READ
RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D7.1
HTR Engine Based on HMMs P1

Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Enrique Vidal

UPVLC

Distribution: http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	READ
Project full title	Recognition and Enrichment of Archival Documents
Instrument	H2020-EINFRA-2015-1
Thematic priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date/duration	01 January 2016 / 42 Months

Distribution	Public
Contract. date of delivery	31.12.2016
Actual date of delivery	31.12.2016
Date of last update	31.12.2016
Deliverable number	D7.1
Deliverable title	HTR Engine Based on HMMs P1
Type	Demonstrator
Status & version	Final
Contributing WP(s)	WP7
Responsible beneficiary	UPVLC
Other contributors	UPVLC
Internal reviewers	Johannes Michael, Max Weidemann, Nathanael Philipp, Günter Mühlberger
Author(s)	Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Enrique Vidal
EC project officer	Martin Majek
Keywords	Handwritten Text Recognition, Hidden Markov models

Contents

Executive summary

This report describes the research developed in the first year of the READ project on Handwritting Text Recognition based on Hidden Markov Models as optical models. Several collections have been researched this year and the current results are described here. The consolidated developments have been integrated in Transkribus.

1 Introduction

Classical Handwritten Text Recognition (HTR) borrows concepts and methods from the field of Automatic Speech Recognition, such as Hidden Markov Models (HMM), n-grams and Neural Networks (NN) [4, 3, 7, 2]. In recent years, pure NN-based methods have achieved impressive results in HTR [10, 9], but HMM-based methods are competitive with these pure NN-based methods [5]. Furthermore, HMM-based techniques have very attractive characteristics that make them very convenient for several problems: there are well-known methods for dealing with language models and integrating them; and there exist efficient techniques for dealing with lattice-based techniques (as it happens in Task 2.2 and Task 2.5 of READ), and the decoding problem is well known for HMM-based HTR.

Task 7.1 - Hidden Markov Model-based HTR

The problem of HTR can be stated formally as follows:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x}) = \arg\max_{\mathbf{w}} P(\mathbf{x} \mid \mathbf{w}) P(\mathbf{w}) \tag{1}$$

where $\hat{\mathbf{w}}$ is the best transcript for the line image $\mathbf{x}$ among all possible transcripts $\mathbf{w}$. $P(\mathbf{x} \mid \mathbf{w})$ represents the optical modelling that is approximated with HMM in this task and $P(\mathbf{w})$ is the language model (LM) that is approximated with n-grams. The models involved in this expression are trained from examples.

Training $P(\mathbf{w})$ is currently easy since only plain text is necessary. This text can be obtained from the web or from linguistic resources. Usually, the more text in order to capture the frequency of word (or character) sequences adequately the better. Language model training is mainly researched in Task 7.4-Language modelling.

Training HMM for computing $P(\mathbf{x} \mid \mathbf{w})$ is more difficult since it is necessary to have line images and their corresponding diplomatic transcripts, each line with its corresponding transcript. So there is no need for segmented words nor characters for training the optical models. An Expectation-Maximization (EM) algorithm [4] is used for training the HMM, provided that this information is given. The necessary line images and their corresponding transcripts are prepared by human experts. The amount of data that is necessary for obtaining a well-trained HMM depends on many factors (different hands, quality of the images, ...) and it may range from 50 to 500 pages, or in other terms, from 10K to 800K words. This process is expensive and time-consuming, and makes the human user necessary in the production/training loop.

Task 7.1 in READ is related with research on HMM-based techniques for HTR. This means that both training techniques and decoding techniques are researched and developed. The most consolidated techniques are integrated in Transkribus. In Y1 in READ, this research was along the lines that are described in the following sections.

2 Research on HMM-based HTR in several collections

During this period some HTR experiments with different READ collections have been carried out. These collections are: the “Lope” collection, the “Girona” collection, the “RSEAPV” collection and the “Konzilsprotokolle” collection. Also discriminative experiments have been carried out with the “Bozen” dataset and in a Bengali dataset. Finally, a study using the MGGI methodology to improve the language model in marriages register books has been carried out.

