Journal of Data Mining and Digital Humanities

OCR17: Ground Truth and Models for 17th century French Prints (and hopefully more)

Simon Gabay1, **Thibault Clérice**2, **Christian Reul**3

1Université de Neuchâtel, Université de Genève (Switzerland)
2École nationale des chartes, Centre Jean Mabillon (EA 3624), PSL University (France)
3Universität Würzburg (Germany)

Corresponding author: Simon Gabay , simon.gabay@unige.ch

Abstract

Machine learning begins with machine teaching: in the following paper, we present the data that we have prepared to kick-start the training of reliable OCR models for 17th century prints written in French. The construction of a representative corpus is a major challenge: we need to gather documents from different decades and of different genres to cover as many sizes, weights and styles as possible. Historical prints containing glyphs and typefaces that have now disappeared, transcription is a complex act, for which we present guidelines. Finally, we provide preliminary results based on these training data and experiments to improve them.

Keywords

OCR, 17th century French, training data, transcription guidelines, corpus building

I INTRODUCTION

OCR engines such as Abbyy[1] or Tesseract [Smith, 2007] today come with models that work perfectly for the most recent documents written in French (i.e. the 19th and 20th centuries)[2], but older periods are still less well handled by the machines. Our objective is therefore to solve this problem, and to propose the tools, as much as the data and the method necessary for the processing of historical documents following the recommendations on open science [Chagué et al., 2020], and especially prints in French of the 17th century. In doing so, we hope to prepare the digitisation of the entire period of the Ancien Régime, the prints of the 18th century and the second half of the 16th century written in French being relatively similar to those of the 17th century.

The reliability of OCR models depends on both the quantity and the quality of training data. On the one hand, quantity needs to be produced and made freely available to other scholars, which is sadly not always the case. On the other hand, quality needs to be properly defined, since philological traditions vary from one place to another [Duval, 2018], but also from one period to another [Gabay, 2014, Duval, 2015]. Both problems need therefore to be addressed to propose a reliable solution for the OCRisation of historical prints written in French.

Following the example of *GT4HistOCR* [Springmann et al., 2018], which mainly focuses on German (and marginally Latin [Springmann et al., 2016]), we have designed a corpus of Ground

[1] https://pdf.abbyy.com

[2] Several models are available at this address: https://github.com/tesseract-ocr/tessdata_best.

Truth (GT) made of c. 30,000 lines taken from 37 prints in French of the 17th century (tab. 9). These documents have been carefully chosen so that they contain different kinds of typefaces (style, weight, size), and thus cover a maximum of the fonts possibly used for this type of document. Because graphetic variants that have now disappeared may have existed (such as the long *s*: *ſ*), transcription is a particularly technical act for historical documents, and we describe our transcription guidelines. Based on these data, we offer robust state-of-the art models for two open source OCR engines, both available to users *via* simple interfaces: *Kraken* [Kiessling, 2019]/*eScriptorium* [Kiessling et al., 2019] and *Calamari* [Wick et al., 2020]/*OCR4all* [Reul et al., 2019].

II CORPUS BUILDING

Several corpora exist today for 17th century documents. We have already mentioned *GT4HistOCR* [Springmann et al., 2018], but others are available such as the *IMPACT Dataset* [Papadopoulos et al., 2013], of which 80% of the documents date however from the 19th and 20th centuries. The first corpus is mainly focusing on German[3], as we previously mentioned, such as the RIDGES dataset [Springmann and Lüdeling, 2017], and the second corpus contains only about 15% of documents in French. In both cases we do not control (or know precisely) the transcription guidelines, which is an important philological problem. We have therefore decided to create our own corpus.

Producing training data in order to kick-start the creation of a generic model for 17th c. documents written in French implies the gathering of various sources, which can be selected in many ways, from piling up data from different projects to the scrupulous association of complementary sources. For our project, due to the paucity of available data, we chose to follow the second option, which is not a simple task. We therefore had to define a method to select the documents that should be included in our corpus.

Table 1: Distribution of the prints in the training corpus per decade

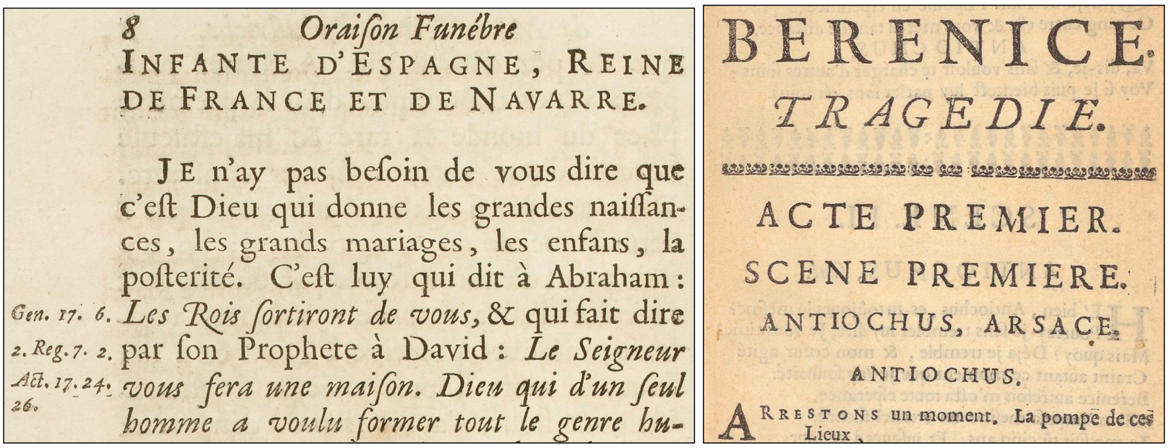
Decade	Total items	Total lines
00's	1	617
10's	1	198
20's	3	2,689
30's	5	3,159
40's	5	3,527
50's	3	2,008
60's	5	5,089
70's	4	3,836
80's	5	3,336
90's	5	3,709

