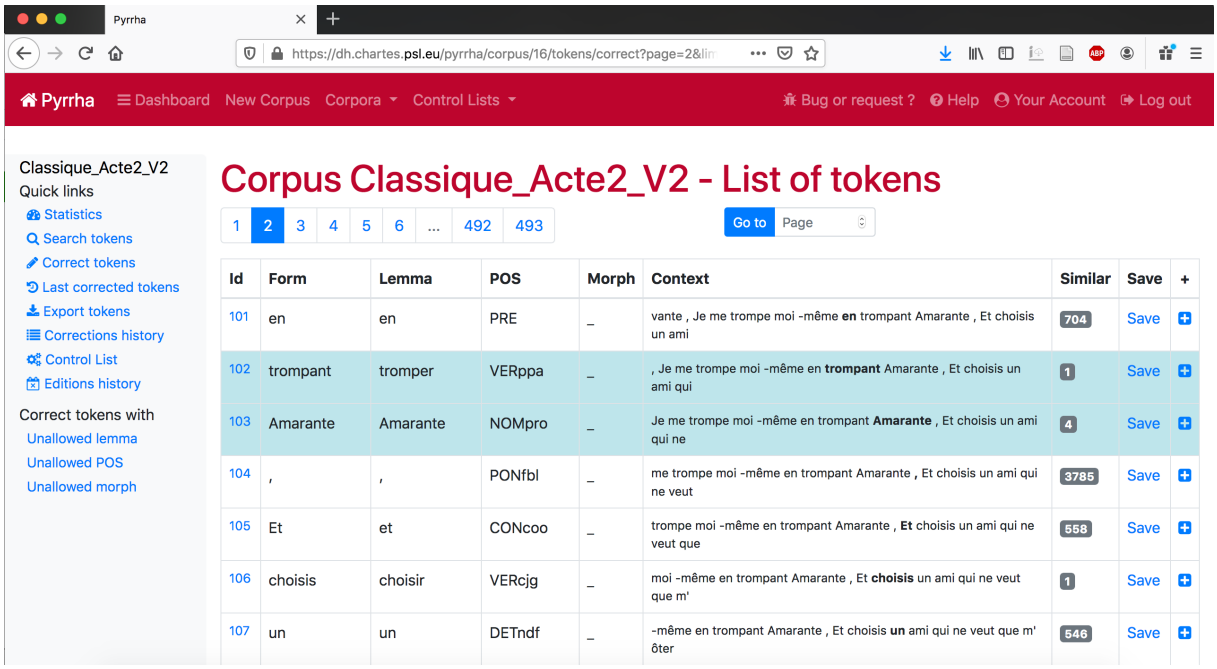


Figure 2 – Screenshot of the *Pyrrha* interface: main correction view

due to the morpho-syntactic nature of our annotation, which, for instance, labels substantive adjective as common nouns (*le **beau*** is `NOMcom`).

