Toselli, AH.; Leiva, LA.; Bordes-Cabrera, I.; Hernández-Tornero, C.; Bosch Campos, V.; Vidal, E. (2018). Transcribing a 17th-century botanical manuscript: Longitudinal evaluation of document layout detection and interactive transcription. Digital Scholarship in the Humanities. 33(1):173-202. https://doi.org/10.1093/llc/fqw064

Transcribing a 17th Century Botanical Manuscript: Longitudinal Evaluation of Document Layout Detection and Interactive Transcription

Alejandro H. Toselli[1]
ahector@prhlt.upv.es

Luis A. Leiva[1]
luileito@prhlt.upv.es

Isabel Bordes-Cabrera[2]
isabel.bordes@bne.es

Celio Hernández-Tornero[1]
cehertor@prhlt.upv.es

Vicent Bosch[1]
viboscam@prhlt.upv.es

Enrique Vidal[1]
evidal@prhlt.upv.es

[1] **PRHLT Research Center**
Universitat Politècnica de València
Camino de Vera, s/n
46022 Valencia (Spain)

[2] **Biblioteca Nacional de España**
Paseo de Recoletos, 20
28071 Madrid (Spain)

February 16, 2017

Transcribing a 17th Century Botanical Manuscript: Longitudinal Evaluation of Document Layout Detection and Interactive Transcription

Abstract

We present a process for cost-effective transcription of cursive handwritten text images that has been tested on a 1 000 pages 17th century book about botanical species. The process comprised two main tasks, namely: (1) preprocessing: page layout analysis, text line detection, and extraction; and (2) transcription of the extracted text line images. Both tasks were carried out with semiautomatic procedures, aimed at incrementally minimizing user correction effort, by means of computer-assisted line detection and interactive handwritten text recognition technologies.

The contribution derived from this work is three-fold. First, we provide a detailed human-supervised transcription of a relatively large historical handwritten book, ready to be searchable, indexable, and accessible to cultural heritage scholars as well as the general public. Second, we have conducted the first longitudinal study to date on interactive handwriting text recognition, for which we provide a very comprehensive user assessment of the real-world performance of the technologies involved in this work. Third, as a result of this process, we have produced a detailed transcription and document layout information (i.e., high-quality labeled data) ready to be used by researchers working on automated technologies for document analysis and recognition.

1 Introduction

Nowadays, due to the proliferation of web-based technologies, digital libraries as well as public and private archives are making a major investment in the mass digitization of their historical handwritten document collections. As a primary goal, this aims at making accessible all their valuable manuscripts to the general public for educational and research purposes. While digitization can be quickly carried out by using the latest scanners and digital cameras currently available on the market, this is not enough to open up the *content* of these manuscripts. In many cases, such content, even after digitization, remains accessible only to researchers with Paleography skills. Then, professional transcription is the natural next step. However, for this step to become realizable, large amounts of time, effort, and investment are necessary. Together, these caveats prevent heritage institutions to initiate such projects at a large scale. And still there is not a single cultural heritage institution that is not willing to offer full text access of their manuscripts, a fact that could enable important possibilities for both researchers and citizens in general.

At present, the transcription of historical handwritten documents is generally carried out by experts in Paleography, who are specialized in reading ancient scripts, characterized, among other things, by different calligraphy and print styles from diverse historical contexts and time periods. Eventually,

how long experts take to carry out a transcription of one of these manuscripts depends on their skills and experience. For example, to manually transcribe many of the relatively simple pages of the book described in this article, they can spend as much as half an hour per page.

One way to speed up the aforementioned transcription process is by means of automated methods that allow the transcribers to transcribe such digitized documents both accurately and efficiently. However, state-of-the-art cursive handwritten text recognition systems can by no means substitute the human expertise in this task. These systems have indeed proven useful for restricted applications involving constrained-form handwriting and/or fairly limited vocabulary (such as postal addresses or bank checks), achieving in this kind of tasks a relatively high recognition accuracy (Dimauro et al., 2002; Srihari and Keubert, 1997). But in the case of difficult, unconstrained handwritten documents such as historical manuscripts, current technology typically achieves only results that are far from being directly acceptable in practice (Romero et al., 2013; Vinciarelli et al., 2004).

By explicitly acknowledging these limitations, the Computer Assisted Transcription of Text Images (CATTI) framework (Romero et al., 2012; Toselli et al., 2010) is intended to speed up the process of manually transcribing such documents in an interactive way. While CATTI primarily aims at producing high-quality, professional transcripts, it is also promising for crowdsourcing transcription projects such as, for example, Transcribe Bentham (Causer et al., 2012; Causer and Wallace, 2012), which originally sought to manually transcribe a large handwritten document collection in a collaborative way. In these cases, CATTI is expected to increase the productivity of the crowdsourcing volunteers and to allow for more accurate and homogeneous results (Leiva et al., 2011; Romero et al., 2009).

In this article, we focus on testing the real use of CATTI technology at a moderately large scale. To this end, a historical monothematic collection entitled "Historia de las plantas" has been chosen. The transcription of such a large manuscript is a real major undertaking task and we have started with the first volume of this collection, which comprises more than 1 000 pages. So far, our results include the following: 1) We have cost-effectively produced a detailed, human-supervised transcription of this volume, ready to be searchable, indexable, and accessible to cultural heritage scholars as well as the general public. 2) We present for the first time a longitudinal case study of the entire transcription task. Our study shows how the CATTI system, using the user feedback obtained at no cost from the proper interactive transcription process, is able to produce increasingly accurate transcription predictions over time. 3) The final transcripts are annotated at different levels and have been reviewed by human experts, thus they can be considered reference data —most commonly known as *ground truth* in Computer Science— which can be used by other researchers working on the development of technologies for automated document analysis and recognition.

It is worth emphasizing that results 1) and 3) constitute an excellent basis for cost-effectively undertaking cutting-edge digital, hypertext editions of manuscripts processed this way. The detailed, labeled data obtained through the proposed process can be easily used to provide advanced features such as page, line, and word alignments of images and transcripts, as well as a rich, layered view of the

transcripts, with different levels of detail such as pure diplomatic transcription, expanded abbreviations, modernized capitalization, diacritics and punctuation, multilingual annotations, and so on.

We should mention that our host institutions, the Universitat Politècnica de València (UPV), and the Biblioteca Nacional de España (BNE), held a collaboration project in which BNE kindly provided the digitized images of the aforementioned manuscript. The BNE also offered invaluable support by providing personnel and resources during the whole transcription process, for which Paleography students from the Universidad Complutense de Madrid (UCM) took part in the project.

This article is organized as follows. Section 1.1 introduces the concept of cursive Handwritten Text Recognition and how it is essentially different from classical Optical Character Recognition. Section 1.2 describes the manuscript we have transcribed. Section 2 introduces the transcription criteria and rules we followed during the transcription process. Section 3 presents in detail the aforementioned transcription process, including the interactive layout analysis and handwritten text recognition subsystems, and the web-based transcription application. Section 4 describes the experimental evaluation, presenting how the manuscript was partitioned for human supervision, and the results of the longitudinal study we followed for both the layout analysis and the transcription process itself. Finally, Section 6 describes the high-quality transcription data that were produced as a result of this process. Section 7 concludes this article and highlights some avenues for future work.

1.1 Background

Handwritten text image transcription, mostly known as "off-line" handwriting text recognition (HTR), is the task of converting handwritten text images into an electronic text format. HTR refers to recognizing *cursive* handwriting (Fig. 1), and is thus a task quite apart from Optical Character Recognition (OCR), which is (only) suitable for *printed* text. In fact, there is no OCR system that supports cursive handwriting recognition as of today (Alabau and Leiva, 2012).

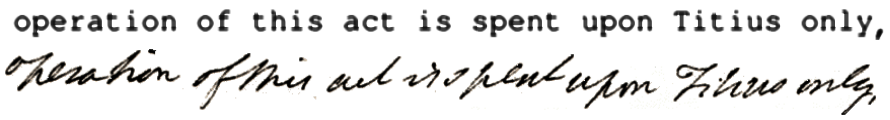


Figure 1: OCR is suitable for documents with predictable inter-word and inter-character spaces, consistent typesetting, etc. (top). HTR, instead, has to deal with slanted and skewed continuous handwriting, irregular calligraphy, inconsistent spacing and so on (bottom). Here, OCR can reach close to 100% character recognition accuracy, while state-of-the-art HTR recently achieved 96% in this example (Sanchez et al., 2015).

For some time in the past decades, the interest in HTR was diminishing under the assumption that modern computer technologies would soon make paper-based text documents useless. However, more recently, the task has become an important research topic, especially because of the increasing number of online archives and digital libraries publishing huge quantities of digitized legacy documents. This fact has fostered the creation of digital libraries by public institutions such as the British Library,[1]

[1] `http://www.bl.uk`

the World Digital Library,[2] the Europeana web portal,[3] the Biblioteca Digital Hispánica[4] (of the BNE) or the Miguel de Cervantes virtual library,[5] among many others. These efforts not only are intented to preserve the cultural heritage, but also aim to offer digital editions of historic documents and document collections that allow to copy, edit, translate, index, and search for textual information in these collections.

Although many historical manuscripts have been manually transcribed as of today, most digital libraries only host scanned images of their (large) original handwritten collections. Therefore, the vast majority of these documents, thousands of terabytes worth of digital image data, remain waiting to be transcribed into an electronic format that would provide publishers, historians, demographers, and other researchers with new, convenient ways of consulting and working with these documents.

While the state of the art in HTR has dramatically advanced in the last few years, current technology is still far from providing sufficiently accurate transcripts for typical applications in a fully automated way. Therefore, rather than relying on inefficient manual procedures, interactive-predictive techniques such as CATTI are ideal candidates to achieve the required quality in a user-friendly way.