Table 2: Distribution of the prints in the training corpus per genre

Genre	Total
Drama	17
Poetry	4
Novel	3
Letter	2
Philosophy	2
Physics	2
Sermon	1
Theology	1
Travel	1
Maxims	1
Medicine	1
Memoirs	1
Mechanics	1

[3]It is important to note that a large part of the German-written corpora are not only printed in *antiqua* but in *fraktur*, which considerably minimises their interest for the OCRisation of documents in French.

Since the advent of corpus pragmatics, linguists have been working on how to associate data to obtain representativeness, i.e. "the extent to which a sample includes the full range of variability in a population" [Biber, 1993], but such a notion is now more and more debated [Raineri and Debras, 2019]. Following the example of corpus linguists using extralinguistic criteria (sociological, demographic…) [Crowdy, 1993], we have decided to select samples mostly according to bibliographical metadata (printing date and place, literary genre, author…), which serve as a proxy for paleographical information – a good diachronic distribution should for instance ensure a correct representation of the very diverse typographical material (e.g. fig. 1). We also took into account digital information (size and resolution of the images), in addition, of course, to a careful philological analysis of the documents.



(a) Bossuet, *Oraison*, 1683 (b) Racine, *Oeuvres*, 1676

Figure 1: Mixing styles, heights and casing in 17th c. French prints

Prints production dates are distributed over the century, with a special attention for books printed between 1620 and 1700 (tab. 1) because it covers one of the most important periods in the literary history of France, that of classical French. Regarding genre, the result can be seen as a two-tier corpus (tab. 2), with a primary one consisting of literary texts (drama, poetry, novels…) and a secondary one made of scientific works (medicine, mechanics, physics…). If the vast majority has been printed in Paris, we have also included books coming from Belgium (Brussels) and Holland (Leiden), which were major production centres at the time [van Eeghen, 1960–1978, Eisenstein, 1992][4].

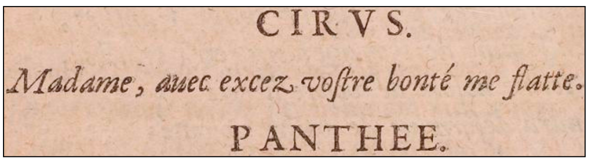


Figure 2: Tristan L'Hermite, *Panthée*, 1639

	Lower	Upper	Total
Dramatic texts	396,984 90.17%	43,295 9.83%	440,279
Non-dramatic Texts	297,527 95.96%	12,544 4.04%	310,071

Table 3: Percentage of uppercase letters in dramatic texts vs. non-dramatic texts in our dataset.

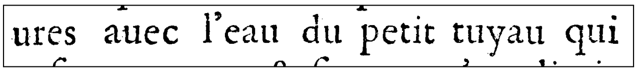
As we can see, the corpus is not balanced, since not only literary texts, but also plays are clearly

[4]A detailed list of the contents of the corpus can be found in the appendix (cf. tab. 9).

over-represented. Such a choice has been made for two reasons. On the one hand, we need GT in italics, and since versified texts use this type of style a lot [Speyer, 2019], we increase the amount of data in italics by selecting plays in verse. On the other hand, we must also add capital letters in the GT, and therefore find examples of this type of character: plays, once again, are an abundant source with the names of the speakers written in capitals (fig. 2 and tab. 3). Such a strategy should help us deal with highly complex layouts (fig. 1).

Regarding the resolution, images used can be divided into three classes: 72 (20 prints), 400 (14 prints) and 600 dpi (1 print) (fig. 3). Indeed, many scans available online are in low resolutions (usually 72 dpi, an older computer standard introduced by Apple in the 1980s), which introduces significant changes in the shape of letters (fig. 4) that our model needs to handle properly.

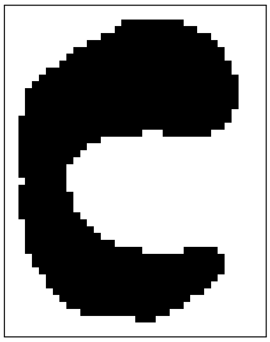
(a) Balzac, *Lettres*, 1624, 72 dpi



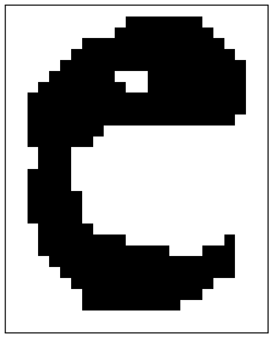
(b) Pascal, *Equilibre*, 1663, 400 dpi

(c) Papin, *Manière*, 1682, 600 dpi

Figure 3: Examples of GT with different resolutions



(a) 72 dpi



(b) 400 dpi

Figure 4: Impact of the resolution on the letter *e*, which can be confused with *c* because of the disappearance of the eye.

