CORPUS17: a philological corpus for 17th c. French

Simon Gabay, Alexandre Bartz, Yohann Deguin

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:
Simon Gabay, Alexandre Bartz, and Yohann Deguin. 2020. CORPUS17: a philological corpus for 17th c. French. In *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC '20), October 15–17, 2020, Online, Tunisia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3423603.3424002

1 INTRODUCTION

Specialists of 17th c. French texts do not have the habit to adopt a philological approach when editing texts [Duval 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. étoit* 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 eſtoit by *eſtoit*, 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.

DTUC '20, October 15–17, 2020, Hammamet, Tunisia
2020. https://doi.org/10.1145/3423603.3424002

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].

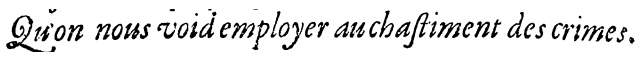


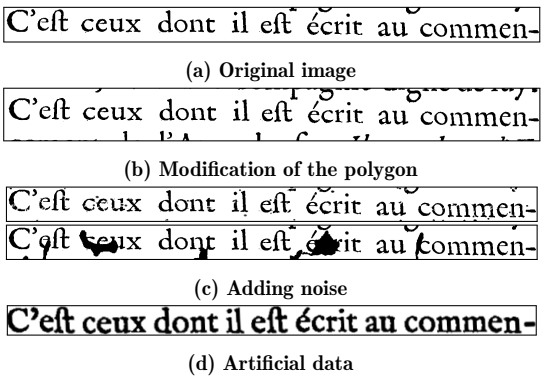
Figure 1: Italics. Tristan L'Hermite, *Panthée*, 1639.

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.* ‹ﬅ›) 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.



(a) Original image

(b) Modification of the polygon

(c) Adding noise

(d) Artificial data

Figure 3: Techniques tested to improve accuracy

Models have been produced with two different engines offering accessible user interfaces for non-specialists: *Kraken* [Kiessling 2019]/*eScriptorium* [Kiessling et al. 2019] and *Calamari* [Wick et al. 2018]/*OCR4all* [Reul et al. 2019]. It has to be noted that scores are not strictly comparable (setups and evaluations are different), but both show extremely good results on the in-domain test set. Out-of-domain data has been prepared with 16th, 18th and 19th c. prints to evaluate the capacity of the model to generalise – which seems to be the case.

	Kraken	Calamari
Test set	97.92%	99.05%
16th c.	98.06%	98.68%
18th c.	97.78%	98.78%
19th c.	95.50%	97.05%

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

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.

Source
Sur tout ie redout**o**is cette Mélanccolie
Où j'a**y** v**eu f**i long-temps vo**f**tre **A**me enseuelie.
Ie craign**o**is que le Ciel, par **v**n cruel fecours,
Target
Surtout je redout**a**is cette Mélanccolie
Où j'a**i** v**u s**i longtemps votre **Â**me ensevelie.
Je craign**a**is que le Ciel, par **u**n cruel secours,

Table 2: Example of normalisation

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 *quoique*?). Such a task being similar to translation (en. *garden* → fr. *jardin*, 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 caracters) using two different tools: *cSMTiser* [Ljubešić 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 and Casacuberta 2018], and the use of new powerful language models such as *CamemBERT* [Martin et al. 2019] or *FlauBERT* [Le et al. 2020].

	BLEU (4-grams)	METEOR	wAcc
SMT	77.67	87.89	86.68
NMT	83.65	90.78	91.42

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...). 17th c. 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 offer (minimal) interoperability of data. Regarding medieval French, we count the *Base Geste* [Camps 2016] or the *Base de français médiéval* [Guillot et al. 2017]. For (early) modern French we find *Presto* [Blumenthal et al. 2017] or those of the *Réseau Corpus Français Préclassique et Classique* [Amatuzzi et al. 2019].

Linguistic resources have been developed for the *Presto* corpus [Diwersy et al. 2017] that are now widely popular among specialists of (early) modern French, especially the extended version of the *LGeRM* [Souvay and Pierrel 2009] authority list of lemmas for modern French (called *mode*) [ATILF-CNRS and Université de Lorraine 2017]. The main interest of this list is that it is related to the *Dictionnaire du Moyen Français* [ATILF-CNRS and Université de Lorraine 2015] and the *Trésor de la Langue Française informatisé* [Pierrel et al. 2004], and therefore allows maximal interoperability with older and more recent state of languages but also major lexicographic resources. Using the *Dictionnaire étymologique de l'ancien français électronique* [Möhren 2002] or the digital version of the *Altfranzösisches Wörterbuch* [Tobler et al. 2002], as medievalists do [Camps et al. 2019; Glessgen and Stein 2005], is not possible, because of the too important lexicographic evolution.

