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

Simon Gabay, Thibault Clérice, Christian Reul

Journal of Data Mining and Digital Humanities

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

Simon Gabay[1], **Thibault Clérice**[2], **Christian Reul**[3]

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

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

Abstract

Machine learning starts with machine teaching: in the following paper, we present the data that we have gathered and created to train reliable OCR models for 17th c. French prints, and preliminary results based on these training data and experiments to improve them.

Keywords

OCR, 17th c French, training data, corpus building

I INTRODUCTION

The reliability of OCR models depends on both the quantity and the quality of training data. So, on the one hand, *quantity* needs to be produced and made freely available to other scholars. 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].

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 Truth (GT) made of c. 30,000 lines taken from 37 French prints of the 17th c. (tab. 6), following strict philological 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., 2018]/*OCR4all* [Reul et al., 2019].

II CORPUS BUILDING

Creating training data 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. Since the apparition 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 extrapaleo-

graphical criteria. Two main categories have been used: bibliographical (printing date and place, literary genre, author) and computational (size and resolution of the images) information.

Regarding dates, prints are diachronically distributed over the century, with a special attention for books printed between 1620 and 1700 (tab. 1) because it is our period of interest. 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.

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

Decade	Total
00's	1
10's	1
20's	3
30's	5
40's	5
50's	3
60's	5
70's	4
80's	5
90's	5

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

As we can see, the corpus is not balanced, since not only literary texts, but also plays are clearly over-represented. Such a choice has been made for two reasons. On the one hand, dramatic texts are versified, which means that they tend to be printed in italics at the beginning of the 17th c. [Speyer, 2019]. On the other hand, they traditionally use capital letters to indicate the name of the speaker, which is an easy way to increase the amount of such rarer glyphs (fig. 1 and tab. 3). Such a strategy should help us deal with highly complex layouts (fig. 2).

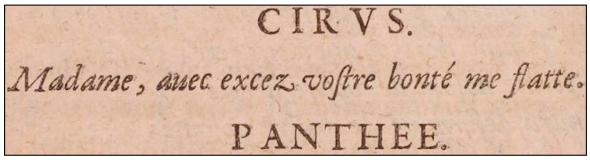
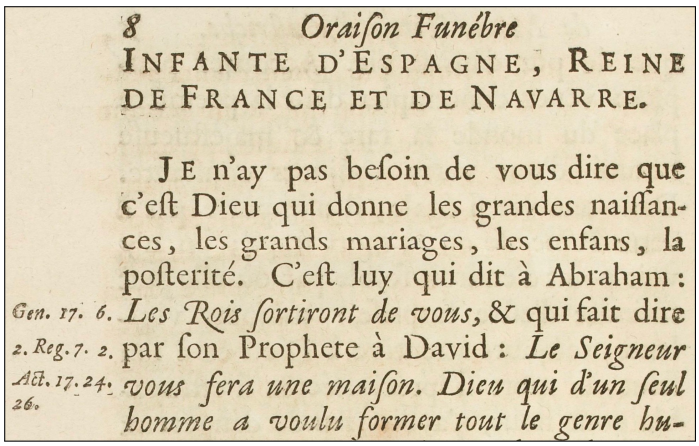

CIRVS.
Madame, auec excez voſtre bonté me flatte.
PANTHEE.

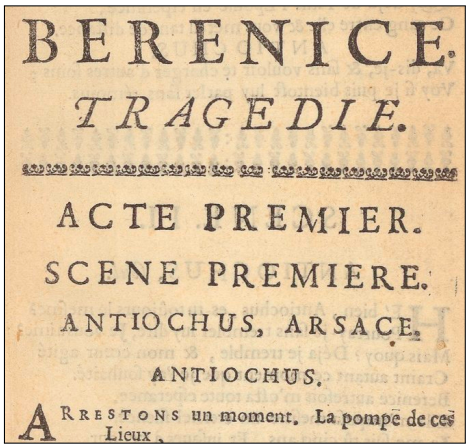
Figure 1 – Tristan L'Hermite, *Panthée*, 1639

	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



(a) Bossuet, *Oraison*, 1683

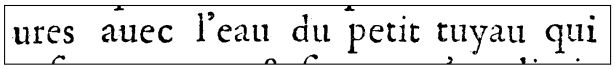


(b) Racine, *Oeuvres*, 1676

Figure 2 – Mixing fonts, heights, weight, layouts in prints

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, which introduces significant changes in the shape of letters (fig. 4) that our model needs to handle properly.