III TRANSCRIPTION GUIDELINES

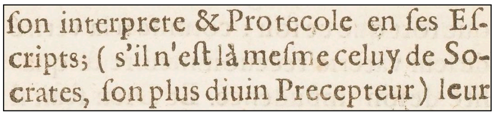
Theoretical background. Transcription is a very delicate matter: more than copying, transcribing has to be understood as an act of translation [Robinson and Solopova, 1993] and, as the saying goes, *traduttore, traditore*. Following Robinson and Solopova, there are four different levels of transcription, rearranged into two categories by D. Stutzmann [Stutzmann, 2011]: those that describe the image (called "graphic" and "allographetic" transcriptions) and those that describe the text (called "graphemic" and "regularised"). For our project, we exclude the two most extreme types of the spectrum: graphic transcription (which retains all the visual richness of the original) would be far too time-consuming, and regularised transcription (which fully aligns the spelling to a standard) would be linguistically too poor. Only the allographetic and the graphemic transcriptions will therefore interest us.

The allographetic transcription aims at keeping all the graphetic richness: it reduces all the graphic variants to extended types, and thus gives access to various forms of each letter or sign. For instance, the distinction between ‹ſ› and ‹s› is kept because they are two graphetic variants of *s*, but not the distinction between ‹ſ› with or without a leftward swash, because they both would be graphic variations of the ‹ſ›. Long impossible for material reasons, such transcriptions are now accessible to all researchers thanks to projects such as the Medieval Unicode Font Initiative (MUFI, Haugen [2015]), that have played an important role by designing and pushing new

Unicode code points to the Unicode standard [Consortium, 2019][5].

The graphemic transcription further limits variation by reducing the different possible types to their meaning in the alphabetic system. Unlike the allographetic approach, graphemic transcription is better known to philologists because it resembles the traditional semi-diplomatic transcription. Scholars have clarified its execution framework, and have been using it in their editions for a long time. In practice, words like *eſtoit* are transposed as *estoit* in the edited version, but not as *était* which would be a regularised transcription. A few points of detail remain however debated, in particular concerning the need to expand the abbreviations when adopting a graphemic approach[6].

Project framework. As our intent for our data is to produce OCR models and to transcribe automatically print to do more research on the French language used in these prints, our transcription guidelines lies on the side of the graphemic transcription, without regularisation. However, we introduced few graphetic concerns, listed below, because graphetic variations can be a linguistic evidence, such as the long *s*, which inform us of the content of the type boxes used by printers[7]. The result is a hypbrid graphemic-allographetic transcription with very punctual regularisations.



ſon interprete & Protecole en ſes Eſ-
cripts; (s'il n'eſt là meſme celuy de So-
crates, ſon plus diuin Precepteur) leur

Figure 5: Excerpt and transcription example of Marie de Gournay, *Egalité*, 1622

Normalisation of spelling. Our choice leads us to keep the original spelling of the source (e.g. fig. 5). We include in spelling the absence of normalisation for letters such as ‹u›/‹v› and ‹i›/‹j›, whose usage was different from the current one (no consonant/vowel distinction): e.g., we transcribe *diuin* where one would normalise it as *divin*. We respect accent absence (e.g. *interprete* and not *interprète*), but we transcribe dotless i (‹ı›) as ‹i›, as it is in many cases a printing problem. To avoid confusion for the machine, commas used as cedillas (*FRANC,OIS* and not *FRANÇOIS*) and apostrophes used as accents (*ARME'ES* and not *ARMÉES*) are kept as they are, and not regularised as accents or cedillas. Historical spellings (e.g. *Eſcripts*, normalised *Ecrits*) and calligraphic letters (e.g. *celuy*, normalised *celui*) are kept.

Variation of letters and ligatures. As mentioned, we keep one allographetic variation: the long *s* (e.g. *meſme* and not *mesme*). Other variations are ignored. Aesthetic ligatures that still exist in French (e.g. ‹œ› vs ‹oe›) have been encoded, but not those that have disappeared despite their possible existence in Unicode (e.g. ‹ﬅ›)[8]. Examples are provided in Table 4[9].

[5]In certain cases, characters which were not accepted (yet) by the Unicode governing bodies might be stored in the private zone of Unicode, being only supporter with MUFI-related font as a consequence.

[6]The *Conseils pour l'édition des textes médiévaux* (which are a reference for rigorous philological editing of texts, medieval or not) suggest the expansion of abbreviations [Bourgain and Vieillard, 2001, p. 61] and the absence of expansion in graphemic transcriptions is presented as a "hybrid" practice by D. Stutzmann [Stutzmann, 2011, p. 251]. However, HTR data production and edition should be seen as two different tasks, and moreover, transcription and abbreviation resolution should be seen as two different computational tasks.

[7]The use of ligatures, often involving a long *s*, has slowed down the use of accents [Biedermann-Pasques, 1992, p. 92].

[8]Not all ligatures are present in the unicode standard or in MUFI: the task would therefore have been too complicated for a very limited interest.

[9]In the dataset, some folders are named with `mufi`: they include a richer use of unicode character. These folders weren't use for training purposes but were used to evaluate the weight of a wider transcription of allographs.

About spacing. Spacing is a problem because the compositor can "pack" the words so that they all fit into the line space. It is therefore typographical information that must be treated with care, but in most cases we follow the graphemic approach, which tends to distinguish units grammatically rather than graphically, while retaining period peculiarities.

IV EXPERIMENTS

General set-up. In order to train and evaluate models, we use a regular 80% of the produced dataset for training, 10% for development purposes and 10% for evaluation. The split is produced at the level of each print, resulting *de facto* as a *in-domain* test.