Regarding POS and morphology, we are more dubious of *Presto*'s choice to follow MULTEXT [Ide and Veronis 1994] and Grace [Adda et al. 1998; Lecomte 1997] recommendations: this choice was made at a time when, on the one hand, the most important French corpus (*FranText* [ATILF-CNRS and Université de Lorraine 2020]) was using a tagger [Crabbé and Candito 2008] trained on a corpus (The French treebank (FTB) [Abeillé et al. 2003]) using a different tagset [Ollinger 2018], and on the other hand the international standard UD-POS (Universal dependencies POS tag set) already existed [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 17th c. French is already annotated with this tag set [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].

Name	Gold	Norm.	Tokens	POS	Morph
CornMol	Yes	Yes	90k	CATTEX	Yes
FranText	No	Yes	2,400k	EAGLES	No
Presto gold	Yes	Yes	60k	MULTEXT	No
Presto core	No	Yes	6,820k	MULTEXT	No

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].

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

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

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>`, `<lb>`), correct typos (`<choice>`, `<sic>`, `<corr>`) and distinguish the text from the peritext (`<fw>`, `<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>` +`<l>`), the text from theatrical (`<sp>`, `<stage>`, `<speaker>`) and epitsolary (`<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 computationnally, 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>`).

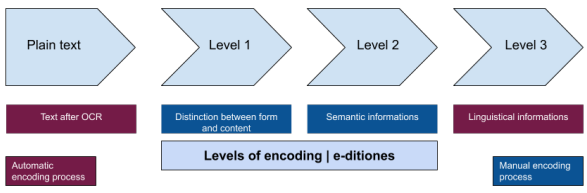


Figure 4: Encoding levels

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 [Bauman 2011], especially with the ELTeC because it contains French-written texts, by following the *TEI Lite* guidelines written by Burnard and Sperberg-McQueen [2012].

Corpus	Markup
DTA	`<div>`, `<pb>`, `<p>`, `<lg>`, `<figure>`, `<cb>`, `<head>`,
ELTeC	`<div>`, `<pb>`, `<p>`, `<l>`, `<front>`, `<back>`, `<hi>`
CORPUS17	`<div>`, `<pb>`, `<p>` (only one), `<front>`, `<back>`, `<fw>`, `<lb>`, `<hi>`, `<sic>`, `<corr>`, `<choice>`

Table 6: Markup allowed for basic encoding

Contrary to the ELTeC [Burnard et al. 2019], no specific vocabulary to our project has been added (`<eltec:sex>`, `<eltec:size>`...): our selection is a strict subset of the TEI, and our final encoding remains therefore fully TEI-compliant.

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>`)

It also has to be noted that, because most of 17th c. 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, focus 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

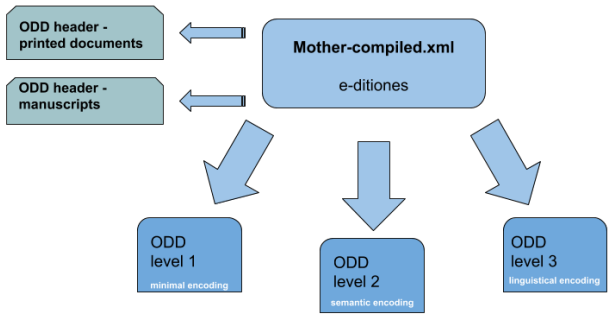


Figure 5: ODD chaining

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[1].

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).

```
<seg>
 <choice>
  <orig>
   <w lemma="je" pos="PROper" msd="NOMB.=s">
    <choice>
     <orig>i'</orig>
     <reg cert="high">j'</reg>
    </choice>
   </w>
   <w lemma="être" pos="VERcjg"
      msd="MODE=ind|TEMPS=pft|PERS.=1|NOMB.=s">
    <choice>
     <orig>estoit</orig>
     <reg cert="high">étais</reg>
    </choice>
   </w>
   ...
  </orig>
  <reg>j'étais</reg>
 </choice>
</seg>
```

Figure 6: Example of level 3 encoding

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:

