READ

RECOGNITION & ENRICHMENT OF ARCHIVAL DOCUMENTS

D7.5

Interactive Predictive Transcription Engine P2

A toolkit for the interactive transcription of handwritten documents

Verónica Romero, Enrique Vidal, Vicente Bosch, 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	31.12.2017
Actual date of delivery	28.12.2017
Date of last update	15.12.2017
Deliverable number	D7.5
Deliverable title	Interactive Predictive Transcription Engine P2
Type	Demonstrator
Status & version	in process
Contributing WP(s)	WP7
Responsible beneficiary	UPVLC
Other contributors	
Internal reviewers	Johannes Michael, Max Weidemann
Author(s)	Verónica Romero, Enrique Vidal, Vicente Bosch, Lorenzo Quirós, Joan Andreu Sánchez
EC project officer	
Keywords	Interactive handwritten transcription, Computer assisted transcription

Contents

Executive Summary

This second 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 the last years for transcribing handwritten documents and aim to help the user in the transcription process. These techniques are being used to 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, 1]. 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 for 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 reads the prediction until an error is found. At this point the user corrects this error, generating a new, extended prefix. Then, 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 step, 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 [1].

1.2 Task T7.2

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.

In this period 2 of the project we have been working on the first level. The word graphs associated to lines or sentences obtained in previous tasks for some READ databases have been used for interactive-predictive HTR studies.

2 Preliminary CATTI results on READ text images

During this period we have been working with two READ collections: "The RSEAPV Collection" and the "Girona Collection".

2.1 The RSEAPV Collection

The first collection we have been working with, called "The RSEAPV Collection", has been provided by the "Real Sociedad Económica de Amigos del País de Valencia" (RSEAPV). It is a partnership that was established in 1776 by King Carlos III from 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 framework for discussion and treatment of the most important and cutting-edge issues of that time. The RSEAPV possesses an archive composed of more than 8,000 documents including the full documented history of the partnership from its foundation to nowadays. More than half of the archive documents belong to the 18th Valencian century in fields as diverse as economy, arts, literature, science and history, mostly written in the cursive style. This archive has been recently digitalized and made available to the public[1].

In this task we have chosen a document of this collection to test the CATTI system 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. These pages were annotated with two different types of annotations. First, a layout analysis of each page was manually done to indicate text blocks and lines, resulting in a dataset of 651 lines. Second, the dataset was completely transcribed line by line by an expert paleographer.

2.1.1 Quantitative results

The experiments carried out to assess the performance of the CATTI system, were performed using the WGs generated in previous tasks (see Deliverable D7.2).

In Table 1 we can see the Word Stroke Ratio (WSR). It is defined as the number of errors that the user must correct during the transcription process using the CATTI system. We can also see the the Word Error Rate (WER). It corresponds with the estimated post-editing human effort and is defined as the number of insertions, deletions

[1] `https://riunet.upv.es/handle/10251/18484`

and substitutions to carry out in the transcription proposed by the system to obtain the perfect one without any kind of assistance. Finally, the table also shows the estimated effort reduction (EFR) computed as the relative difference between WER and WSR. This value gives us a good estimate of the reduction in human effort that can be achieved by using CATTI with respect to using a conventional HTR system followed by human post-editing.

Table 1: WER, WSR and EFR using the RSEAPV Collection

WER	WSR	EFR
55.7	45.8	18.1

According to these results, to produce 100 words of a correct transcription, a CATTI user should only have to type 46 words; the remaining 54 would be automatically predicted by the system. On the other hand, if interactive transcription is compared with post-edition approach: for every 100 word errors corrected in post-edition approach the CATTI user would interactively correct only 82 . The remaining 18 words would be automatically corrected by CATTI, thanks to the feedback derived from other interactive corrections.

A detailed description of the work carried out with this dataset has been published in the 2017 Iberian Conference on Pattern Recognition and Image Analysis [2].

2.2 The "Oficio de Hipotecas de Girona" Collection

The second 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 17th 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. So far, around 400 pages have been transcribed. These pages are available on the READ platform.

2.2.1 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. The first model was trained using 48 pages perfectly annotated at two levels: layout analysis and transcription.

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 word graphs used in the CATTI process were generated in previous tasks (see Deliverable D7.2).

2.2.2 Quantitative results

As mentioned before, transcription is carried out in batches of around 50 pages. Table 2 shows the batches names, the number of pages and number of accumulated pages at the moment of processing each batch. Except the first batch, which was manually transcribed, all the batches were transcribed using the CATTI system.