1.2 Manuscript Overview

When selecting a manuscript to develop and test HTR tools, we must ensure that our research interests actually match those of our target audience, i.e., humanists and, more specifically, paleographers. For that reason, we decided to select an extensive collection that, on the one hand, would allow a paleographer investigate "all questions that must be done" (Petrucci, 1985) (What, Who, When, Where, How, Why?) and, on the other hand, would be large and interesting enough to conduct possible studies about its content. The manuscript collection chosen for the present work, entitled "Historia de las plantas" (Mss 3357–3363, Biblioteca Nacional de España), was written in Medieval Spanish language by Bernardo de Cienfuegos, one of the most outstanding Spanish botanists of the 17th century. We will use the short name PLANTAS to refer to this manuscript throughout this article.

Cienfuegos was born c. 1580 in Tarazona, a little village of Zaragoza (Aragon, Spain). Son and grandson of alchemists, he had always repudiated this discipline because it led his family to bankruptcy. He studied Medicine at the University of Alcalá de Henares (Madrid, Spain), and later he worked as a Humanities teacher in the same institution in 1599. But his actual passion was Botanics, to which he had devoted most part of his life: traveling through different Spanish regions, looking for different plants; studying classical and modern botanists in their original languages; and creating a definitive compendium that included almost all of the botanical knowledge until the beginning of the 17th century. Bernardo de Cienfuegos started to write his manuscripts in 1627, when he was already living in Madrid. However, the lack of sponsors and a disease that led him to the death (c. 1640), precluded the completion and publication of his work (Arévalo, 1935; Pardo-de Santanya et al., 2014).

[2] http://www.wdl.org
[3] http://www.europeana.eu
[4] http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio
[5] http://www.cervantesvirtual.com

Nevertheless, later botanists such as Antonio José Cavanilles or Ignacio Jordán de Asso studied, admired, and recognized his work. Cavanilles, for example, dedicated the genus "Cienfuegosia" to him, whereas Willdenou created the "Cienfuegia" and Jussieu, the "Fugosia" (Cavanilles y Centi and Lagasca y Segura, 2000).

Plantas is an unpublished pre-linnaean masterpiece that contains descriptions of plant species by ancient and contemporary authors, practical use cases (e.g., medical or esoteric), and illustrations. Cienfuegos carefully acknowledges the sources of his work, not only text-wise, but also illustration-wise. Among the illustrations, one can find not only different versions of depictions of plant species, but also original pictures of the author, many of which are watercolored. All these peculiarities make the Plantas collection a valuable source of information for the specialized public beyond basic research, especially for those interested in reviewing the botanical knowledge of that historical period.

It is worth remarking that a multilingual vocabulary is thoroughly used in Plantas; namely: Latin, Portuguese, Catalan, French, German, English, Flemish, Polish, and Bohemian, among others. It is so because the author frequently identifies (and describes in a summarized way) different plant species in different languages. Of course, this peculiarity makes the manuscript an even more challenging and interesting test-bed for assessing HTR technologies. Further, Plantas contains a huge amount of possible descriptions of plant species, places, and people names. This turns the content of this work into a potentially interesting historical glossary, adding more value to the mere transcription (and later release) of the content of this work. In other words, if this work had seen the light at that time, it could have stood out as one of the most important botanical texts so far. A more detailed information about this masterpiece and the author can be found in Blanco Castro et al. (1994).

In general, the Plantas collection suffered from typical degradations (stains, tears, ink bleeding-through, etc.) although it is in a reasonably good state of preservation, facilitating thus its readability. His author used a quill-pen and black ink on paper support (see Fig. 2), and the writing style is classified as "humanist italic writing", a script-type very similar to today's calligraphy style. The collection is composed of seven volumes and is currently archived at the Biblioteca Nacional de España. A medium-resolution PDF collection is available online at Biblioteca Digital Hispánica (BDH).[6]

In this article, the first volume of Plantas (Mss 3357) was considered for complete transcription, leaving the remaining ones for future transcription tasks or for automatic indexing procedures by means of keyword spotting approaches (Toselli et al., 2016). The first volume has 49 pages at the beginning comprising indices, reference tables, a botanical glossary in different languages, and a 36-page preface written by Cienfuegos. This is followed by 887 numbered pages that contain 152 chapters about cereals and related plants, including 126 botanical illustrations. All in all, the first volume has 1 035 pages, containing about 20 000 handwritten text lines. This volume was digitized at 300 ppi in 24 bit RGB color and saved as TIFF images by using an i2S SuprascanII scanner with the software Digibook.[7]

[6] See the BDH bibliographic work (`http://bdh.bne.es/bnesearch/detalle/bdh0000140162`) and their digitized work (`http://bdh-rd.bne.es/viewer.vm?id=0000140162`).

[7] `http://www.digibook.com/en/index.html`

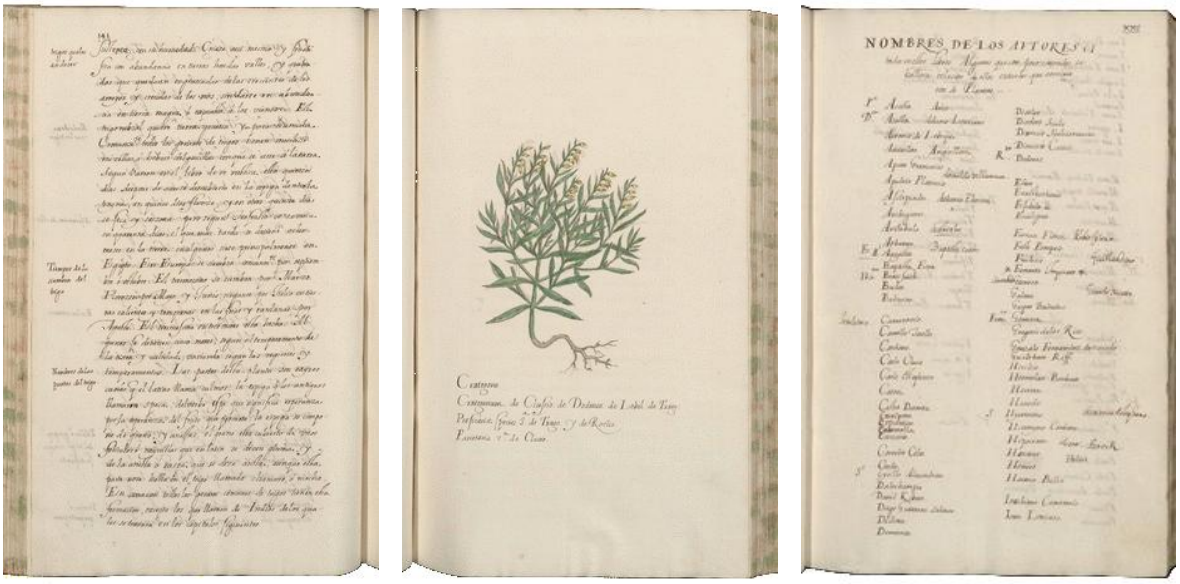


Figure 2: Examples of pages from the first volume of PLANTAS.

2 Transcription Criteria

Before conducting the actual transcription process, a work methodology was established, considering that it was necessary to unify some computer science and humanistic criteria. It was agreed that the primary transcription target should be a transliteration of the text images as close as possible to the original text. For this reason, we decided to adopt the so-called *diplomatic* (or *paleographic*) transcription style (Ortí Cárcel, 1997), which is frequently used in diplomatic editions. This transcription style aims to be as accurate as possible, preserving all significant characteristics of the original manuscript, including e.g. spelling, deletions, insertions, and other text alterations, while also adapting the transcripts to the currently accepted linguistic and orthographic norms. Actually, we essentially adopted the rules established by the *Comission Internationale de Diplomatique* (Bautier, 1984), with some minor changes. These changes aimed to meet the requirements raised by the HTR technology to be used (see Section 3.2) and also to allow incorporating feature-rich information into the transcripts by means of word tagging (see Section 2.1), as a step towards creating future digital hypertext editions.

Among the most important transcription conventions applied, we can mention the following:

- Transcripts must respect to the maximum extent the original document. In case of doubt, contemporary grammatical criteria must be used.
- Writing lines and text blocks must preserve the original format as much as possible.
- Uppercase and lowercase letters are transcribed as they appear in the original text image.
- Abbreviations (Fig. 3) are transcribed as such, including tags that specify how they should be expanded into the corresponding full forms.
- Special characters such as large “s”, “i” or “f”, among others, must be transcribed as normal characters; i.e., using modern characters (see Fig. 4 and Fig. 5).

- If two consecutive but differently recognizable words are connected (as "*dela*" and "*deM*$^{d}_{\cdot}$" in Fig. 3), these must be separated (as "*de la*" and "*de M*$^{d}_{\cdot}$"), unless they make a contraction. Conversely, if a blank space appears within a recognizable word (as in "*algo don*" in Fig. 3 and "*es piga*" in Fig. 6), both parts must be joined (as in "*algodon*" and "*espiga*").
- Hyphenated words must be tagged including both the starting part of the word and its continuation in the next line (Fig. 5).
- Crossed-out words are tagged, and also transcribed, if they are readable (Fig. 6).

2.1 Transcribing with Word Tags

As commented in Section 1.2, PLANTAS contains (among other features) numerous abbreviations, multilingual words, crossed-out and added text, etc. Therefore, in order to mark these features and thereby enrich the manuscript's ground truth data, a word tagging mechanism was necessary. To this end, a set of tags was defined using the standard XML TEI format.[8] For most of these tags, adequate shortcuts were defined based on two special *escape* characters ("$" and "#"). These shortcuts can be easily converted into TEI and have proved very convenient to avoid handling otherwise cumbersome fully expanded TEI transcripts, both for the HTR/CATTI systems and the human transcribers. The most frequent tags are listed in Table 1.