HTR in the “Lope” collection

The “Lope” collection has been provided by the Biblioteca Nacional de España. It is composed by around 250 documents of Lope de Vega. Lope de Vega was a Spanish writer, poet and novelist. He was one of the key persons in the Spanish Golden Century of Baroque literature.

In this task we chose a document called “La contienda de García Paredes” and carried out some experiments to test the HMMs-based HTR approach. The selected document is composed by around 200 pages (see Fig. 1 for examples). This collection is included in Transkribus.

These pages were manually annotated with the layout analysis of each page to indicate text blocks and lines and also they were completely transcribed line by line by an expert paleographer. The recognition error of this document is around 50% at word level and 25% at character level. These results and experiments have not been published yet.

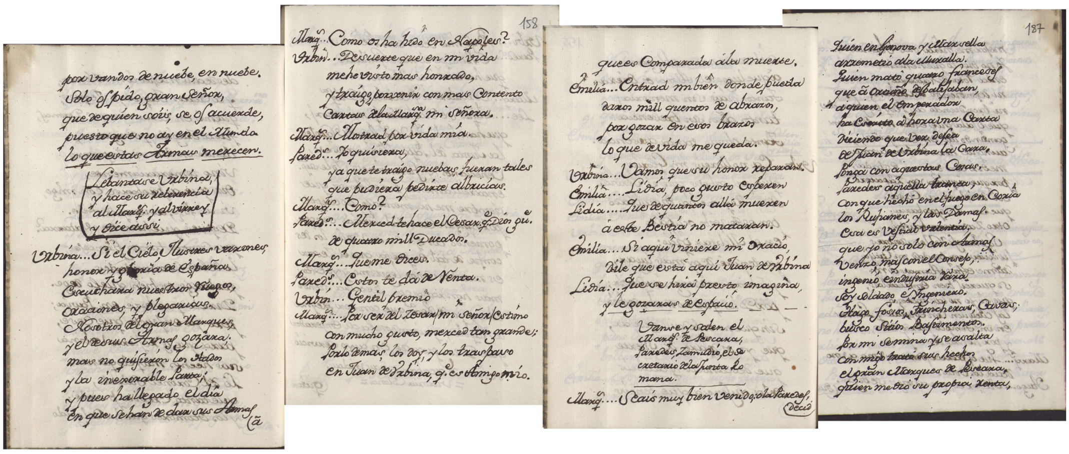


Figure 1: Pages of the Lope collection.

HTR in the “RSEAPV” collection

The “RSEAPV” collection has been provided by the “Real Sociedad Económica de Amigos del Pais de Valencia” (RSEAPV). RSEAPV is a partnership that was established in 1776 by

King Carlos III of Spain. The RSEAPV was, since its foundation, and especially during the 18th century, a reference center for all the Valencian society, for which it established a frameworl for discussion and treatment of the most important and cutting-edge issues of that moment. RSEAPV has a very large collection of documents that is digitized and it can be used in READ.

In this task we chose a document of this collection and we studied the ability of the HMM-based HTR approach on it. This document was written by a single writer in Spanish in 1905 and it is composed of 170 pages. To carry out the experiments we used a small set of the document composed by the first 42 pages. Fig. 2 shows some examples of this collection.

The automatic processing of this document is difficult, due to the typical degradation and heterogeneity of these kind of documents, and mainly to the few number of pages available to train the statistical models. However, taking into account these difficulties, the obtained results are quite encouraging. In a closed-vocabulary experiment, the recognition error of this document was around 36% at word level and 16% at character level. A detailed description of the work carried out with this dataset has been submitted to the next Iberian Conference on Pattern Recognition and Image Analysis [11].

