CORPUS17: a philological corpus for 17th c. French
Simon Gabay, Alexandre Bartz, Yohann Deguin

To cite this version:

HAL Id: hal-03041871
https://hal.science/hal-03041871
Submitted on 11 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CORPUS17: a philological corpus for 17th c. French

Simon Gabay
prenom.nom@unine.ch
Universités de Neuchâtel et de Genève
Neuchâtel and Genève, Switzerland

Alexandre Bartz
prenom.nom@chartes.psl.eu
Ecole des Chartes
Paris, France

Yohann Deguin
prenom-nom@univ-rennes2.fr
Université de Rennes
Rennes, Breizh

ABSTRACT
We investigate the creation of a 17th c. French literary corpus. We present the main options regarding available standards, the training data we created and the efficiency of the models produced for OCR, spelling normalisation and lemmatisation always with open-source solutions. We also present our encoding choices and the global logic of a corpus designed as a virtuous circle, enhancing automatically the tools that are used for its construction.

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

KEYWORDS
17th c. French, OCR, normalisation, lemmatisation, POS-tagging, named entities, digital humanities, XML-TEI

ACM Reference Format:

1 INTRODUCTION
Specialists of 17th c. French texts do not have the habit to adopt a philological approach when editing texts [Dubov 2015; Gabay 2014, 2019]. The recent development of digital tools have not triggered more reflection on their practices, especially regarding transcriptions which are still heavily (and silently) normalised [Schöch 2018], despite the new opportunities offered by computers and the standards used in the digital humanities [Burnard 2014].

Until now, transcriptions have been produced and normalised manually, which has allowed researchers to bypass an important linguistic problem: the persistence of graphic polymorphism. Indeed, the existence of various spellings for one single word (e.g. était vs estoit) prevents even the most simple query on the data, and this problem is likely to grow with the existence of always more powerful OCR engines and robust models retaining increasing amounts of typographical information. If not for philological reasons, manual normalisation will soon be discarded for practical ones (especially time), and we need to engage now the question of the transformation of classical texts into structured information.

The main challenge is, using only open-source and efficient tools that require minimal infrastructure, to design a workflow that converts image scans into usable data while keeping as much information as possible along the way, and to link all the versions of a same information to enrich the mining options. In other words, we need to transcribe efoit by efoit, normalise the spelling (→ était), provide linguistic annotation (être|VERcjg) and link all these information. Convinced that such a project is about data as much as tools, we have conceived training sets with precise philological standards, in order to create state-of-the-art models. These models, which tackle the problem of OCRisation, lemmatisation and normalisation, are designed as general solutions able to deal with heterogeneous (early) modern sources, and do not have a limited capacity on very specific prints, literary genres...

A particular attention has been paid to the nesting of all these solutions one into the another, and thus create a functional workflow. It is indeed important that the lemmatiser and the normaliser take as a source texts that are similar to those produced by the OCR engine to enhance the efficiency of the system. These tools are indeed used to produce a multi-layered corpus for humanists, organised into clear philological strata, with minimal noise and a rich linguistic and semantic annotation, but also for computer scientists, for which we will produce large amounts of high-quality data for further computational exploration (named entity recognition, language identification...).

2 DATA PRODUCTION
Three main datasets have been created to carry the three main tasks: OCR, linguistic normalisation and lemmatisation/POS tagging. All of them have been gathered from sources as representative as possible of 17th c. French literary material, in order to propose general rather than specific solutions. These datasets are used to train machine learning-based models, since it appears to be either the only (OCR) or the most efficient (spelling normalisation, lemmatisation/POS tagging) technique.

2.1 OCR
Following the example of Springmann et al. [2018], we have created a dataset of ground truth (GT) (c. 30,000 lines) [Gabay et al. 2020b]. In order to maximise the efficiency of the training data, we have unbalanced the corpus in two different ways.
On the one hand, capital letters being under-represented compared to lower-case letter, we have decided to over-represent plays in the training data, because this literary genre uses more than others this kind of glyph. On the other hand, in order to have enough GT in italic (cf. fig. 1), we have over-represented texts in verse, which traditionally use this font in the first half of the century [Speyer 2019].

