The birth of French orthography. A computational analysis of French spelling systems in diachrony

Simon Gabay, Thibault Clérice

The birth of French orthography. A computational analysis of French spelling systems in diachrony*

Simon Gabay[1,*], Thibault Clérice[1,*]

[2]*Université de Genève*
[1]*Inria Centre de Recherche de Paris*

Abstract
The 17th c. is crucial for the French language, as it sees the creation of a strict orthographic norm that largely persists to this day. Despite its significance, the history of spelling systems remains however an overlooked area in linguistics for two reasons. On the one hand, spelling is made up of micro-changes which requires a quantitative approach, and on the other hand, no corpus is available due to the interventions of editors in almost all the texts already available. In this paper, we therefore propose a new corpus allowing such a study, as well as the extraction and analysis tools necessary for our research. By comparing the text extracted with OCR and a version automatically aligned with contemporary French spelling, we extract the variant zones, we categorise these variants, and we study their frequency to study the (ortho)graphic change during the 17th century.

Keywords
Computational linguistics, History of orthography, Information extraction, Corpus building

1. Introduction

The grapho-phonetic aspects of French during the in 17th c. paradoxically remain very poorly known, despite the importance of the graphematic question at this period, which saw the appearance of the French orthography[1]. Rather than the actual practice of scriptors, it is the depth of theoretical debates on spelling that has until now concentrated most of research (e.g. [1] or [2]), and the notebooks of Mezeray [3] or the *Remarques* of Vaugelas [4, 5] still remain among the main sources used, rather than statistical surveys on vast corpora.

If the various dialects and other *scriptae* populating Old and Middle French have been abundantly described (e.g. in [6]), just like the "orthographie" of the Renaissance (to quote the term used by Baddeley [7]), the slow imposition of an orthographic norm thourghout modern times, although a major phenomenon in the history of a language as prescriptive as French, remains a blind spot in diachronic linguistics. How has the French that we know today supplanted its various modern variations?

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark
*Corresponding author.
simon.gabay@unige.ch (S. Gabay); thibault.clerice@inria.fr (T. Clérice)
https://cv.hal.science/simon-gabay (S. Gabay); https://cv.hal.science/thibault-clerice (T. Clérice)
0000-0001-9094-4475 (S. Gabay); 0000-0003-1852-9204 (T. Clérice)

[1]In this article, we will distinguish between "spelling systems" and "orthography". To simplify, the first are coherent and competing logics of spelling words (as for the manuscripts with dialectal traits of the Middle Ages), the second is a strict norm which is recognised as a standard.

One of the main technical challenges for carrying out such a study relies on the existence of important amounts of data, in order to guarantee quantitatively the reliability of the results. Unfortunately, such corpora of classical French do not exist for two reasons. On the one hand, as Cl. Vachon [8, p. 32, n. 31] bitterly experienced, text editors got into the habit of standardising the language of that era [9, 10], which makes its study particularly complicated, if not impossible to use for graphematic studies. On the other hand, the few corpora that have been created, like that of Cl. Vachon, but also others like that of the *Réseau Corpus Français Préclassique et Classique* (RCFC) [11] are not, or not in full, available to researchers. Given the ever-increasing quantities of data necessary for computational studies, it is however dubious that these two corpora, even freely accessible, would in any case remain insufficient for the most recent approaches proposed in NLP.

This paper proposes to return to the history of the French *vêtement graphique* ("graphic clothing") in a computational way. We introduce a two-step approach: first, a unique corpus creation pipeline meticulously extracts spelling information from digital facsimiles. This pipeline includes a layout analysis model to distinguish text from paratext on the page, an OCR model that retains the historical character ‹ſ› pivotal to written French, and a linguistic normalizer that "translate" historical French into its contemporary counterpart. In the second step, we analyze the created corpus using a comparison algorithm that matches the extracted historical text with its modern equivalent. This enables us to pinpoint significant variations, categorize these differences, and uncover detailed trends throughout the 17th century. This methodological framework not only enhances our understanding of historical French orthography but also proposes a new approach for computational linguistic studies of spelling variations.

2. State of the art

Corpus building from OCR has long been a task in digital humanities and corpus linguistics. Initially deemed unsuitable for historical sources in 1993 [12], OCR gained credibility in the late 1990s for corpus building, including XML TEI formalisation in commercial projects such as the *Patrologia Latina Database*, and for Ancient Greek scripts in the 2010s [13]. Most project using TEI, such as the First1KGreek project, relied on manual formalisation of the text's logical structure [14], as manual work was considered essential for accuracy. The advent of user-friendly OCR and HTR technologies has spurred interest in automatic document formalisation (ADF), primarily focused on facsimile formalisation [15] and noisy text removal with tools based on vocabularies such as SegmOnto [16], which standardises the identification of paratextual zones (running titles, footnotes, etc.). Few projects, however, have utilised font, geometric, and textual features to reconstruct or emulate the original text structure from born-digital PDFs or OCR outputs. *PaperXML* [17] demonstrated such transformation but was limited to the ACL Anthology structure. Grobid [18] and Grobid Dictionaries [19, 20] employed geometric, font, and textual features to produce XML TEI output, though they were specific to scientific papers and dictionaries. In 2022, visual features outperformed linguistic ones in document formalisation, with YOLO models using the SegmOnto controlled vocabulary surpassing LayoutLM models in multilingual settings [21]. Recently, research has started on OCR output formalisation for corpus building with a controlled vocabulary and a training dataset

for models [22, 23]. Lastly, the Layout Analysis Dataset with SegmOnto (LADaS) [24] allowed a much finer granularity in the analysis, and a significant improvement of the entire pipeline for the automatic creation of files encoded in XML-TEI that goes beyond fac-simile approach and closer to reproducing the logical structure of the text.

Linguistic Normalisation (LN) has a long history, dating back to the 80's [25], but has developed itself as derived task from Machine Translation (MT) in the beginning of the 2010's, usually to improve downstream tasks in the pipeline such as linguistic annotation [26]. LN share important similarities with MT, and therefore relies on the same methods, but with a slightly different objective: to "translate" a source into another state of the language, usually more recent (16th c. German→ contemporary German), rather than into another language (Italian → German). Resources existed first for Slovene, German, English, Hungarian, Spanish, Swedish, Portuguese [27], but several studies have recently improved both resources [28] and techniques for historical French, first comparing rule-based, statistical and neural methods [29], and then alignment-based and neural MT-approaches [30].

Computational scriptology is based on the notion of *scripta*, coined by Remacle [31] and widely used in Romanistics to to distinguish a spoken language (the dialect) and a written language (the *scripta*). The first studies on dialectometry date back from the early 70's with the pioneer work of Jean Séguy, who invented the term *dialectométrie* [32], on the distance between dialects in vast corpora [31]. Since then, two main schools, based in Salzburg [33] and Groningen [34], have advanced research on the topic, but relying mainly on geographical data to localise dialects. In parallel to these research, Cl. Vachon has changed the approach, switching to corpus-based research, using historical data to study semi-automatically the spelling [8], and more recently, J.-B. Camps has shifted the method, using unsupervised stylometry to categorise medieval *scriptae* [35]. Regarding modern French, alternative studies have proposed alignement-based approaches to compare the historical source and an automatically normalised version to detect the evolution of spellings [36] or to categorise documents [37].