- Letters with Faret, *Recueil de lettres nouvelles*, Paris: T. du Bray, 1627; or Puget, *Le bouquet des plus belles fleurs de l'eloquence*, Paris: Pierre Billaine/Nicolas Bessin, 1624.
- A novel with Marcassus, *Amadis de Gaule*, Paris: P. Rocolet, 1629.
- An essay with Gournay, *Egalité des hommes et des femmes*, N.d.: N.p., 1622.
- Tales with La Fontaine, *Deuxiesme partie des Contes et nouvelles en vers*, Paris: C. Barbin, 1666.
- Comedies with Molière, *George Dandin*, Paris, J. Ribou, 1669; or Corneille, *L'Amour à la mode*, Rouen: L. Maurry, Paris: G. de Luynes, 1653
- Tragedies with Pradon, *Scipion*, Paris: J. Ribou, 1700; or Campistron, *Achile et Polixene*, Paris, Academie royale de musique, 1687
- Poetry with Tristan L'Hermite, *Ode*, 1641.

[1] https://e-ditiones.github.io/ODD/ODD-1.html, ODD-2.html, ODD-3.html, ODD-header_MS.html and ODD-header_printed.html.

- Manuscripts with excerpts of the MS Harvard, Lowell collection 282 and the MSS Princeton, C0710, vol. 3 and 4.

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.

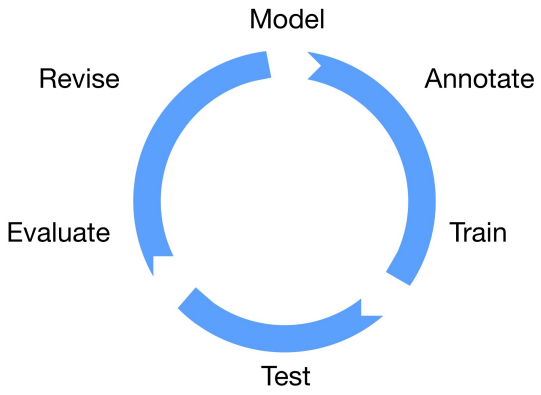


Figure 7: MATTER workflow

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 availale 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 Clérice, contributed to the creation of models with its GPU.

REFERENCES

Anne Abeillé, Lionel Clément, and Loïc Liégeois. 2019. Un corpus arboré pour le français : le French Treebank. *Traitement Automatique des Langues* 60, 3 (2019), 19–43. https://www.atala.org/content/un-corpus-arbor%C3%A9-pour-le-fran%C3%A7ais-le-french-treebank

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a Treebank for French. In *Treebanks: Building and Using Parsed Corpora*, Anne Abeillé (Ed.). Springer Netherlands, Dordrecht, 165–187. https://doi.org/10.1007/978-94-010-0201-1_10

G. Adda, J. Mariani, J. Lecomte, P. Paroubek, and M. Rajman. 1998. The GRACE French Part-Of-Speech Tagging Evaluation Task. In *Proc. of LREC'98 (1st International Conference on Language Resources and Evaluation)*. Granada, Spain. https://infoscience.epfl.ch/record/98004

Antonella Amatuzzi, Carine Skupien Dekens, Wendy Ayres-Bennett, Annette Gerstenberg, and Lene Schoesler. 2019. Améliorer et appliquer les outils numériques. Ressources et approches pour l'étude du changement linguistique en français préclassique et classique. In *Le français en Diachronie*. Editions de linguistique et de philologie, Strasbourg, 337–364.

Anila Angjeli, Andrew MacEwan, and Vincent Boulet. 2014. ISNI and VIAF – Transforming ways of trustfully consolidating identities. In *World Library and Information Congress: IFLA General Conference and Assembly – Libraries, Citizens, Societies: Confluence for Knowledge*. Lyon, France. https://doi.org/10.13140/RG.2.1.1350.8640

ATILF-CNRS and Université de Lorraine. 1998–2020. Base textuelle Frantext (En ligne). https://www.frantext.fr

ATILF-CNRS and Université de Lorraine. 2015. Dictionnaire du Moyen Français (1330-1500). http://www.atilf.fr/dmf

ATILF-CNRS and Université de Lorraine. 2017. Lemmes Graphies et Règles Morphologiques (LGeRM). https://hdl.handle.net/11403/lgerm/v1 ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

ATILF-CNRS and Université de Lorraine. 2019. Morphalou. https://hdl.handle.net/11403/morphalou/v3.1 ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Syd Bauman. 2011. Interchange vs. Interoperability. Montréal, Canada. https://doi.org/10.4242/BalisageVol7.Bauman01