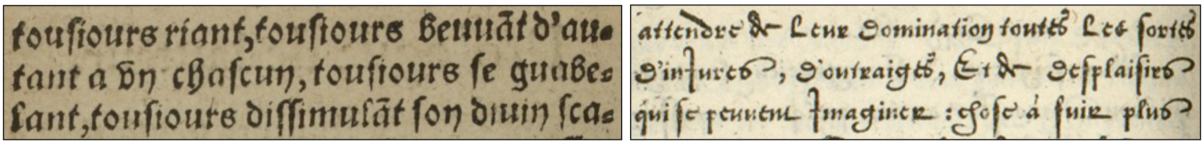
Category	Description	Status	Transcription	Example
Ligature	Ligature O+E ‹œ›	Graphetic	U+0153/U+0152	cœur
Ligature	Ligature A+E ‹æ›	Graphetic	U+00E6/U+00C6	Cæ- Ægyſte
Ligature	Ligature long S+T ‹ſt›	Graphemic	No ligature	Chreſtienté.
Ligature	Ligature L+L ‹ll›	Graphemic	No ligature	eternelle
Ligature	Ligature C+T ‹ct›	Graphemic	No ligature	Edict?
Ligature	Ligature S+P ‹sp›	Graphemic	No ligature	l'eſprit
Ligature	Ligature long S+L ‹ſl›	Graphemic	No ligature	meſlent
Ligature	Ligature U+S ‹us›	Graphemic	No ligature	indiuidus
Ligature	Ligature S+I ‹ſi›	Graphemic	No ligature	quaſi
Ligature	Ligature long S+long S ‹ſſ›	Graphemic	No ligature	ieuneſſe
Ligature	Ligature F+F+I ‹ffi›	Graphemic	No ligature	ſuffiſance
Ligature	Ligature I+S ‹is›	Graphemic	No ligature	puis
Allograph	Capital E	Graphemic	U+0045	Et Et
Allograph	Capital A	Graphemic	U+0041	A Au
Allograph	Capital M	Graphemic	U+004D	Moy Mais Mais
Allograph	Small E with long finial	Graphemic	U+0065	eſpece
Allograph	Tittle as tilde or dot	Regularised		je jugé
Allograph	Small long and short S	Graphetic	U+017F	ſans
Abbreviation	Combining tilde ‹◌̃›	Graphetic	U+0303	d'autãt deuõs
Abbreviation	Combining Macron ‹◌̄›	Graphemic	U+0303	vãteric
Abbreviation	Ampersand ‹&›	Graphetic	U+0026	&
Diacritics	Combining vertical line ‹◌̍›	Regularised		à
Diacritics	Apostrophe	Graphetic	U+0027	A RMÉES
Diacritics	Comma	Graphetic	U+002C	F R A N Ç O I S
Hyphenation	Hyphen	Codified ‹¬›	U+00AC	ſoigneu¬ imbecile, di-

Table 4: Main transcription choices

We additionally produced 4 others small samples for out-of-domain testing based on centuries, from the 16th c. to the 19th (*cf.* tab. 10,11, 12 for details, tab. 5 otherwise). We specifically designed these out-of-domain samples to exclude gothic and other special fonts such as *civilités* ones (*cf.* fig. 6), as our training corpora only include roman or italic typefaces.

Dataset	Caracters
17th c.	91,104
16th c.	18,542
18th c.	16,691
19th c.	13,103

Table 5: Description of test sets, character counts are in NFC.



(a) Rabelais, *Garguantua*, 1535, gothic typeface (b) Trissino, *Sophonisba*, 1559, civilité typeface

Figure 6: Non-selected typfaces

Two separate open-source OCR engines are used for training OCR models, namely *Kraken* [Kiessling, 2019] and *Calamari* [Wick et al., 2020]. Both tools were used in order to leverage their various differences in order to produce the best model possible. Default engine model architecture as well as hyperparameters were used for the baseline model.

Kraken Experiment: artificial lines vs. synthetic data. Texts were normalised using unicode's decomposition normalisation (NFD). This results in splitting characters such as ‹é› into two characters ‹e› + ‹◌́› (combining acute accent, `U+0301`). This has become in the French DH community of Kraken the *de facto* choice for French language OCR.

To improve the efficiency of the engine, two additional experiments have made. On the one hand, we tested a larger model architecture than the base one[10], doubling the filter size of each convolutional layer, respectively from 32 to 64 and from 64 to 128, to handle the heterogeneity of the training data. On the other hand, we used a synthetic training set on top of the manually compiled one with 27 different fonts[11].

Calamari Experiment: Multiple voters and data augmentation. Regarding *Calamari*, we have tested another type of unicode normalisation (NFC) making sure that diacritics are combined[12]. We replicated here the successful protocol from Reul et al. [2018] by combining model fine-tuning (FT) – *i.e.* building from existing models (historical non-French *Antiqua*) instead of starting the training from scratch –, voting (VT) – *i.e.* training five models instead of one and combining their outputs during predictions –, and data augmentation (DA), – *i.e.* generating modified images of the input lines by blurring them, stretching them, etc.

[10] Base VGSL architecture of Kraken recognition model: `[1,48,0,1 Cr3,3,32 Do0.1,2 Mp2,2 Cr3,3,64 Do0.1,2 Mp2,2 S1(1x0)1,3 Lbx100 Do]`.

[11] Namely: IM FELL English SC, IM FELL English, IM FELL Great Primer, IM FELL Double Pica, IM FELL Double Pica SC, IM FELL DW Pica, 1592 GLC Garamond, 1689GLCGaramondW00SC-Norm, Garamond, EB Garamond, EB Garamond 12 All SC, 1689 Almanach, Fournier MT Std, Bodoni 72 Oldstyle, Didot, Chapbook, DTLElzevirS, DTL Elzevir, P22 Operina Romano, Hultog, JSL Ancient, Old Claude LP Std, Chapbook, 1756DutchW01-Normal, 1726RealEspanolaW01-Rg, 1776_Independence, Palatino.