3. Corpus building

3.1. Data

For practical reasons, a first corpus of limited size (c. 600 texts) spanning from the 17th c. was produced with our pipeline. The data comes from the *Gallica* digital library and contains only French-language documents. For our experiment, we have selected only plays, which offer medium size documents (compared to novels, potentially much longer), and linguistically homogeneous data (spelling can influenced by the type of document, such as legal documents which tend to use more "archaic" traits and may involve Latin phrases).

3.2. Method

Our pipeline allows us to extract data, enrich it and store it in a standard format (cf. fig. 1). Firstly we apply a layout analysis model specialised in theatrical data trained for the occasion,

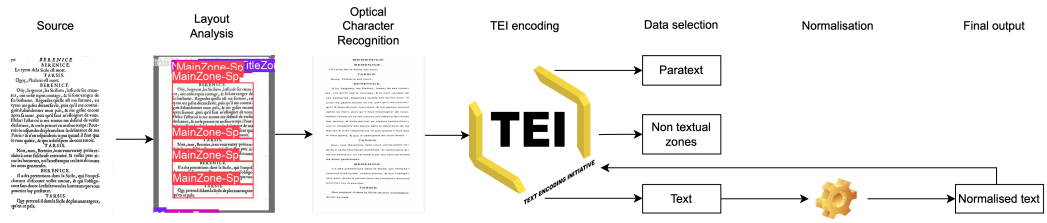


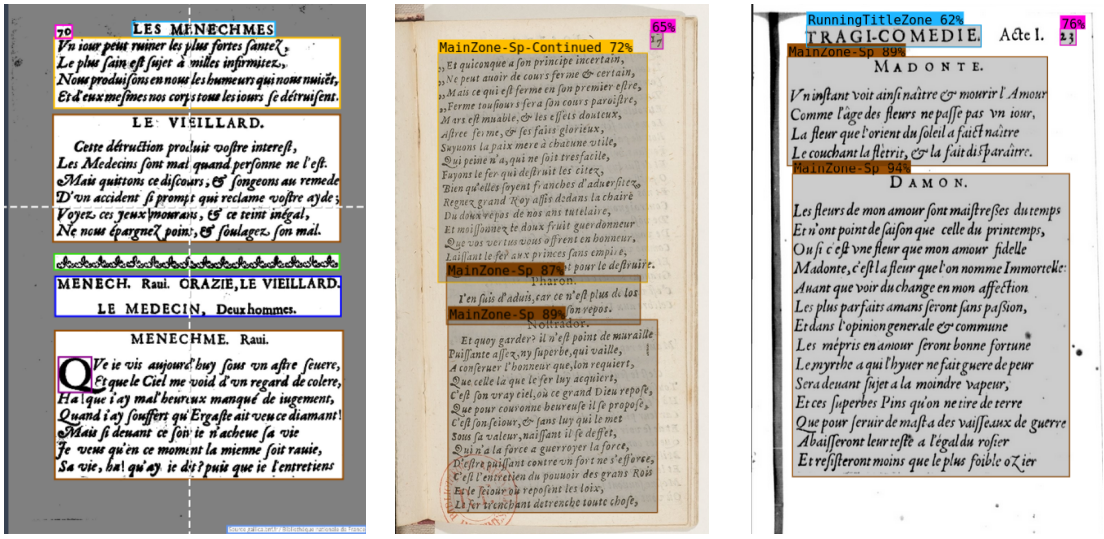
Figure 1: Data production pipeline.

then we use an OCR model prepared for this study which preserves the long *s* (‹ſ›). Based on the layout analysis we convert ALTO files to TEI files. Only textual data which contains text of the work (paragraph, speech, verse, etc.), and not linked to the organisation of the book (running title, page number, notebook number, etc.) is extracted and normalised automatically, before to be reintroduced into the TEI file.

Layout analysis. Based on the results of Najem-Meyer and Romanello [21] and the initial evaluation of YOLO region segmenter against Kraken's [38] as a region segmenter with YALTAi [39], we proposed to evaluate the ability of YOLOv8 [40] to detect regions in our 17th c. print corpus. For this purpose, we annotated one random image from each digitised version of our corpus, which could include empty pages (e.g., bookbinding, cover) and full pages. This resulted in a corpus of 620 images for training, evaluation, and testing. Our final corpus comprises 32 null pages (without annotations) and a variety of annotations, with a majority of speech-related tags (`MainZone:SP`, `MainZone:SP#Continued`), paratextual-related objects (e.g., `NumberingZone`, `RunningTitleZone`), a smaller number of logical structuring features such as scene titles (`MainZone:Head`) and cast lists (`MainZone:Entry`), as well as a few paragraphs and poetic excerpts, mainly found in incipits or prefaces of the books (`MainZone:P`), as seen in tab. 1.

Table 1
Training and evaluation data for Layout Analysis across the datasets. Each image represent a single document in the dataset.

	Train	Dev	Test
Images	497	61	62
MainZone-Sp	1738	219	187
NumberingZone	384	49	45
RunningTitleZone	373	43	41
DigitizationArtefactZone	189	23	24
QuireMarksZone	183	29	15
MainZone-Head	159	28	33
MainZone-Sp-Continued	154	16	18
DropCapitalZone	136	23	23
GraphicZone-Decoration	130	16	22
MainZone-Entry	89	6	19
MainZone-Lg	58	5	1
MainZone-P	41	10	12
MainZone-P-Continued	28	1	2
MarginTextZone-ManuscriptAddendum	21	3	7
MarginTextZone-Notes	30	1	0
StampZone-Sticker	24	1	2
MainZone-Other	19	3	2
StampZone	13	3	5
TitlePageZone	8	2	4
GraphicZone	5	0	1
MainZone-Incipit	4	1	0
MainZone-Signature	1	0	0



(a) Groundtruth. (b) Prediction. (c) Prediction.

Figure 2: Three page examples with zone objects.

Optical character recognition. Since YOLO is well integrated within YALTAi, which in turn works seamlessly with Kraken, we decided to use the latter to train a new OCR model that includes the long *s* (‹ſ›). Kraken, unlike other OCR system, avoids the integration of a strong language model which in turns, for our purpose, allows for keeping more variations. This new model, derived from CATMuS Print [41], uses three datasets for fine-tuning [42, 43, 44] and one evaluation dataset [45] (cf. tab. 2). We evaluate on a test set that includes data spanning three centuries (from the 16th to the 18th) and comprises one page from 10 different documents for each period.

Table 2
Training and evaluation data for OCR.