Table 1: Most frequent and relevant tags.

Description	**Tagged transcript**
Initial *and* final hyphen marks	`$-` *and* `-$`
Deletion	`$/word`
Expansion	`$.word`
Catchword	`$>word`
Underline	`$_word`
Superscript	`$^word`
Latin, French, Portuguese, etc.	`$l:word`, `$f:word`, `$p:word`, etc.
Crossed out legible *and* illegible words	`#word` *and* `#`

These tags are used to prefix transcribed words according to their corresponding text features, as illustrated in the figures 3-6 commentd above.

[8] `http://www.tei-c.org/index.xml`

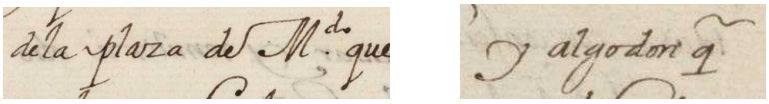


Figure 3: Examples of the handwritten abbreviations $M^{d.}$ (*Madrid*) and $\tilde{q}$ (*que*). The corresponding tagged transcripts are: `M.$.Madrid` and `q$.que`.

Figure 4: Example of text handwritten in Latin. The corresponding tagged transcript is: `$l:Atsi $l:triticeam $l:messem $l:robustaque $l:farra`.

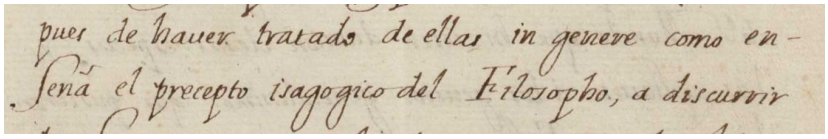


Figure 5: Example of two adjacent text lines with the hyphenated word *en-seña*. The corresponding tagged transcripts of its initial and final parts are: `en$-` and `-$seña`.

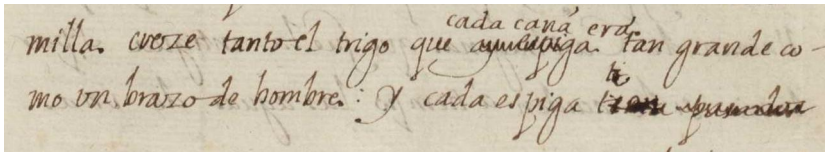


Figure 6: Example of added text along with two lines with readable and unreadable crossed-out words, transcribed as: `<add>cada caña era</add> ... que #ay #espiga tan ... y cada espiga # #`.

3 Transcription Process

As mentioned in Section 1, transcribing the first volume of Plantas was carried out in collaboration with BNE, which supplied the aforementioned manuscript, together with the cooperation of UCM, which provided us with Paleography expertise and four students in History who would perform the main transcription work. More specifically, the students were assigned the task of testing for several months the HTR tools developed by UPV. It is worth mentioning that the recruited students were actually not experienced in Paleography before enrolling in this work. Therefore, a qualified paleographer was additionally hired in order to supervise the transcripts produced by the students and also to help them by solving their doubts and answering their comments and questions. From now on, such a role will be referred to as "supertranscriber".

The whole transcription workflow is summarized in Fig. 7. It mainly comprises five tasks, the first two belonging to the so-called preprocessing phase, while the last three belong to the transcription phase: (a) layout analysis, for detecting text regions and text lines; (b) user supervision, for checking and amending (if necessary) block/line detected regions; (c) line preprocessing and feature extraction; (d) decoding and wordgraph generation; and (e) transcription itself. For the last step, two different, mutually-excluding transcription modalities were considered: manual transcription (baseline condition) or transcription assisted by CATTI, the interactive-predictive transcription system (experimental condition). This allowed us to compare both conditions in terms of efficiency and effectiveness. In

Fig. 7, tasks highlighted with the "person" symbol are the ones where paleographers can intervene directly (black color) or in a semi-supervised way (gray color).

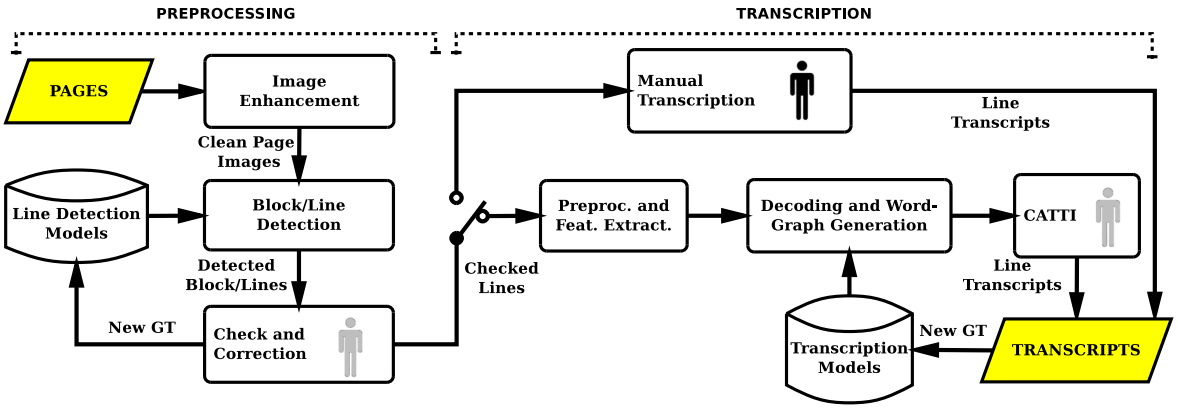


Figure 7: Workflow diagram of the overall transcription process. Checked Lines are selected for transcription, either in manual or interactive (CATTI) mode. **GT** stands for ground truth (fully supervised, labeled data). The stages with the "person" symbol are the ones where paleographers can influence the system directly (black color) or in a semisupervised way (gray color).

Layout analysis, line detection, text decoding, and wordgraph generation (to be explained later in Section 3.2) were performed by means of systems that make use of statistical models. As initial dataset to train a first version of these models, we used the Prologue chapter (38 pages), which was completely transcribed by the supertranscriber. Then, the models were periodically retrained when new batches of transcribed pages became available. This allowed our CATTI system to learn from user feedback and thereby improve its recognition and prediction performance over time.

To fairly balance the transcription tasks among the four participants, chapters were assigned in a way that each participant would transcribe a similar number of pages. In addition, considering both transcription modalities to be tested, a balanced number of tasks were designated for the baseline (manual) and experimental (CATTI) conditions. Section 4.1 provides more details about the transcription task assignments.

3.1 Layout Analysis

Layout analysis consisted in the segmentation of text blocks and lines, which was carried out as a three-step top-down sequential procedure. First, usual image enhancement techniques were applied to each page in the current batch of digitized images. Then, for each enhanced page image, text blocks were automatically detected and manually reviewed. Finally, text lines for each detected block were also automatically extracted and interactively reviewed. These steps, illustrated in Fig. 8, are described in more detail below.

3.1.1 Image Enhancement

Three usual image processing procedures were applied to correct geometric distortions and to enhance the overall quality of each page image: global skew correction, noise reduction, and contrast enhancement (Baird, 1995).

3.1.2 Text Block Detection

After image enhancement, text and no-text areas were detected and labeled. Specifically, the areas of interest were: page, number, heading, main text block, catchword, and drawings. Some examples can be seen in Fig. 8 (a–b).

Block detection and labeling were performed semiautomatically. First, an automatic identification of text areas was performed by means of horizontal and vertical border detection methods, based on the use of enhanced profiles (Malleron and Eglin, 2011; Ramel et al., 2007). After this automatic process, all detected areas were verified, labeled and (if necessary) manually corrected by means of a specialized layout editing tool.

3.1.3 Text Line Detection and Extraction

After block detection, the text lines found in the text blocks are detected by means of a semiautomatic iterative process (Bosch et al., 2014).

First, a simple projection profile technique is used to obtain an initial line segmentation of a first batch of images. This is reviewed by a user, to ensure that the resulting lines have actual ground truth quality. Once a batch has been processed and supervised, the system uses this information to (re-)train its statistical models before processing the next batch of page images. By retraining the models after each batch, the system accuracy improves, thereby reducing effectively the overall human effort, as will be shown in Section 4.

In a final step, dynamic programming techniques are used to actually extract (segment) the detected lines. Fig. 8 (b–c) show examples of detected and segmented text lines.

3.2 Text Line Transcription System

After line detection and extraction, transcribers performed their tasks on a line-by-line basis, using a web-based system (to be described in Section 3.3) which supported either manual transcription or interactive (CATTI) assistance. In this section, we outline the HTR technology behind CATTI and the CATTI system itself.

