Data Diversity in handwritten text recognition. Challenge or opportunity?

Jean-Baptiste Camps, Chahan Vidal-Gorène, Dominique Stutzmann, Marguerite Vernet, Ariane Pinche

Data Diversity in handwritten text recognition: challenge or opportunity?

Jean-Baptiste Camps[1], Chahan Vidal-Gorène[1], Dominique Stutzmann[2], Marguerite Vernet[1], and Ariane Pinche[1]

[1]École nationale des chartes, Université PSL
[2]Institut de recherche et d'histoire des textes/CNRS

1 Introduction

In this paper, we wish to show approaches in handling diversity in larger collections of training data for text acquisition pipelines, specifically handwritten text recognition for medieval manuscripts in Latin and French. Present throughout medieval Europe, Latin is one, if not the most used written language of the time on this continent, while French has known from a relatively early date (around the 12th century judging from preserved manuscripts) a vernacular production that soon became one of the most prominent of Western Europe, influencing the written culture of its neighbours from its central position. Combined, they provide a case study whose diversity and general scope could, we hope, allow to provide results with broader applicability, even beyond medieval Western manuscripts.

Heterogeneity or diversity in the collections can result from intrinsic features (e.g. linguistic, palaeographic, diachronic variation in the sources), but also from extrinsic features (aim and provenance of transcriptions, idiosyncrasies of transcribers...). We propose to approach both types of diversity by reusing several open data sets from various research projects in diverse fields and involving many collaborators. We add a double focus, linguistic (Latin vs. French manuscripts) and graphic (abbreviated vs. normalised transcriptions). We hope to be able to overcome, to some extent, the issue of linguistic diversity and propose a common, modular pipeline for different languages, related but different in their inner structure and declension mechanisms.

When, on the one hand, recent studies focus on "hyperdiplomatic" digital editions to study the production of specific items, the implementation of natural language processing and text mining is commonly based on a normalised text. Instead of aiming at defining a single universal translinguistic transcriptional standard to merge all existing standards – an utopic endeavour, and perhaps even not desirable –, and instead of designing a unified pipeline supported by dedicated libraries (e.g. image > hyperdiplomatic > normalised > lemmatised+POS-tagged > critical text) to constrain all existing editions, we applied a more modular approach to reuse and pool datasets to train multiple models and design paths more fitted to the variety of goals encountered in medieval studies.

In this attempt, we will strive to answer more specifically the following questions:

1. To what extent can we (and should we) mutualise HTR training material between preexisting datasets and even related languages? (and is it worth the effort?)

2. Are approaches that decompose image to text prediction and further linguistic normalisations (abbreviation expansion for instance) better performing for that goal than straightforward "image to normalised text" approaches?

2 Diversity in our corpus

2.1 Extrinsic diversity: variation in data production

The most obvious source of diversity is artificial, in the sense that it is the result of the production of the data (and particularly of transcription choices) and not of the sources itself.

For this research three macro-datasets, themselves mostly aggregates of smaller micro-datasets, have been used, one French and two Latin.

The French dataset is `Cremma-medieval`, composed of 17431 lines from eleven Old French manuscripts written between the 13th and 14th centuries (Table 1). It is made from pre-existing transcriptions, and sample size is very different from one source manuscript to the other. A graphemic[1] transcription method has been chosen to maintain a many to one mapping between signs in the source and the transcription (abbreviations and their expansions are both kept, *u/v* or *i/j* are not dissimilated), but allographs are normalised (*e.g.*, round and long *s* are both transcribed *s*). Finally, spaces are not homogeneously represented in the ground truth text annotation, with transcribers reproducing the manuscript spacing while others are using lexical spaces. It must be stressed that spaces are the most important source of errors in medieval HTR models (see for instance the model Bicerin, where spaces represent 33.9% of errors Pinche, 2021). In this `cremma-medieval` macro-dataset, several transcriptions from different transcribers, coming from different projects, have been collected.

This diversity is also very present in the `Oriflamms` macro-dataset, containing 120 111 lines from no less than 779 manuscripts (Table 1). This dataset has been composed along several different projects over a substantial interval of time, and is a mix of aligned preexisting normalised editions (without abbreviations) and graphemic transcriptions (including abbreviations and their expansion). It is composed of both French, Latin and bilingual texts.