Dataset	Century	Language	Books	Lines
Train/Dev	16	French	7	17817
Train/Dev	17	French	19	20267
Train/Dev	16	Latin	12	10648
Test	16	French	10	
Test	17	French	10	
Test	18	French	10	

TEI Document production. Document formalisation follows a logical approach based on the ALTO output produced by Kraken and YALTAi, rather than a neural one. Each region is processed in reading order, with regions not matching `MainZone` being ignored, except for the "default" region, which handles orphan lines. The default region is placed into a `<fw>` ("forme work") tag, which is typically excluded from our text export processes. Regions marked as `#Continued` are logically merged with previous ones. Each line is prepended by a TEI `<lb/>` (line beginning) tag to facilitate back-to-document correction capabilities. Hyphenisation is resolved by removing hyphen but keeping the `<lb/>` tag at its place. While machine learning is employed for initial region detection, the formalisation process itself does not involve any

learned behaviour. Metadata are systematically integrated in the <teiHeader>, using information automatically retrieved from the catalogue of the French National Library via the ark ID.

Linguistic normalisation. All documents are processed via a normaliser previously trained[2]. Only text contained in <p> and <sp> ("speech") elements are kept for normalisation, because a specific spelling variation occurring in the running title, for instance, would be repeated every two page and potentially alter artificially the result of the scriptometric analysis. The text is split into sentences (ending by a full stop, an exclamation or a question mark) or subsentences (ending by a colon or a semicolon), all stored in a <seg> ("arbitrary segment") element, with the source text in <orig> ("original form") and the automatically normalised text in <reg> ("regularization"). The normalised version is evaluated against a dictionary of modern French to control the quality of the final product.

3.3. Experimental Setup and Evaluation

Layout analysis. We evaluate two possible setups: both use fine-tuning with the original YOLOv8L models and an input image size of 960 pixels (higher than the default). One setup uses only the dataset produced in the context of this paper, while the other merges this dataset with the larger LADaS dataset (5,000 images). We train both setups for 100 epochs with otherwise default parameters.

Table 3
Results of the YOLO model on modern plays.

			Theatrical corpus				Theatrical and LADaS corpus			
Class	Images	Instances	Box(P	R	mAP50	mAP50-95)	Box(P	R	mAP50	mAP50-95)
all	62	463	0.824	0.705	0.768	0.626	0.739	0.738	**0.8**	0.666
MainZone-Entry	2	19	0.456	0.048	0.463	0.244	0.857	0.316	**0.76**	0.486
MainZone-Head	24	33	0.915	0.697	**0.825**	0.698	0.854	0.532	0.722	0.587
MainZone-Lg	1	1	0.74	1	**0.995**	0.895	0.807	1	**0.995**	0.895
MainZone-Other	2	2	1	0	**0.174**	0.139	0.0427	0.107	0.105	0.0732
MainZone-P	5	12	0.549	0.711	**0.66**	0.474	0.655	0.25	0.55	0.49
MainZone-P-Continued	2	2	0.92	1	**0.995**	0.946	0.385	1	**0.995**	0.995
MainZone-Sp	41	187	0.967	0.979	**0.988**	0.941	0.955	0.973	0.982	0.924
MainZone-Sp-Continued	18	18	1	0.891	**0.995**	0.93	0.978	1	0.955	0.969

Since our study focuses exclusively on the MainZone, which contains the primary text and excludes all paratextual elements (such as decorations, page numbers, and running titles), we have concentrated our evaluation on this specific zone. Overall, when considering all classes, we found that integrating our data with the LADaS corpus yields improved results (0.768 vs. 0.8). However, for the most critical classes (Sp and Sp-continued), the model trained exclusively with theatrical data produces slightly better outcomes. As previously mentioned, these are the classes essential for our study.

[2]https://huggingface.co/rbawden/modern_french_normalisation.

Text recognition. To fine-tune and adapt the CATMuS Print OCR model to the allographic variation of s/long-s, we modified the classifier codec (`--resize new` mode) and used a standard learning rate of 0.0001, along with a batch size of 32. This logical approach ensures the model is fine tuned to the specific typographic variations without relying on any learned behaviour during the formalisation process. We compare this approach to a model without fine-tuning, trained from scratch, with the same architecture (cf. tab. 4), revealing the superiority of the approach with fine tuning.

Table 4
Character and word error rates for both models.

Models	Characters	Errors	CER	WER
No fine-tuning	38394	924	2.41	11.06
Fine-tuning	38394	649	1.69	8.34

Table 5
Character and word error rates for both models.

% errors	CER (part)	Errors	Correct	Generated
8.78%	0.14%	57	SPACE	
7.55%	0.13%	49	'	’
6.62%	0.11%	43	s	ſ
3.23%	0.05%	21	–	Ø
2.77%	0.05%	18	ſ	f
2.62%	0.04%	17	’	'
2.16%	0.04%	14	Ø	SPACE
2%	0.03%	13	1	l
2%	0.03%	13	.	Ø
1.85%	0.03%	12	◌	Ø
1.69%	0.03%	11	,	.
1.54%	0.03%	10	0	o
1.54%	0.03%	10	t	r
1.54%	0.03%	10	◌	Ø

Most of the errors are errors related to poor segmentation of the text (cf. tab. 5), in which there should be a space that is missing from the prediction – a classic error for ancient prints. The prediction errors regarding two types of apostrophes (curved or straight) are of little concern because they do not affect the result from a linguistic point of view and are due to poor data preparation that is easily correctable. The confusion between the rounded *s* and the long *s* is likely attributable to the fine-tuning process and the absence of the long *s* in the base model.

Linguistic normalisation To evaluate the results of the normalisation, we compare the prediction of the normaliser with a dictionary of contemporary French to obtain a Word Accuracy (WAcc). Results are satisfactory (cf. fig. 3), with a median above 90%. Texts with a WAcc under 80% are removed to avoid using unreliable data.

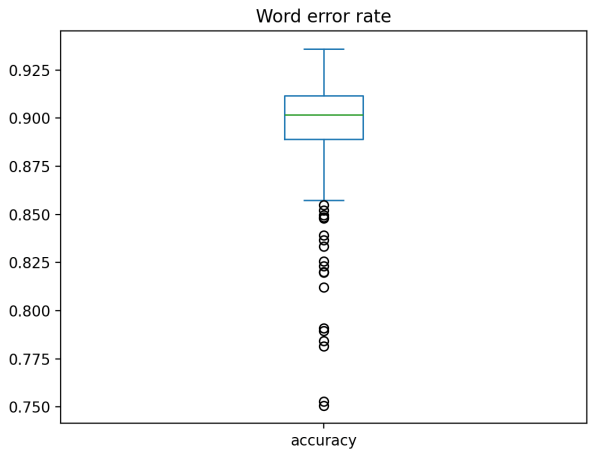


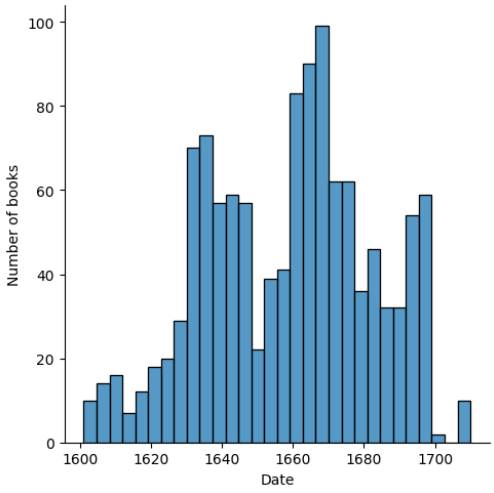
Figure 3: Word error rate for the corpus.