## VI 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. 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.

## 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.

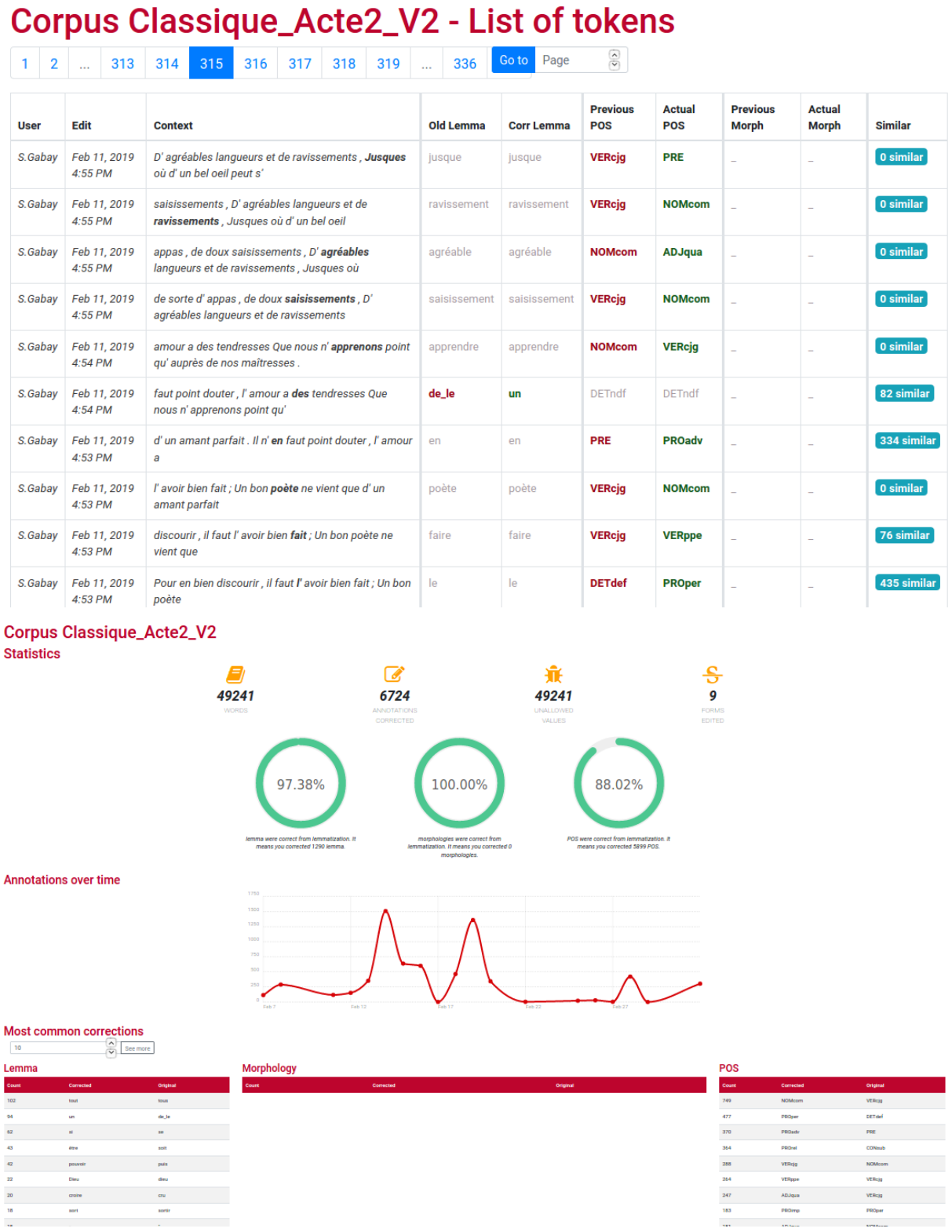


Figure 3 – Screenshot of the *Pyrrha* interface: correction history and corpus statistics

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`.

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).

## 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.

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.

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.

Sascha Diwersy, 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.

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

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

Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. Principes d'annotation cattex09. Technical report, École normale supérieure de Lyon, Lyon, 2013. 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.

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.

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.

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/.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, 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.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1272. URL https://www.aclweb.org/anthology/D15-1272.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. A universal part-of-speech tagset. *CoRR*, abs/1104.2086, 2011. URL http://arxiv.org/abs/1104.2086.

Sophie Prévost, Céline Guillot, Alexei Lavrentiev, and Serge Heiden. Jeu d'étiquettes morphosyntaxiques CATTEX2009. Technical report, École normale supérieure de Lyon, Lyon, 2013. version 2.0. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf.

Alain Riffaud. Répertoire du théâtre français imprimé au xviie siècle, 2014. URL https://repertoiretheatreimprime.yale.edu/.

Laurent Romary, Susanne Salmon-Alt, and Gil Francopoulo. Standards going concrete: from lmf to morphalou. In *The 20th International Conference on Computational Linguistics (COLING 2004) - ElectricDict '04 Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 22–28, 2004. URL https://hal.inria.fr/inria-00121489.

Benoît Sagot. External Lexical Information for Multilingual Part-of-Speech Tagging. Research Report RR-8924, Inria Paris, June 2016. URL https://hal.inria.fr/hal-01330301.

Benoît Sagot. The le*fff*, a freely available and large-coverage morphological and syntactic lexicon for french. In *Proceedings of the 7th international conference on Language Resources and Evaluation*, 2010. URL https://hal.archives-ouvertes.fr/inria-00521242.

Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, page 47–50, 1995.

Christof Schöch. *Théâtre Classique*, paul fièvre (ed.), 2007-2018. *RIDE*, 8, 2018. URL https://ride.i-d-e.de/issues/issue-8/theatre-classique.

Gilles Souvay and Jean-Marie Pierrel. Lgerm lemmatisation des mots en moyen français. *Traitement Automatique des Langues*, 50(2):149–172, 2009. URL https://halshs.archives-ouvertes.fr/halshs-00396452.

Isabelle Tellier, Yoann Dupont, and Arnaud Courmet. Un segmenteur-'etiqueteur et un chunker pour le français (a segmenter-pos labeller and a chunker for french)[in french]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 5: Software Demonstrations*, pages 7–8, 2012.

Assaf Urieli. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université Toulouse le Mirail - Toulouse II, 2013. URL https://tel.archives-ouvertes.fr/tel-00979681/.

Denis Vigier and Peter Blumenthal. Presto - l'évolution du système prépositionnel du français, 2013-2017. URL http://presto.ens-lyon.fr.