The last macro-dataset `Saint-Victor` is the most homogeneous, composed of transcriptions from two Victorine mss, i.e., BnF latin 14588 and BnF latin 14525 written by no less than twelve scribes at the end of the 12th century and the first part of the 13th century (Table 1). Both mss have the same type of writing. It has been created during a master's thesis. It is divided into two sub-corpus. A first corpus is transcribed without abbreviations. The transcription uses lexical spaces. It is the most important of the two sub-corpus with 10736 lines. The second sub-corpus consists of a small part of the first (1860 lines), which has been transcribed with abbreviations.

Early tests have shown the tremendous variations in the choice of signs used to transcribed medieval graphemes, in particular abbreviations, including MUFI and out of MUFI characters. For example, the common abbreviative marker has been transcribed alternatively as U+0303 COMBINING TILDE, U+0304 COMBINING MACRON, U+0305 COMBINING OVERLINE, F00A COMBINING HIGH MACRON WITH FIXED HEIGHT (PART-WIDTH), and even, in composition, U+1EBD LATIN SMALL LETTER E WITH TILDE, U+0113 LATIN SMALL LETTER E WITH MACRON, etc. Even when using MUFI (Medieval Unicode Font Initiative), different types of Tironian *et* or *p* flourish can be used. To facilitate machine learning, a conversion table was used to apply a first level of normalisation, and to reduce the 262 preexisting character class to around 30 (Clérice and Pinche, 2021).

2.2 Intrinsic diversity: variation in language, script and scribal practice

Diversity is also due to linguistic differences inside the corpus, with a main distinction between Latin and French texts, the latter in a variety of regional *scriptae*, including Anglo-French, Eastern (Lorrain) and Picard, and also diachronic variation, from 12 to 14th century.

[1] We use the terminology graphemic (*graphématique*) and graphetic (*allographétique*) following Stuzmann (2011).

Corpus	Editors	Manuscripts	Pages	Lines	DOI
CREMMA-MEDIEVAL					
French Corpus					
Otinel	Camps	2	75	**13568**	10.5281/zenodo.7506657
Wauchier	Pinche	1	49	**6148**	10.5281/zenodo.7506657
Maritem	Mariotti	1	18	1026	
CremmaLab	Pinche et al.	7	55	13568	
—					
Total		11	149	17431	
ORIFLAMMS					
Bilingual Corpus					
Reg.chancell. Poitou	Guérin	200	1217	30015	
Reg.chancell. Paris	Viard	2	29	474	
Morchesne	Guyotjeannin et al.	1	189	10394	
Cartulaire de Nesle	Hélary	1	117	3899	
Latin Corpus					
Chartes Fontenay	Stutzmann	104	104	**1384**	10.5281/zenodo.6507963
Psautiers	Oriflamms	27	48	5793	
PsautierIMS	Stutzmann	48	132	**3145**	10.5281/zenodo.6507973
MSS dat. lat.	Oriflamms / ECMEN	101	101	**2299**	10.5281/zenodo.6507965
French Corpus					
Queste del saint Graal	Marchello-Nizia, Lavrentiev	1	130	10725	
BnF fonds fr.	ECMEN	159	189	13510	10.5281/zenodo.6507975
Mss dat. fr.	ECMEN	45	55	3355	
Album XIIIe.	Careri, et al.+ECMEN	52	52	1992	
Légende dorée	IRHT+ECMEN	18	679	31742	
Pèlerinage	OPVS+ECMEN	20	56	1384	10.5281/zenodo.6507981
—					
Total		779	3098	120111	
ST-VICTOR					
Latin Corpus					
Saint-Victor	Vernet	2	54	**12596**	10.5281/zenodo.7510410

Table 1: Composition of the `cremma-medieval`, `Oriflamms` and `st-victor` macro-datasets [For this abstract, only corpora in bold have been used]

The variety is also in the writing styles. Copyists used different script types according to their place and date of activity (e.g. *praegothica*, *textualis*, *cursiva*, *semitextualis*[2]). Some script types were used preferentially according to the genre of the text under copy (e.g. liturgy, literature, diplomatic and pragmatic texts). Conversely, textual genres could influence some specific scribal practices (layout, abbreviations, etc.).