Peter Beal. 2005. Catalogue of English Literary Manuscripts 1450–1700. https://celm-ms.org.uk/

Emmanuelle Bermès, Vincent Boulet, and Céline Leclaire. 2016. Améliorer l'accès aux données des bibliothèques sur le web : l'exemple de data.bnf.fr. In *World Library and Information Congress: IFLA General Conference and Assembly – Connections. Collaboration. Community*. Columbus, OH. https://hal-bnf.archives-ouvertes.fr/hal-01393255

Peter Blumenthal, Sascha Diwersy, Achille Falaise, Marie-Hélène Lay, Gilles Sourvay, and Denis Vigier. 2017. Presto, un corpus diachronique pour le français des XVIe-XXe siècles. In *Actes de la 24^e conférence sur le Traitement Automatique des Langues Naturelles - TALN'17*. Association pour le traitement automatique des langues, Orléans, France. https://halshs.archives-ouvertes.fr/halshs-01585010

Marcel Bollmann. 2019. A Large-Scale Comparison of Historical Text Normalization Systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 3885–3898. https://doi.org/10.18653/v1/N19-1389

Marie-Laurence Bonhomme. 2018. *Répertoire des Notaires parisiens Segmentation automatique et reconnaissance d'écriture*. report. Inria. 1–10 pages. https://hal.inria.fr/hal-01949198

Lou Burnard. 2014. *What is the Text Encoding Initiative? : How to add intelligent markup to digital resources*. OpenEdition Press, Marseille. http://books.openedition.org/oep/426

Lou Burnard. 2019. What is TEI Conformance, and Why Should You Care? *Journal of the Text Encoding Initiative* Issue 12 (Jan. 2019). https://doi.org/10.4000/jtei.1777 Number: Issue 12 Publisher: Text Encoding Initiative Consortium.

Lou Burnard, Christof Schoch, and Carolin Odebrecht. 2019. In search of comity: TEI for distant reading. In *What is text, really? TEI and beyond*. Graz, Austria. https://doi.org/10.5281/zenodo.3552489

Lou Burnard and C. Michael Sperberg-McQueen. 2012. TEI Lite: Encoding for Interchange: an introduction to the TEI. https://tei-c.org/release/doc/tei-p5-exemplars/html/tei_lite.doc.html

Declan Butler. 2012. Scientists: your number is up. *Nature* 485, 7400 (May 2012), 564. https://doi.org/10.1038/485564a

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. NMTPY: A Flexible Toolkit for Advanced Neural Machine Translation Systems. *Prague Bull. Math. Linguistics* 109 (2017), 15–28. https://doi.org/10.1515/pralin-2017-0035

Jean-Baptiste Camps. 2016. Geste: un corpus de chansons de geste. https://doi.org/10.5281/zenodo.1744918

Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. 2019. *Deucalion, Modèle Ancien Francais (0.2.0)*. École nationale des chartes. https://doi.org/10.5281/zenodo.3237455

Jean-Baptiste Camps, Simon Gabay, Paul Fièvre, Thibault Clérice, and Florian Cafiero. 2020. Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre. (2020). https://halshs.archives-ouvertes.fr/halshs-02591388 working paper or preprint.

Thibault Clérice, Julien Pilla, Jean-Baptiste Camps, Vincent Jolivet, and Ariane Pinche. 2019. *Pyrrha, A language independant post correction app for POS and lemmatization*. École nationale des chartes. https://doi.org/10.5281/zenodo.2325427

Benoît Crabbé and Marie Candito. 2008. Expériences d'analyse syntaxique statistique du français. In *Actes de la 15^{e} conférence sur le Traitement Automatique des Langues Naturelles - TALN'08*. Association pour le traitement automatique des langues, Avignon, France, pp. 44–54. https://hal.archives-ouvertes.fr/hal-00341093

Sascha Diwersy, Achille Falaise, Marie-Hélène Lay, and Gilles Souvay. 2017. Ressources et méthodes pour l'analyse diachronique. *Langages* 206, 2 (2017), 21–43. https://halshs.archives-ouvertes.fr/halshs-01581141

Miguel Domingo and Francisco Casacuberta. 2018. A Machine Translation Approach for Modernizing Historical Documents Using Back Translation. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*. Bruges, Belgium, 39–47. https://workshop2018.iwslt.org/downloads/Proceedings_IWSLT_2018.pdf

Stefan Dumont. 2016. correspSearch – Connecting Scholarly Editions of Letters. *Journal of the Text Encoding Initiative* Issue 10 (Dec. 2016). https://doi.org/10.4000/jtei.1742 Number: Issue 10 Publisher: Text Encoding Initiative Consortium.