3.4. Result dataset

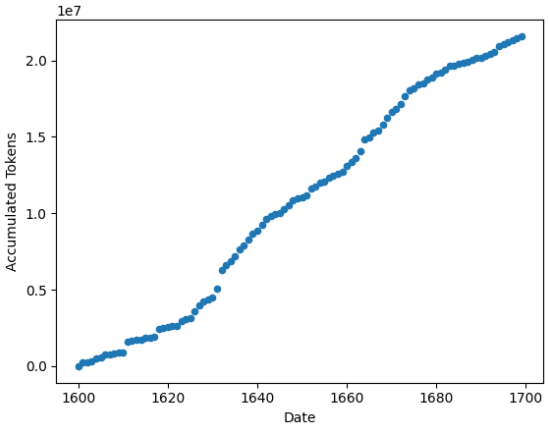
The final dataset is made of around 80,000 pages for 620 documents. While the number of unit is uneven over the years (cf. fig. 5a), the accumulated tokens are progressing evenly (cf. fig. 5b). An example of our TEI encoding is presented in fig. 4

```
<sp>
   <ab>
      <seg>
         <orig>SGANARELLE.</orig>
         <reg>SGANARELLE.</reg>
      </seg>
      <seg>
         <orig>Promettez-moy donc, Seigneur Geronimo, de me parler avec toute forte de franchife.</orig>
         <reg>Promettez-moi donc, Seigneur Geronimo, de me parler avec toute sorte de franchise.</reg>
      </seg>
   </ab>
</sp>
<sp>
   <ab>
      <seg>
         <orig>GERONIMO.</orig>
         <reg>GERONIMO.</reg>
      </seg>
      <seg>
         <orig>Ie vous le promets.</orig>
         <reg>Je vous le promets.</reg>
      </seg>
   </ab>
</sp>
```

Figure 4: Example of TEI encoding with normalisation.



(a) Number of bibliographical units per year. The bin around 1720 represent printed books from within the 17th century but with unclear or imprecise printing dates.



(b) Accumulated tokens over the years. Year of printing is used for the date, any document without a precise date are removed from the plot. Tokens are taken from the original OCR documents, only from the `MainZones`.

Figure 5: Description of the OCRised corpus.

4. Evaluation of spelling variation

4.1. Method

We use the ABA [46] tool to precisely identify the portions of words which differ between the original version and the normalised version, and group similar differences, for example having the same historical-linguistic origin, or the same type of operations in terms of addition, deletion or modification of characters. Each `<orig>` and `<reg>` of the corpus is split into words, the punctuation is removed, and then the original and normalised versions are aligned at the word level using the Needleman-Wunsch [47] algorithm, using the Levenshtein distance [48] between each pair of words in the same `<seg>` in the original and normalised version[3].

Secondly, for each of the aligned word pairs, the original version and the normalised version are aligned at the character level, still using the Needleman-Wunsch algorithm, but using a specific substitution matrix to allow not only identical letters to be aligned, but also letters considered close in (pre)classical French and contemporary French (presence/absence of diacritic, ligatures...). For example, while identical letters benefit from a substitution score of 4, letters differing only in accent or cedilla benefit from a score of 2, as do ‹ſ› and ‹s› or ‹s› and ‹ß› for example. Other pairs of letters benefit from a score of 1, such as ‹u› and ‹v›, ‹s› and ‹z› or even ‹n› and ‹m ›. Conversely, a score of -1 is assigned to pairs of distinct letters not subject to such exceptions, as well as to the deletion or insertion of a character.

This execution of the Needleman-Wunsch algorithm to obtain character-level alignment is illustrated in the matrix in tab. 6, where each number represents the similarity score of the best alignment found between the prefix of ‹Apoſtre› and ‹Apôtre› up to this box. It is preceded by an arrow indicating which box to come from to obtain this best alignment. For example, to obtain the best alignment between ‹Apoſ› and ‹Apô›, we must consider the best alignment between ‹Apo› and ‹Apô› (which has a score of 10) then make an insertion of *ſ*, which has a score of -1, which provides a total score of 9. If we had preferred to first consider the best alignment between ‹Apoſ› and ‹Ap›, which has a score of 6, then delete the *ô*, which has a score of -1, we would have obtained an alignment with a score of 5, therefore lower than optimal. In case of insertion or deletion during this alignment step, we use the ¤ character in order to obtain two words of the same length in both the original and normalised version. Thus, at the end of this second alignment step, the word *Apoſtre* in the original version is matched with *apô¤tre* in a normalised version to obtain character-by-character alignment.

Table 6: Prefix similarity matrix for the original and normalised version of ‹Apoſtre›. The arrows indicate the previous box on the optimal path to calculate the similarity between two prefixes, one from the word on the first row, the other from the word in the first column. On this optimal path, green indicates equality, red indicates substitution, and blue indicates deletion.

	A	p	o	ſ	t	r	e
A	↘ 4	→ 3	→ 2	→ 1	→ 0	→ -1	→ -2
p	↓ 3	↘ 8	→ 7	→ 6	→ 5	→ 4	→ 3
ô	↓ 2	↓ 7	↘ 10	→ 9	→ 8	→ 7	→ 6
t	↓ 1	↓ 6	↓ 9	↓ 8	↘ 13	→ 12	→ 11
r	↓ 0	↓ 5	↓ 8	↓ 7	↓ 12	↘ 17	→ 16
e	↓ -1	↓ 4	↓ 7	↓ 6	↓ 11	↓ 16	↘ 21

Finally, for each word in the corpus, its original and normalised versions are analysed, char-

[3]Some subtleties are brought to this adjustment, such as *et* and *&* which are considered equivalent.

acter by character, to detect, in the case of different characters at the same position, the normalisation rule that applies, or to signal that no existing rule was identified when appropriate. 72 rules were defined based on the bibliography and the differences observed in the gold FrEEM$_{norm}$ parallel corpus [28]. For example, the rule *Ramist letter* is detected if an ‹i›, a ‹j›, an ‹u› or a ‹v› is present in the associated original word respectively to a ‹j›, an ‹i›, a ‹v› or an ‹u› in the normalised version.

4.2. Results

Based on the alignments obtained using the Needleman-Wunsch algorithm and the detections of the 72 rules mentioned earlier, our analysis reveals four distinctive patterns of historical spelling changes. The principle underlying this analysis is straightforward: if a normalization rule is detected less frequently, it indicates that the historical spelling it targets is becoming less prevalent in the corpus. To examine its evolution throughout the century, we normalize the total number of rule applications to its percentage within each text. For instance, the etymological spelling ‹gn›, found in form *cognoitre* (<lat. cognoscere), is less and less replace by ‹nn› (today *connaître*, eng. “to know”), signifying the slow disappearance of this spelling (cf. fig. 6).

