READ
RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D7.6
Interactive Predictive Transcription Engine P3

A toolkit for the interactive transcription of handwritten documents

Verónica Romero, Enrique Vidal, Lorenzo Quirós, Joan Andreu Sánchez
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	20.12.2018
Actual date of delivery	20.12.2018
Date of last update	20.12.2018
Deliverable number	D7.6
Deliverable title	Interactive Predictive Transcription Engine P3
Type	Demostrator
Status & version	in process
Contributing WP(s)	WP7
Responsible beneficiary	UPVLC
Other contributors	
Internal reviewers	
Author(s)	Verónica Romero, Enrique Vidal, Lorenzo Quirós, Joan Andreu Sánchez
EC project officer	Christophe Doin
Keywords	Interactive handwritten transcription, Computer assistive transcription

Contents

Executive Summary

This third year deliverable describes the work carried out in the Task T7.2 *Interactive-predictive process for transcription and line detection.* Interactive techniques have been proposed in recent years for transcribing handwritten documents and aim to help the user in the transcription process. These techniques are being used for perfectly transcribe some historical collections. In this deliverable, the state of the art of the interactive transcription approach is reviewed. Then, the work carried out in READ in the transcription of some documents using these techniques is described and some qualitative and quantitative results are presented and shortly explained.

1 Introduction

The work carried out in T7.2 *Interactive-predictive process for transcription and line detection* is briefly described in this deliverable.

1.1 Review of state of the art

Interactive HTR techniques have been proposed in the last years for transcribing handwritten documents. In this approach the user and the system work jointly in tight mutual collaboration to obtain perfect transcripts of the text images. The interactive handwritten text transcription system used here was recently introduced by the UPVLC team and presented in [3, 2]. It is referred to as "Computer Assisted Transcription of Text Images" (CATTI). In the CATTI framework, the human transcriber is directly involved in the transcription process since he/she is responsible of validating and/or correcting the HTR output.

The interactive transcription process starts when the HTR system proposes a full transcript of a given text line image. In each interaction step the user validates a prefix of the transcript which is error free and keys in new information. At this point, the system, taking into account the feedback of the user, suggests a suitable continuation. This process is repeated until a complete and correct transcript of the input signal is reached. A key point of this interactive process is that, at each user-system interaction, the system can take advantage of the prefix validated so far to attempt to improve its prediction. In order to make the interaction process fast, in the recognition stage, a Word Graph (WG) is obtained for each recognized line. A WG represents all the transcriptions with high probability of the given text image. It can be represented as a weighted directed acyclic graph, where each edge is labelled with a word and a score, and each node is labelled with a point of the handwritten image. Then, during the CATTI process, the system makes use of these word graphs in order to complete the prefixes accepted by the human transcriber. A detailed description of the CATTI system can be found in [2].

1.2 Task 7.2

The goal of this task is to research the interactive-predictive process for correcting recognition errors at two levels: first, interactive-predictive HTR techniques, and second, interactive-predictive HTR techniques in combination with interactive-predictive line detection.

2 Preliminary CATTI results on READ text images

During this period we have been working with in two READ collections: the "'Oficio de Hipotecas de Girona collection" and the "Spanish golden age collection".

2.1 The "Oficio de Hipotecas de Girona" Collection

The collection we have been working with is provided by the *Centre de Recerca d'Història Rural* (CRHR) from the Universitat de Girona (MOU partner of the READ project). The collection, called *"Oficio de Hipotecas de Girona"*, is composed of a large number of notarial documents from the XVII century. The CRHR is interested in the perfect transcription of this collection. UPVLC and CRHR are working together in order to carry out this transcription using the CATTI system. Until this moment around 600 pages have been transcribed. These pages are available in the READ platform.

2.1.1 Transcription Workflow

The transcription is carried out in batches of around 50 pages each, where previously transcribed batches are used to (re-)train system models in order to improve the accuracy of the models on each iteration. Also, an external vocabulary of anthroponyms is used to improve the language model.

Additionally to the diplomatic transcription of the document, some tags are added to the corresponding words in order to provide a more rich information about the content of the document: toponyms, anthroponyms, trades, registry typology, abbreviations and hyphenated words.

The correct transcripts of each page are interactively obtained using the CATTI system. The workflow is depicted in Figure 1.

At the begining the CATTI technology was based on traditional Gaussina Mixture Hidden markov Models (GMM-HMM) and word n-grams fro optical and language modeling, respectively. Several batches were processed in this way. During this year, CATTI was adapted to also work with convolutional-recurrent neural networks and character n-grams for optical and language modelling, respectively. Although much less mature, these new modeling approaches have shown to significantly outperform traditional ones. For this reason, at some temporal point of the process, a new CATTI version based on the new models was adopted.