# 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_DONSANCHE | CORNEILLE, Pierre | DON SANCHE D'ARAGON, COMÉDIE HÉROÏQUE | 1649 | Comédie héroïque | moeurs espagnoles | 5 act. | vers | 1641-1650 |
| CP_GALERIEDUPALAIS | CORNEILLE, Pierre | LA GALERIE DU PALAIS ou L'AMIE RIVALE | 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 | MÉLITE OU LES FAUSSES LETTRES, 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 |
| CP_PULCHERIE | CORNEILLE, Pierre | PULCHÉRIE, COMÉDIE HÉROÏQUE | 1673 | Comédie héroïque | histoire chrétienne | 5 act. | vers | 1671-1680 |
| CP_SUITEMENTEUR | CORNEILLE, Pierre | LA SUITE DU MENTEUR, COMÉDIE | 1645 | Comédie | moeurs françaises | 5 act. | vers | 1641-1650 |
| CP_SUIVANTE | CORNEILLE, Pierre | LA SUIVANTE, COMÉDIE | 1637 | Comédie | moeurs françaises | 5 act. | vers | 1631-1640 |
| CP_TITE | CORNEILLE, Pierre | TITE ET BÉRÉNICE, COMÉDIE HEROÏQUE | 1671 | Comédie héroïque | histoire romaine | 5 act. | vers | 1661-1670 |
| CP_VEUVE34 | CORNEILLE, Pierre | LA VEUVE OU LE TRAÎTRE TRAHI, COMÉDIE | 1634 | Comédie | moeurs françaises | 5 act. | vers | 1631-1640 |
| CT_AMOURALAMODE | CORNEILLE, Thomas | L'AMOUR À LA MODE, COMÉDIE. | 1651 | Comédie | moeurs espagnoles | 5 act. | vers | 1651-1660 |
| CT_CHARMEDELAVOIX | CORNEILLE, Thomas | LE CHARME DE LA VOIX, COMÉDIE | 1658 | Comédie | moeurs italiennes | 5 act. | vers | 1651-1660 |
| CT_COMTESSEORGUEIL | CORNEILLE, Thomas | LA COMTESSE D'ORGUEIL, COMÉDIE | 1690† | Comédie | moeurs françaises | 5 act. | vers | 1651-1660 |
| CT_DOMBERTRANDECIGARRAL | CORNEILLE, Thomas | DON BERTRAN DE CIGARRAL, COMÉDIE | 1709* | Comédie | moeurs espagnoles | 5 act. | vers | 1651-1660 |
| CT_DOMCESARDAVALOS | CORNEILLE, Thomas | DON CÉSAR D'AVALOS, COMÉDIE. | 1661 | Comédie | moeurs espagnoles | 5 act. | vers | 1671-1680 |
| CT_FEINTASTROLOGUE | CORNEILLE, Thomas | LE FEINT ASTROLOGUE, COMÉDIE | 1651 | Comédie | moeurs françaises | 5 act. | vers | 1651-1660 |
| CT_FESTINPIERRE | CORNEILLE, Thomas | LE FESTIN DE PIERRE, COMÉDIE | 1677 | Comédie | moeurs espagnoles | 5 act. | vers | 1671-1680 |
| CT_GALANTDOUBLE | CORNEILLE, Thomas | LE GALANT DOUBLÉ, COMÉDIE. | 1659 | Comédie | moeurs espagnoles | 5 act. | vers | 1651-1660 |
| CT_GEOLIERDESOISMEME | CORNEILLE, Thomas | LE GEÔLIER DE SOI-MÊME, COMÉDIE. | 1655 | Comédie | moeurs italiennes | 5 act. | vers | 1651-1660 |
| CT_ILLUSTRESENNEMIS | CORNEILLE, Thomas | LES ILLUSTRES ENNEMIS, COMÉDIE | 1657 | Comédie | moeurs espagnoles | 5 act. | vers | 1651-1660 |
| CT_INCONNU | CORNEILLE, Thomas | L'INCONNU, COMÉDIE. | 1675 | Comédie | moeurs françaises | 5 act. | vers | 1671-1680 |
| M_AMPHITRYON | MOLIÈRE | AMPHITRYON, COMÉDIE | 1668 | Comédie | mythe grec | 3 act., prol. | vers | 1661-1670 |
| M_DEPITAMOUREUX | MOLIÈRE | LE DÉPIT AMOUREUX | 1656 | Comédie | moeurs françaises | 5 act. | vers | 1651-1660 |
| M_DOMGARCIEDENAVARRE | MOLIERE | DON GARCIE DE NAVARRE, COMÉDIE | 1682‡ | Comédie | moeurs espagnoles | 5 act. | vers | 1661-1670 |
| M_ECOLEDESFEMMES | MOLIÈRE | L'ÉCOLE DES FEMMES, COMÉDIE. | 1663 | Comédie | moeurs françaises | 5 act. | vers | 1661-1670 |
| M_ETOURDI | MOLIÈRE | L'ÉTOURDI ou LES CONTRE-TEMPS, COMÉDIE | 1663 | Comédie | moeurs françaises | 5 act. | vers | 1661-1670 |