Table 2: Number of pages and number of accumulated pages

Name	Pages	Accumulated Pages
b001	48	48
b002	50	98
b003	50	148
b004	50	198
b005	50	248
b006	50	298
b007	50	348

The system models used to recognize each batch are trained with the previously transcribed batches. In Table 3 we can see the estimated human effort (WSR), the corresponding estimated post-editing effort (WER) and the estimated effort reduction (EFR) for the different batches with (WER_T, WSR_T, EFR_T) and without (WER, WSR, EFR) taking the tags into account. The first row does not take tagger transcripts into account because tagging was first introduced in b003.

Table 3: WER, WSR and EFR using the Girona Collection

Train	Test	WER	WSR	EFR	WER_T	WSR_T	EFR_T
b001	b002	31.1	18.7	39.9	-	-	-
+b002	b003	37.1	24.6	33.7	47.7	36.2	24.1
+b003	b004	25.0	11.7	53.2	30.5	16.9	44.6
+b004	b005	21.1	10.2	51.6	23.6	14.0	40.7
+b005	b006	23.2	11.3	51.2	26.2	15.3	41.6
+b006	b007	23.5	11.8	49.7	26.7	14.9	44.2
+b007	b008	24.1	12.8	46.9	27.1	15.4	43.2

2.2.3 Qualitative results

As previously commented, the correct transcripts of each page are generated by a human expert using an implementation of the CATTI system. The interface used in this transcription was `http://transcriptorium.eu/demots/htr/index.php/ui/chapters/RH_Girona_1769`.

After each batch transcription process the human expert reported some feedback in order to improve the CATTI experience. The main concerns during the transcription

of the first batches were related with the interface implementation. For example, the user indicates that it would be very useful to be able to visualize all the line transcripts at the same time, or it would be very useful to have the pages numerated. After the improvement of the points indicated by the user, and according to the user comments, the CATTI experience was very comfortable. The user comments that in general the system works well, it is nimble and simple to use and faster than manual transcription. In addition, the tagging process is very fast and comfortable.

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 Fig. 1 the web version of the CATTI engine is shown.

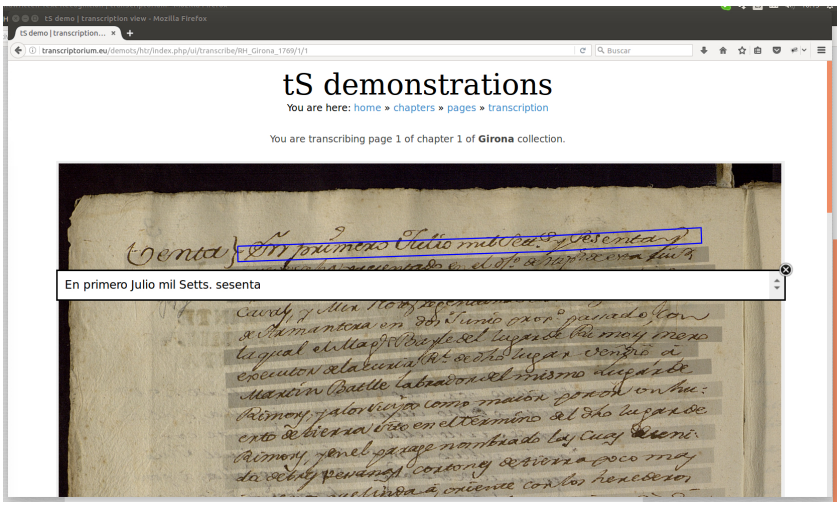

Figure 1: Example of the web implementation of the CATTI engine.

4 Plans for next period

Plans for the following period include:

- Improve the CATTI integration in Transkribus.
- Computer-Assisted transcription of untranscribed manuscripts of the Spanish Theater golden Age BNE collection.
- Investigate ways to combine both WG and Character Lattices (CL) for interactive-predictive transcription.

- Integration of the WGs and/or CLs into "Line Graphs" for integrated interactive-predictive correction of transcription and line detection.

References

[1] 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.

[2] Verónica Romero, Vicente Bosch Campos, Celio Hernández, Enrique Vidal, and Joan Andreu Sánchez. *A Historical Document Handwriting Transcription End-to-end System*, pages 149–157. Springer International Publishing, 2017.

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