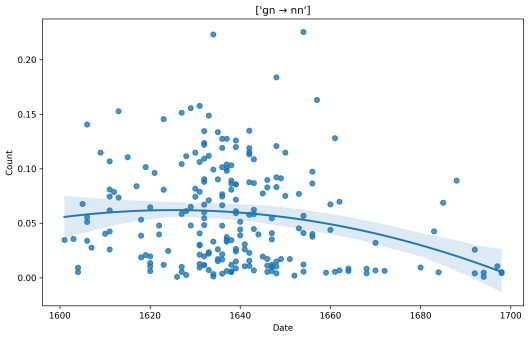
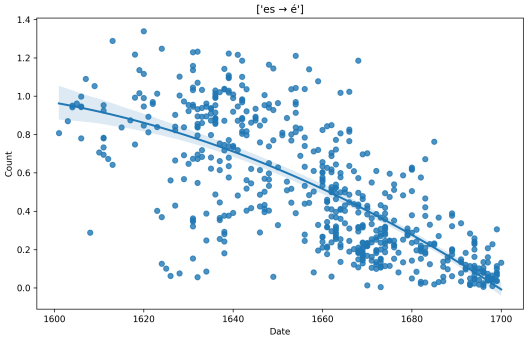
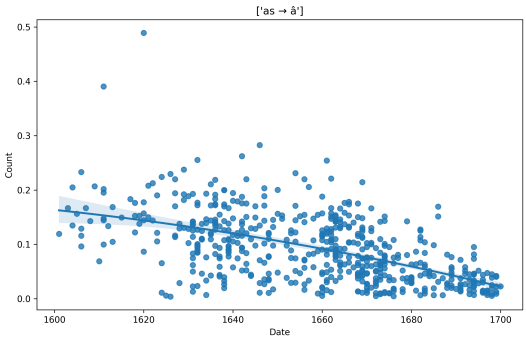


Figure 6: Disappearance of ‹gn›.



(a) Substitution of ‹es› by ‹é›.

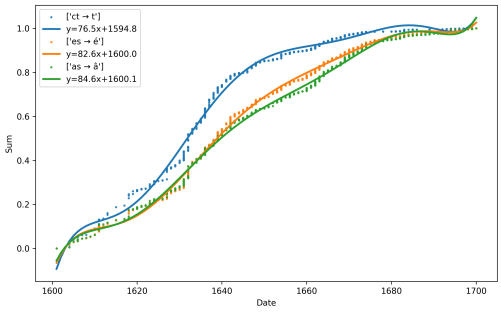


(b) Substitution of ‹as› by ‹â›.

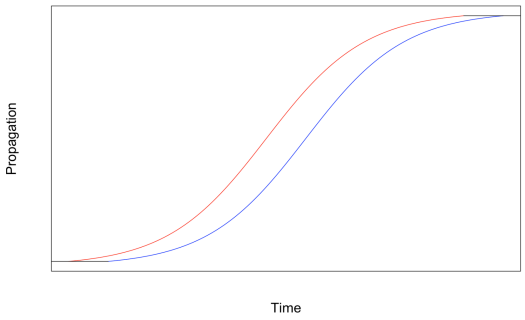
Figure 7: Disappearance of ‹s› as a diacritical letter.

Pattern A: constant rate. Using ABA, it is possible to detect more complex traits of historical graphics systems than the specific use of a single letter (e.g. ‹u› vs ‹v› as a vowel) or a group of letters (‹gn› vs ‹nn›), such as the presence of a diacritical letter to change the sound-value of the letter to which it is added (e.g. vowel + ‹s›). In historical French, the phoneme [e] is thus regularly noted with the grapheme ‹es› where today we use ‹é› (*estat* vs *état*, eng. “state”), and the phoneme [ä] is noted ‹as› where we now find ‹â› (*pasturage* vs *pâturage*, eng “pasture”). If counting the presence of ‹v› followed by consonant (‹vne› = [yn]) to identify the historical use

of ‹v› is enough, it is impossible to count the occurrences of ‹es› to measure the presence of a diacritical *s* (‹esponge› → ‹éponge›, eng. “sponge”, but ‹espagnol› → ‹espagnol› and not ‹épagnol›, eng. “spanish”): the transition from a complex grapheme (such as a digraph) to a simple grapheme requires an alignment at the character level of the original text and its normalised version, and then the deduction of the spelling change from the difference between the two.



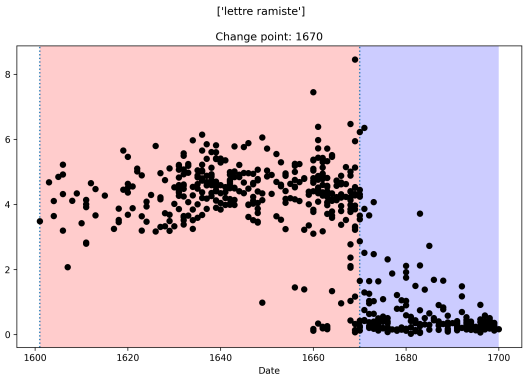
(a) Accumulation of occurrences of different spellings: two similar (es→é, as→â) and one different (ct→t). Data are scaled to base 100 to be comparable.



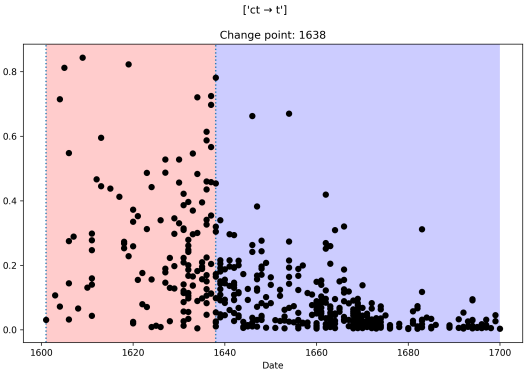
(b) Theoretical progression of two similar variants over time, which start at different times, but progress at the same speed, according to the constant rate hypothesis.

Figure 8: The constant rate hypothesis in practice (left) vs in theory (right).

In our corpus, we detect a clear decrease in the use of complex graphemes with a diacritical *s*, whether the latter is combined with ‹e› (cf. fig. 7a) or with ‹a› (cf. fig. 7b). Interestingly, the propagation of these two new spellings (vowel+accent) does occur at a very similar speed (cf. fig. 8a), recalling Kroch’s constant rate hypothesis (cf. fig. 8b)[4], of which researchers have already found traces in syntactic [50] and phonological [51] change.



(a) Apparition of the contemporary use of ramist letters.