3 Pipeline description

Our aim is to evaluate the impact of data heterogeneity to build models for Latin and medieval French. Our corpus contains two levels of heterogeneity: it contains documents in one of two different languages (including internally some diatopic variation)[3], and variety of specifications for transcriptions. Each sentence of our corpus includes both abbreviated forms and expanded forms of words, thanks to the original encoding of the editions, that followed the Guidelines of the Text Encoding Initiative, and used a combination of `<choice>`, `<abbr>` and `<expan>` (TEI

[2] For classification criteria and the lack of consensus among palaeographers, see Derolez (2003), Stutzmann et al. (2020).

[3] We excluded documents with mixed contents (i.e., parts in French intertwined with parts in Latin), except for the ECMEN corpus which only contains small quotations or single words in Latin.

Consortium, 2022).

Corpus have endured varying types of normalisations, sometimes contradictory (combined or decombined, etc.), to smooth discrepancies between transcriptions. The normalisation step follows this pipeline:

1. lowercasing;
2. normalising unicode (NFKD);
3. making substitutions based on an equivalence table and the use of "chocoMUFIn" (Clérice and Pinche, 2021). In particular,
    - normalisation of allographs (hypernormalised);
    - suppression of ramist distinctions (*u/v* and *i/j*);
    - removal of punctuation;
    - suppression of editorial marks (diacritics, apostrophes, cedillas, ...).

We have divided our corpus into four training datasets to perform our evaluations and see potential benefits of fine-tuning for such an approach, on Latin or French texts and on abbreviated or expanded texts. The distribution of each corpus is described in table 2.

	Abbr (Lines)	**Exp (Lines)**
LAT	**TOTAL : 8,528**	**TOTAL :17,404**
	Fontenay : 1,365	Fontenay : 1,365
	MsDat : 2,217	MsDat : 2,217
	PsautierIMS : 3,086	PsautierIMS : 3,086
	StVictorLite : 1,860	StVictorFull : 10,736
FRO	**TOTAL : 19,532**	**TOTAL : 19,530**
	ECMEN : 9,831	ECMEN : 9,831
	otinel bodmer : 1,977	otinel bodmer : 1,977
	otinel vaticane : 1,758	otinel vaticane : 1,758
	wauchier : 4,582	wauchier : 4,580
	Pelerinage : 1,384	Pelerinage : 1,384

Table 2: Distribution of corpora into the four main datasets

Based on the experiments made by Camps et al. (2021b) on abbreviated manuscripts, two approaches have been considered. Training on abbreviated data has been carried out with `Kraken` (Kiessling, 2019, Kiessling et al., 2019), an OCR and HTR system previously used with success on a wide range of manuscripts (Camps et al., 2021a, Scheithauer et al., 2021, Thompson et al., 2021), and training on expanded data with `Calfa`, an OCR and HTR system originally developped for highly abbreviated Oriental manuscripts (Vidal-Gorène et al., 2021). These two architectures use an encoder-decoder approach, the first one trained at the character level, the second one at the word level. If we keep the same hyperparameters defined previously (Camps et al., 2021b), we use a deeper architecture for the first one, architecture capable of high recognition rate in CREMMA (Pinche and Clérice, 2021).

4 Preliminary results and discussion

Current results show, perhaps counter intuitively, a better performance for expanded models, at least for Latin (fig. 1 and table 3), while, for French, the abbreviated model seem to perform slightly better (fig. 1). Perhaps more importantly, they show important variation in the distribution of the character error rates per page inside each test set and between test sets (fig. 2). Apart