[12] As a result, score between both engines are not comparable, as they do not use the same unicode normalisation which results in a different number of evaluated characters.

V ANALYSIS

Augm. Architect.	Artif. Data	17th c.	16th c.	18th c.	19th c.
-	-	97.47%	97.74%	**97.78%**	94.50%
Yes	-	**97.92%**	**98.06%**	**97.78%**	94.23%
-	Yes	96.65%	97.26%	97.74%	**95.50%**
Yes	Yes	97.26%	97.68%	97.84%	94.84%

Table 6: Accuracy (1-CER) for the experiment with *Kraken*.

Pretrain.	Voters	Data Augm.	17th c.	16th c.	18th c.	19th c.
-	1	-	98.47%	98.14%	98.27%	93.11%
Yes	1	Yes	98.76%	98.49%	96.47%	**97.05%**
Yes	5	Yes	**99.05%**	**98.68%**	**98.78%**	**97.05%**

Table 7: Accuracy (1-CER) for the experiment with *Calamari*

Considering the (deliberately) extreme heterogeneity of our data, such scores are promising (cf. tab. 6 &7). However, it is clear that, regarding Kraken, synthetic data did not improve results at all, except for 19th c., and might actually in some cases lowered the score (specifically for the in-domain test). Kraken however benefited from a larger model, and this change impacted also out of-domain results except from later one (18th and 19th c.). *Calamari* shows again that the protocol from Reul et al. [2018] is beneficial to the results and incremental (multiple voters enhance the results of the already better ones from data augmentation and pretraining).

Despite being focused on the 17th c., the dataset is able to produce model resistant to changes in neighbouring centuries. We see that in both case, the accuracy drops by less than one percentage point. This is definitely due to the filtering of gothic fonts and special typefaces of the 16th c. prints, but also to the limited changes in common typefaces between these centuries. As for 19th century, the score dropped more for Kraken (-2 to -3 percentage points) than for Calamari (-1.7 points) for their best performing models on other centuries. Only the use of artificial data allowed for performance gains on 19th c. for Kraken, most probably due to the regularity provided by them.

Confusion table from *Calamari* (cf. tab. 14, 13, 15, 16, 17) shows an important issue with spacing recognition, and as such, word segmentation. This could be linked to both the density of the composition in early modern prints or because of the instability of the graphic segmentation, some words being sometimes welded (*puisque*) and sometimes not (*puis que*, cf. tab. 8).

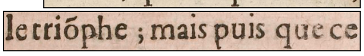
Spacing error	Example	Transcription
Composition		lemẽt , qu'il oſe `[NO SPACE]` pretendre
Graphic segmentation		le triõphe ; mais puis `[SPACE]` que ce

Table 8: Possible sources of word segmentation errors. The token `SPACE` indicates the problematic zone.

Another important source of error is linked to the ‹ſ›, which, once again, can be linked to paleographic problems (confusion ‹ſ›/‹f›). The confusion ‹ſ›/‹s› might be related to the language model overtaking the OCR, or more simply, input errors on the side of the GT[13].

Both issues can be treated with post-processing steps. Segmentation or ‹ſ›/‹s› confusion can be approached as a character classification from a pure natural language processing point of view, as shown per Clérice [2020]. In this paper, content were encoded at the character level with a per-character binary classification (word boundary vs. in-word content) which resulted in very high

[13]As a reviewer kindly said, "muscle memory" is sometimes quite strong.

accuracy. The same process could be applied to both type of confusions. However, regarding the ‹ſ›/‹s› confusion, another option would be to drop the differentiation of the allographs at training and testing time, until enough GT has been produced to avoid this kind of issues.

VI FUTURE WORK

The diachronic efficiency of the model can be improved by adding data for more recent prints: c. 20,000 additional lines will be added, to carry further tests on the creation of a model for French prints in general, and not only modern prints. Out-of-domain tests sets composed of non-francophone prints should also be created to test the efficiency of the model on similar prints in other languages.

While creating the GT, we have corrected the layout of each image. Alto and PageXML will be used to train a segmenter, the importance of which must not be underestimated since it is on its result that the OCR is performed.

AUTHOR CONTRIBUTIONS

S.G. designed the research project, built the corpus and prepared the data for training. T.C. helped all along the process providing advice, scripts and feedback. C. R. performed the experiments with *Calamari*. All authors discussed the results and contributed to the final manuscript.

DATA

Training data is available online (10.5281/zenodo.3826894): it contains all the GT used to train models, and it is distributed with a CC-BY licence. Ongoing research on OCR, with additional data and scripts, is available on Github (`https://github.com/e-ditiones/OCR17`).

ACKNOWLEDGMENTS

This paper would not have been possible without the help of J.-B. Camps (PSL-ENC).

ADDENDUM

This article has been originally written and submitted in 2020, and it was corrected in 2023. While we received reviews early, we were not able to complete the correction proposed in due time. We still think that this paper paints an interesting state of OCR and HTR at the time.

However, since then, tools have evolved, and the dataset has evolved has well. It became the OCR17+ dataset[14], using ALTO-XML representation instead of line based segmentation. It has been largely completed since then through project such as GalliCorpora[15]. Kraken has since then adopted augmentation of images, and uses larger line input rather than higher convolution filters to reach better results.

