READ

RECOGNITION & ENRICHMENT OF ARCHIVAL DOCUMENTS

D7.4

Interactive Predictive Transcription Engine P1

A toolkit for the interactive transcription of handwritten documents

Verónica Romero, Joan Andreu Sánchez, 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	28.12.2016
Date of last update	15.12.2016
Deliverable number	D7.4
Deliverable title	Interactive Predictive Transcription Engine P1
Type	Demostrator
Status & version	in progress
Contributing WP(s)	WP7
Responsible beneficiary	UPVLC
Other contributors	
Internal reviewers	Günter Mühlberger, Johannes Michael, Max Widemann, Nathanael Philipp
Author(s)	Verónica Romero, Joan Andreu Sánchez, Enrique Vidal
EC project officer	Martin Majek
Keywords	Interactive handwritten transcription, Computer assistive transcription

Contents

Executive Summary

The first 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 recently for transcribing handwritten documents and aim to help the user in the transcription process. In this deliverable, the state of the art of the interactive transcription approach is presented. Then, 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 recently 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 [2, 1]. It is refered 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 [1].

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.

In this first period 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 results

The first collection used in the experiments, 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 posseses 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 dare documents from 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 the recognition step.

In Table 1 we can see the estimated human effort (WSR). It is defined as the number of errors that the user must correct during the transcription process using the CATTI system. In the table the corresponding estimated post-editing effort (WER) is also shown. It is defined as the number of insertion, deletion and substitutions to carry out in the transcription proposed by the system to obtain the perfect one without any kind of assistence. 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

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

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 submitted to the next Iberian Conference on Pattern Recognition and Image Analysis [3].

2.2 The Girona Collection results

The second collection used in the experiments, called "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 have chosen a document of this collection to test the CATTI system. We have carried out both, laboratory experiments as weel as real user transcription experiments.

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. The next 50 pages of the document were also annotated with the layout analysis, resulting in a datased of more than 2000 lines. Then these lines were automatically transcribed (using the previous 48 pages to train the statistical models) and finally, an expert paleographer was in charge to correct the automatic transcription using the CATTI system. These two sets are available in the READ platform with the IDs 5146 (RH Girona 1769) and 5448 (RH Girona 1769(2)).

2.2.1 Quantitative results

The quantitative experiments have been carried out using the first 48 pages of the document. These experiments were performed using the WGs generated in previous tasks (see Deliverable D7.1).

In Table 2 we can see the estimated human effort (WSR). and the corresponding estimated post-editing effort (WER) is also shown. Finally, the table also shows the estimated effort reduction (EFR).

Table 2: WER, WSR and EFR using the Girona Collection

WER	WSR	EFR
32.8	18.4	43.9

According to these results, to produce 100 words of a correct transcription, a CATTI user should only have to type 19 words; the remaining 81 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 57 . The remaining 43 words would be automatically corrected by CATTI, thanks to the feedback derived from other interactive corrections.

2.2.2 Qualitative results

As previously commented, the next 50 pages were automatically transcribed using a system trained with the previous 48 pages. Then, this automatic transcription was corrected 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 the transcription process the human expert reported some points in order to improve the CATTI experience. The main concerns were related with the interface implementation. Next we summarize the different points commented by the user:

- In general the system works well, it is nimble and simple to use and faster than manual transcription.
- It would be very useful to be able to visualize all the line transcriptions at the same time.
- It would be very useful to have the pages numerated.
- The automatic system does not recognize the crossed out words.

3 A toolkit for the interactive transcription of handwritten documents.

A first working version of CATTI engine has been implemented and is available at `https://github.com/PRHLT/CATTI`.

This first version of CATTI has been integrated in the READ platform "Transkribus". Fig. 1 shows a page of the Girona database in the Transkribus platform is shown and the option CATTI transcription is activated (lower right corner)

A web version of CATTI has also been implemented and isshown in Fig. 2.

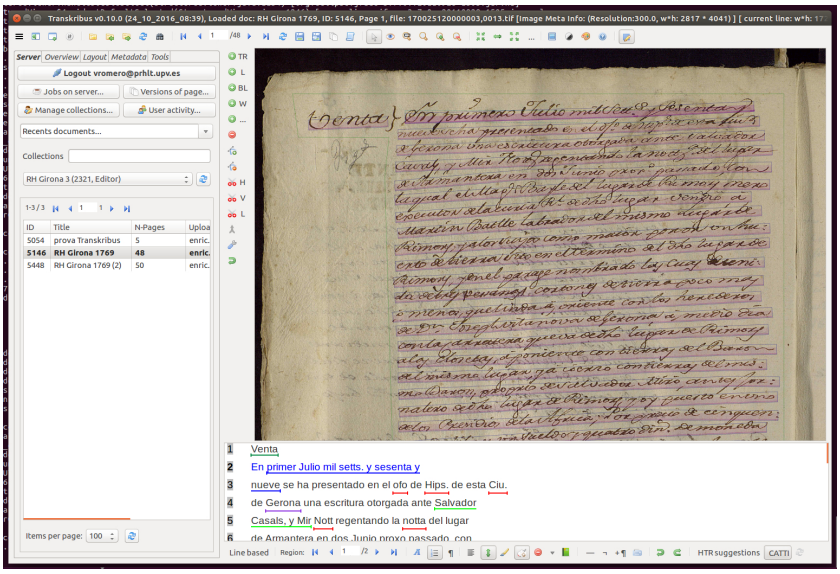

Figure 1: Example of the CATTI engine in the Transkribus platform.

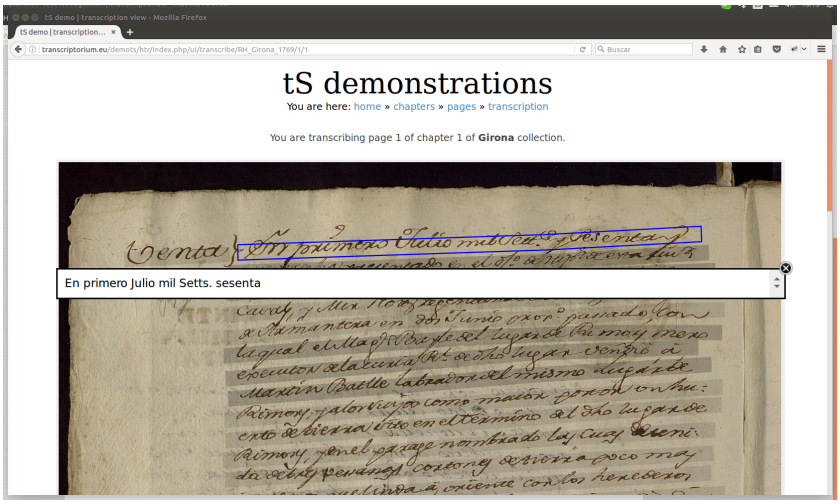

Figure 2: Example of the web implementation of the CATTI engine.

4 Plans for next period

The work on this task has just started, but thanks to the use of existing background, progress has been fast. Plans for the next period include:

- Improve the CATTI integration in Transkribus.
- Computer-Assisted transcription of untranscribed manuscripts of the Spanish Theatre 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 intereactive-predictive correction of transcritpion 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] A.H. Toselli, V. Romero, M. Pastor, and E. Vidal. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1824–1825, 2010.

[3] Celio Hernández Enrique Vidal Verónica Romero, Vicente Bosch Campos and Joan Andreu Sánchez. A historical document handwriting transcription end-to-end system. In *Iberian Conference on Pattern Recognition and Image Analysis*, 2017.