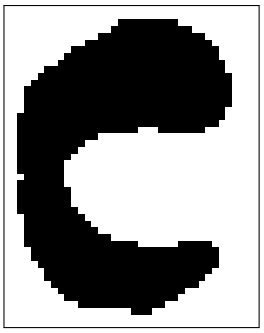
(a) Balzac, *Lettres*, 1624, 72 dpi



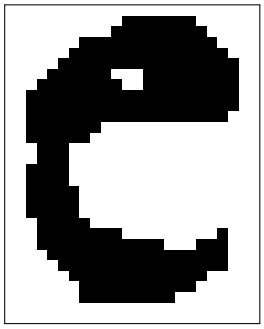
(b) Pascal, *Equilibre*, 1663, 400 dpi

(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*

III TRANSCRIPTION RULES

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]:

- Graphetic (or allographetic) transcription
  - Graphic: every information (space, layout, letters. . .) in the manuscript is preserved.
  - Graphetic: every distinct letter-type is distinguished (*e.g.* two glyphs ‹s› and ‹ſ› for one single grapheme *s*).
- Graphemic transcription
  - Graphemic: manuscript spelling is preserved (*estoit* and not *était*).

– Regularized: all manuscript spellings are regularized to a particular norm (in our case, contemporary French).

Our transcription guidelines have been the following: encode as much information as possible, as long as it is available in unicode [Consortium, 2019]. The result is therefore a mix between graphetic and graphemic transcription.

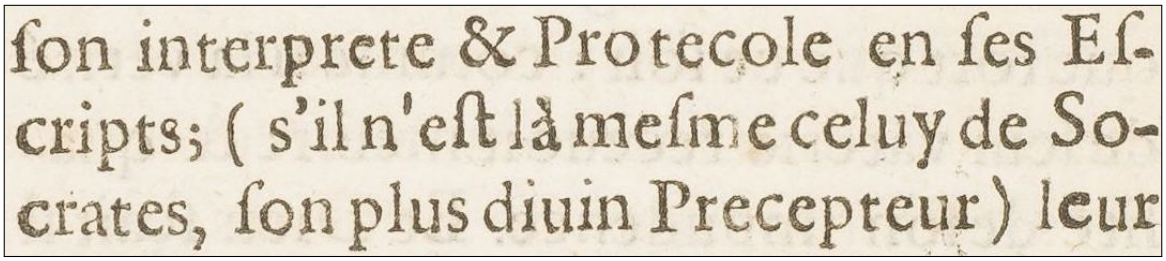


Figure 5 – Excerpt of Marie de Gournay, *Egalité*, 1622

In practice (fig. 5), it means that we do not dissimilate ‹u›/‹v› (*diuin*) or ‹i›/‹j›, we do not normalise accents (*interprete* and not *interprète*), we keep historical, diactritical (*Efcripts* and not *Ecrits*) or calligraphic letters (*celuy* and not *celui*). We keep the long *s* (*meſme* and not *mesme*), but most of the other allographetic variations are not encoded (cf. fig 6).

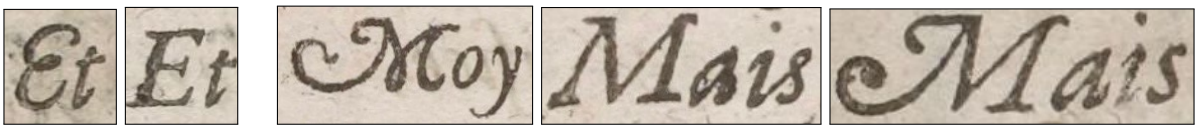
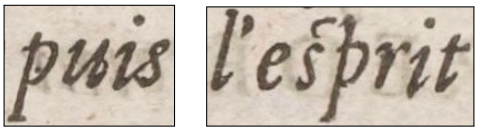