| M_FEMMESSAVANTES | MOLIÈRE | LES FEMMES SAVANTES, COMÉDIE | 1672 | Comédie | moeurs françaises | 5 act. | vers | 1671-1680 |
| M_MISANTHROPE | MOLIÈRE | LE MISANTHROPE ou L'ATRABILAIRE AMOUREUX, COMÉDIE | 1667 | Comédie | moeurs françaises | 5 act. | vers | 1661-1670 |
| M_TARTUFFE | MOLIÈRE | LE TARTUFFE ou L'IMPOSTEUR, COMÉDIE | 1669 | Comédie | moeurs françaises | 5 act. | vers | 1661-1670 |
| O_ABSENTCHEZSOI | OUVILLE, Antoine le Métel | L'ABSENT CHEZ SOI | 1643 | Comédie | moeurs françaises | 5 act. | vers | 1641-1650 |
| O_FAUSSESVERITES | OUVILLE, Antoine le Métel | LES FAUSSES VÉRITÉS, COMÉDIE | 1643 | Comédie | moeurs françaises | 5 act. | vers | 1641-1650 |
| O_SOUPCONS | OUVILLE, Antoine le Métel | LES SOUPÇONS SUR LES APPARENCES, COMÉDIE | 1650 | héroïco-comédie | moeurs françaises | 5 act. | vers | 1641-1650 |
| R_BELLEALPHREDE | ROTROU, Jean | LA BELLE ALPHRÈDE, COMÉDIE | 1639 | Comédie | moeurs arabes | 5 act. | vers | 1631-1640 |
| R_CAPTIFS | ROTROU, Jean | LES CAPTIFS OU LES ESCLAVES, COMÉDIE | 1640 | Comédie | moeurs françaises | 5 act. | vers | 1631-1640 |
| R_SOSIES | ROTROU, Jean | LES SOSIES, COMÉDIE | 1638 | Comédie | mythe grec | 5 act. | vers | 1631-1640 |
| S_DOMJAPHETDARMENIE | SCARRON, Paul | DON JAPHET D'ARMÉNIE, COMÉDIE. | 1653 | Comédie | moeurs françaises | 5 act. | vers | 1651-1660 |
| S_GARDIENDESOIMEME | SCARRON, Paul | LE GARDIEN DE SOI-MÊME | 1654 | Comédie | moeurs italiennes | 5 act. | vers | 1651-1660 |
| S_HERITIERRIDICULE | SCARRON, Paul | L'HÉRITIER RIDICULE OU LA DAME INTÉRESSÉE.. | 1650 | Comédie | moeurs espagnoles | 5 act. | vers | 1641-1650 |
| S_JODELET | SCARRON, Paul | LE JODELET OU LE MAÎTRE VALET, COMÉDIE. | 1648 | Comédie | moeurs françaises | 5 act. | vers | 1641-1650 |
| S_JODELETDUELISTE | SCARRON, Paul | LE JODELET DUELLISTE, COMÉDIE. | 1646 | Comédie | moeurs françaises | 5 act. | vers | 1641-1650 |
| S_MARQUISRIDICULE | SCARRON, Paul | LE MARQUIS RIDICULE, OU LA COMTESSE faite à la hâte. COMÉDIE. | 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 |
|---|---|---|---|---|---|---|
| K639:1890:52916 | LOTI, Pierre | Le Roman d'un enfant | 1891 | Paris | Calmann-Levy | 61,417 |
| K934:1898:35797 | ROSTAND Edmond | Cyrano de Bergerac | 1898 | Paris | Fasquelle | 47,213 |
| K999:1899:45606 | GOURMONT Remy de | Esthétique de la langue française… | 1899 | NA | Soc. Mercure de Fr. | 53,640 |
| L233:1895:13272 | JARRY Alfred | Ubu Roi, *Œuvres complètes*, t. 4. | s.d. | Monte-Carlo | éd. Du Livre | 17,672 |
| L266:1905:59455 | POINCARÉ Henri | La Valeur de la science | 1905 | Paris | Flammarion | 66,125 |
| L433:1882:31745 | BECQUE Henry | Les Corbeaux, *Théâtre complet*, t. 2 | 1922 | s.p. | Fasquelle | 38.813 |
| L486:1884:68155 | HUYSMANS Joris-Karl | À rebours | 1907 | Paris | Fasquelle | 79,058 |
| L499:1901:63336 | JAURÈS Jean | Études socialistes | 1902 | Paris | Ollendorf | 71,159 |
| L784:1908:100993 | LEROUX Gaston | Le Parfum de la dame en noir | 1908 | Paris | L'Illustration | N.tok |
| L846:1857:68636 | ABOUT Edmond | Le Roi des montagnes | 1857 | s.p. | Hachette | 79,720 |
| L884:1867:20871 | MEILHAC Henri et Ludovic HALÉVY | La Vie parisienne | 1867 | Paris | M. Levy | 27,381 |
| M223:1873:70999 | VERNE Jules | Le Tour du monde en quatre-vingts jours | s.d. | Paris | Hetzel | 84,925 |