3.2.1 Outline of HTR Technology and Wordgraphs

The HTR technology used in this work is based on *character* hidden Markov models (HMMs) and *language* models (N-grams), following the fundamentals presented, e.g., by Bazzi et al. (1999); Toselli

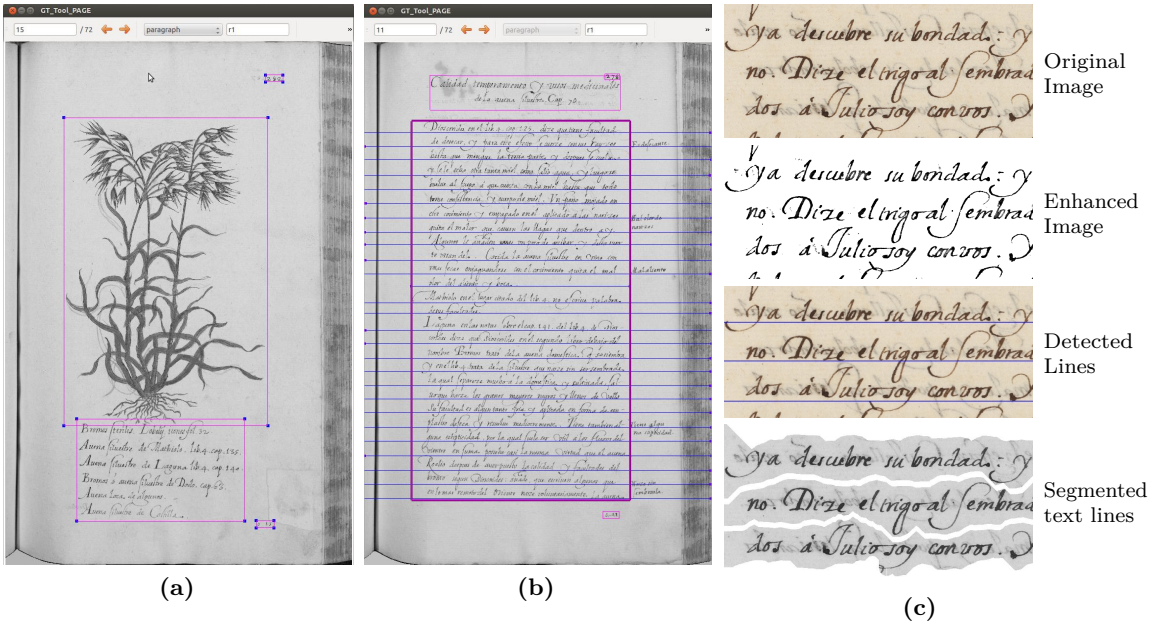


Figure 8: Examples of text block detection (a–b) and text lines detection (b). Details of text line detection and extraction are shown in (c).

et al. (2004); Vinciarelli et al. (2004). Both HMMs and N-grams, to be described below, are statistical models trained from examples of transcribed handwritten text line images.

In sum, an HTR system takes as input a handwritten text line image and outputs the most likely word sequence. Such a system generally comprises two main building blocks: one for image preprocessing and feature extraction, and other for image decoding.

On the one hand, image preprocessing aims at normalizing text size and other handwriting style attributes, such as slope and slant. Then, the feature extraction module converts each preprocessed line image into a sequence of numerical vectors; i.e., a sequential numerical representation that can be understood and processed by computers. Details of these two modules can be seen in Romero et al. (2012); Toselli et al. (2004).

The decoding process, on the other hand, relies on models at three hierarchical perception levels: optical, lexical and syntactic. The optical level is modeled by character HMMs, which probabilistically account for the possible shapes of each character, punctuation mark, etc. The syntactic level is modeled by an N-gram language model, which provides statistical information about which words are likely to follow each other. The lexical level consists in a vocabulary containing the possible spellings of each word, according to the characters modeled by the HMMs. Using well-known statistical estimation methods (Jelinek, 1998; Romero et al., 2012), HMMs are trained with examples of transcribed line images, while N-grams are learned using only the plain text from line transcripts, possibly supplemented with additional texts obtained from previously transcribed documents. The vocabulary is finally compiled as a byproduct of N-gram training.

For a given input image, represented by a feature vector sequence, the decoding process itself is

performed with the Viterbi search algorithm (Jelinek, 1998; Romero et al., 2012). Moreover, rather than obtaining just a single best hypothesis (i.e., a single machine-decoded transcript), a set of alternative hypotheses can be obtained and represented as a "wordgraph" structure. Put it simply, a wordgraph stores in a compact way a huge set of most likely HTR transcripts, along with the associated recognition probabilities, word image segmentations, and other useful informations of the given text line image.

3.2.2 CATTI System

The CATTI framework has already been presented in full detail by Romero et al. (2012), therefore we only provide here a succinct overview. In a nutshell, a human transcriber (hereafter called "user"), who is responsible of the correctness of the output transcripts, is assisted by a HTR-based interactive system. The transcription process begins when the system proposes a full transcript of a line image. In each interaction step, the user validates a prefix of the transcript which is error-free and introduces some amendments in order to correct the erroneous text that follows the validated prefix, thereby producing a new prefix. At this point, the system takes into account the new prefix and tries to automatically fix possible remaining word-level errors, by searching for a suitable suffix or prediction. This process, which can be seen as a kind of multi-word auto-completion mechanism, is repeated until a fully correct transcript of the input line image is reached. A key point of this interactive process is that, at each step, the system can take advantage of the user-consolidated prefix to produce hopefully improved results. Fig. 9 provides a graphical example of a CATTI session. It is worth pointing out that, in this example, the system was able to provide the right transcript, including a correct prediction of a tagged abbreviation, after just one user intervention.

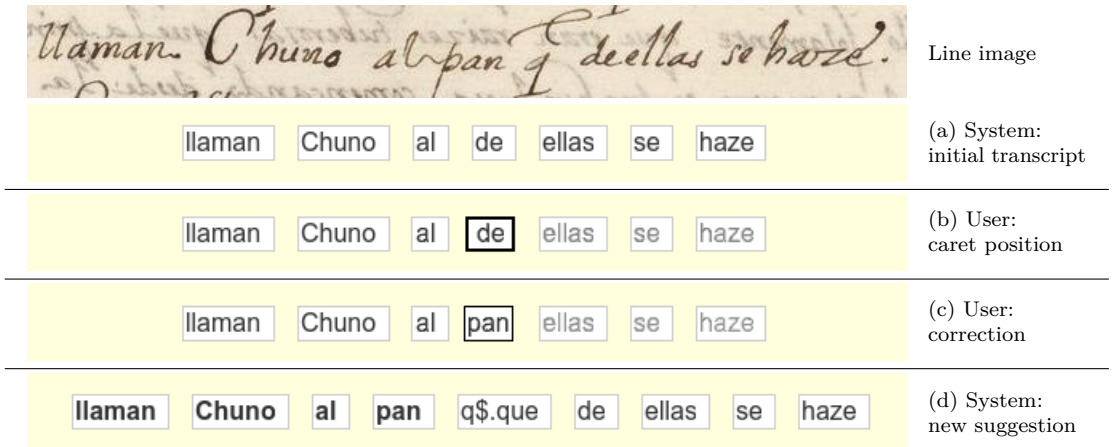


Figure 9: A CATTI session example. The system begins suggesting an initial transcript (a). Then the user moves the caret (b), to amend the first wrong word (c). This implicitly validates a prefix and thus invalidates the corresponding suffix (in lightgray color). The system uses this information to provide a (hopefully better) suitable continuation (d) of the user prefix (which becomes now printed in bold typeface).

To achieve the computing speed demanded by such an interactive-predictive operation, CATTI relies on the aforementioned wordgraph structure. Using a wordgraph, the initial transcription hypothesis

can be obtained easily. Then, at each interaction step, the wordgraph helps CATTI to perform the fast computations needed for real-time prediction of the most likely continuation of the transcript parts previously entered, validated, or corrected by the user.

3.3 Web-based Transcription Application and Interaction Protocol

A decoupled client-server architecture was adopted for our CATTI prototype (Fig. 10). On the one hand, a web browser (the client) implements a graphical User Interface (UI) which pulls data from a remote web server. The hardware requirements in the client are very low, since all tasks related to HTR and CATTI are carried out remotely on a server, so virtually any computer (including netbooks, tablets or 3G mobile phones) should be capable of operating the UI. On the other hand, the client uses a custom API to communicate with an HTR engine through binary TCP sockets. This way, a dedicated, persistent connection is established between the client and the HTR engine. In addition, because all architecture components are loosely coupled, the server does not depend of the implementation of either the HTR engine or the client.

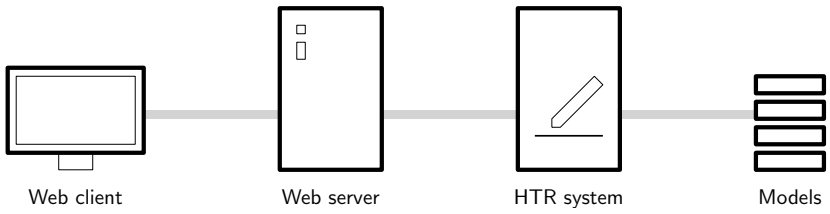


Figure 10: System architecture.

The UI shows the book divided into chapters (Fig. 11a) on the home page. Then, each chapter is divided into pages (Fig. 11b). The user can transcribe one page at a time, on a line-by-line basis. A login module allows the users to identify themselves, in such a way that the application can provide dedicated content to each user.

On the main transcription area (Fig. 11c), whenever the user clicks on a text line, the main application unfolds a keyboard input area (Fig. 11d). In the left margin of each page, lines may be color-marked as validated (all the text was reviewed), pending (only a fraction of the current line image has been transcribed), or locked (someone else is working on the same text line image).

The same UI supports the two transcription modalities of our study; namely, manual and interactive (CATTI) assistance. In both modalities, special editing operations were implemented in order to better assist the user in the transcription process:

Substitute An erroneous word is replaced with a correct word.

Insert A new word is inserted between two words that are both assumed to be correct.

Delete An incorrect word between two correct words is removed.

Merge Two consecutive (incorrectly split) words are concatenated to generate a correct word.

Split An incorrect word is split into two different (correct) words.

Validate The full transcript is accepted.

Reject The user indicates an erroneous word in the proposed transcript and the system automatically proposes a new continuation, in which the first word is different from the erroneous one.

All these operations, except the last one, are common to both transcription modalities (manual and CATTI). In CATTI operation mode, "Reject" allows the user to obtain alternative predictions in a fast and (hopefullly) comfortable way, by simply clicking on the erroneous word.

These edit operations together constitute the so-called interaction protocol; that is, a set of operations with which the user is expected or allowed to interact with the system. They are meant to be used in a left-to-right scan over each line image, and the application of each edit operation has the immediate effect of establishing a new, increasingly validated transcript prefix, as needed by CATTI to determine its best prediction for a suitable transcript continuation.