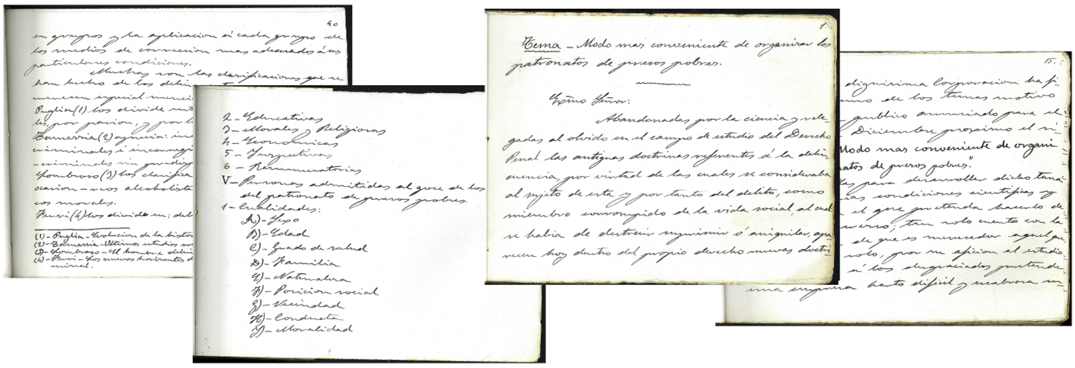


Figure 2: Pages of the RSEAPV collection.

HTR in the “Girona” collection

The “Girona” collection has been provided by the “Centre de Recerca d’Història Rural” of the “Facultat de lletres de la Universitat de Girona” and it is composed by notarial documents.

In this task we chose a document of this collection to test the HMM-based HTR approach. The selected document is a mortgage register. The first 48 pages of the document were annotated at two levels. First, a layout analysis of each page was manually done to indicate text blocks and lines, resulting in a dataset of 1882 lines. Second, the pages were completely transcribed line by line by an expert paleographer. This set is available in the Transkribus platform with the identification number 5146 (RH Girona 1769). Fig. 3 shows some examples of this collection.

In spite of the typical degradation and heterogeneity problems present in this document, the recognition error obtained was around 30% at word level. This result is quite encouraging. This research is ongoing and the results will be published along the following period.

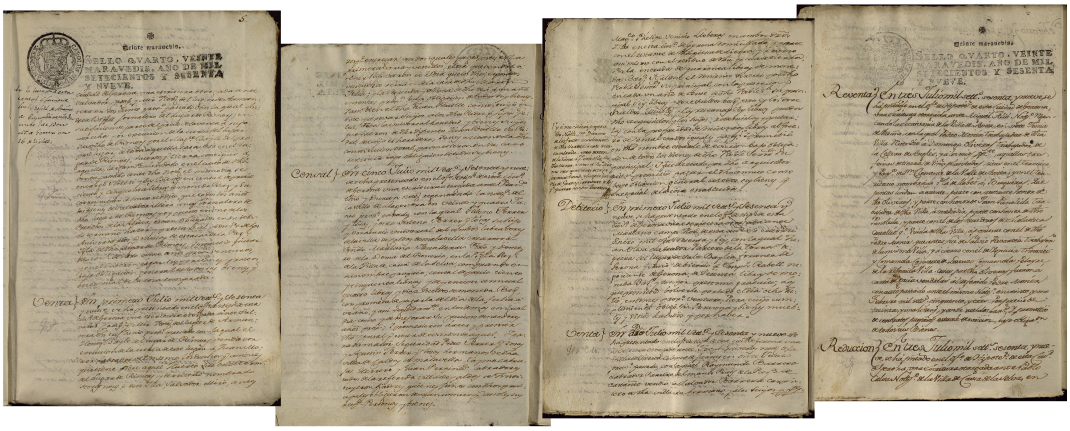


Figure 3: Pages of the Girona collection.

HTR in the “Konzilsprotokolle” collection

During this period some experiments with the Alvermann "Konzilsprotokolle" collection were carried out. The collection was provided by the University of Greifswald and they cover a time period of 30 years with 3 different writers. The documents used in the experiments are available in Transkribus (ID 3678, ID 3679, ID 3680 and ID 3681). Fig. 4 shows some examples of this collection.