References

Douglas Biber. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257, January 1993. URL `https://academic.oup.com/dsh/article/8/4/243/928942`.

Liselotte Biedermann-Pasques. *Les grands courants orthographiques au XVIIe siècle et la formation de l'orthographe moderne. Impacts matériels, interférences phoniques, théories et pratiques (1606–1736)*. De Gruyter, Berlin, Boston, 1992. URL `https://doi.org/10.1515/9783110938593`.

[14]`https://github.com/Heresta/OCR17plus`.

[15]`https://github.com/Gallicorpora`

Pascale Bourgain and Françoise Vieillard. *Conseils pour l'édition des textes médiévaux*, volume 3, Textes littéraires. Ecole nationale des chartes - Comité des travaux historiques et scientifiques, Paris, 2001.

Alix Chagué, Thibault Clérice, and Floriane Chiffoleau. HTR united, a centralization effort of htr and ocr ground-truth repositories for french languages, 2020. URL https://github.com/HTR-United/htr-united.

Thibault Clérice. Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin. *Journal of Data Mining and Digital Humanities*, 2020. URL https://hal.archives-ouvertes.fr/hal-02154122.

Unicode Consortium. Unicode 12.0.0, 2019. URL http://www.unicode.org/versions/Unicode12.0.0.

Steve Crowdy. Spoken Corpus Design. *Literary and Linguistic Computing*, 8(4):259–265, January 1993. URL https://academic.oup.com/dsh/article/8/4/259/928943.

Frédéric Duval. Les éditions de textes du XVIIe siècle. In *Manuel de la philologie de l'édition*, pages 369–394. De Gruyter, Berlin, Boston, 2015. URL https://www.degruyter.com/view/books/9783110302608/9783110302608-017/9783110302608-017.xml.

Frédéric Duval, editor. *Pratiques philologiques en Europe*. Études et rencontres. Publications de l'École nationale des chartes, Paris, September 2018. URL http://books.openedition.org/enc/692.

Elizabeth L. Eisenstein. *Grub Street Abroad: Aspects of the French Cosmopolitan Press from the Age of Louis XIV to the French Revolution*. Clarendon Press, Oxford, 1992.

Simon Gabay. Pourquoi moderniser l'orthographe? Principes d'ecdotique et littérature du XVIIe siècle. *Vox Romanica*, 73(1):27–42, 2014. URL https://elibrary.narr.digital/article/99.125005/vox201410027.

Odd Einar Haugen. *MUFI character recommendation v. 4.0*. Medieval Unicode Font Initiative, December 2015. URL https://bora.uib.no/handle/1956/10699.

Benjamin Kiessling. Kraken - an Universal Text Recognizer for the Humanities. In *Digital Humanities Conference 2019 - DH2019*, Utrecht, The Netherlands, July 2019. ADHO. URL https://dev.clariah.nl/files/dh2019/boa/0673.html.

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

Christos Papadopoulos, Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. The impact dataset of historical document images. In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, HIP '13, page 123–130, New York, NY, USA, 2013. Association for Computing Machinery. URL https://doi.org/10.1145/2501115.2501130.

Sophie Raineri and Camille Debras. Corpora and Representativeness: Where to go from now? *CogniTextes. Revue de l'Association française de linguistique cognitive*, 19, 2019. URL http://journals.openedition.org/cognitextes/1311.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. Improving OCR Accuracy on Early Printed Books by combining Pretraining, Voting, and Active Learning. *Journal for Language Technology and Computational Linguistics*, 33(1: Special Issue on Automatic Text and Layout Recognition), 2018. URL https://doi.org/10.21248/jlcl.33.2018.216.

Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences*, 9(22), January 2019. URL https://www.mdpi.com/2076-3417/9/22/4853.

Peter Robinson and Elizabeth Solopova. Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue. In *The Canterbury Tales Project Occasional Papers 1*, 1993.

Ray Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, Curitiba, Brazil, 2007. URL https://doi.org/10.1109/ICDAR.2007.4376991.

Miriam Speyer. Les dieux écrivent-ils en italiques ? Typographie et mise en livre de pièces en vers et en prose. *L'Habillage du livre et du texte aux XVIIe et XVIIIe siècles*, 9, 2019. URL https://hal-normandie-univ.archives-ouvertes.fr/hal-02184237.

Uwe Springmann and Anke Lüdeling. OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly*, 11-2, 2017. URL http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html.

Uwe Springmann, Florian Fink, and Klaus U. Schulz. Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings. *arXiv:1606.05157 [cs]*, June 2016. URL http://arxiv.org/abs/1606.05157.

Uwe Springmann, Christian Reul, Stefanie Dipper, and Johannes Baiter. Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. *Journal for Language Technology and Computational Linguistics*, 33:97–114, 2018. URL https://arxiv.org/abs/1809.05501.

Dominique Stutzmann. Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? *Kodikologie und Paläo-graphie im digitalen Zeitalter = Codicology and Palaeography in the Digital Age*, 2:247–277, 2011. URL https://halshs.archives-ouvertes.fr/halshs-00596970.

Isabella H. van Eeghen. *De Amsterdamse Boekhandel 1680-1725*. Publicaties van de Gemeentelijke Archiefdienst. Scheltema & Holkema, Amsterdam, 1960–1978.

Christoph Wick, Christian Reul, and Frank Puppe. Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *Digital Humanities Quarterly*, 14(1), 2020. URL http://www.digitalhumanities.org/dhq/vol/14/2/000451/000451.html.