| M289:1874:69484 | FROMENTIN Eugène | Un été dans le Sahara | 1877 | Paris | Plon | 80,612 |
| M362:1825:111416 | BRILLAT-SAVARIN Jean-Anthelme | Physiologie du goût… | 1847 | Paris | Charpentier | 127,968 |
| M374:1830:28471 | FOURIER Charles | Le Nouveau monde industriel… | 1830 | Paris | Bossange | 32,566 |
| M425:1794:4621 | CHÉNIER André | L'Invention, *Œuvres complètes*, t. 2. | 1910 | Paris | Delagrave | 5,473 |
| M433:1807:57291 | STAËL Germaine de | Corinne ou l'Italie : t. 1 | 1807 | s.p. | Peltier | 65,012 |
| M464:1801:72408 | DESTUTT DE TRACY | Élémens d'idéologie, 1 | 1804 | Paris | Courcier | 80,433 |
| M468:1803:62746 | KRÜDENER Barbara Juliane von | Valérie | 1840 | Paris | Charpentier | 73,463 |
| M473:1809:82847 | LAMARCK Jean-Baptiste | Philosophie zoologique : t. 1 | 1809 | s.p. | Dentu | 95,111 |
| M492:1805:123259 | CUVIER Georges | Leçons d'anatomie comparée : t. 1 | 1805 | Paris | Baudouin | 138,674 |
| M528:1798:12861 | GUILBERT DE PIXERÉCOURT | Victor ou l'Enfant de la forêt | 1798 | Paris | Barba | 16,878 |
| M548:1832:199497 | SAY Jean-Baptiste | Traité d'économie politique | 1841 | Paris | Guillaumin | 223,038 |
| M622:1789:24158 | SIEYÈS Emmanuel | Qu'est-ce que le Tiers état ? | 1888 | Paris | Soc. Hist. Rév. Fr. | 27,002 |
| M629:1792:31251 | FLORIAN Jean-Pierre | Fables | 1792 | Paris | Didot | 37,172 |
| M893:1869:61374 | GONCOURT Edmond et Jules de | Madame Gervaisais | 1876 | s.p. | Charpentier | 71,174 |
| M939:1852:84555 | COMTE Auguste | Catéchisme positiviste… | 1909 | Paris | Garnier | 95,589 |
| N245:1838:32486 | HUGO Victor | Ruy Blas, *Œuvres complètes. Théâtre*, 3 | 1905 | Paris | Ollendorff | 41,487 |
| N268:1802:63535 | BONALD Louis de | Législation primitive… t. 1 | 1802 | Paris | Le Clère | 72,539 |
| N429:1778:64388 | BUFFON Georges-Louis de | Des époques de la nature, *Hist. Natur.*, t. 5. | 1778 | Paris | Impr. Royale | 70,866 |
| P556:1872:177870 | VIOLLET-LE-DUC Eugène | Entretiens sur l'architecture : t. 2 | 1872 | Paris | A. Morel | 199,695 |
| Q454:1784:53428 | RÉTIF DE LA BRETONNE Nicolas | La Paysanne pervertie…t. 1 | 1784 | s.p. | Vve 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) par Jean Donneau de Visé*, éd. Claude Bourqui et 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 |
|---|---|---|---|---|---|---|---|---|
| ANONYME_PARDONNEUR | anonyme | FARCE NOUVELLE TRÈS BONNE ET FORT JOYEUSE À TROIS PERSONNAGES, FARCE | 1500 c. | Farce | Sermon joyeux | 1 act. | vers | 1501-1600 |
| ANONYME_PONTAUXANES | anonyme | FARCE NOUVELLE FORT JOYEUSE DU PONT AUX ÂNES | 1500 c. | Farce | Sermon joyeux | 1 act. | vers | 1501-1600 |
| ANONYME_RESURRECTION-JENINLANDORE | anonyme | FARCE NOUVELLE TRÈS BONNE ET FORT JOYEUSE DE LA RESURRECTION DE JENIN LANDORE, FARCE | 1500 c. | Farce | Sermon joyeux | 1 act. | vers | 1501-1600 |
| ANONYME_SERMONJOYEUX | anonyme | SERMON JOYEUX DE BIEN BOIRE, FARCE | 1500 c. | Farce | Sermon joyeux | 1 act. | vers | 1501-1600 |
| BEZE_ABRAHAM | BEZE, Théodore | ABRAHAM SACRIFIANT, TRAGÉDIE. | 1550 | Tragédie | bible | 1 act. | vers | 1541-1550 |
| GRINGORE_SAINT-LOUIS | GRINGORE, Pierre | Mystère de Saint Louis | 1514 | mystère | histoire médiévale | 8 livres | vers | 1511-1520 |