Finally, for analytics purposes, an event logging mechanism was included in the web application. It allowed us to register all user interactions at a fine-grained level (e.g., keyboard and mouse activity, client/server messages, etc.). The generated log files were stored in XML format for later analysis, in order to obtain the results presented in the next section.

4 Evaluation and Results

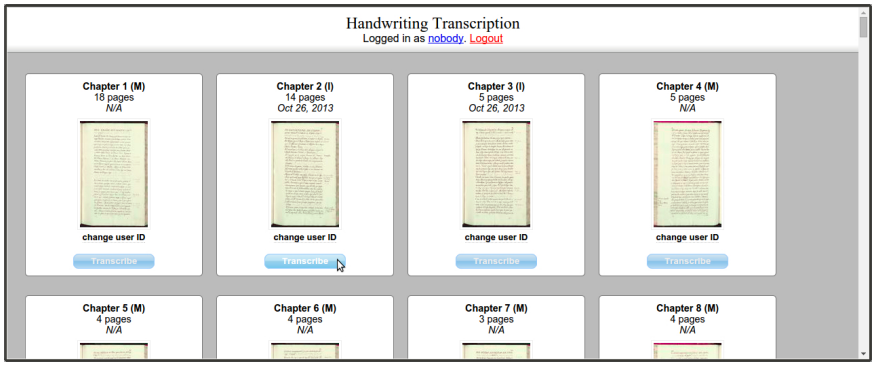
In this section we describe how Plantas was partitioned for transcription and how partitions were distributed among the participants. Then, we provide results on the longitudinal study, including the evaluation of layout analysis and text transcription, both from the users' and the system's perspective. To conclude this section, we report qualitative observations and discuss the comments raised by the participants, aimed at shedding more light about the observed results.

We should note that the principles of a longitudinal study were adopted, by following the users over a given number of sessions (four months), monitoring their interactions with the system to evaluate their progress, as well as the system's, over time. To the best of our knowledge, this is the first evaluation of its kind in the HTR literature.

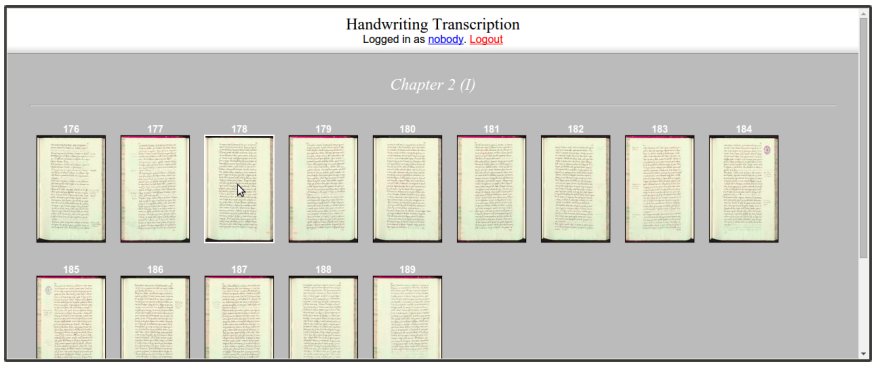
4.1 Volume Partition and Transcription Task Assignments

Plantas was divided into 11 batches in addition to the Prologue (which is considered "batch 0", for initial system training). Each batch comprises a set of chapters that were evenly assigned to the participants. The underlying idea behind this organization of the transcription work is that, once the transcription of one batch is finished, their transcripts are added to the training set. This way, the statistical HTR models (HMMs and N-grams) are improved for the transcription of the next batch. Essentially the same batch division was adopted for incremental layout analysis and line detection.

Table 2 shows the composition of each batch (left) and the overall distribution of work among transcribers (right), including the amount of manual and interactive text transcripts submitted in each

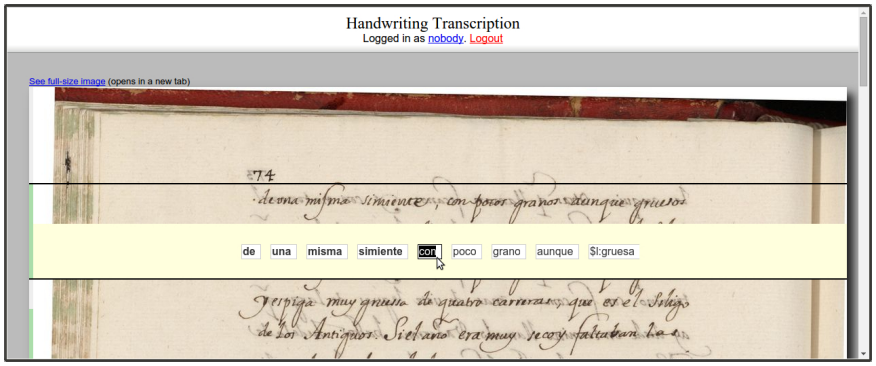


(a)



(b)

(c)



(d)

Figure 11: CATTI user interface. The book is divided into chapters (a). The number of pages is indicated above each page thumbnail, together with the last accessed date. Each chapter is available for transcription (b) either manually (M) or interactively (I). Under the hood, each page is divided into lines (c). Upon clicking on each line, the application unfolds a keyboard input area below it (d).

case. As observed, both the batches and the assignments are fairly balanced, according to the number of pages that would be transcribed in both modalities by each participant. A more detailed description of these batches and the transcriber assignments is shown in Appendix: Task Assignments.

Table 2: Book partition and transcription task assignment, both per batch (1,...,11) and per user (P01,...,P04).

Batch	**No. Chapters**	**No. Pages**	
		Manual	Interact.
1	16	73	0
2	16	0	77
3	16	73	0
4	12	0	72
5	12	56	10
6	17	20	64
7	13	60	20
8	11	22	55
9	12	48	32
10	10	30	52
11	16	48	42
Total	151	430	424

User	**No. Chapters**	**No. Pages**	
		Manual	Interact.
P01	33	117	113
P02	35	92	100
P03	46	137	117
P04	37	84	94
Total	151	430	424

4.2 Longitudinal Evaluation of Layout Analysis

After the standard page image preprocessing and enhancement steps, text blocks were automatically detected using simple image segmentation techniques which did not need to be trained from previously annotated page images. The results were then manually reviewed and the block segmentation errors were amended. Since this only required little human effort, the whole block segmentation review task was assigned to a single user; namely, the supertranscriber.

Once reliable text blocks were available, lines were automatically detected within these blocks and then manually reviewed following the interactive approach described in Section 3.1.3. As with text block segmentation, this process is also preparatory and much less time consuming than the transcription process itself. Therefore, this task was also assigned to the supertranscriber, who was in charge of text block detection.

The performance of this interactive, semisupervised text line detection process was evaluated by means of the user review time (URT), which is the time required to review and correct the automatically detected text lines of one page, averaged over a batch of pages.

Fig. 12 shows the evolution of the review time for the successive batches. Once each batch was reviewed and corrected as needed, the statistical models for line detection were retrained with the new supervised data. As observed in the figure, thanks to this incremental model retraining approach, the system improved over time, leading to significant reductions in the supertranscriber's effort required (the review time was lowered by half after processing 10 batches). We therefore conclude that the proposed process provides high-quality text line detection in a cost-effective way.

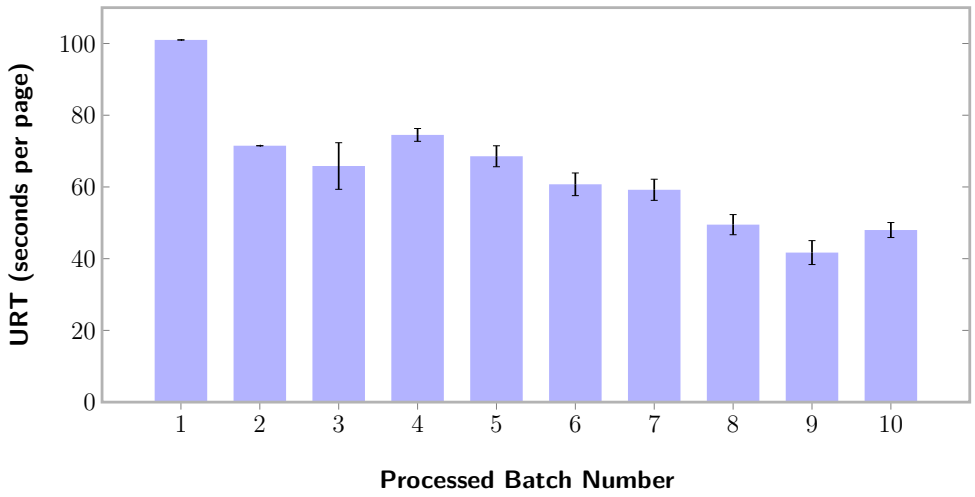


Figure 12: Computer assisted line segmentation: evolution of user review time (URT) with the successive batches of page images processed. Thanks to incremental model retraining, supertranscriber's reviewing effort significantly decreased over time.

4.3 Longitudinal Evaluation of Text Transcription

In the following we focus on measuring how a CATTI system can assist the user while transcribing, and we compare it against a conventional, non-interactive, manual approach. As previously mentioned, four participants (aged 22–28) were recruited. They were students in History in their last academic year and worked with our HTR prototype for about four months. Each participant was randomly assigned the manual or CATTI versions of our prototype on a weekly basis. There were thus two conditions: manual (baseline) and interactive-predictive (experimental).

The data provided by the participants was aggregated (averaged) into a single time series on a weekly basis, in order to provide a general overview of the longitudinal evaluation. Again, we want to stress out the fact that the evaluation is focused on measuring the system's performance, not that of the users'. Nevertheless, the reader can refer to Appendix B for a subject-by-subject results. In addition, we will discuss later some interesting qualitative observations that allowed us to understand better e.g. *when* the CATTI system works as intended (and when it does not) and *why*, how the interactive assistance can be improved, and so on.