Some internal benchmarks were defined in the READ project to work with this dataset. All the information related to the benchmark can be found in the READ wiki[1]. The dataset is available through Zenodo [1]. The work carried out by the UPVLC is also presented in the READ wiki[2].

This research was partially performed in a collaborative activity that was held to test different combinations of both the URO’s & UPVLC’s page/line image preprocessing and feature extraction for using with the HMM-based HTR system. Results showed that at this point there is no significant difference on the final overall HTR performance by using such combinations. The recognition error of this document is around 28% at word level and 11% at character level in an open vocabulary experiment. This study is outlined in details in the READ wiki[3].

3 Advanced research on HMM-based HTR

In addition to the previous research activitiy performed on HMM-based HTR, other challenging research activities were carried out. These research activities are described below.

[1] http://read02.uibk.ac.at/wiki/index.php/Benchmarks

[2] http://read02.uibk.ac.at/wiki/index.php/Benchmarks:Alvermann_Konzilsprotokolle:_UPVLC

[3] http://read02.uibk.ac.at/wiki/index.php/Technical_Meetings:AlexandroAtURO#UPV_HTR_System:_Testing_URO_preprocessing_modules

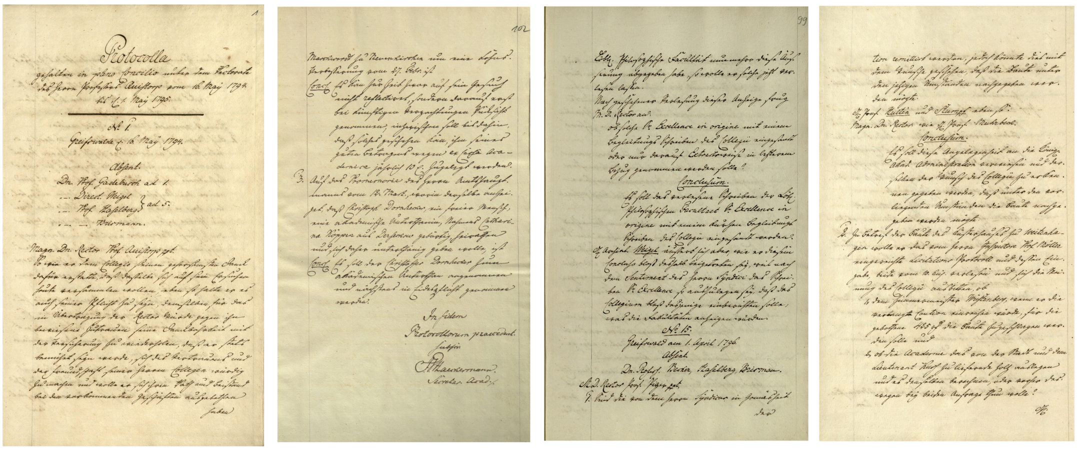


Figure 4: Example of the Konzilsprotokolle collection.

Discriminative training

A further refinement to the acoustic models using a discriminative training approach to HMM parameter estimation has been studied during this period. These techniques obtained competitive results compared with NN-based HTR systems [5].

Some initial experiments have been carried out using the "Bozen" dataset (see Fig. 5). The benchmark used in the experiments is the same used in the HTR competition presented in the ICFHR 2016 [9]. The obtained results showed that, using discriminative training, some improvements in recognition accuracy can be obtained.