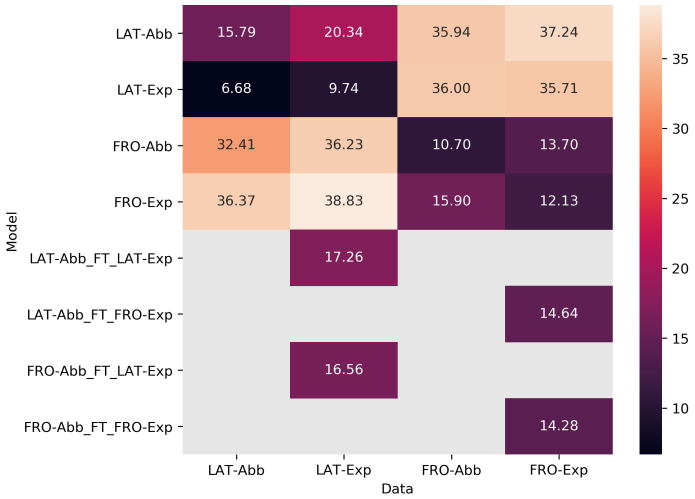


Figure 1: Matrix of the cross evaluation of models

from a few strong outliers on the Latin corpora, with CER between 40 and 90% (due to issues in the test material), they show a situation that varies according mostly to the origin of the data. For some subcorpora, the CERs display very limited variation, with a very small interquartile range (CREMMA corpora for instance), while the results obtained for corpora such as ECMEN could reflect the larger variety of material they contain.

Nevertheless, among various observations, the following cases can be noted. On the one hand, on LAT-Exp predictions, the efficiency of the model is especially linked to the script used. Thus, the particularly angular *textualis quadrata*, widely used in `PsautierIMS` and some manuscripts of `MSS dat. lat`, is poorly recognised. We find a lot of issues related to the stems **ii** / **u** / **n** / etc. In the most extreme cases a significant difficulty in differentiating **c** and **e** occurs. For these scripts, tildes are seldom understood and abbreviations are therefore badly expanded. Meanwhile, in diplomatic texts of `Fontenay`, although the form of the letters is often sophisticated and flourished - especially in the first line of the charters - the model is able to recognise tildes and abbreviations. We also observe that the quality of the ink greatly influences the efficiency of the model. On the other hand, this multi-level heterogeneity seems to affect benefits we could expect of fine-tuning. We do not notice any gain in recognition by fine-tuning abbreviated models with expanded data yet. Nevertheless we can already observe that cross-lingual fine-tuned models achieve similar recognition rates, even though abbreviations are widely different for these languages.

All of this is deserving of further investigations, particularly to evaluate the impact of training set size versus training set diversity, and to measure the robustness of models trained with and applied to mixed language corpora. . Moreover, further normalisation of the training sets, and a direct inspection of outliers could allow to increase performance and intelligibility of the results.

Data and materials availability

The datasets used for this paper are available on Zenodo, under different licensing and access conditions (see the DOI in Table 1).

Two models, trained on abbreviated and expanded data are available at the following DOIs:

Latin and Old French Abbreviated 10.5281/zenodo.7516310.