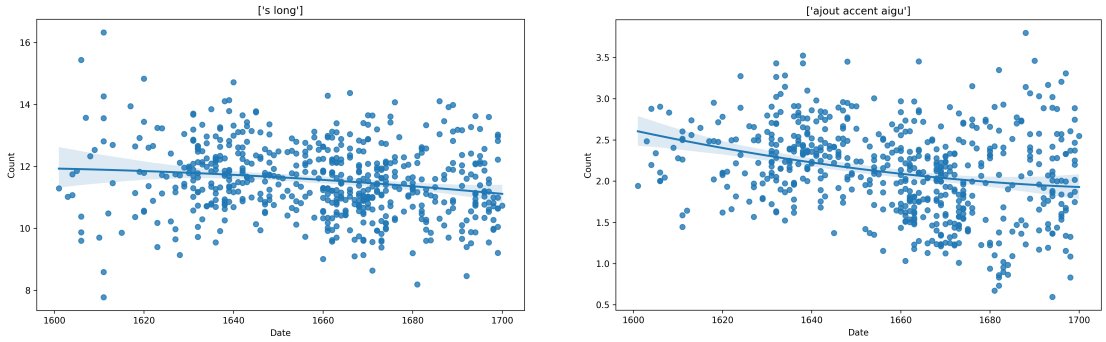
(b) Disappearance of the etymological combination ‹ct›.

Figure 9: Computing the change-point of two changes of spelling.

[4]“When one grammatical option replaces another with which it is in competition across a set of linguistic contexts, the rate of replacement, properly measured, is the same in all of them.” [49]

Pattern B: abrupt change. On the basis of such observations, it is however possible to go further and date the moment when a break occurs in the scribal practice, to date the moment when the spelling changes. To do so, we can use binary segmentation (BS) [52, 53], an algorithm using a forward stepwise method, to identify change-point detection. This method has already been used in diachronic linguistic to study the sudden introduction of new lexical items [54].

One of the main discoveries of our study is the extremely abrupt nature of certain changes, which take place at very high speed, such as the disappearance between 1668 and 1672 of Ramist letters (cf. fig. 9a), as proposed by Christophe Plantin in the 16th c. [55] and defended by Pierre Corneille in his foreword *au lecteur* of 1663 [56]. A similar phenomenon, although slightly less abrupt, exists for the disappearance of the etymological ‹c› followed by ‹t› (e.g. ‹faict›<FACTUM, today *fait*, eng. “fact”) at the end of the 1630s (cf. fig. 9b).



(a) Slow decrease of the long *s* (‹ſ›).

(b) Increase of the acute accent.

Figure 10: Long *s* vs acute accent.

Pattern C: correlation. L. Biedermann-Pasques proposed as one of the parameters for spelling change the type case available: “the typographical use of the ligature has slowed down, in our opinion, the regular replacement of silent *s* by an accent” [57, p. 92]. It is indeed faster to compose the word *eſtoit* with the ligature (e+ſt+o+i+t=4 characters) than without (e+ſ|s+t+o+i+t=5 characters). One could argue that switching to the accented letter also requires only four characters (é+t+o+i+t), but if the ligatures are present in number in the printer’s type case, the accented characters are less so.

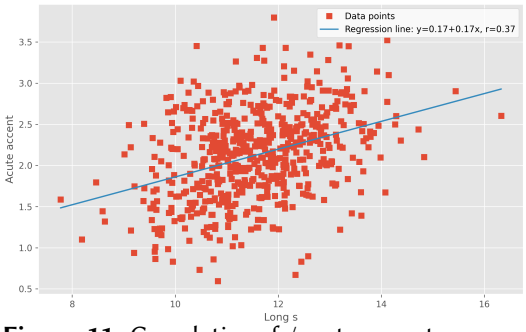
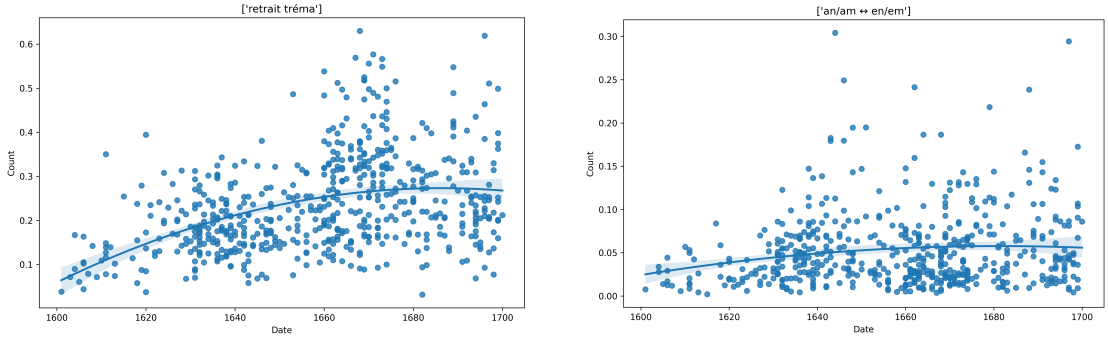


Figure 11: Correlation ‹ſ›/acute accent.

Our working hypothesis is as follows: as ligatures are largely composed of a long *s* (‹ſ›), we should obtain a correlation between the use of this *s* (cf. fig. 10a) and the acute accent (cf. fig. 10b). We evaluate the correlation between the evolution of the two phenomena over time, and obtain a Pearson product-moment correlation coefficient of 0.365 with a *p*-value of 4.88e-20, which indicate a good correlation (cf. fig. 11).



(a) Increase of the diaeresis.

(b) Increase of the confusion ‹en|m›/‹an|m›.

Figure 12: Apparition of new phenomena.

Pattern D: innovation. Finally, it is important to note that, in this slow movement of standardisation that we are drawing, innovations also appear. These innovations concern a lot diacritics, some of which are exploding in number like the diaeresis (cf. fig. 10a): scriptors tend to add them more and more on one of the two hiatus vowel, especially with the sequence ‹ue› (*louër* or *loüer*, today *louer*, eng. “to rent”). We also note a great hesitation regarding the notation of nasal vowels (cf. fig. 12b), especially [ã], for which we can use ‹en|m› or ‹an|m› such as *aventure* vs *avanture* (today *aventure*, eng. “adventure”).

5. Conclusion and further work

The spelling of the 17th c. is changing throughout the century, and at the beginning of the 18th century, if the main features inherited from the past tend to have disappeared (etymological letters, use of diacritical letters, historical use of ‹u› and ‹i›, etc.), we still note a certain instability, which concern more minor hesitations than anything else (notation of nasal vowels, hiatus vowel).

Among all these changes in the spelling, some obey external influences, such as the limitations imposed by the type case of printers (appearance of the accent), others to (at least partially) internal logics, with parallel trajectories of similar phenomena. The velocity of change varies greatly from one phenomenon to another, with sometimes slow shifts over decades, or sometimes abrupt ruptures whose cause is not entirely clear. It is therefore, in the end, difficult to identify a clear moment of change, and instead a long change is emerging, spread throughout the century (cf. tab. 7).

Table 7
Main spelling change and their dating using change-point detection.