Figure 1: Italics. Tristan L’Hermite, Panthé, 1639.

Images used are in 72, 400 and 600 dpi (cf. fig. 2) to be able to deal with both high and low resolution images.

(a) e, 72 dpi       (b) e, 400 dpi

Figure 2: Impact of the image resolution after binarisation

Transcriptions are graphemic with graphematic traits: no spelling normalisation is introduced, abbreviations are not developed, typos are not corrected, the long s is kept... A model handling aesthetic ligatures (e.g. ⟨ › ⟩) existing in unicode and MUFI [Haugen 2007] has been conceived with Kraken (c.1,400 lines for the train set for a Character Error Rate (CER) of 2.84%). Data augmentation with artificial GT has been tested, without significant impact.

Table 1: Scores on the test set and out-of-domain data

<table>
<thead>
<tr>
<th></th>
<th>Kraken</th>
<th>Calamari</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test set</td>
<td>97.92%</td>
<td>99.05%</td>
</tr>
<tr>
<td>16th c.</td>
<td>98.06%</td>
<td>98.68%</td>
</tr>
<tr>
<td>18th c.</td>
<td>97.78%</td>
<td>98.78%</td>
</tr>
<tr>
<td>19th c.</td>
<td>95.50%</td>
<td>97.05%</td>
</tr>
</tbody>
</table>

Further research has to be carried on the segmentation on the image, which is now the new front for OCR research [Bonhomme 2018]. Because the layout has been also encoded while transcribing the GT, we already have the XML-ALTO files of 1,000 images that could be used to train an efficient segmenter.

2.2 Normalisation

Raw transcriptions coming out of the OCR engine are improper for mining purposes and need to be somehow normalised. Since readers of classical French texts traditionally expect linguistically normalised versions, we do propose to align the historical spelling with the contemporary one. Such a normalisation not only eases the reading process, but also allows researchers to retrieve more information with a simple query: reducing many variants (estoit, estoit, étoit...) to one single form (était) can help improving the precision of the results.

Table 2: Example of normalisation

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sur tout je redoutais cette Mélancolie</td>
<td>Tout je redoutais cette Mélancolie</td>
</tr>
<tr>
<td>Où j’ai vu si longtemps votre Ame ensevelie.</td>
<td>Où j’ai vu si longtemps votre Âme ensevelie.</td>
</tr>
<tr>
<td>Je craignois que le Ciel, par un cruel secours,</td>
<td>Je craignais que le Ciel, par un cruel secours,</td>
</tr>
<tr>
<td>TARGET</td>
<td></td>
</tr>
<tr>
<td>Sur tout je redoutais cette Mélancolie</td>
<td></td>
</tr>
<tr>
<td>Où j’ai vu si longtemps votre Ame ensevelie.</td>
<td></td>
</tr>
<tr>
<td>Je craignais que le Ciel, par un cruel secours,</td>
<td></td>
</tr>
</tbody>
</table>

For obvious reasons, such a task has to be automated, and cannot be done with a simple correspondence table. Indeed, if estoit is always normalised était, in many cases we have to take into account the context to find the correct normalised form. It is the case for the spelling, with words like vostre (possessive determiner votre or pronoun vôtre?), but also word segmentation, with series of tokens like quoi que (relative pronoun quoi que or coordinate conjunction quoi que?). Such a task being similar to translation (en. to the → fr. à, en. to the → fr. au(x)), we have decided to opt for automatic translation tools to tackle the problem.

Following the conclusion of M. Bollmann [2019], we have decided not to use a rule-based system and to focus our effort on Statistical Machine Translations (SMT) and Neural Machine Translation (NMT). If such solutions are more efficient,
they do require important amount of data to be trained on: we have therefore decided to create a parallel corpus (Corpus17). Transcriptions are (mainly) produced with our OCR model and are pre-normalised with a rule-based system, before being manually corrected [Gabay et al. 2019]. Our corpus is a two tier one, with a core version composed of literary texts, and a secondary corpus with peripheral documents dealing with medicine, theology, philosophy, physics... to extend the lexicon. Because spelling evolves with time, samples are distributed diachronically all over the century, and because they vary diatopically, they do not come only from Parisian prints.