| JODELLE_CLEOPATRE | JODELLE, Étienne | CLÉOPÂTRE CAPTIVE, TRAGÉDIE. | 1574 | Tragédie | histoire romaine | 5 act. | vers | 1571-1580 |
| JODELLE_DIDON | JODELLE, Étienne | DIDON SE SACRIFIANT, TRAGÉDIE. | 1574 | Tragédie | mythe grec | 5 act. | vers | 1571-1580 |
| JODELLE_EUGENE | JODELLE, Étienne | L'EUGÈNE, COMÉDIE. | 1574 | Comédie | moeurs françaises | 5 act. | vers | 1571-1580 |
| TURNEBE_CONTENTS | TURNÈBE, Odet de | LES CONTENTS, COMÉDIE NOUVELLE EN PROSE FRANÇAISE | 1584 | Comédie | moeurs françaises | 5 act. | prose | 1581-1590 |
| AUBIGNAC_PUCELLE | AUBIGNAC, François Hédelin | La Pucelle d'Orléans | 1642 | Tragédie | histoire médiévale | | prose | 1641-1650 |
| CHAMPREPUS_ULYSSE | CHAMPREPUS, Jacques de | ULYSSE, TRAGÉDIE. | 1603 | Tragédie | mythe grec | 5 act. | vers | 1601-1610 |
| CHAPPUZEAU_ARMETZAR | Samuel Chappuzeau (1625-1701) | ARMETZAR OU LES AMIS ENNEMIS, TRAGICOMÉDIE. | 1656 | Tragi-comédie | histoire turque | 5 act. | vers | 1651-1660 |
| DESHOULIERES_GENSERIC | DESHOULIÈRES, Antoinette du Ligier de la Garde | GENSERIC, TRAGÉDIE | 1680 | Tragédie | histoire romaine | 5 act. | vers | 1671-1680 |
| DESJARDINS_FAVORI | DESJARDINS, Marie-Catherine-Hortense dite de Villedieu | LE FAVORI, TRAGICOMÉDIE. | 1665 | Tragi-comédie | moeurs espagnoles | 5 act. | vers | 1661-1670 |
| DURANT_OISIVETE | DURANT, Catherine | OISIVITÉ EST MÈRE DE TOUS LES VICES, PROVERBE. | 1699 | Proverbe | moeurs françaises | 1 act. | prose | 1691-1700 |
| MATHIEU_MAGICIENNE-ETRANGERE | MATTHIEU, Pierre | LA MAGICIENNE ÉTRANGÈRE, TRAGÉDIE. | 1617 | Tragédie | histoire française | 4 act. | vers | 1611-1620 |
| SCUDERY_LIGDAMON-ELIDIAS | SCUDERY, Georges de | LIGDAMON ET LIDIAS, TRAGICOMÉDIE | 1631 | Tragi-comédie | histoire médiévale | 5 act. | vers | 1631-1640 |
| URFE_SYLVANIRE | URFÉ, Honoré d' | SYLVANIRE ou la MORTE VIVE, FABLE BOCAGÈRE | 1627 | Pastorale héroïque | pastorale | 5 act., un prologue | vers | 1621-1630 |
| VONDREBECK-ALARD_FORCESDELAMOUR | VONDREBECK, Maurice et ALARD, Charles | LES FORCES DE L'AMOUR ET DE LA MAGIE, DIVERTISSEMENT | 1678 | Divertissement | moeurs françaises | 3 act. | prose | 1671-1680 |
| BERGASSE_JOURNEE-DESDUPES_1790 | BENSERADE, Isaac de | LA JOURNÉE DES DUPES, TRAGI-COMÉDIE | 1790 | Tragi-comédie | histoire française | 4 act. | prose | 1781-1790 |
| BIEVRE_VERCINGENTORIXE | BIÈVRE, François-Georges Mareschal de | VERCINGENTORIXE, TRAGÉDIE. | 1770 | Tragédie | histoire française | 1 act. | vers | 1761-1770 |
| BOISSY_VIEESTUNSONGE | BOISSY, Louis de | LA VIE EST UN SONGE, COMÉDIE-HÉROÏQUE. | 1732 | Comédie héroïque | moeurs polonaises | 5 act. | vers | 1731-1740 |
| DANCOURT_SANCHO | DANCOURT, Florent CARTON dit | SANCHO PANÇA, GOUVERNEUR, COMÉDIE | 1712 | Comédie | moeurs françaises | 5 act. | vers | 1711-1720 |
| DIDEROT_FILSNATUREL | DIDEROT, Denis | LE FILS NATUREL ou Les ÉPREUVES DE LA VERTU. | 1757 | Tragi-comédie | moeurs françaises | 5 act. | prose | 1751-1760 |
| DUFRESNY_MARIAGEFAIT-ETROMPU | DUFRESNY, Charles | LE MARIAGE FAIT ET ROMPU, COMÉDIE | 1721 | Comédie | moeurs françaises | 3 act. | vers | 1721-1730 |
| GUDIN_LOTHAIRE | GUDIN de la BRENELLERIE, Paul-Philippe | LOTHAIRE, ROI DE LORRAINE, TRAGÉDIE | 1759 | Tragédie | histoire française | 5 act. | vers | 1751-1760 |
