# tranScriptorium: A European Project on Handwritten Text Recognition

Joan Andreu Sánchez
Universitat Politècnica de
València, ITI
46022 Camí de Vera s/n
València, Spain
jandreu@dsic.upv.es

Günter Mühlberger
Universitaet Innsbruck
6020 Innrain 52
Innsbruck, Austria
guenter.muehlberger@
uibk.ac.at

Basilis Gatos
National Center for Scientific
Research "Demokritos"
15310 Patriarchou Gregoriou
Agia Paraskevi
Athens, Greece
bgat@iit.demokritos.gr

Philip Schofield
University College London
WC1E6BT Gower Street 1
London, UK
p.schofield@ucl.ac.uk

Katrien Depuydt
Instituut voor Nederlandse
Lexicologie
2300RA Matthias de Vrieshof
2-3, 2311 BZ
Leiden, Netherlands
Katrien.Depuydt@inl.nl

Richard M. Davis
University of London
WC1E 7HU Malet Street
London, UK
r.davis@ulcc.ac.uk

## ABSTRACT

The tranScriptorium project aims to develop innovative, efficient and cost-effective solutions for annotating handwritten historical documents using modern, holistic Handwritten Text Recognition (HTR) technology. Three actions are planned in tranScriptorium: i) improve basic image preprocessing and holistic HTR techniques; ii) develop novel indexing and keyword searching approaches; and iii) capitalize on new, user-friendly interactive-predictive HTR approaches for computer-assisted operation.

## Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; I.5.4 [**Pattern Recognition**]: Applications—*Text Processing*; I.7.5 [**Document and Text Processing**]: Document Capture—*Document analysis*

## Keywords

Interactive handwritten text recognition, digital libraries, document image analysis, crowd-sourcing

## 1. INTRODUCTION

tranScriptorium[1] is a three years project that started on January 2013 and it is funded by the European Union's Seventh Framework Programme. tranScriptorium project aims to develop innovative, efficient and cost-effective solutions for the indexing, search and full transcription of historical handwritten document images, using modern, holistic Handwritten Text Recognition (HTR) technology.

The tranScriptorium consortium is composed by experts in HTR, Document Image Analysis (DIA), linguistic resources developers, content providers, crowd-sourcing experts and integration experts. The tranScriptorium partners are: Universitat Politècnica de València (UPVLC), Spain; University of Innsbruck (UIBK), Austria; National Center for Scientific Research "Demokritos" (NCSR), Greece; University College London (UCL), UK; Institute for Dutch Lexicology (INL), Netherlands; University of London Computer Centre (ULCC), UK.

tranScriptorium will address the following specific objectives: enhancing HTR technology for efficient transcription; bringing the HTR technology to users; and, integrating the HTR results in public web portals. For achieving these objectives, tranScriptorium will develop HTR tools that will be tested in two real-life scenarios: in the first scenario, the HTR technology will be made available through a content provider site for individual users; in the second scenario, the developed technology will be integrated in an existing crowd-sourcing platform.

## 2. HTR TECHNOLOGY

Current state-of-the-art transcription products rely on technology for isolated character recognition (OCR) developed in the last two decades. But character segmentation is just impossible in unconstrained handwritten text images like those encountered in most old documents of interest (see Figure 1) to the project. tranScriptorium will use the new segmentation-free *off-line* HTR technology [2] for such transcription tasks.

In contrast with OCR, this new HTR technology does not need the characters or even the words of a handwritten text image to be previously segmented or isolated. To some extent, the transcription of (old) handwritten text images is

[1] http://transcriptorium.eu

*DocEng'13,* September 1013, 2013, Florence, Italy.
ACM 978-1-4503-1789-4/13/09
http://dx.doi.org/10.1145/2494266.2494294.

comparable with the task of recognising continuous speech in a (significantly degraded) audio file. And, in fact, recent technology for HTR borrows concepts and methods from the field of Automatic Speech Recognition, such as Hidden Markov Models (HMMs) and N-grams [1].

Currently available HTR technologies are far from offering satisfactory solutions. To obtain correct transcripts, heavy human-expert correction work is needed; but this "post-editing" process is inefficient and uncomfortable for the users and is not generally accepted by expert transcribers. As an alternative, computer assisted interactive predictive solutions [3] offer significant improvements in practical performance and user acceptance. In these approaches, the user and the system work interactively in tight mutual collaboration to obtain the perfect transcript of the given data [3].