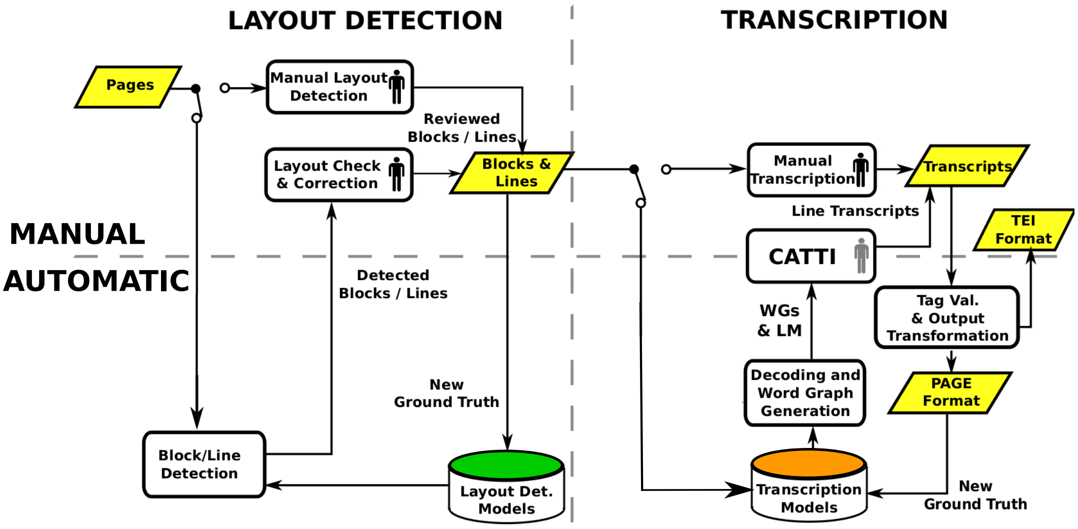

Figure 1: Semi-automatic system workflow.

2.1.2 Results

In general, users have been very satisfied with the results and the reduction in time and effort on the whole transcription process.

As explained previously, we use GMM-HMM optical model and n-gram language models to transcribe the first 10 batches (*b002* to *b011*). At that point we change the optical model to RNN. Results are reported in terms of *Character Error Rate* (CER) and *Word Error Rate* (WER) to assess the quality of the transcription and an estimation of pot-editing user effort; and *Words Stroke Ratio* (WSR) as an estimator of the effort needed by a human transcriber to produce correct transcriptis using CATTI. The relative difference between WER and WSR (called Estimated Effort-Reduction - EFR) gives an estimation of the reduction in human effort that can be achieved using CATTI. This metrics are used with and without tags, since one of the main goals is to produce diplomatic transcripts. On Fig. 2 CER and WER measures are presented per test batch, to assess the quality of the transcripts.

The WSR and EFR are reported on Fig. 3 and Tab. 1 respectively, to show the effect of the CATTI system in the user effort.

Notice that at each batch the test set is different and has not been previously seen by the system.

Firs two batches do not have tags at the moment of the semi-automatic transcription. Although tags were added a posteriori in order to train new models the results are presented only without tags.

Transition from HMM-GMM to HMM-CRNN was pretty straight forward, due the workflow previously defined. Improvement respect to previous model was significant, with an improvement in CER over 34% respect to the best result obtained using HMM-GMM model (*b005* vs. *b012*).

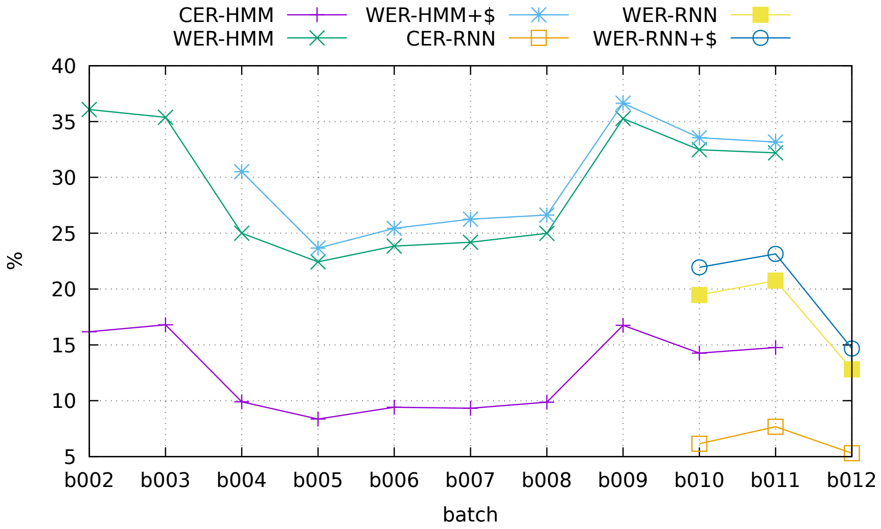

Figure 2: Results of Levenshtein distance metrics.*-HMM stands for HMM-GMM optical model, *-RNN stands for HMM-CRNN optical model and *+$ stands for tagged transcripts.