| MARIVAUX_PEREPRUDENT | MARIVAUX, Pierre de | LE PÈRE PRUDENT ET ÉQUITABLE, COMÉDIE | 1712 | Comédie | moeurs françaises | 1 act. | vers | 1711-1720 |
| REGNARD_MENECHMES | REGNARD, Jean-François | LES MÉNECHMES, ou LES JUMEAUX, COMÉDIE | 1705 | Comédie | moeurs françaises | 5 act. | vers | 1701-1710 |
| VOLTAIRE_MORTDECESAR | VOLTAIRE | LA MORT DE CÉSAR, TRAGÉDIE EN TROIS ACTES | 1736 | Tragédie | histoire romaine | 3 act. | vers | 1731-1740 |
| ALLAIS_BONBOUGRE | ALLAIS, Alphonse | LE PAUVRE BOUGRE ET LE BON GÉNIE, FÉÉRIE EN UN ACTE. | 1889 | Féérie | moeurs françaises | 1 act. | prose | 1881-1890 |
| AUDE_ECOLETRAGIQUE | AUDE, Joseph | L'ÉCOLE TRAGIQUE, OU CADET ROUSSEL MAITRE DE DÉCLAMATION COMÉDIE | 1802 | Comédie | moeurs françaises | 1 act. | mixte | 1801-1810 |
| BANVILLE_ANCIENPIERROT | BANVILLE, Théodore de | ANCIEN PIERROT, MONOLOGUE. | 1877 | Monologue | moeurs françaises | 1 act. | prose | 1871-1880 |
| BONNETAIN_APRESLE-DIVORCE | BONNETAIN, Paul | APRÈS LE DIVORCE, PIÈCE | 1890 | Comédie | moeurs françaises | 1 act. | prose | 1881-1890 |
| CONSTANT_WALLSTEIN | CONSTANT, Benjamin | WALLSTEIN, TRAGÉDIE | 1809 | Tragédie | histoire allemande | 5 act. | vers | 1801-1810 |
| DUMAS_DONJUAN | DUMAS, Alexandre | DON JUAN DE MARANA, MYSTÈRE | 1836 | Drame | moeurs espagnoles | 5 act. | mixte | 1831-1840 |
| GENLIS_BELLEETLABETE | GENLIS, Stéphanie-Félicité Du Crest de | LA BELLE ET LA BÊTE, COMÉDIE | 1829 | Comédie | moeurs françaises | 2 act. | prose | 1821-1830 |
| HUGO_HERNANI | HUGO, Victor | HERNANI, OU L'HONNEUR CASTILLAN | 1830 | Drame | histoire espagnole | 5 act. | vers | 1821-1830 |
| JARRY_UBUROI | JARRY, Alfred | UBU ROI, DRAME | 1896 | Drame | Fantaisie historique | 5 act. | mixte | 1881-1890 |
| SAND_MOLIERE | SAND, George | MOLIÈRE, DRAME. | 1851 | Comédie | histoire littéraire | 2 act. | prose | 1851-1860 |
| ANOUILH_MEDEE | ANOUILH, Jean | Médée | 1946 | tragédie | mythe grec | NA | prose | 1941-1950 |
| BERNARDT_PARTIEDE-BRIDGE | BERNARD, Tristan | LA PARTIE DE BRIDGE, COMÉDIE | 1930 | Comédie | moeurs françaises | 1 act. | prose | 1921-1930 |
| BERTON_HOMMEQUIA-TUELAMORT | BERTON, René | L'HOMME QUI A TUÉ LA MORT, DRAME | 1928 | Pièce dramatique | moeurs françaises | 2 act. | prose | 1921-1930 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CESAIRE_CHRISTOPHE | CÉSAIRE, Aimé | La Tragédie du roi Christophe | 1963 | tragédie | histoire haïtienne | 3 act. | prose | 1961-1970 |
| COURTELINE_MONSIEUR-BADIN | COURTELINE, Georges | MONSIEUR BADIN, COMÉDIE. | 1904 | Saynète | moeurs françaises | 1 act. | prose | 1891-1900 |
| GIRAUDOUX_ELECTRE | GIRAUDOUX, Jean | Électre | 1937 | tragédie bourgeoise | mythe grec | 2 act. | prose | 1931-1940 |
| HUGUES_CENDRILLON | HUGUES, Clovis | CENDRILLON, COMÉDIE. | 1906 | Comédie | conte de fées | 1 act. | vers | 1901-1910 |
| HUGUES_TYL | HUGUES, Clovis | TYL L'ESPIEGLE, COMÉDIE. | 1906 | Comédie | moeurs flamande | 1 act. | vers | 1901-1910 |
| LESENNE_REVEIL-CORNEILLE | LE SENNE, Camille | LE RÉVEIL DE CORNEILLE, POÈME DRAMATIQUE. | 1916 | Dialogue des morts | histoire littéraire | 1 act. | vers | 1911-1920 |
| RENARD_MAITRESSE | RENARD, Jules | LA MAÎTRESSE, COMÉDIE. | 1927 | Comédie | moeurs françaises | 2 act. | prose | 1921-1930 |