Author	Title	Date	Place	Publisher	Printer	Lines	Ligatures	DPI	Size	Library	ID
Ellain	Advis sur la peste	1606	Paris	D. Douceur		617	No	400	1496×2560	BNF	cb303981499
Regnier	Les Satyres	1612	Paris	Toussaint Du Bray		198	Yes	400	1562×2580	BNF	cb31189430j
Gournay	Egalité des hommes et	1622				824	No	400	1666×2634	BNF	cb30529274x
de Viau	Œuvres	1623	Paris	J. Quesnel		851	No	400	1334×2600	BNF	cb34804166g
Balzac	Lettres	1624	Paris	Toussaint Du Bray		1014	No	72	4267×6667	BNF	cb300515241
Descartes	Discours de la méthode	1637	Leiden		J. Maire	431	No	72	2479×3508	BNF	cb30328384x
Scudéry	L'Amour tirannique	1639	Paris	A. Courbé	M. Brunet + J. de La Coste	860	No	72	4267×5513	BNF	cb31341723p
L'Hermite	La Mariane	1639	Paris	A. Courbé	M. Brunet	673	No	72	3796×5860	BNF	cb39333461s
L'Hermite	Panthée	1639	Paris	A. Courbé	M. Brunet	897	No	72	4267×6100	BNF	cb314972698
Rotrou	La Belle Alphrede	1639	Paris	A. de Sommaville	A. Coulon	298	Yes	400	2570×3695	BNF	cb31251853x
Scudéry	Ibrahim	1641	Paris	A. de Sommaville		1671	No	72	2479×3508	BNF	cb31341849n
Fr. de Sales	Introduction à la vie	1641	Paris	Imprimerie royale		617	No	400	4213×6084	BNF	cb30460001n
Scarron	Typhon	1644	Paris	T. Quinet		196	Yes	400	2746×3608	BNF	cb31308401d
Scarron	Le Jodelet	1645	Paris	T. Quinet	A. Coulon	268	Yes	400	2652×3424	BNF	cb31308475z
Pascal	Expériences nouvelles	1647	Paris	P. Margat		775	No	400	1684×2637	BNF	cb31062878c
Voiture	Oeuvres	1650	Paris	A. Courbé		359	Yes	400	2794×3729	BNF	cb31600370j
Scudéry	Clélie	1656	Paris	A. Courbé		897	No	72	2479×3508	BNF	cb31341819q
Chapelain	La Pucelle	1656	Paris	A. Courbé		752	No	400	4504×6589	BNF	cb365764947
Pascal	Traictez de l'équilibre	1663	Paris	G. Desprez		972	No	600	2083×3634	BNF	cb31081848m
Molière	L'Escole des femmes	1663	Paris	L. Billaine	J. Hénault + Cl. Blageart	1074	No	72	4058×6923	BNF	cb30958651f
Bussy-Rabutin	Histoire amoureuse	1665	Bruxelles	Fr. Foppens		876	No	72	4267×7542	BNF	cb36117831r
Molière	George Dandin	1669	Paris	J. Ribou	Cl. Audinet	1323	No	72	4042×7200	BNF	cb30958651f
Racine	Les Plaideurs	1669	Paris	Cl. Barbin	Cl. Blageart	894	No	72	4109×7643	BNF	cb311693885
Racine	Oeuvres, t. 2	1676	Paris	J. Ribou	J.-B. (I) Coignard	1309	No	72	4267×7783	BNF	cb31168676r
Racine	Oeuvres, t. 1	1676	Paris	J. Ribou	J.-B. (I) Coignard	561	No	72	4267×7821	BNF	cb31168676r
La Fayette	Princesse de Clèves	1678	Paris	Cl. Barbin		948	No	72	4267×7186	BNF	cb307135973
Racine	Oeuvres, t. 1	1679	Paris	D. Thierry	D. Thierry	1810	No	72	3767×6583	BSB	BV012474970
Pradon	Statira	1680	Paris	J. Ribou	Cl. Blageart	1053	No	72	4085×6956	BNF	cb311463583
Papin	La Manière d'amolir	1682	Paris	E. Michallet		547	No	400	1468×2426	BNF	cb31056545b
Bossuet	Oraison funebre	1683	Paris	S. Mabre-Cramoisy		769	No	400	3320×4584	BNF	cb36575655n
Donneau de Vizé	Voyage des ambassadeurs	1686	Paris	Au Palais		161	Yes	415	1500×2416	BNF	cb303484582
La Bruyère	Caractères	1688	Paris	E. Michallet		806	No	72	4267×7258	BNF	cb31452154x
Molière	Dom Garcie de Navarre	1694	Bruxelles	G. de Backer		723	No	72	1006×1768	ÖNB	AC10132063
Boyer	Méduse	1697	Paris	Académie de mus.	C. Ballard	886	No	72	3854×5485	BNF	cb30152139c
Pradon	Oeuvres	1697	Paris	Th. Guillain	Ch. Journel	932	No	72	4080×6924	BNF	cb38652730w
Racine	Oeuvres, t. 1	1697	Paris	D. Thierry	D. Thierry	1046	No	72	2457×2149	BNW	7805546
Bussy-Rabutin	Mémoires, t. 1	1698	Paris	J. Anisson		122	Yes	400	3128×4036	BNF	cb393648983

Table 9: Training data

Author	Title	Date	Place	Publisher	Lines	Ligatures	DPI	Size	Library	ID
Bartas	La Sepmaine	1578	Paris	M. Gadoulleau	62	No	415	2840×3880	BNF	cb303572930
Beroalde	Avantures de Floride	1594	Tours	J. Mettayer	63	No	400	1174×2186	BNF	cb30092726b
Calvin	Institution de la religion	1562	Geneve	G. Bourgeois	54	No	400	2633×4078	BNF	cb365761545
Du Bellay	La Defence et illustration	1549	Paris	A. l'Angelier	58	No	400	1589×2445	BNF	cb11968311h
Du Fail	Discours d'Eutrapel	1585	Rennes	N. Glamet	65	No	416	1596×2576	BNF	cb30367435k
Rabelais	Tiers Livre	1546	Paris	Ch. Wechel	46	No	400	1573×2647	BNF	cb31167405f
Ronsard	Les Amours	1552	Paris	Vve de la Porte	59	No	400	1678×2711	BNF	cb432409623

Table 10: Testing data, 16th c.