Rule	Example	Change-point
‹cque›→‹c›	*avecque*→*avec*	1632
‹ct›→‹t›	*exploict*→*exploit*	1638
tilde → vowel	*hõme*→*homme*	1637
‹gn› → ‹nn›	*incognu*→*inconnu*	1654
‹és›→‹é›	*estat*→*état*	1660
‹as›→‹â›	*pasle*→*pâle*	1669
Distinction of ramist letters	*vniuers*→*univers*	1670
‹eu›→‹u›	*asseurance*→*assurance*	1675
Suppression of etymological letter	*nopce*→*noce*	1683
Suppression of calligraphic letter	*vray*→*vrai*	1688

If corpus linguistics is not new, the rapid and automatic creation of corpora responding to an *ad hoc* question has, to our knowledge, never been tested in diachronic linguistics. Our results show that, based on recent improvements in available tools, it is now possible to conduct effective studies on large datasets rapidly created. The rest of our work will therefore lead us towards the expansion of the corpus, notably by adding data from the 16th and 18th centuries to obtain more precise results, and including the supposed moment of the completion of the standardisation of French, and the appearance of spelling, not in theory, but in practice.

Acknowledgments

TO BE ADDED

References

[1] N. Catach, Histoire de l'orthographe française, Honoré Champion, 2001.

[2] L. 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), Max Niemeyer Verlag, 1992. doi:10.1515/9783110938593.

[3] Académie française, Cahiers de remarques sur l'orthographe françoise pour estre examinez par chacun de Messieurs de l'Academie, avec des observations de Bossuet, Pellisson, etc., éd. charles joseph marty-laveaux ed., Jules Gay, 1863. URL: https://books.google.ch/books?id=u5Y5AQAAIAAJ.

[4] C. F. d. Vaugelas, Remarques sur la langue françoise, Droz, Geneva, 2009. éd. Marzys, Zygmunt.

[5] C. F. d. Vaugelas, Remarques sur la langue françoise, utiles à ceux qui veulent bien parler et bien escrire, Vve J. Camusat et P. Le Petit, 1647.

[6] A. Dees, Dialectes et scriptae à l'époque de l'ancien français, Revue de Linguistique Romane 49 (1985) 87–117.

[7] S. Baddeley, L'Ortographie française au temps de la Réforme, Droz, Geneva, 1993.

[8] C. H. Vachon, Le Changement linguistique au XVIe siècle: une étude basée sur des textes littéraires français, ELiPhi, Éditions de linguistique et de philologie, 2010.

[9] S. Gabay, Pourquoi moderniser l'orthographe? principes d'ecdotique et littérature du XVIIe siècle, Vox Romanica 73 (2014) 27–42. doi:99.125005/vox201410027.

[10] F. Duval, Les éditions de textes du XVIIe siècle, in: D. Trotter (Ed.), Manuel de la philologie de l'édition, De Gruyter, 2015, pp. 369–394. doi:10.1515/9783110302608-017.

[11] A. Amatuzzi, W. Ayres-Bennett, A. Gerstenberg, L. Schøsler, C. Skupien-Dekens, Changement linguistique et périodisation du français (pré)classique: deux études de cas à partir des corpus du RCFC, Journal of French Language Studies 30 (2020) 301–326. doi:10.1017/S0959269520000058.

[12] Netherlands Historical Data Archive, Nijmegen Institute for Cognition & Information, Optical Character Recognition in the Historical Discipline: Proceedings of an International Workshop, Halbgraue Reihe zur Historischen Fachinformatik, St. Katharinen, 1993.

[13] B. Robertson, F. Boschetti, Large-scale optical character recognition of ancient greek,

Mouseion: Journal of the Classical Association of Canada 58 (2017) 341–359. URL: https://muse.jhu.edu/article/679181.

[14] L. Muellner, Digital Classical Philology Ancient Greek and Latin in the Digital Revolution, De Gruyter Saur, 2019, pp. 7–17. doi:10.1515/9783110599572-002.

[15] H. Scheithauer, A. Chagué, L. Romary, Which TEI representation for the output of automatic transcriptions and their metadata? An illustrated proposition, 2022. URL: https://inria.hal.science/hal-04001303, working paper or preprint.

[16] S. Gabay, J.-B. Camps, A. Pinche, C. Jahan, SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more), in: 1st International Workshop on Computational Paleography (IWCP@ICDAR 2021), Lausanne, Switzerland, 2021. URL: https://hal.science/hal-03336528.

[17] U. Schäfer, B. Weitz, Combining OCR outputs for logical document structure markup. technical background to the ACL 2012 contributed task, in: R. E. Banchs (Ed.), Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 104–109. URL: https://aclanthology.org/W12-3212.

[18] L. Romary, P. Lopez, GROBID - Information Extraction from Scientific Publications, ERCIM News 100 (2015). URL: https://inria.hal.science/hal-01673305.

[19] M. Khemakhem, L. Foppiano, L. Romary, Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields, in: electronic lexicography, eLex 2017, Leiden, Netherlands, 2017. URL: https://hal.science/hal-01508868.

[20] M. Khemakhem, Standard-based Lexical Models for Automatically Structured Dictionaries, Ph.D. thesis, Université de Paris, Paris, 2020. URL: https://theses.hal.science/tel-03274454.

[21] S. Najem-Meyer, M. Romanello, Page layout analysis of text-heavy historical documents: a comparison of textual and visual approaches, in: Proceedings of the Computational Humanities Research Conference 2022, Antwerp, Belgium, 2022, pp. 36–54. doi:10.48550/arXiv.2212.13924.

[22] J. Janes, A. Pinche, C. Jahan, S. Gabay, Towards automatic TEI encoding via layout analysis, in: Fantastic future 21, 3rd International Conference on Artificial Intelligence for Librairies, Archives and Museums, AI for Libraries, Archives, and Museums (ai4lam), Paris, France, 2021. URL: https://hal.science/hal-03527287.

[23] A. Pinche, K. Christensen, S. Gabay, Between automatic and manual encoding, in: TEI 2022 conference : Text as data, Newcastle, United Kingdom, 2022. URL: https://hal.science/hal-03780302. doi:10.5281/zenodo.7092214.

[24] T. Clérice, J. Janes, H. Scheithauer, S. Bénière, L. Romary, B. Sagot, Layout Analysis Dataset with SegmOnto, in: DH2024 - Annual conference of the Alliance of Digital Humanities Organizations, ADHO, Washington, D.C., United States, 2024. URL: https://inria.hal.science/hal-04513725.

[25] H. Fix, Automatische Normalisierung - Vorarbeit zur Lemmatisierung eines diplomatischen altisländischen Textes, Max Niemeyer Verlag, 1980, pp. 92–100. doi:doi:10.1515/9783111438788.92.

[26] E. Tjong Kim Sang, M. Bollmann, R. Boschker, F. Casacuberta, F. Dietz, S. Dipper, M. Domingo, R. van der Goot, M. van Koppen, N. Ljubešić, R. Östling, F. Petran, E. Petters-

son, Y. Scherrer, M. Schraagen, L. Sevens, J. Tiedemann, T. Vanallemeersch, K. Zervanou, The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation, Computational Linguistics in the Netherlands Journal 7 (2017) 53–64. URL: https://clinjournal.org/clinj/article/view/68/61.