Four evaluation measures were adopted to assess performance, on a per-line basis: word error rate (WER), character error rate (CER), number of keypresses, and transcription time. Both WER and CER are well-established measures of HTR performance. WER measures the percentage of word changes that are needed to transform a given text into its reference text, with respect to the length of the reference text. CER is analogous to WER, but defined at the character level. On the other hand, both the number of keypresses and transcription time provide an overall picture of the user's typing and thinking effort.

WER (and CER) were computed according to three different references:

WERs is the word error rate of the initial HTR transcripts in interactive mode but just before

start using CATTI assistance, with respect to the final result provided by the supertranscriber (considered to be the reference). WERs thus indicates how accurate were the system's initial transcripts.

WERu is the word error rate of the initial HTR transcripts with respect to those produced by the users in CATTI mode. WERu thus indicates how accurate were the system's transcripts with regard to the users'.

WERr is the "residual" word error rate of the transcripts produced by the users, either manually or assisted by CATTI, with respect to the final result provided by the supertranscriber. WERr thus indicates how accurate were the final transcripts produced by the users.

Ideally, it is expected that WERs will converge over time to a certain minimum value as more supervised training data are acquired, since the HTR system would be able to recognize more accurately the text images. Also, it is expected that the difference between WERs and WERu will decrease as well, since the HTR system would be able to make more suitable predictions requiring less user iterations to correct miss-recognized words. Similarly, it is expected that WERr will decrease towards zero over time, since the instructions provided by the supertranscriber to the users would help to strengthen their transcription criteria to produce the final transcripts.

As it can be observed in figures 13–16, there were interesting improvements over time. The x-axis displays consecutive working weeks, usually corresponding to Tuesday to Friday, with breaks on the weekends and Mondays. We denote each week in the ISO week-numbering format, i.e., the first week of the year is 01 and the last week of the year is 52.

Both WERs and WERu (which only apply to the CATTI condition) notably decreased; starting with a daily error rate of more than 50% and ending with close to 30%. This decrease indicates that both the user and the system were learning over time from each other, and cooperated together toward their ultimate goal of producing better results. In addition, the initially high values of WERs and WERu were expected, since the first HTR models were poorly trained. On the other hand, the WERr (the "residual" error) was similar for the manual and CATTI conditions and did not significantly varied over time. This indicates that the transcribers carefully performed their job independently of the used modality. For example, with CATTI assistance, when the initial HTR hypotheses had many errors, the transcribers generally corrected all of them, but when the system was already producing accurate predictions, the transcribers were attentive enough to fix most of the fewer wrong predictions. In any case, WERr remained more or less the same over time, as the users always encountered some transcription issue that was never seen before. As such, they did not know how to deal with those unseen issues correctly, and had to rely on the supertranscriber's expertise, who instructed them in order to avoid future, similar cases.

A very similar trend can be observed in Fig. 14, which shows the results regarding CER. Here we should clarify that CER is a less pessimistic measure than WER, since CER takes into account mismatching characters within misrecognized words, while WER accounts only for mismatching words. We should insist that CERs and CERu (and WERs and WERu as well) can only be computed in the

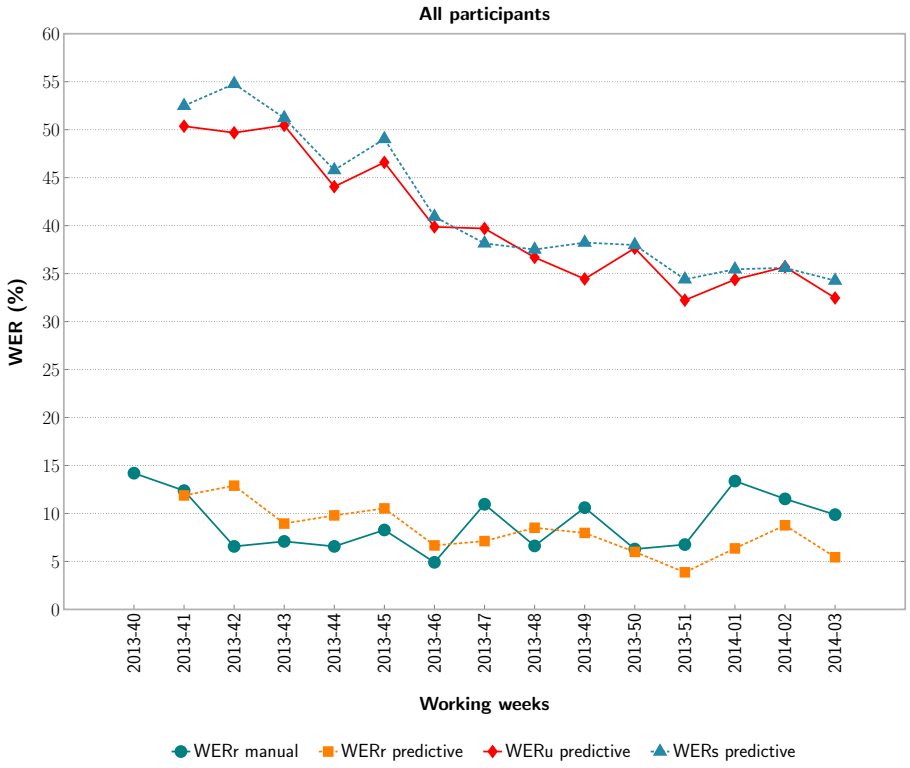


Figure 13: Evolution of the average word error rate per week. WERs: initial HTR system errors; WERu: words in the initial system prediction that were modified by the users; WERr: residual user errors. Working weeks correspond to the period: week 40 (from September 30 to October 6) of 2013 to week 03 (from January 13 to January 19) of 2014.

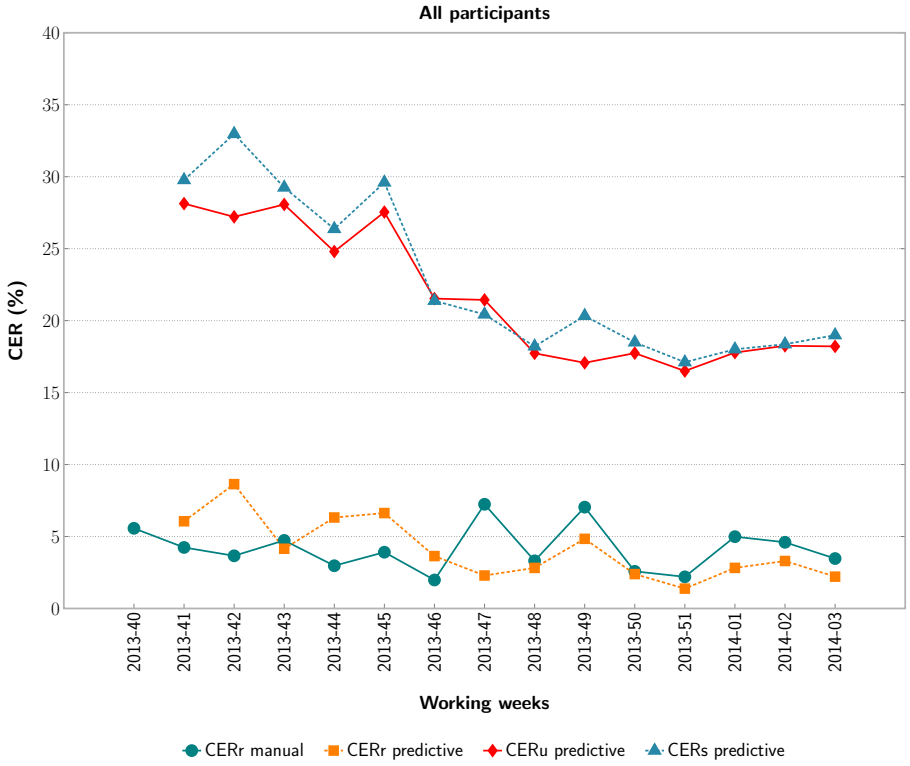


Figure 14: Evolution of the average character error rate per week. CERs: initial HTR system errors; CERu: characters in the initial system prediction that were modified by the users; CERr: residual user errors. Working weeks correspond to the period: week 40 (from September 30 to October 6) of 2013 to week 03 (from January 13 to January 19) of 2014.

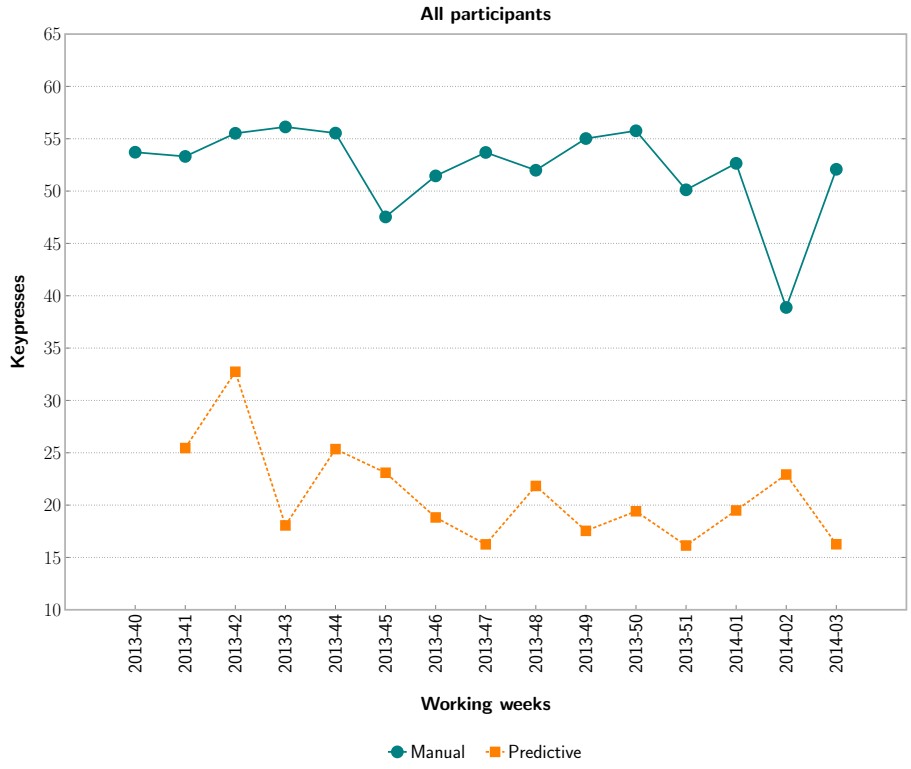


Figure 15: Evolution of the average number of keypresses required to transcribe a line. Working weeks correspond to the period: week 40 (from September 30 to October 6) of 2013 to week 03 (from January 13 to January 19) of 2014.