## 3.2 Varia

| ID | auteur | titre | genre | forme | date | source | ed. |
|---|---|---|---|---|---|---|---|
| ANONYME_PAIX-BERGERAC | Chancellerie royale | Paix de Bergerac. Édit de Poitiers. Poitiers, septembre 1577. | légal | prose | 1577 | ELEC | éd. Bernard Barbiche, École nationale des chartes. |
| ANONYME_REITRES | anonyme | Le vray discours sur la route et admirable desconfiture des Reistres… | discours | prose | 1587 | WS | éd. Éd. Fournier, Var. hist. et litt., t. 9 |
| ARBEAU_BELLE | Thoinot Arbeau | Belle qui tiens ma vie | chanson | vers | 1588 | WS | manq. |
| FRANCOISPREMIER-_CORRESP | François Ier | La correspondance du chancelier Antoine Du Bourg (1535-1538) | correspondance | prose | 1537 | ELEC | éd. Olivier Poncet |
| LENONCOURT-MARTIN_HYPNEROTO-MACHIE | Robert de Lenoncourt, Jean Martin (trads) | Hypnérotomachie, ou Discours du songe de Poliphile… | discours | prose | 1546 | WS | Jacques Kerver, 1546 |
| LIEBAUT_MISERES | Liébaut,Nicole Estienne | Les Misères de la Femme mariée… | poésie morale | vers | 1587 | WS | éd. Éd. Fournier, Var. hist. et litt., 3, 1855. |
| LOUISELABE_ELEGIE | Louise Labé | Élégie I | poésie lyrique | vers | 1555 | WS | éd. Charles Boy, 1887. |
| MARGUERITENAVAR-RE_MIROIR | Marguerite de Navarre | Miroir de l'âme pécheresse | miroir | vers | 1529 | WS | éd. Félix Frank, 1873 |
| RABELAIS_PANTAGRUEL | Rabelais | Pantagruel | roman | prose | 1532 | WS | éd. Marty-Laveaux, 1868 |
| RONSARD_SONNETS | Ronsard, Pierre de | Je plante en ta faveur cet arbre de Cybèle | poésie | vers | 1578 | WS | éd. Roger Sorg,1921 |
| ANNEROHAN_PORTRAIT | Anne de Rohan-Soubise | Portrait de feue la duchesse de Nevers… | poésie | vers | 1629 | GALL | éd. Éd. de Barthélemy, Paris, 1862 |
| ANONYME_AIRABOIRE | anonyme | Air à boire | chanson | vers | 16… | WS | La Chanson française du XVe au XXe siècle, éd. Jean Gillequin, 1910. |
| ANONYME_ARREST | anonyme | Arrest notable donné au profit des femmes contre l'impuissance des maris | légal | prose | 1626 | WS | éd. Éd. Fournier, Var. hist. et litt., 6, 1856 |
| BOILEAU_SATIRE | Nicolas Boileau | satires (satire I) | satire | vers | 1660 | WS | Paris, 1872 |
| DONNEAU_NOUVELLES | Jean Donneau de Visé | Nouvelles nouvelles | actualité | prose | 1663 | other | éd. Bourqui & Schuwey |
| ELISABETHBOHEME-_CORRESPONDANCE | Élisabeth de Bohême | Correspondance avec René Descartes | correspondance | prose | 1643 | WS | manq. |
| PERRAULT_PETIT-CHAPERON | Charles Perrault | Le petit Chaperon rouge | conte | prose | 1697 | WS | Paris, 1902 |
| POULLAIN_EGALITE | François Poullain de La Barre | De l'Égalité des deux sexes, Discours phisique et moral | discours | prose | 1679 | GALL | Paris, 1679 |
| SCUDERY_ARTAMENE | Madeleine de Scudéry | Artamène ou le Grand Cyrus | roman | prose | 1654 | WS | Auguste Courbé, 1654. |
| VERVILLE_UN-JOUR | Béroalde de Verville | « Un jour reconnaissant que je suis incapable » | poésie | vers | av. 1626 | WS | manq. |
| ANONYME_DECLARATION | anonyme | Déclaration des Droits de l'Homme et du Citoyen de 1793 | légal | prose | 1793 | WS | manq. |
| CHARRIERE_LETTRE | Isabelle de Charrière | lettre | correspondance | prose | 1755 | other | éd. S. van Dijk et M. van Strien-Chardonneau |
| CHATELET_DISCOURS | Émilie du Châtelet | Discours sur le bonheur | discours | prose | 1744 | WS | éd. Bourlet de Vauxcelles, 1796 |
| CHENIER_ODE | André Chénier | Ode IX (à Charlotte Corday) | poésie | vers | 1793 | WS | éd. H. de Latouche, 1819 |
| FREDERICDEUX_DIVINE-EMILIE | Frédéric II de Prusse | À la Divine Émilie | poésie | vers | 1737 | FG | Œuvres de Frédéric le Grand, Berlin, 1850 |
| MERICOURT_DISCOURS | Méricourt, Théroigne de | Discours devant la Société fraternelle des Minimes, 25 mars 1792 | discours | prose | 1792 | GALL | Paris, 1792 |
| MONTEGUT_ISMENE | Jeanne de Montégut-Ségla | Ismene, élégie | poésie | vers | 1739 | other | Paris, 1768 |
| ROUGET_MARSEILLAISE | Rouget de Lisle | La Marseillaise | chanson | vers | 1792 | WS | manq. |
| TENCIN_SIEGE | Madame de Tencin | Le Siège de Calais | roman | prose | 1739 | WS | éd. L.-S. Auger, Paris, 1820. |
| VOLTAIRE_ZADIG | Voltaire | Zadig | nouvelle | prose | 1747 | WS | Paris, 1877. |
| CMH_PV | Commission des monuments historiques | PV 29 février 1884 | administratif | prose | 1884 | ELEC | éd. J.-D. Pariset. |
| DELESCLUZE_LETTRE | Henri Delescluze | Lettre à Charles Delescluse (Carnets de prison) | correspondance | prose | 1851 | ELEC | éd. Chr. Nougaret et Fl.Clavaud |
| DEROULEDE_CHANTS | Paul Déroulède | Nouveaux chants du soldat | chanson | vers | 1883 | GALL | Paris, 1883 |
| DESBORDESVALMORE_IN-SOMNIE | Marceline Desbordes-Valmore | L'Insomnie | poésie | vers | 1860 | WS | Paris, 1860. |
| HEREDIA_EPEE | Hérédia, José Maria de | L'Épée | poésie | vers | 1893 | WS | Paris, 1893. |
| MERIMEE_ACADEMIE | Prosper Mérimée | Discours de réception à l'Académie française | discours | prose | 1845 | other | Site de l'Académie française |
| POTTIER_INTERNATIONALE | Eugène Pottier | L'Internationale | chanson | vers | 1871 | WS | Paris, 1908. |
| SAND_MARIANNE | George Sand | Marianne | roman | prose | 1877 | WS | Paris, 1877 |
| SEGUR_BOSSU | Comtesse de Ségur | François le Bossu | roman | prose | 1864 | WS | Paris, 1901 |
| STAEL_ALLEMAGNE | Mme de Staël | De l'Allemagne | essai | prose | 1810 | WS | Paris, 1814 |
| ANONYME_CONSCRITS | anonyme | Les Conscrits insoumis | chanson | vers | 1902 | WS | manq. |
| BLUM_FRONTPOP | Léon Blum | « Nous sommes un gouvernement de front populaire » | discours | prose | 1936 | GALL | J. O.de la Rép. fr. 7 juin 1936 |
| BRASILLACH_JEANNE | Robert Brasillach | Le Procès de Jeanne d'Arc | étude historique | prose | 1941 | GALL | Paris, 1941 |
| BRIMONT_MIRAGES | Renée de Brimont | Mirages | poésie | vers | 1919 | WS | Paris, 1919 |
| COLETTE_MAISON | Colette | La Maison de Claudine | nouvelle | prose | 1922 | WS | manq. |
| LAHIRE_ROUE | Jean de La Hire | La Roue fulgurante | feuilleton SF | prose | 1908 | GALL | Le Matin, 11 avril 1908 |
| LASTEYRIE_TESTAMENT | Gaspard Louis Guy de Lasteyrie marquis du Saillant | Testament | légal | prose | 1915 | other | Testaments de Poilus |
| LONDRES_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_GRENOUILLES | Rachilde | Le Tueur de Grenouilles | nouvelle | prose | 1900 | WS | Mercure de France |