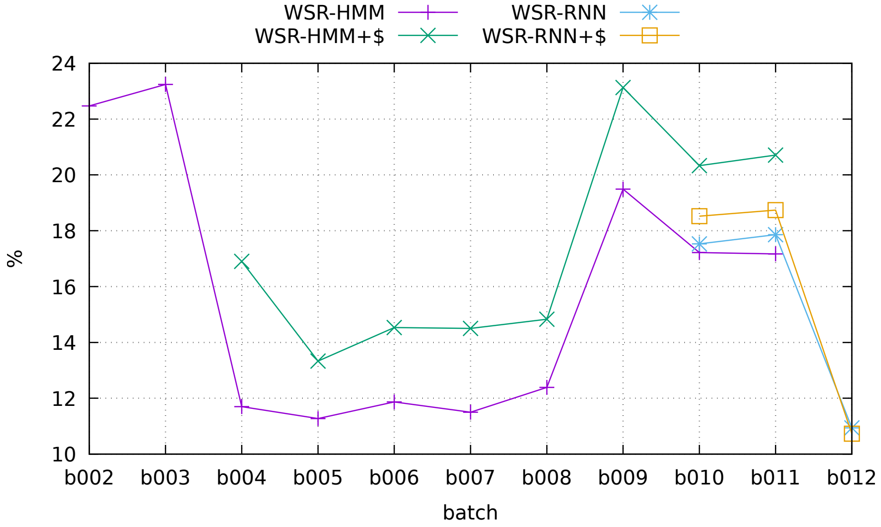

Figure 3: WSR Results. *-HMM stands for HMM-GMM optical model, *-RNN stands for HMM-CRNN optical model and *+$ stands for tagged transcripts.

Table 1: EFR results summary. *-HMM stands for HMM-GMM optical model, *-RNN stands for HMM-CRNN optical model and *+$ stands for tagged transcripts.

batch	EFR-HMM	EFR-HMM+$	EFR-RNN	EFR-RNN+$
b002	32.41	—	—	—
b003	29.20	—	—	—
b004	53.20	44.59	—	—
b005	49.75	43.68	—	—
b006	50.23	42.88	—	—
b007	52.46	44.78	—	—
b008	50.42	44.31	—	—
b009	44.72	36.87	—	—
b010	46.97	39.40	26.86	32.68
b011	46.66	37.54	13.92	19.01
b012	—	—	14.59	20.85

2.1.3 Conclusions

In this work we present a longitudinal study of a production scenario HTR end-to-end system. Throught out the iterative transcription of a corpus we evaluate our worklow. The study shows that our process does not only adapt to the variability typical of a production scenario, but also allows us to improve it with the inclusion of new technologies without major issues.

This method allows us to produce ground-truth quality transcripts of more than 550 pages of the *Oficio de Hipotecas de Girona* collection, along with the structural information of the document (layout annotation). The process is also capable of automatically providing very useful enriched text which will be used by historians to extract easily the valuable information recorded in the documents.

The inclusion of HNN-CRNN model gives a step up improvement in transcription quality, and therefore the effort of the user.

A detailed description of the work carried out in this collection can be found in [1]

2.2 The “Spanish golden age” Collection

We have been working with the documents of the Spanish Golden Age available at the Biblioteca Nacional de España (BNE)[1]and collaboration of the ProLope research group[2]. Both the BNE and ProLope are MOU partners.

Lope de Vega was a Spanish writer, poet and novelist. He was one of the key figures in the Spanish Golden Century of Baroque literature. Around 3,000 sonnets, 3 novels, 4 novellas, 9 epic poems and about 500 plays are attributed to him. At the BNE there are around 250 registers whose authorship is Lope de Vega. However, not all of these documents are autographs, and there are documents written by several copyists.