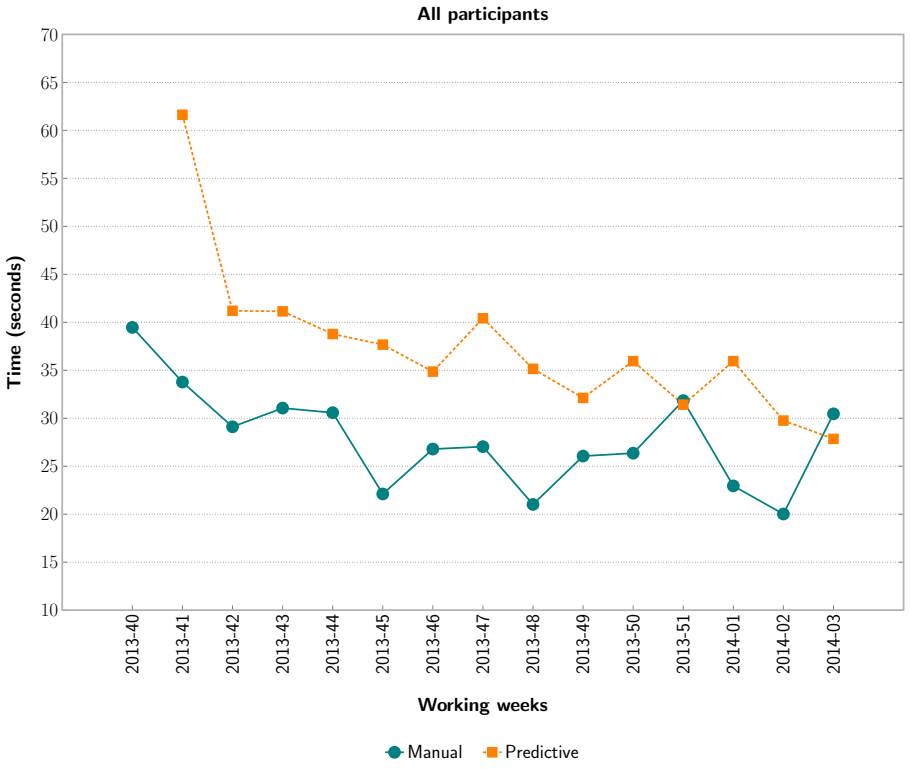


Figure 16: Evolution of the average time spent transcribing a line. Working weeks correspond to the period: week 40 (from September 30 to October 6) of 2013 to week 03 (from January 13 to January 19) of 2014.

interactive-predictive condition, since those measures account for the performance of the HTR system (CATTI). In the manual condition there is no HTR system involved. An example showing that CER is less pessimistic than WER is shown in Fig. 14, where it can be observed that CERs and CERu have the same decreasing trend for WERs and WERu, but with corresponding lower values: from about 35% to less than 20%.

Fig. 15 shows the evolution of the keypresses per line over time. Only the keypresses corresponding to printed characters were considered and outlier observations were excluded to obtain the average values per week shown in the plot. As expected, in the manual approach, where the user has to type everything (including edit shortcuts), the number of keypresses per line is more or less constant over time. On the other hand, with CATTI assistance, the typing effort is much lower and moderately decreasing over time – thanks to the increasingly improved predictions provided by the CATTI system. Therefore, CATTI allows the user to notably save typing effort, i.e., it avoids to type all words from scratch, even at the early stages of the system's learning process. This shows a clear potential of CATTI to boost the efficiency of human transcribers. However, such a significant typing effort reduction did not result in net user effort savings, as shown in Fig. 16.

As expected, the efficiency of CATTI-assited transcription did significantly improve over time – due both to the increasingly better CATTI predictions and to the growing familiarity of the users with the transcripition task and with the CATTI system. Similarly, but much less significantly, manual transcription also became more efficient over time, in this case thanks only to the increasing familiarity of the users with the transcription task and with the editing system. Still, the transcription time per line was found to be generally higher with CATTI assistance (albeit with a slightly decreasing difference, thanks to the faster improvement of CATTI assisted transcription). This unexpected and counterintuitive finding can be explained by the additional amount of time the user may need to read and understand each system prediction, which may change after each interaction step. In the next section we provide a detailed discussion about this observation, which has to be carefully considered to develop improved CATTI systems and correspondingly better transcription interfaces.

5 Qualitative Observations

In addition to the above empirical assessment results, now we provide an overall summary of the comments our participants made over the 4-month evaluation period, which have been clustered into the following six major areas of interest.

5.1 First-time Sessions

The first working days most of the participants had major doubts about the transcription criteria, and so they left many image lines untranscribed. In addition, because participants had not been exposed to the UI before, they operated it at a slower pace in comparison as to how they proceed in subsequent

sessions. This is reflected in the decreasing tendency of the time spent per line, which can be seen in Fig. 16.

During the first sessions, all participants left catchwords untranscribed. However, the supertranscriber quickly instructed them again —he had done so before starting the longitudinal study—and as a result they improved considerably afterward. Overall, participants stated that the transcription criteria were easy to learn. Further, because many of the events relevant to these criteria appear very frequently throughout the book, eventually they quickly became familiarized with them. Anecdotally, neither the system or the participants did recognize the (rather unusual, medieval form of the) letter "ñ" in the first sessions. Then, participants were re-instructed and as a result the system learned it later, too.

One participant stated that, until the third transcription session, they did not like the autocompletions at all, and thus they decided to write everything in the same text field, including already-correct words in the predictions. It is clear that the CATTI assistance did not work for this user for the first time, though they became familiarized after minimal training on the UI usage. In fact, we believe that an adequate UI is probably the most important asset to make the most of the CATTI technology.

5.2 User Interface Usage

As previously commented, it was interesting to notice that all participants quickly became familiarized with the web-based UI. For example, they used the keyboard rather than the mouse most of the time (even when switching from line to line) as well as the shortcuts to speed up the transcription editing process. The most commonly observed transcription strategy was the following: (1) read the handwritten image in the first place, (2) select it, (3) read the system suggestion (only available in interactive mode), (4) interactively edit the erroneous words, and (5) finally validate the full sentence. Half of the participants validated everything, including empty lines, in order to put on record what they did. As instructed, all participants did not come back to revise previously validated transcripts, since the supertranscriber would verify all submitted transcripts at a later time.

One issue most participants complained about was the fact that when inserting a new word the focus moved to the next text field plus one. This was so because the insert operation implicitly assumed that the next word is correct and therefore moving the focus automatically "to the next text field plus one" theoretically would allow the participants to save time. However, in the end participants seemed not to like this UI behavior, and preferred entering multiple words in a single text field instead of using the insertion shortcut. This preference was reinforced over time, because of the fact that in the manual mode participants initially had to enter text on a single text field, and so they began entering multiple words also in the predictive condition, in an attempt to speed up editing time. While this typing behavior is perfectly legitimate for the manual mode, it precludes taking full advantage of the system predictive behavior in interactive mode, since predictions occur only when pressing the Enter key or when changing the focus to the next text field. The lesson learned here is that any UI devoted

to HTR tasks, whether interactive or not, should present the user with a single text field to enter text, rather than one text field per word as in the UI used in this work (Fig. 11d).

Finally, we should mention that, in the current version of our web application, if a user decides to go some text fields back and re-edit a word, the CATTI system may overwrite some of the previously entered text if it is considered statistically optimal. To avoid this UI behavior, the user can press the ESC key and disable the interactive predictions. However, we noticed that only one participant did so, and its use was anecdotal (6 out of their 126 335 logged keypress events). Further, we recommend that future HTR UIs incorporate more intelligence and be able to automatically disable predictions if the user goes back to re-edit some previously entered text, as in the CASMACAT workbench (Alabau et al., 2014).

5.3 Interactive Predictions

We observed that some typographic/paleographic details do adversely affect the predictive system. For example, ornaments, ligatures, flourishes, dots that are actually patches or quill stains, etc. are details that participants might notice based on their own criteria and previous experience when working with the application in manual mode. However, when using the predictive mode, one might go with the system flow and blindly trust the predictions. This was especially true in the first few weeks, when participants considered that the system worked notably well. Eventually, participants became more proficient with CATTI and were more aware of the predictions' quality.

Participants reported that the predictive system tended to err while transcribing chapter titles. We also observed that concordances such as noun (singular, plural), gender (masculine, feminine, neutral), and some affixes were poorly supported by the predictive system to begin with. These errors caused an initially negative impact on participants' perception toward the system. In fact, two participants complained quite often about the transcription errors of the predictive system. They stated that if they are shown a lot of transcription errors, then it is better to completely delete the text and start transcribing from scratch. "If there are a couple of errors at most, then I'm happy with the predictive system. Otherwise, it would be worth disabling the predictions." However, as previously discussed, only one participant actually did so, and it was of anecdotal use. Eventually, all participants agreed on the fact that the system improved considerably over time, and so did the predictions.