Author	Title	Date	Place	Publisher	Lines	Ligatures	DPI	Size	Library	ID
Buffon	Histoire naturelle	1750	La Haye	P. De Hondt	112	No	400	3756×4582	BNF	cb301741874
Laclos	De la Monarchie	1791	Paris	Impr. nationale	79	No	400	2274×3430	BNF	cb302389989
Diderot	Essais sur la peinture	1785	Paris	Fr. Buisson	56	No	400	1881×2903	BNF	cb44312299p
Marivaux	Le jeu de l'amour et du hazard	1730	Paris	Briasson	59	No	400	1440×2609	BNF	cb30886471g
Montesquieu	Lettres persanes	1721	Amsterdam		48	No	400	1452×2588	BNF	cb119437548
Rousseau	Les Pensées	1764	Amsterdam		60	No	400	1664×2904	BNF	cb31257216h
Voltaire	Zadig	1748	Paris/Nancy	L.-Fr Prault/A. Leseure	44	No	600	2022×3676	BNF	cb316044160

Table 11: Testing data, 18th c.

Author	Title	Date	Place	Publisher	Lines	Ligatures	DPI	Size	Library	ID
Chateaubriand	Atala	1801	Paris	Migneret/Dupont	41	No	600	1914x3280	BNF	cb30227639h
Constant	Adolphe	1816	Paris/Londres	Treuttel et Würtz/H. Colburn	40	No	400	1560×2653	BNF	cb319643212
Flaubert	Salammbô	1863	Geneve	M. Lévy frères	61	No	400	2264×3348	BNF	cb304403988
Gautier	Le Roman de la momie	1858	Paris	L. Hachette	57	No	400	1605×2739	BNF	cb30490246s
Hugo	Odes	1823	Paris	Persan/Pélicier	42	No	600	2208×3656	BNF	cb32263200h
Musset	A quoi rêvent les jeunes filles	1833	Paris	E. Renduel	59	No	400	1858×3193	BNF	cb30999539c
Nerval	Scènes de la vie orientale	1848	Paris	F. Sartorius	53	No	400	2300×3643	BNF	cb32482331k

Table 12: Testing data, 19th c.

GT	PRED	COUNT	PERCENT
{ }	{}	39	4.36%
{é}	{e}	22	2.46%
{,}	{.}	22	2.46%
{l}	{}	21	2.35%
{f}	{ſ}	20	2.23%
{}	{ }	19	2.12%
{i}	{}	14	1.56%
{t}	{}	13	1.45%
{’}	{}	12	1.34%
{c}	{e}	10	1.12%

Table 13: Confusion table for the best Calamari models, in-domain test, 17th c. prints

GT	PRED	COUNT	PERCENT
{ }	{}	33	12.69%
{’}	{}	14	5.38%
{f}	{ſ}	11	4.23%
{}	{ }	11	4.23%
{é}	{e}	5	1.92%
{s}	{ſ}	5	1.92%
{,}	{}	5	1.92%
{e}	{é}	5	1.92%
{.}	{}	4	1.54%
{a}	{à}	3	1.15%

Table 14: Confusion table for the best Calamari models, out-of-domain test, 16th c. prints

GT	PRED	COUNT	PERCENT
{ }	{}	60	25.75%
{s}	{ſ}	14	6.01%
{è}	{e}	9	3.86%
{è}	{é}	7	3.00%
{u}	{n}	4	1.72%
{c}	{e}	4	1.72%
{è}	{ê}	4	1.72%
{f}	{ſ}	4	1.72%
{-}	{¬}	3	1.29%
{à}	{a}	3	1.29%

Table 15: Confusion table for the best Calamari models, out-of-domain test, 18th c. prints

GT	PRED	COUNT	PERCENT
{ }	{}	51	13.18%
{t}	{l}	32	8.27%
{f}	{ſ}	21	5.43%
{»}	{n}	19	4.91%
{è}	{é}	18	4.65%
{è}	{e}	13	3.36%
{e}	{o}	12	3.10%
{— }	{}	11	5.68%
{c}	{e}	8	2.07%
{» }	{}	7	3.62%

Table 16: Confusion table for the best Calamari models, out-of-domain test, 19th c. prints

GT	PRED	COUNT	PERCENT
{ }	{}	144	16.36%
{f}	{ſ}	36	4.09%
{t}	{l}	33	3.75%
{è}	{é}	25	2.84%
{è}	{e}	23	2.61%
{»}	{n}	20	2.27%
{s}	{ſ}	19	2.16%
{}	{ }	18	2.05%
{’}	{}	14	1.59%
{e}	{o}	12	1.36%

Table 17: Confusion table for the best Calamari models, out-of-domain test, 16th c., 18th c., 19th c. prints