Preliminary tests have been carried [Gabay and Barrault 2020] on 160k tokens (c. 600k characters) using two different tools: cSMTiser [Ljubošič et al. 2016] for SMT and NMTPyTorch [Caglayan et al. 2017] for NMT. In spite of its qualities, the former has proven a limited capacity to provide models able to generalise efficiently, while NMT has shown better results despite limited training data. The most reliable indicator, word accuracy (wAcc), should easily be improved with additional training data, potentially produced using back-translation [Domingo et al. 2018], and the use of new powerful language models such as CamemBERT [Martin et al. 2019] or FlauBERT [Le et al. 2020].

<table>
<thead>
<tr>
<th>BLEU (4-grams)</th>
<th>METEOR</th>
<th>wAcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT</td>
<td>77.67</td>
<td>87.89</td>
</tr>
<tr>
<td>NMT</td>
<td>83.65</td>
<td>90.78</td>
</tr>
</tbody>
</table>

Table 3: Evaluation of the best models for each system

2.3 Linguistic annotation

Along with normalisation, we offer linguistic annotation of texts: lemma, Part Of Speech (POS) and morphology (gender, number, mood, tense...). 17thc. French being pre-orthographic, we have decided to prioritise the compatibility of our data with other old states of language – i.e. mainly 18th and 16th c., but also middle and ancient French – to allow deep diachronic research across centuries. Several corpora of historical French already exist, which share (more or less) common annotation practices that we take into account to describe compatibility with medieval data, and a first corpus of normalised 17thc. French is already annotated with this tagset [Petrov et al. 2011]. If the latter is now receiving the favours of the NLP community (the FTB is now using it [Abeillé et al. 2019]), we have decided to use CATTEX-max [Prévost et al. 2013] because it allows basic compatibility with medieval data, and a first corpus of normalised 17thc. French is already annotated with this tagset [Camps et al. 2020]. Detailed annotation guidelines have been produced to document our choices [Gabay et al. 2020a], following closely those written for the BFM [Guillot et al. 2013].

<table>
<thead>
<tr>
<th>Name</th>
<th>Gold</th>
<th>Norm.</th>
<th>Tokens</th>
<th>POS</th>
<th>Morph</th>
</tr>
</thead>
<tbody>
<tr>
<td>CornMol</td>
<td>Yes</td>
<td>Yes</td>
<td>90k</td>
<td>CATTEX</td>
<td>Yes</td>
</tr>
<tr>
<td>FranText</td>
<td>No</td>
<td>Yes</td>
<td>2,400k</td>
<td>EAGLES</td>
<td>No</td>
</tr>
<tr>
<td>Presto gold</td>
<td>Yes</td>
<td>Yes</td>
<td>60k</td>
<td>MULTTEXT</td>
<td>No</td>
</tr>
<tr>
<td>Presto core</td>
<td>No</td>
<td>Yes</td>
<td>6,820k</td>
<td>MULTTEXT</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 4: Available training data

The data used for training mix several heterogeneous train sets (cf. tab. 4) which have all been aligned on our standards with Pyrrha [Clérice et al. 2019]. Two models have been trained with Pie [Manjavacas et al. 2019]: one for lemmatisation with all the available data, another one for POS and morphology (CornMol+Presto gold only) [Gabay et al. 2020c]. In-domain and out-of-domain testing has been carried to evaluate their performance:

3 DATA STRUCTURE

To follow our logic of interoperability, like many other literary corpora, we have decided to encode our corpus in XML-TEI P5 [Burnard 2014]. Because documenting the encoding choices is (sadly according to Burnard [2019]) not common in France, our decisions are inspired by two non-French projects: the Deutsches Textarchiv (DTA) [Haaf et al. 2014] and the European Literary Text Collection (ELTeC) [Odebrecht et al. 2019].
3.1 Markup

Following the examples of the DTA and the ELTeC, we have designed a corpus organised in three layers. If the overall structure is the same, details do differ because of different scientific and institutional situations. Contrary to the ELTeC, which deals with more recent texts that are easy to OCRIse (when they are not already available online) and to process, and contrary to the DTA, which has benefited from long term institutional funding, we need to extract and structure quickly data out of rare and old prints while maintaining minimal ecdotical standards and with limited money. To do so, we have decided to organise our three layers so that it mimics the philological process (cf. fig. 4).