Latin and Old French Expanded 10.5281/zenodo.7516057

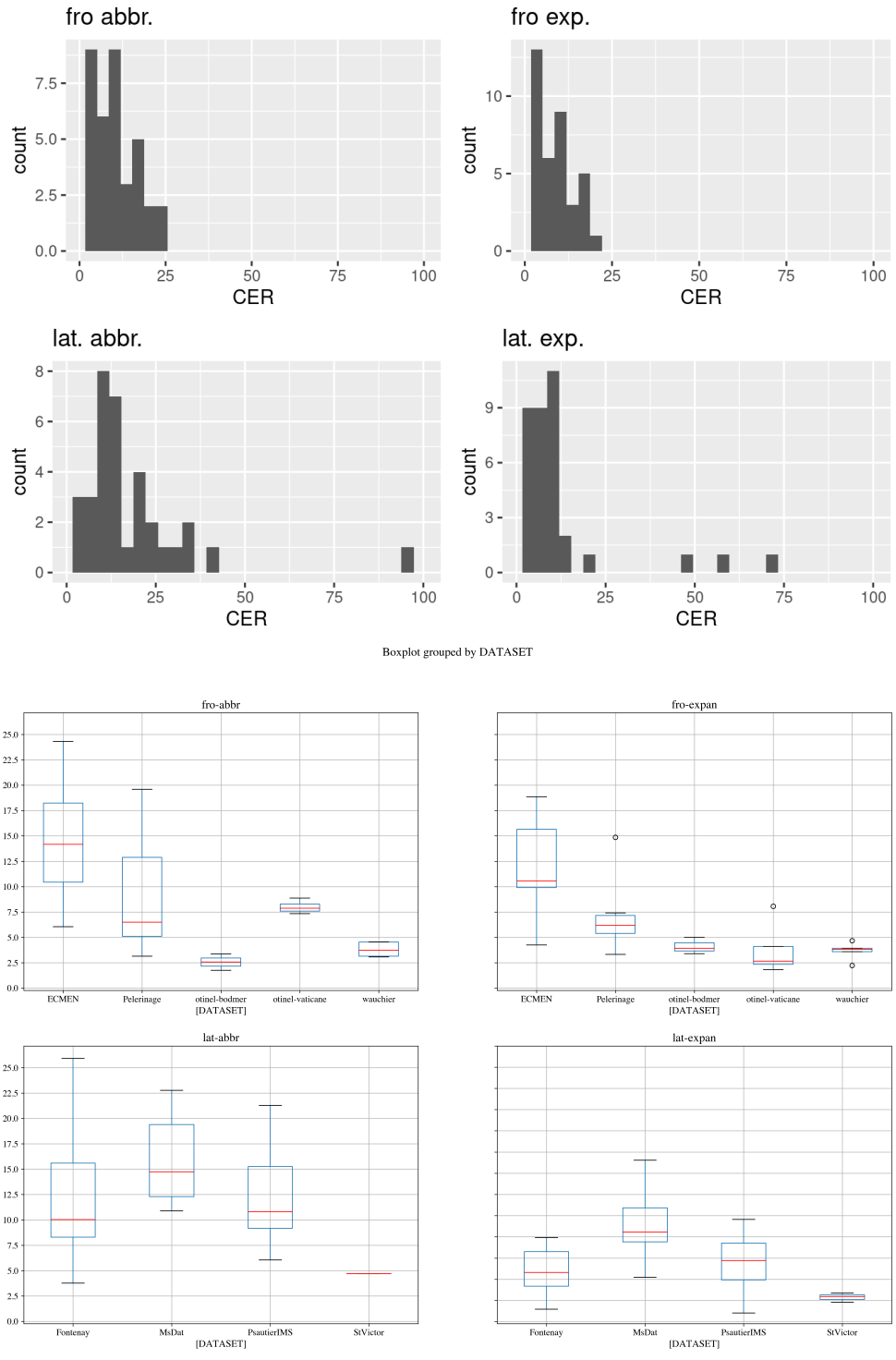
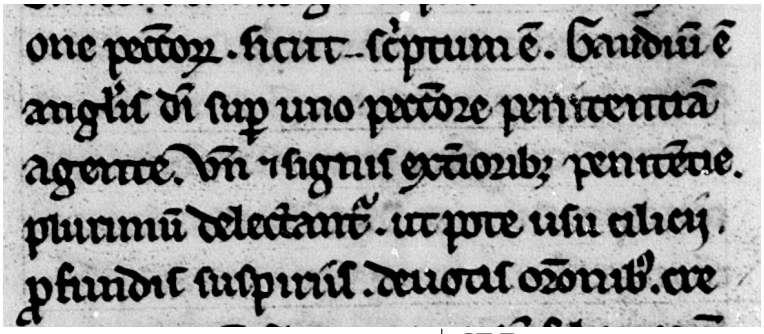


Figure 2: Distribution of character error rates per page in the test sets; histograms (top) and boxplots (with outliers above 25% removed)



GT Abbr	GT Expan
one pecc̃oꝝ sicut sc̣ptum ẽ gaudiũ ẽ angl̃is d̃i sup uno pecc̃ore penitentiã agente uñ ⁊ signis ex̃tiorib penitẽtie plurimũ delectan̂ ut pote usu cilicii ꝓfundis suspiriis deuotis oro̊nib̊ cre	one peccatorum sicut scriptum est gaudium est angelis dei super uno peccatore penitentiam agente unde et signis exterioribus penitentie plurimum delectantur ut pote usu cilicii profundis suspiriis deuotis orationibus cre
Prediction Abbr	**Prediction Expan**
one **pecc̃oꝝ** sicut sc̣ptum ẽ **sudiũ** ẽ angl̃is dī sup uno pecc̃ore penitentiã agente uñ ⁊ signis ex̃tiorib penitẽtie plurimũ delectan̂ ut pote usu cilicii ꝓfundis **suspinis** deuotis oro̊nib̊ cre	one peccatorum sicut scriptum est **saudium** est angelis dei super uno peccatore penitentiam agente unde et signis exterioribus penitentie plurimum delectantur ut pote usu cilicii profundis suspiriis deuotis orationibus cre

Table 3: Facsimile with ground truth abbreviated and extended and abbreviated and extended predictions from an extract of latin corpus of St Victor (BnF, Latin 14525, fol. 41va).