To achieve good (plain or interactive) HTR accuracy, a combination of techniques is needed, such as layout analysis, text line extraction, preprocessing operations, lexical and language modelling, HMMs, etc. Although these technologies are already providing useful results in some cases, much remains to be developed, especially for historical documents, which suffer from typical degradations.

The models used in segmentation-free HTR are trained using already well known, powerful learning techniques, most of them based on the Expectation-Maximisation algorithm. tranScriptorium intends to make progress in automatic training techniques in order to achieve satisfactory accuracy.

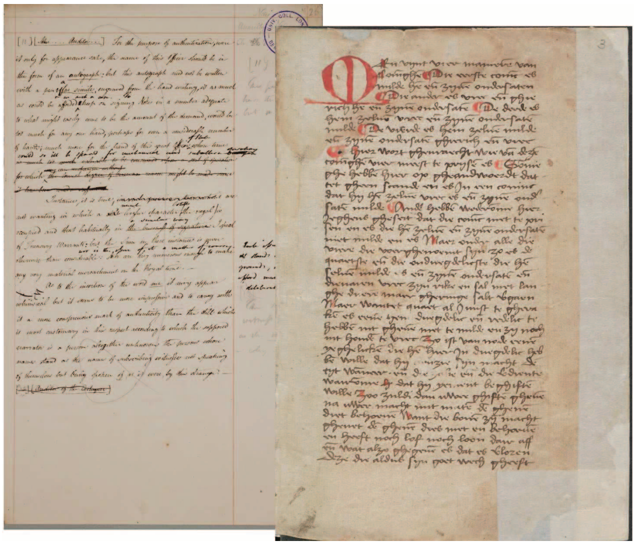

**Figure 1: Document samples to be processed in tranScriptorium.**

## 3. PROJECT OBJECTIVES

Despite recent significant improvements, currently available HTR technologies are still far from offering fully automated solutions for transcription. In this project, we will turn HTR into a mature technology by addressing the following objectives:

1. *Enhancing HTR technology for efficient transcription.* Departing from state-of-the-art HTR approaches, tranScriptorium will capitalise on interactive-predictive techniques for effective and user-friendly computer-assisted transcription [3].
2. *Bringing the HTR technology to users.* Expected users of the HTR technology belong mainly to two groups: a) individual researchers with experience in handwritten documents transcription interested in transcribing specific documents. For this kind of users, the HTR tools will be available through handwritten text image content provider portals. Archives and libraries will be also benefited from these users since they will be able to integretate the obtained transcripts in their collections; b) volunteers which collaborate in large transcription projects. For this kind of users, the HTR tools will be available through a specialised crowd-sourcing web portal which provides support for structured collaborative work.
3. *Integrating the HTR results in public web portals.* The HTR technology will become a support in the digitisation of the handwritten materials. Most digital libraries nowadays attach the output of modern OCR to the digitised pages of printed text documents. In a similar way, the outcomes of the tranScriptorium tools will be attached to the published handwritten document images. This includes not only full, correct transcriptions produced with the interactive HTR transcription techniques, but also partially correct transcriptions and other kinds of automatically produced metadata, useful for indexing and searching based on Key Word Spotting (KWS) techniques.

Within the tranScriptorium project span, it is intended to apply the developed HTR technology to historical documents in cursive handwriting, for which only HTR technology can offer appropriate solutions. tranScriptorium will focus on four languages: Spanish, German, English and Dutch.

## 4. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 600707 - tranScriptorium.

## 5. ADDITIONAL AUTHORS

Enrique Vidal, Universitat Politècnica de València, ITI, 46022 Camí de Vera s/n, València, Spain, email: `evidal@iti.upv.es` . Jesse de Does, Instituut voor Nederlandse Lexiicologie, 2300RA Matthias de Vrieshof 2-3, 2311 BZ, Leiden, Netherlands email: `jesse.dedoes@inl.nl`

## 6. REFERENCES

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

[2] A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta. Integrated Handwriting Recognition and Interpretation using Finite-State Models. *Int. Journal of PRAI*, 18(4):519–539, June 2004.

[3] A.H. Toselli, E. Vidal, and F. Casacuberta. *Multimodal Interactive Pattern Recognition and Applications.* Springer, 1st edition edition, 2011.