Stefan Dumont, Susanne Haaf, Sabine Seifert, Stefan Dumont, Susanne Haaf, and Sabine Seifert. 2019. Introduction. In *Encoding Correspondence. A Manual for TEI-XML-based Encoding of Letters and Postcards in TEI-XML and DTABf* (1 ed.). Berlin-Brandenburg Academy of Sciences and Humanities, Berlin.

Frédéric Duval. 2015. Les éditions de textes du XVIIe siècle. In *Manuel de la philologie de l'édition*, David Trotter (Ed.). De Gruyter, Berlin, Boston, 369–394. https://doi.org/10.1515/9783110302608-017

Simon Gabay. 2014. Pourquoi moderniser l'orthographe? Principes d'ecdotique et littérature du XVIIe siècle. *Vox Romanica* 73 (2014), 27–42. http://periodicals.narr.de/index.php/vox_romanica/article/view/2254

Simon Gabay. 2019. Éditer le Grand Siècle au XVIIIe s. Remarques sur les choix (ortho)graphiques de quelques éditeurs. *Book Practices & Textual Itineraries* 9 (2019), 133–148. https://hal.archives-ouvertes.fr/hal-01900036 Version avant relecture.

Simon Gabay and Loïc Barrault. 2020. Machine Translation for the Normalisation of 17th c. In *Actes de la 27^{e} conférence sur le Traitement Automatique des Langues Naturelles - TALN'20*. Association pour le traitement automatique des langues, Nancy, France, 213–222. https://hal.archives-ouvertes.fr/hal-02784770

Simon Gabay, Jean-Baptiste Camps, and Thibault Clérice. 2020a. Guidelines for linguistic annotation of modern French (16th-18th c.). https://hal.archives-ouvertes.fr/hal-02571190 Manuel d'annotation en vue de la création de modèle de lemmatisation et d'annotation morpho-sytaxique et morphologique du français des XVI-XVIII^{e} s.

Simon Gabay, Thibault Clérice, and Christian Reul. 2020b. OCR17: Ground Truth and Models for 17th c. French Prints (and hopefully more). (May 2020). https://hal.archives-ouvertes.fr/hal-02577236 working paper or preprint.

Simon Gabay, Thibault Clérice, Jean-Baptiste Camps, Jean-Baptiste Tanguy, and Matthias Gille Levenson. 2020c. Standardizing linguistic data: method and tools for annotating (pre-orthographic) French. Hammamet (Tunisia).

Simon Gabay, Marine Riguet, and Loïc Barrault. 2019. A Workflow For On The Fly Normalisation Of 17th c. French. In *DH2019*. Alliance of Digital Humanities Organizations (ADHO), Utrecht, Netherlands. https://hal.archives-ouvertes.fr/hal-02276150

Gérard Genette. 1997. *Paratexts. Thresholds of interpretation*. Cambribge University Press, Cambridge.

Martin Glessgen and Achim Stein. 2005. Resources and tools for analyzing Old French texts. In *ScriptOralia*. Narr, Tübingen, 135–145. https://doi.org/10.5167/uzh-33587 Issue: 130 Number: 130.

Céline Guillot, Celine Guillot, Sophie Prévost, and Alexei Lavrentiev. 2013. *Manuel de référence du jeu Cattex09*. Technical Report. http://bfm.ens-lyon.fr/spip.php?article323

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

Susanne Haaf, Alexander Geyken, and Frank Wiegand. 2014. The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources. *Journal of the Text Encoding Initiative* Issue 8 (Dec. 2014). https://doi.org/10.4000/jtei.1114 Number: Issue 8 Publisher: Text Encoding Initiative Consortium.

Odd Einar Haugen. 2007. Medieval Unicode Font Initiative (MUFI): Coordinating Medieval characters in the Latin alphabet. *Sprache und Datenverarbeitung* 31, 1-2 (2007), 91–99.