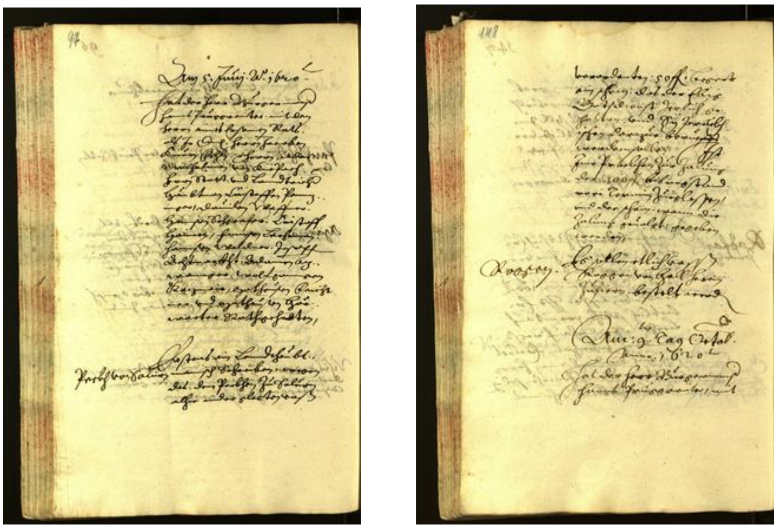


Figure 5: Examples of the Bozen dataset.

In addition to these experiments with the "Bozen" dataset, the discriminative training approach was researched in other tasks [8], and the obtained results were encouraging.

Using the MGGI methodology for category-based language modeling

Handwritten marriage licenses books have been used for centuries by ecclesiastical and secular institutions to register marriages. The information contained in these historical documents is useful for demography studies and genealogical research, among others. Despite the generally simple structure of the text in these documents, automatic transcription and semantic information extraction is difficult due to the distinct and evolutionary vocabulary, which is composed mainly of proper names that change along the time. Figure 6 shows some examples of the collection used in the experiments.

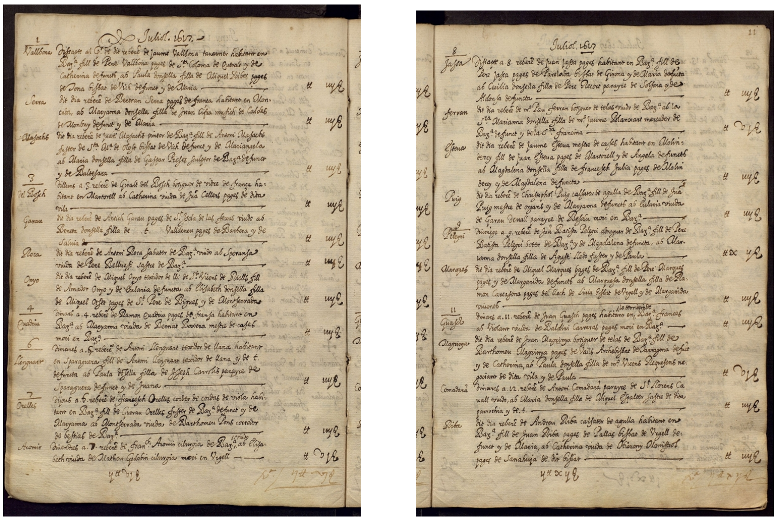


Figure 6: Example of marriage license pages.

During this period we have studied the use of category-based language models to both improve the automatic transcription accuracy and make the extraction of semantic information easier. Then, the main causes of the observed semantic errors were analyzed and a Grammatical Inference technique known as MGGI was applied, to improve the semantic accuracy of the language model obtained. Using this language model, full handwritten text recognition experiments have been carried out with results supporting the interest of the proposed approach. The recognition error of this document is around 10% at word level in an open vocabulary experiment. The precision and recall for obtaining categories in this dataset was 85% and 76%, respectively. More details about this study can be found in [6].

HTR non-Latin scripts

Handwritten text recognition has been evaluated on other non-Latin scripts. Bengali script was researched this time which has been served to face up with a variety of problems that are not present in Latin scripts: more complex character shapes, involving a large alphabet, etc. The dataset used for these experiments was provided by researchers not involved in READ. Figure 7 shows some examples of the collection used in the experiments.