- First we establish the text. We need to offer three main options: describe the layout (<pb>, <1b>), correct typos (<choice>, <sic>, <corr>) and distinguish the text from the peritext (<fu>, <front>, <back>) – to use a Genettian concept [Genette 1997].
- Second we annotate the text. We need to distinguish prose from verse (<p> and its equivalent <lg> + <1>), the text from theatrical (<sp>, <stage>, <speaker>) and epistolary (<opener>, <closer>) peritexts, and additional information such as lists (<list>, <item>), notes (<note>), headings (<header>) and basic named entities (<persName>, <placeName>).
- Third we add all the information that can be produced computationally, such as spelling normalisation (<orig>, <reg>) at the the word (<w>) or sub-phrase (<seg>, delimited with specific punctuation marks such as ::?? or .) levels, or linguistic annotation (<lemma @pos, @msc on <w>).

Because the encoding levels are not only organised semantically, but follow the different steps of the encoding process, some problems arise. Typos that have been forgotten while preparing the first level can be corrected while encoding the second level, which creates two versions of the same text. To solve this problem, a script converts any text encoded in level 2 back into a text encoded in level 1. It is from the level 2 version, which therefore serves as a basis format, that the level 3 is automatically produced.

This logic implies minor differences between the encoding of our corpus with the one of the DTA and the ELTeC (cf. tab. 6), which prevents any direct interoperability. Efforts, however, have been made to maintain basic interchange [Bau- man 2011], especially with the ELTeC because it contains French-written texts, by following the TEI Lite guidelines written by Burnard and Sperberg-McQueen [2012].

3.2 Metadata

The final corpus is planned to have printed texts, but also manuscript transcriptions: a specificity of our corpus is therefore to have two different <teiHeader>, one for each type of document. It is indeed complicated to describe a manuscript like a print: the description of the former is usually based on its conservation (library, shelfmark...), and the latter on its production (printing date and place, publisher...).

It has to be noted that, contrary to other literary traditions, there is no catalogue of (early) modern French manuscripts (mss) such as the one published by Beal [2005] for English writers. Metadata must therefore offer, on top of the simple location of the manuscript, basic information about the document such as:

- its binding (<bindingDesc>, <binding>)
- its paper (<material>, <watermark>)
- its hand (<handDesc>, <handNote>)
- its decoration (<decoDesc>, <decoNote>, <sealDesc>)
- its history (<accMat>, <history>, <provenance>, <acquisition>)
- its content (<incipit>, <explicit>)

Table 5: Lemmatisation accuracies of the best model on out-of-domain data.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>16th</th>
<th>17th</th>
<th>18th</th>
<th>19th</th>
<th>20th</th>
<th>All cent.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test on normalised data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lemma</td>
<td>97.08</td>
<td>97.83</td>
<td>97.99</td>
<td>97.66</td>
<td>97.15</td>
<td>97.55</td>
</tr>
<tr>
<td>POS</td>
<td>93.12</td>
<td>95.09</td>
<td>95.12</td>
<td>92.74</td>
<td>93.5</td>
<td>93.92</td>
</tr>
<tr>
<td>Test on non-normalised data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lemma</td>
<td>95.12</td>
<td>96.8</td>
<td>97.59</td>
<td>97.66</td>
<td>97.15</td>
<td>96.89</td>
</tr>
<tr>
<td>POS</td>
<td>89.36</td>
<td>92.66</td>
<td>95.01</td>
<td>92.74</td>
<td>93.5</td>
<td>92.69</td>
</tr>
</tbody>
</table>

Table 6: Markup allowed for basic encoding
It also has to be noted that, because most of 17thc. French mss are letters, we have decided to take into account the recommendations of the TEI Correspondence SIG [Dumont et al. 2019] and use a `<correspDesc>` to enable data sharing via `correspSearch` [Dumont 2016].