Nancy Ide and Jean Veronis. 1994. MULTEXT: Multilingual Text Tools and Corpora. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*. https://www.aclweb.org/anthology/C94-1097

Benjamin Kiessling. 2019. Kraken - an Universal Text Recognizer for the Humanities. Alliance of Digital Humanities Organizations (ADHO), Utrecht, The Netherlands. https://dev.clariah.nl/files/dh2019/boa/0673.html

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2. 19–19. https://doi.org/10.1109/ICDARW.2019.10032

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2479–2490. https://www.aclweb.org/anthology/2020.lrec-1.302

Josette Lecomte. 1997. *Codage Multext -GRACE pour l'action GRACE*. Technical Report. INALF, Nancy.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. Bochum, Germany, 146–155. https://www.linguistics.rub.de/konvens16/pub/19_konvensproc.pdf

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving Lemmatization of Non-Standard Languages with Joint Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1493–1503. https://doi.org/10.18653/v1/N19-1153

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a Tasty French Language Model. *arXiv e-prints*, Article arXiv:1911.03894 (Nov 2019), arXiv:1911.03894 pages. arXiv:cs.CL/1911.03894

Ruth Mostern. 2016. *World-Historical Gazetteer*. Technical Report. http://dev.whgazetteer.org

Frankwalt Möhren. 2002. *Dictionnaire étymologique de l'ancien français électronique (DEAFél)*. http://www.deaf-page.de

Carolin Odebrecht, Lou Burnard, Borja Navarro Colorado, Maciej Eder, and Christof Schöch. 2019. The European Literary Text Collection (ELTeC). In *Digital Humanities 2019 Conference Abstracts*. Alliance of Digital Humanities Organizations (ADHO), Utrecht, The Netherlands. https://dataverse.nl/dataset.xhtml?persistentId=hdl:10411/ALRFLG

Sandrine Ollinger. 2018. *Modèle Talismane pour textes littéraires en français moderne*. ATILF. https://www.ortolang.fr/market/tools/talismane-frantext-modern

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2011. A Universal Part-of-Speech Tagset. *CoRR* abs/1104.2086 (2011). arXiv:1104.2086 http://arxiv.org/abs/1104.2086

Jean-Marie Pierrel, Jacques Dendien, and Pascale Bernard. 2004. Le TLFi ou Trésor de la Langue Française informatisé. In *Proceedings of the 11th EURALEX International Congress* (6-10), Geoffrey Williams and Sandra Vessier (Eds.). Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, Lorient, 165–170.

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

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media.

Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019. OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences* 9, 22 (Jan. 2019), 4853. https://doi.org/10.3390/app9224853

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

Karen Smith-Yoshimura, Janifer Gatenby, Grace Agnew, Christopher Brown, Kate Byrne, Matt Carruthers, Peter Fletcher, Stephen Hearn, Xiaoli Li, Marina Muilwijk, Chew Chiat Naun, John Riemer, Roderick Sadler, Jing Wang, Glen Wiley, and Kayla Willey. 2020. Addressing the Challenges with Organizational Identifiers and ISNI. https://www.oclc.org/research/publications/2016/oclcresearch-organizational-identifiers-and-isni-2016.html

Gilles Souvay and Jean-Marie Pierrel. 2009. LGeRM Lemmatisation des mots en Moyen Français. *Traitement Automatique des Langues* 50, 2 (2009), 21. https://halshs.archives-ouvertes.fr/halshs-00396452

Miriam Speyer. 2019. Les dieux écrivent-ils en italiques ? Typographie et mise en livre de pièces en vers et en prose. In *L'Habillage du livre et du texte aux XVIIe et XVIIIe siècles*, Nicolas Brucker, Nathalie Collé, Pierre Degott, and Anne-Elisabeth Spica (Eds.). Number 9. PUN - Éditions Universitaires de Lorraine, Nancy. https://hal-normandie-univ.archives-ouvertes.fr/hal-02184237

Uwe Springmann, Christian Reul, Stefanie Dipper, and Johannes Baiter. 2018. Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. *CoRR* abs/1809.05501 (2018). arXiv:1809.05501 http://arxiv.org/abs/1809.05501

Adolf Tobler, Peter Blumenthal, Achim Stein, Erhard Lommatzsch, and Hans Helmut Christmann. 2002. *Tobler-Lommatzsch : altfranzösisches Wörterbuch, édition électronique*. Franz Steiner Verlag, Stuttgart. https://trove.nla.gov.au/version/215526332

Raffaele Viglianti. 2019. One Document Does-it-all (ODD): a language for documentation, schema generation, and customization from the Text Encoding Initiative. In *Proceedings of the Symposium on Markup Vocabulary Customization*. Washington, DC. https://doi.org/10.4242/BalisageVol24.Viglianti01

Christoph Wick, Christian Reul, and Frank Puppe. 2018. Calamari-A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *arXiv preprint arXiv:1807.02004* (2018).

Marc Wick. 2005. GeoNames. http://www.geonames.org.