[1] http://www.bne.es/es/Inicio/index.html

[2] http://prolope.uab.cat/

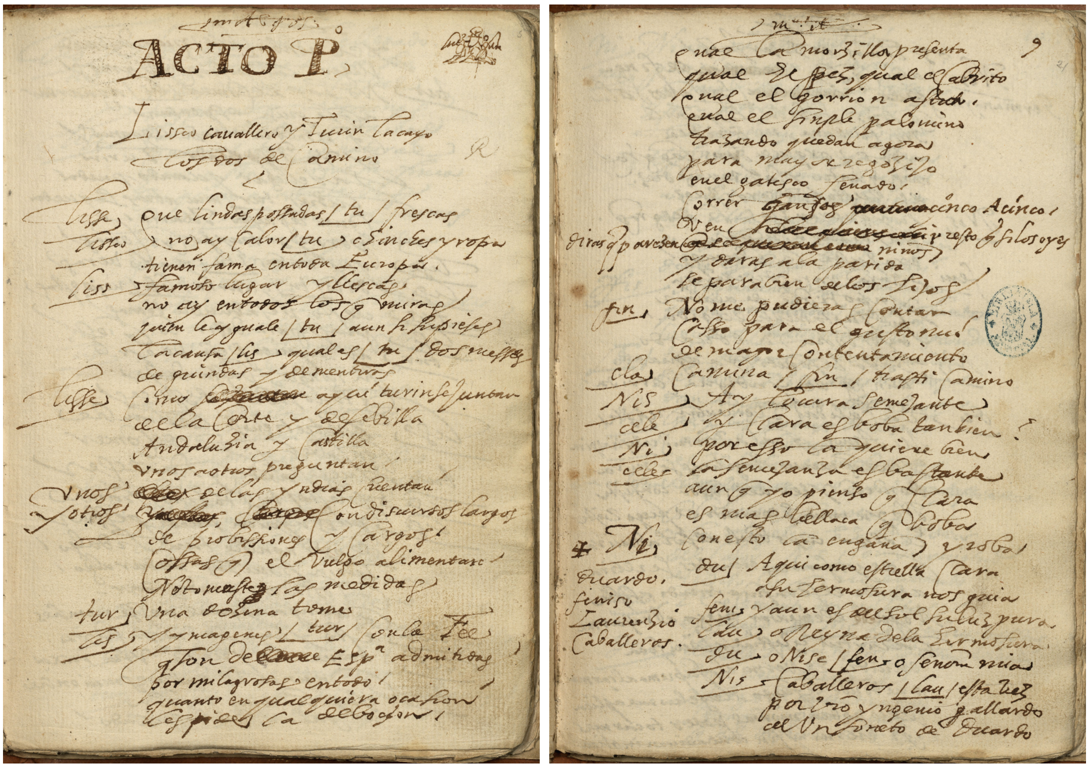

Figure 4: Pages of the *"La Dama Boba"* manuscript.

During P1 and P2 of the READ project, some HTR experiments with different documents to test the CATTI engine were carried out (see Deliverable D7.5). During this P3, we have processed the whole collection, more than 320 documents, and they have been indexed (see Deliverable D8.12). In addition, one autograph of Lope de Vega has been processed to be transcribed using the CATTI system in a crowdsourcing way.

The BNE has organized an exposition about Lope de Vega and the Teather in the Spanish Golden Age that will take place from November 28th to March 17th of 2019. During this time the visitants to the exposition can help in the transcription of the *"La Dama Boba"* manuscript using the CATTI system (`http://prhlt-carabela.prhlt.upv.es/tso/`). *"La Dama Boba"* is composed by 143 page images. These pages were automatically annotated with the layout analysis of each page to indicate the text blocks and lines and then, for each line, a word graph was obtained to carry out the intereactive transcription. In Figure 4 we can see some page examples of this manuscript

From the beginning of the exposition the transcription system has had a good acceptance by the users. The average access per day is of 18 pages, algthoug only a few lines are really transcribed. Until the moment 175 lines have been corrected by the users. With respect to the places where this work is carried out, it is important to remark, that altought the majority of the access are from the BNE, there are other access from Madrid, Valencia or Tenerife.

3 A toolkit for the interactive transcription of handwritten documents.

The improvements carried out in the CATTI engine during this period have been uploaded to the CATTI implementation available at `https://github.com/PRHLT/CATTI`

This version of the CATTI has been integrated in a web platform. In `http://prhlt-carabela.prhlt.upv.es/tso/` the web version of the CATTI engine that the visitants to the exposition can use to transcribe *“La Dama Boba”* manuscript is available.

References

[1] Lorenzo Quirós, Vicente Bosch, Lluis Serrano, Alejandro Toselli, and Enrique Vidal. From hmms to rnns: Computer-assited transcription of a handwritten notarial records collection. In *International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018*. IEEE, 2018.

[2] V. Romero, A.H. Toselli, and E. Vidal. *Multimodal Interactive Handwritten Text Transcription*. Series in Machine Perception and Artificial Intelligence (MPAI). World Scientific Publishing, 1st edition edition, 2012.

[3] A.H. Toselli, V. Romero, M. Pastor, and E. Vidal. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1824–1825, 2010.