[27] M. Bollmann, Normalization of Historical Texts with Neural Network Models, Ph.D. thesis, Ruhr-Universität Bochum, Bochum, 2018. URL: https://www.linguistics.rub.de/forschung/arbeitsberichte/22.pdf.

[28] S. Gabay, FreEM-corpora/FreEMnorm: FreEM norm Parallel corpus, 2022. doi:10.5281/zenodo.5865428.

[29] S. Gabay, L. Barrault, Traduction automatique pour la normalisation du français du XVIIe siècle, in: Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles, Nancy, France, 2020, pp. 213–222. URL: https://aclanthology.org/2020.jeptalnrecital-taln.20.

[30] R. Bawden, J. Poinhos, E. Kogkitsidou, P. Gambette, B. Sagot, S. Gabay, Automatic Normalisation of Early Modern French, in: LREC 2022 - 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3354–3366. URL: https://inria.hal.science/hal-03540226.

[31] L. Remacle, Le Problème de l'ancien wallon, Presses universitaires de Liège, 1948. URL: http://books.openedition.org/pulg/338.

[32] J. Séguy, La dialectométrie dans l'atlas linguistique de la gascogne, Revue de linguistique romane 37 (1973) 1–24.

[33] H. Goebl, Dialektometrie, in: Quantitative Linguistik/Quantitative Linguistics. Ein internationales Handbuch/An International Handbook, De Gruyter, 2005, pp. 498–531.

[34] J. Nerbonne, W. Heeringa, 31. measuring dialect differences, in: Theories and Methods: An International Handbook of Linguistic Variation, volume 1, De Gruyter Mouton, 2010, pp. 550–567.

[35] J.-B. Camps, Manuscripts in Time and Space: Experiments in Scriptometrics on an Old French Corpus, in: A. U. Frank, C. Ivanovic, F. Mambrini, M. Passarotti, C. Sporleder (Eds.), Corpus-Based Research in the Humanities CRH-2, Proceedings of the Second Workshop on Corpus-Based Research in the Humanities CRH-2, 25-26 January 2018, Vienna, Austria, Vienna, Austria, 2018, pp. 55–64. URL: https://hal.science/hal-01695899.

[36] S. Gabay, R. Bawden, P. Gambette, J. Poinhos, E. Kogkitsidou, B. Sagot, Le changement linguistique au XVIIe s. : nouvelles approches scriptométriques, in: CMLF 2022 - 8e Congrès Mondial de Linguistique Française, volume 138 of *SHS Web of conferences*, EDP Sciences, Orléans, France, 2022, pp. 02006.1–14. doi:10.1051/shsconf/202213802006.

[37] S. Gabay, P. Gambette, R. Bawden, B. Sagot, Ancien ou moderne ? Pistes computationnelles pour l'analyse graphématique des textes écrits au XVIIe siècle, Linx 85 (2023). doi:10.4000/linx.9346.

[38] B. Kiessling, Kraken - an Universal Text Recognizer for the Humanities, in: Digital Humanities Conference 2019 - DH2019, ADHO, Utrecht, The Netherlands, 2019. doi:10.34894/Z9G2EX.

[39] T. Clérice, You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine, Journal of Data Mining & Digital Humanities Documents historiques et reconnaissance automatique de texte (2023). doi:10.46298/jdmdh.9806, v. 4.
[40] D. Reis, J. Kupec, J. Hong, A. Daoudi, Real-time flying object detection with yolov8, CoRR abs/2305.09972 (2023). doi:10.48550/ARXIV.2305.09972.
[41] S. Gabay, T. Clérice, P. Jacsont, E. Leblanc, M. Jeannot-Tirole, S. Solfrini, S. Dolto, F. Goy, C. C. Luján, M. Zaglio, M. Perregaux, J. Janes, B. Sagot, R. Bawden, R. Dent, O. Nédey, A. Chagué, Reconnaissance des écritures dans les imprimés, in: Humanistica 2024, OCR, Association francophone des humanités numériques, Meknès, Morocco, 2024. URL: https://hal.science/hal-04557457.
[42] S. Gabay, FONDUE-FR-PRINT-16, 2024. doi:10.5281/zenodo.11526150.
[43] S. Gabay, FONDUE-FR-PRINT-17, 2024. doi:10.5281/zenodo.11526040.
[44] S. Gabay, M. Jeannot-Tirole, F. Goy, FONDUE-LA-PRINT-16, 2024. doi:10.5281/zenodo.11526160.
[45] S. Gabay, T. Clérice, J. Janès, FONDUE-MLT-PRINT-TEST-longS, 2024. doi:10.5281/zenodo.11526316.
[46] J. Poinhos, ABA (Alignment-Based Approach), 2020. URL: https://github.com/johnseazer/aba.
[47] S. B. Needleman, C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology 48 (1970) 443–453. doi:10.1016/0022-2836(70)90057-4.
[48] V. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet physics doklady 10 (1966) 707–710. URL: https://www.mathnet.ru/eng/dan31411.
[49] A. S. Kroch, Reflexes of grammar in patterns of language change, Language Variation and Change 1 (1989) 199–244. doi:10.1017/S0954394500000168.
[50] R. Zimmermann, An improved test of the constant rate hypothesis: late modern american english possessive have, Corpus Linguistics and Linguistic Theory 19 (2023) 323–352. doi:doi:10.1515/cllt-2021-0038.
[51] J. Fruehwald, J. Gress-Wright, J. Wallenberg, Phonological rule change: The constant rate effect, in: Proceedings of the 40th Annual Meeting of the North East Linguistic Society, GLSA Publications, 2013, pp. 219–230. URL: https://www.research.ed.ac.uk/files/14416788/Fruewald_Gress_Wright_Wallenberg_Phonological_Rule_Change.pdf.
[52] L. Vostrikova, Detection of the disorder in multidimensional random-processes, Doklady Akademii Nauk SSSR 259 (1981) 270–274. URL: http://mi.mathnet.ru/dan44582.
[53] J. Bai, Estimating multiple breaks one at a time, Econometric Theory 13 (1997) 315–352. doi:10.1017/S0266466600005831.
[54] C. Klaussner, C. Vogel, A. Bhattacharya, Detecting linguistic change based on word co-occurrence patterns, in: Proceedings of the 4th International Workshop on Computational History, Singapore, 2017, pp. 14–21. URL: https://ceur-ws.org/Vol-1992/paper_4.pdf.
[55] N. Catach, J. Golfand, L'orthographe plantinienne, De Gulden Passer 50 (1973) 19–69. URL: https://www.dbnl.org/tekst/_gul005197301_01/_gul005197301_01_0003.php.
[56] P. Corneille, Le théâtre de P. Corneille, G. de Luyne, Paris, 1663. URL: https://gallica.bnf.

fr/ark:/12148/bpt6k71442p.
[57] L. 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), Max Niemeyer Verlag, 1992. doi:10.1515/9783110938593.