Figure 6 – Examples of ignored allographetic variants, Rotrou, *Alphrede*, 1639

One exception has been made to our unicode rule: aesthetic ligatures that still exist in French (‹œ› vs ‹oe›) have been encoded, but not those that have disappeared despite their existence in unicode (*e.g.* ‹ﬅ›). For the latter, a subsidiary model, using both unicode and MUFI [Haugen, 2015] and based on a limited amount of data, has been trained for testing purposes – excluding, for compatibility reasons, non-MUFI and non-unicode ligatures (fig. 7 and tab. 4).

Ligature	Description	Unicode	Mufi
œ	Latin Small Ligature O E	U+0153	-
st	Latin Small Ligature S T	U+FB06	-
ft	Latin Small Ligature Long S T	U+FB05	-
ct	Latin Small Ligature C T	-	EEC5
sp	Latin Small Ligature S P	-	-
is	Latin Small Ligature I S	-	-

Table 4 – Ligatures

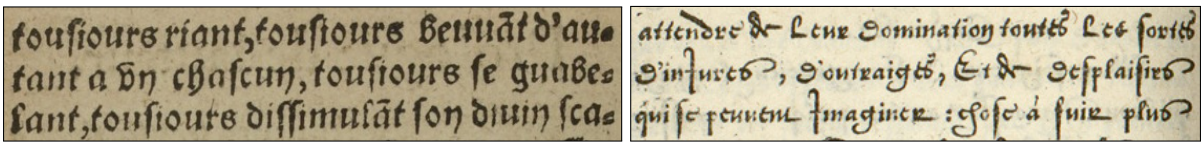


(a) ‹is› (b) ‹sp›

Figure 7 – Examples of ignored ligatures, Rotrou, *Alphrède*, 1639

IV EXPERIMENTS

A general model has been trained on the entire set (tab. 6). Our basic setup is the following: black and white binarisation, NFD unicode normalization, 80% of the data for training, 10% to guide the training. 10% of each print has been reserved for in-domain testing, and additional (yet limited) GT has been produced for out-of-domain testing with small samples of 16th c (tab. 7), 18th (tab. 8) and 19th c. (tab. 9) prints to test the generality of our model – only with roman or italic typefaces, and no gothic (tab. 8(a)) or other special fonts such as *civilité* ones (tab. 8(b)).



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

Figure 8 – Non-selected typfaces

As previously mentioned, two open-source OCR engines have been used, *Kraken* [Kiessling, 2019] and *Calamari* [Wick et al., 2018]. On top of training a model using the default setup regarding the network structure, training parameters. . . , several modifications, have been tested to maximize the final scores.

With *Kraken*, we try to double 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. We have also tested data augmentation by creating artificial GT with the following fonts:

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

With *Calamari*, we have first tested another type of unicode normalization (NFC). Repeating a successful protocol [Reul et al., 2018], we have also tried combining pretraining, *i.e.* building from existing models (historical non-French *Antiqua*) instead of starting the training from scratch, voting, *i.e.* training five models instead of one and combining their outputs during predictions, and data augmentation, which appears to be the most successful setup, not only with more than 99% for the in-domain tests, but also similar results for modern prints.

Model	Test	16th c. prints	18th c. prints	19th c. prints
KRAKEN				
Basic model	97.47%	97.74%	**97.78**%	94.50%
with enlarged network	**97.92**%	**98.06**%	**97.78**%	94.23%
+ artificial data	96.65%	97.26%	97.74%	**95.50**%
with enlarged network	97.26%	97.68%	97.84%	94.84%
CALAMARI				
Basic model	98.38%	98.19%	97.91%	92.7%
Basic model (NFC)	98.47%	98.14%	98.27%	93.11%
Single voter+PT+DA (NFC)	98.76%	98.49%	96.47%	**97.05**%
5 voters+PT+DA (NFC)	**99.05**%	**98.68**%	**98.78**%	**97.05**%

Table 5 – Scores

A small model (1,355 lines for training, cf. tab. 6), trained with *Kraken* on transcriptions including Unicode and MUFI (tab. 4) ligatures reached 99.23% of accuracy.