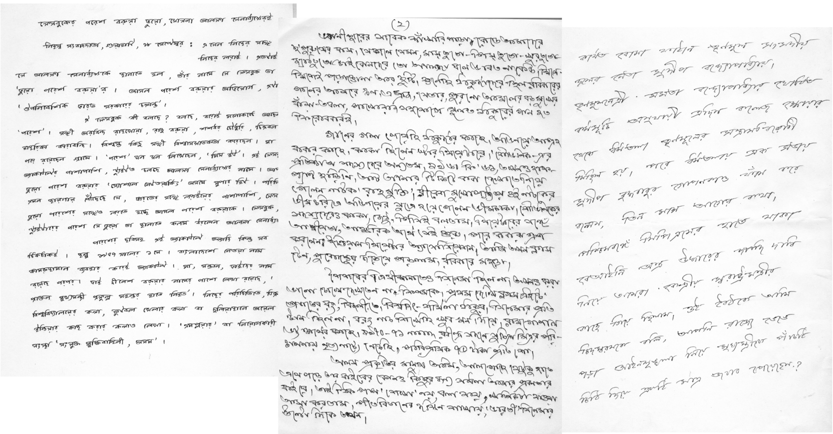


Figure 7: Example of marriage license pages.

Experiments with this dataset had to cope with new codification problems, language modelling problems, etc. The recognition error on these documents is around 58% at word level and 28% at character level in an open vocabulary experiment. More details about this evaluation are given in [8].

Exploiting existing transcripts for HTR

Existing transcripts for historic manuscripts are a very valuable resource for training useful models for automatic recognition, aided transcription, and/or indexing of the remaining untranscribed parts of these collections. However, these existing transcripts generally exhibit two main problems which hinder their convenience: a) the text of the transcripts is seldom aligned with manuscript lines, and b) the text often deviate very significantly from what can be seen in the manuscript, either because writing style has been modernized or abbreviations have been expanded, or both.

A study has been carried out to analyze the problems of aligning transcripts with manuscript lines and, how to deal with what is actually written in the image (e.g. abbreviated words) and the corresponding available (modernized) transcripts. To deal with such problems, developing semi-automatic procedures for minimizing human effort needed to adapt existing transcripts in order to render them usable is required. This work has been carried out using a case study of the Alcaraz dataset (see example pages in Figure 8), which corresponds to written records from an Inquisition process in the 16th century.

Figure 9 illustrates some of the issues that have to be dealt with when using modern transcripts: unaligned text, expansion of abbreviations, modernization of spelling (capitalization, accents), exclusion of striked-out text. In general the modern transcript is available but for HTR the diplomatic transcript is required. More details about this work can be found in [12].

References

[1] https://doi.org/10.5281/zenodo.215383.

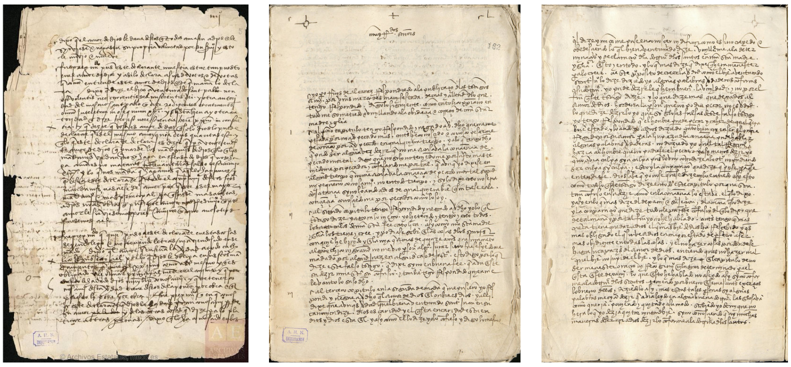


Figure 8: Example of page images from the Alcaraz dataset.

[2] T. Bluche. *Deep Neural Networks for Large Vocabulary Handwritten Text Recognition*. PhD thesis, Ecole Doctorale Informatique de Paris-Sud - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, may 2015. Discipline : Informatique.

[3] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Tr. PAMI*, 31(5):855–868, 2009.

[4] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[5] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Interspeech 2016*, pages 2751–2755, 2016.