Regarding named entities, we use as much as possible standardised identifiers. For places we use `geoNames` [Wick 2005] because it is the most comprehensive database – until the completion of the promising World-Historical Gazetteer (WHG) [Mostern 2016]. Regarding people, we have decided to use the `International Standard Name Identifier` (ISNI) [Smith-Yoshimura et al. 2020] rather than the Virtual International Authority File (VIAF). The ISNI is indeed the only persistent identifier while the VIAF, a sort of stock exchange for identifiers between libraries, focuses on authority control [Angjeli et al. 2014]. VIAF is therefore used as a secondary choice, as well as other resources such as ORCID [Butler 2012] for editors without an ISNI, or `DATA.bnf.fr` [Bermès et al. 2016] for French data.

### 3.3 Implementation

![Figure 5: ODD chaining](image)

Our choices are both described and enforced thanks to an ODD (One Document Does-it-all [Viglianti 2019]). In order to tailor the markup scheme to our need, we have decided to use ODD-chaining and produce multiple sub-schemas for each encoding level, but also for the two different types of `<teiHeader>` (cf. fig. 5). ODDs are not limited to the simple selection of the necessary elements among all those available: significant work has been carried to control the attributes and, when needed, their possible values. Schematron rules have also been added to refine as much as possible the RNG schemas. HTML documentation is produced out of the ODD and available online.

All the necessary scripts for the automation of tasks are distributed with the corpus, such as the python script, including both the linguistic normaliser and the lemmatiser/POS tagger previously described, that automatically creates the level 3 out of the level 2. Because our NMT-based normalised operates at the (sub)phrase level (to take the context into account), we have decided to add, in the very last step, another layer of information: based on the result of the lemmatisation and the POS tagging, we try to fetch the equivalent of each token in a lexicon of French inflected forms (Morphalou [ATILF-CNRS and Université de Lorraine 2019]) and offer a non-contextualised linguistic normalisation at the token level (cf. fig. 6).

![Figure 6: Example of level 3 encoding](image)

A degree of certainty (@cert) for the normalisation of each token is given: if the script finds one answer in Morphalou the level is high, if there are several answers the level is medium, and if it finds none the level is low (the token is just copied and pasted.)

### 3.4 Corpus design

The very first wave of encoding includes:

- Poetry with Tristan L’Hermite, *Ode*, 1641.

Texts that will be encoded will not be strictly selected in order to provide a representative image of 17th c. literature. First because such an idea probably is impossible, but also because our text collection has been thought as a shell that should be able to welcome various texts, depending on our needs as well as those of researchers, and not a perfectly balanced and representative corpus. Any text is welcome – we will just try, loosely, to avoid major imbalance.

This idea is the basis of a more important one, that brings us back to the beginning of our presentation. Our workflow is massively using machine learning-based tools, which all require important amount of training data: any text added (no matter its printing date, its genre, its author…) enters the MATTER workflow (cf. fig. 7) and eventually improves the overall efficiency of the system [Pustejovsky and Stubbs 2012].

Doing so, the corpus becomes at the same time a literary collection available for reading, a linguistic data bank easily minable, but also a computational resource that serves as a forge for future improvement of digital tools.

4 FURTHER WORK
Most of our future work should concern the stabilisation of the overall workflow with the finalisation of our first wave of texts. On top of the various metrics offered in this article, it will be the opportunity to control manually the efficiency of each system, and potentially try to correct mistakes.

5 CONTRIBUTIONS
The project is lead by S. G., who has coordinated the previous studies and prepared this final article. A. B. is the engineer for the actual creation of the corpus, with the help of S. G. and Y. D. for XML-TEI encoding. All authors discussed and contributed to the final manuscript.

6 DATA
All the data used is CC-BY-SA, and, on top of those distributed with our previous articles, are available on the Github of the E-ditiones project: https://github.com/e-ditiones.

ACKNOWLEDGMENTS
All this work has been carried at the university of Neuchâtel, which supported this project during two years. The Ecole Nationale des Chartes, thanks to Jean-Baptiste Camps and Thibault Cléric, contributed to the creation of models with its GPU.

REFERENCES


Syd Bauman. 2011. Interchange vs. Interoperability. Columbus, OH. https://hal-bnf.archives-ouvertes.fr/hal-01393255