References

Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis. *Digital Scholarship in the Humanities*, 36(Supplement_2):ii49–ii71, October 2021a. ISSN 2055-7671. doi: 10.1093/llc/fqab033. URL `https://doi.org/10.1093/llc/fqab033`.

Jean-Baptiste Camps, Chahan Vidal-Gorène, and Marguerite Vernet. Handling heavily abbreviated manuscripts: Htr engines vs text normalisation approaches. In *International Conference on Document Analysis and Recognition*, pages 306–316. Springer, 2021b.

Thibault Clérice and Ariane Pinche. Choco-Mufin, a tool for controlling characters used in OCR and HTR projects, 9 2021. URL `https://github.com/PonteIneptique/choco-mufin`.

Albert Derolez. *The Palaeography of Gothic Manuscript Books from the Twelfth to the Early Sixteenth Century*. Number 9 in Cambridge studies in palaeography and codicology. Cambridge University Press, Cambridge, 2003. ISBN 0-521-80315-2.

Benjamin Kiessling. Kraken - an Universal Text Recognizer for the Humanities. In *Proceedings of the DH2019 Conference - Digital Humanities: Complexities, Utrecht, The Netherlands, 9–12 July 2019*, Utrecht, July 2019. CLARIAH. 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. IEEE, 2019.

Ariane Pinche. CREMMA Medieval, an Old French dataset for HTR and segmentation, 8 2021. URL https://github.com/htR-United/cremma-medieval.

Ariane Pinche and Thibault Clérice. Htr-united/cremma-medieval: 1.0.1 bicerin (doi), August 2021. URL https://doi.org/10.5281/zenodo.5235186.

Hugo Scheithauer, Alix Chagué, Rostaing Aurélia, Lucas Terriel, Laurent Romary, Marie-Françoise Limon-Bonnet, Benjamin Davy, Gaetano Piraino, Franck Beltrami, Danis Habib, et al. Production d'un modèle affiné de reconnaissance d'écriture manuscrite avec escriptorium et évaluation de ses performances. In *Les Futurs Fantastiques-3e Conférence Internationale sur l'Intelligence Artificielle appliquée aux Bibliothèques, Archives et Musées, AI4LAM*, 2021.

Dominique Stutzmann, Christopher Tensmeyer, and Vincent Christlein. Writer Identification and Script Classification. Two Tasks for a Common Understanding of Cultural Heritage. *manuscript cultures*, 15:11–24, 2020. ISSN 1867–9617. URL https://www.csmc.uni-hamburg.de/publications/mc/files/articles/mc15-02-stutzmann.pdf.

Dominique Stuzmann. Paléographie statistique pour décrire, identifier, dater. . . normaliser pour coopérer et aller plus loin ? In Franz Fischer, Christiane Fritze, and Georg Vogeler, editors, *Kodikologie und Paläographie im digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*, volume 3, pages 247–277. Books on Demand (BoD), Norderstedt, 2011.

TEI Consortium. 3.6.5 Abbreviations and Their Expansions. In *TEI P5: Guidelines for Electronic Text Encoding and Interchange, v4.4.0*. Text Encoding Initiative Consortium, 2022. URL https://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#CONAAB.

Walker Thompson et al. Using handwritten text recognition (htr) tools to transcribe historical multilingual lexica. *Scripta & e-Scripta*, 21:217–231, 2021.

Chahan Vidal-Gorène, Boris Dupin, Aliénor Decours-Perez, and Thomas Riccioli. A modular and automated annotation platform for handwritings: evaluation on under-resourced languages. In *International Conference on Document Analysis and Recognition*, pages 507–522. Springer, 2021.