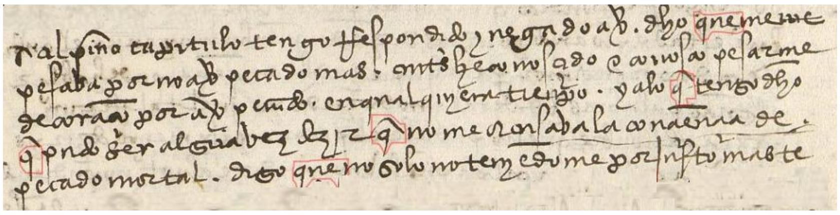
[6] V. Romero, A. Fornés, J.A. Sánchez, and E. Vidal. Using the MGGI methodology for category-based language modeling in handwritten marriage licenses books. In *Proceedings of the 2016 International Conference on Frontiers in Handwriting Recognition*, pages 331–336, 2016.

[7] V. Romero, A.H. Toselli, and E. Vidal. *Multimodal Interactive Handwritten Text Transcription*. Series in Machine Perception and Artificial Intelligence (MPAI). World Scientific Publishing, 2012.

[8] J.A. Sánchez and U. Pal. Hanwrittent text recognition for bengali. In *Proceedings of the 2016 International Conference on Frontiers in Handwriting Recognition*, pages 542–547, 2016.

[9] J.A. Sánchez, V. Romero, A.H. Toselli, and E. Vidal. ICFHR2016 competition on handwritten text recognition on the READ dataset. In *Proceedings of the 2016 International Conference on Frontiers in Handwriting Recognition*, pages 630–635, 2016.

[10] J.A. Sánchez, A.H. Toselli, V. Romero, and E. Vidal. ICDAR 2015 competition HTRtS: Handwritten text recognition on the tranScriptorium dataset. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015.



Diplomatic transcript (needed, but unavailable):

§ al pmo capitulo tengo respondido y negado *avr dho* **que** me ~~*me*~~
pesava por no *avr* pecado *mas* . *ants* he conoscido y conosco pesarme
de *coraço* por *avr* pecado en qualquiera tienpo . *y* a lo **q** tengo *dho*
q *pud* ser alguna vez *dzir* **q** no me acusava la conciencia de
pecado mortal . *digo* **que** no *solo* no *teniedome* por justo mas *te*

Aligned modernized transcript

Iten. Al primero capítulo tengo respondido y negado *aver dicho* **que** *me*
pesava por no *aver* pecado *más*. *Antes* he conoscido y conosco pesarme
de *coraçón* por *aver* pecado en qualquiera tienpo. *Y* a lo **que** tengo *dicho*
que *pudo* ser alguna vez *dezir* **que** no me acusava la conciencia de
pecado mortal. *Digo* **que** no *sólo* no *teniéndome* por justo, mas *teniéndome*

Original modernized transcript (partially available):

Iten. Al primero capítulo tengo respondido y negado aver dicho que me pesava por no aver pecado más. Antes he conoscido y conosco pesarme de coraçón por aver pecado en qualquiera tienpo. Y a lo que tengo dicho que pudo ser alguna vez dezir que no me acusava la conciencia de pecado mortal. Digo que no sólo no teniéndome por justo, mas teniéndome [...]

Figure 9: Excerpt from the Alcaraz dataset showing the original image and its diplomatic and modernized transcripts. The words in blue-italic font are different in the diplomatic and the aligned modernized versions. The appearances of the word “que” are marked in the image surrounded by a red polygon and in the transcripts in red-bold font. This is an example of a modernized word that may or may not appear abbreviated in the image – and therefore in the diplomatic transcript.

[11] C. Hernández E. Vidal V. Romero, V. Bosch Campos and J.A. Sánchez. A historical document handwriting transcription end-to-end system. In *Iberian Conference on Pattern Recognition and Image Analysis*, 2017. (submitted).

[12] M. Villegas, A.H. Toselli, V. Romero, and E. Vidal. Exploiting existing modern transcripts for historical handwritten text recognition. In *Proceedings of the 2016 International Conference on Frontiers in Handwriting Recognition*, pages 66–71, 2016.