V ANALYSIS

Considering the (deliberately) extreme heterogeneity of our data, such scores are promising. When looking at the confusion tables (cf. tab. 11, 12, 13) we first realise that a model trained on 17th c. prints performs reasonably well on texts printed during other centuries. Such results open the possibility to create a general model for French prints (but not incunabula), whatever the printing date.

Regarding the errors produced, it is interesting to note that the most important one, by far, regards the segmentation of words, *i.e.* the incapacity of the model to predict spaces where they should be. This could be linked to both the quality of the composition in early modern prints (cf. 9(a): *lemẽt, qu'il* [SPACE] *oſe* [SPACE] *pretendre*) or because of the instability of the graphic system (cf. 9(b): *puis* [SPACE] *que* vs *puisque*).

(a) Scudéry, *Ibrahim*, 1641 (b) Balzac, *Lettres*, 1624

Figure 9 – Possible sources of segmentation errors

Another important source of error is linked to the ‹ſ›, which, ones again, can be linked to paleographic problems (confusion ‹ſ›/‹f›) as well as linguistic ones (confusion ‹ſ›/‹s›).

For both problems, simple solutions exist. Regarding segmentation, the problem is comparable to the medieval one of the *scripta continua*, for which a solution already exists [Clérice, 2020]. Regarding ‹ſ›, we could use additional training data drawn from more contemporary prints (without ‹ſ›), or just simplify the training by using only ‹s›.

VI FURTHER WORK

The diachronical 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-SA 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).

References

Douglas Biber. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257, January 1993. ISSN 0268-1145. doi: 10.1093/llc/8.4.243. URL https://academic.oup.com/dsh/article/8/4/243/928942. Publisher: Oxford Academic.

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, April 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. ISSN 0268-1145. doi: 10.1093/llc/8.4.259. URL https://academic.oup.com/dsh/article/8/4/259/928943. Publisher: Oxford Academic.

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. ISBN 978-3-11-030260-8. doi: 10.1515/9783110302608-017. 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. ISBN 978-2-35723-123-8. URL http://books.openedition.org/enc/692.

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

Odd Einar Haugen. *MUFI character recommendation v. 4.0*. Medieval Unicode Font Initiative, December 2015. ISBN 978-82-8088-413-8. URL https://bora.uib.no/handle/1956/10699. Accepted: 2015-12-04T09:05:29Z.

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. doi: 10.1109/ICDARW.2019.10032.

Sophie Raineri and Camille Debras. Corpora and Representativeness: Where to go from now? *CogniTextes. Revue de l'Association française de linguistique cognitive*, 19, June 2019. ISSN 1958-5322. URL http://journals.openedition.org/cognitextes/1311. Number: Volume 19 Publisher: Association française de linguistique cognitive (AFLiCo).

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. Improving OCR Accuracy on Early Printed Books by utilizing Cross Fold Training and Voting. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 423–428, April 2018. doi: 10.1109/DAS.2018.30. URL http://arxiv.org/abs/1711.09670. arXiv: 1711.09670.

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):4853, January 2019. doi: 10.3390/app9224853. 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.

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

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. arXiv: 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(1):97–114, 2018. ISSN 2190-6858. 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.

Christoph Wick, Christian Reul, and Frank Puppe. Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *Computing Research Repository*, July 2018. URL http://arxiv.org/abs/1807.02004. arXiv: 1807.02004.

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

Table 6 – 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 7 – 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 8 – 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 9 – 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 10 – Confusion table, 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 11 – Confusion table, 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 12 – Confusion table, 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 13 – Confusion table, 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 14 – Confusion table, out-of-domain test, 16th c., 18th c., 19th c. prints

Dataset	caracters	hidden errors
17th c.	91104	78.55%
16th c.	18542	63.08%
18th c.	16691	51.93%
19th c.	13103	45.74%
all Ood c.	48336	60.91%

Table 15 – Description of train sets