Interestingly, other participants commented that their concentration usually decreased as the number of errors increased: "I had to leave some words untranscribed, especially when the predictions changed a lot with regard to the previous one." Although we observed that the predictions usually did not change too much from one suggestion to another, even when there is just one error in the suggested transcript (or in subsequent ones), participants felt it was necessary to carefully read the whole text to ensure that everything was right. This was found to be cognitively demanding and especially costly in terms of time, and explains why task completion times are similar or even sometimes slower in the interactive mode, in comparison to the manual mode.

Participants commented that there were two situations, which are rather frequent in Plantas, that usually led the system to commit more errors:

1. When there are prominent ascenders or descenders,[9] which happens in all chapters.

2. When there is multilingual text, which happens especially in the first chapters.

These situations not withstanding, the predictions were considered to be generally very good (Median score of 4 in a 1–5 Likert scale). This was especially true near the end of the evaluation, as the system had learned enough examples. Anecdotally, it was found that in the first sessions the system did not suggest punctuation symbols; then, as long as the system was retrained, those symbols started to appear in the predictions.

One participant played a lot with the Reject operation: "it is so entertaining that sometimes I was distracted away from the main [transcription] task". One participant (rightly) observed that rejection should not be used over the last word of the line being transcribed. "In this case it would be even better not using the predictive mode."

Finally, it was interesting to notice that, when there are two words separated only by a tiny or imperceptible blank space, the system often suggested them separately; e.g. "delas" became "de las" (c.f. Fig. 3). This also happened with flourishes and punctuation symbols, a behavior that is attributed to the language model. Then, over time, it managed to learn the transcription criteria adopted for these subtle situations (c.f. Section 2).

5.4 Typing Behaviors

One participant felt their performance was much slower with the predictive mode, "as slow as twice the time spent on the manual mode." However, as shown in the figures of the previous section, the differences between both editing modes are not that large. We should mention that this participant is not a professional typist but actually writes really fast, that is why they tended to prefer manual over predictive mode. They also found that manual mode was more entertaining than predictive mode, because this participant likes writing. "In the predictive system one just has to look at the words that are correct and then validate the transcription. This is theoretically easy, but in practice it is difficult because you have to read some text that you actually did not write, so you have to make a critical review of the system suggestions."

Another participant, instead of relying on the memorized position of keys for typing, proceeded by finding each key by eyesight. This had a negative influence on the predictive assistance, as they did not usually look at the system suggestions. It turned out that this participant does not use the computer very often and also suffered from asthenopia (eyestrain). However, this participant is notably perfectionist, ensuring that their transcripts were totally error-free.

[9] An ascender is the portion of a letter that extends above the baseline of the text font. Some examples can be noticed in Fig. 4. Historically, together with descenders, they have been used to increase the recognizability of words.

Interestingly, it was observed that some errors were propagated from participant to participant. For example, in the last sessions all participants committed the same couple of errors: they swapped the order of the hyphenation tags (-$ instead of $-) and wrote the character U instead of V in the chapter titles.

5.5 Opinions and Suggestions

Many of the participant's opinions did support the interest of the proposed CATTI framework and the features of the corresponding UI we provided to support their work in the present study.

Regarding the web application, they found it to be particularly useful that once a line was validated, the interface automatically opened the next editing box. "It allows you to move quickly to the next line and thus to save time." An important advice derived form this observation is thus that, to allow users to be more productive over time, such standard shortcuts (and possibly others) should be kept in future interactive HTR interfaces.

Similarly, all participants stated that placing the editing area just below the corresponding text line image was indeed a good choice — different from other possible UI layouts where the editing area is on the left or on the right side of the page image being transcribed. Also, having a visible link to the original, high resolution full page image (for consulting details when necessary) was considered a very useful choice.

On the other hand, when asked for alternate UI designs, participants found it counterintuitive and even bothersome, to have to enter each word in a dedicated text field. "People are really accustomed to enter text in a single text field." "It was particularly annoying when the whole sentence is crossed out." Therefore, as previously commented, future interactive HTR interfaces should be equipped with a single text field, both to show the system predictions and to enter the user transcripts or any other text amendments. Together with the previous user comments and suggestions for improvement, we found all these to be really valuable, and they are in fact shaping our new generation of HTR/CATTI sytems, from a practical point of view.

6 Manuscript Ground Truth

The produced reference data of the first volume of Plantas can be considered ground truth for machine learning purposes and is available upon request, namely by contacting any of the authors of this article by email . It has been produced with two different types of annotations. First, layout analysis of each page indicates main text blocks, lines, headings, drawing regions, etc. Second, the manuscript is completely transcribed and tagged according to abreviations, underlined words, added text, and so on.

The ground truth annotation of layout analysis for each page image involves the following structural

components: the bounding box of the main text block, the coordinates of text lines inside the main text block, the bounding boxes of text headings, page numbers, catchwords, drawings (if there was any), and marginalia zones. The basic statistics of this ground truth are reported in Table 3. On the other hand, the produced transcripts contain around 200k words in over 20k lines, transcribed from nearly 900 pages. Table 4 summarizes these details.

Table 3: Statistics of the ground truth for layout analysis.

No. of Main Text Blocks	815
No. of Drawings	101
No. of Catchword Regions	158
No. of Heading Regions	158
No. of Page-Number Regions	792
No. of Signature-Mark Regions	233
No. of Marginalia Zones	1 274

Table 4: Statistics of the produced transcripss.

No. of Pages	881
No. of Lines	19 764
No. of Words	196 858
No. of Unique Words	20 931
No. of Chars*	756 122
No. of Unique Chars*	100

* Excluding tagging characters.

Notice that, since many words have been tagged and, moreover, Plantas is multilingual, there are more than the usual 2×27 Spanish characters. Regarding to word tagging, which was performed along with the transcription process, in Table 5 each tagged word is listed along with its frequency. It is worth noting that foreign words (i.e., non-Spanish words) is the most frequent word tag. Finally, Table 6 provides a breakdown of the proportion of the foreign words by language.

Table 5: Tagging statistics of the produced transcriptions.

No. of Abbreviations	3 966
No. of Underlined Words	198
No. of Added Lines	32
No. of Deleted Words (crossed-out)	679
No. of Illegible Words	123
No. of Hyphenated Words	3 985
No. of Catchwords	191
No. of Superscript Words	519
No. of Foreign Words (non-Spanish)	5 744

Table 6: Tagging statistics of the different foreign words.

Latin	4 132
German	433
Greek	326
French	313
Flemish	139
Arabic	109
Hebrew	41
Portuguese	27
English	24
Catalan	14
Polish	11
Others	17

7 Conclusion and Future Work

We have presented a cost-effective process to transcribe a 1 000 pages handwritten book of the 17th century about botanical species. As a result, we have produced, together with the full transcription, valuable high-quality reference data. Moreover, transcripts are fully tagged with rich information about language, abbreviation expansions, hyphenation, underlined words, etc. Further, as a byproduct of having used HTR technology to assist the transcription process, perfect alignments between transcripts and images have been produced. Using these alignment data, advanced features can be easily implemented, such as a fully synchronized presentation of the images and the transcripts, down to the word level. Together with the feature-rich tagging scheme, this will pave the way for the creation of advanced forms of digital editions.

This article is the first to date that reports insights on the use of an interactive HTR technology via a longitudinal study involving real transcribers. From these insights, we have derived some recommendations to shape the next generation of interactive HTR systems, from a practical point of view.

For the near future, we plan to use all the produced reference data to train high-quality HTR models that will be used for subsequent transcription and/or full indexing of the remaining volumes of the Plantas collection. At the time of this writing, the first steps in this direction have already been taken. More specifically, using optical and language models trained on the produced ground truth data, the second volume of Plantas (about 800 pages) has been automatically indexed at the word level. This means that the whole set of *untranscribed* page images of this volume is now fully searchable under the so called "precision-recall tradeoff model".[10] Moreover, a keyword search UI has been developed for studing both the capabilities of and interest in this model, again with the help of volunteer students from the Universidad Complutense de Madrid. An upcoming article will report the results of such study.

Looking forward, we believe this work enables new research opportunities in HTR and we hope it

[10] An actual demonstrator of this result can be tried at: [URL to be published with this article].

will be useful to other professionals working in this topic, as well as to the potential audience that will access the transcribed texts.

The HTR/CATTI system, together with the prototype used in this work, is publicly available for demonstration purposes at [URL to be published with this article].

Acknowledgments

This work has only been possible thanks to the collaboration of the Biblioteca Nacional de España (BNE) and the Universidad Complutense de Madrid (UCM) through Prof. Paloma Cuenca and the excellent work of her Paleography students: Roberto Alonso, Lucía Sánchez, David Rey, and Andrea Tatiana Zivny.

Funding

This work is supported by the European Commission through the EU projects HIMANIS (JPICH program, Spanish grant Ref. PCIN-2015-068) and READ (Horizon-2020 program, grant Ref. 674943); and the Universitat Politècnica de València [grant number SP20130189]. This work was also part of the Valorization and I+D+i Resources program of VLC/CAMPUS and has been funded by the Spanish MECD as part of the International Excellence Campus program.

References

Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., Germann, U., González-Rubio, J., Hill, R., Koehn, P., Leiva, L., Mesa-Lao, B., Ortiz, D., Saint-Amand, H., Sanchis, G. and Tsoukala, C. (2014), Casmacat: A computer-assisted translation workbench, *in* 'Proc. European Chapter of the Association for Computational Linguistics (EACL)', pp. 25–28.

Alabau, V. and Leiva, L. A. (2012), Transcribing handwritten text images with a word soup game, *in* 'Proc. Extended Abstracts on Human Factors in Computing Systems (CHI EA)', pp. 2273–2278.

Arévalo, C. (1935), 'Bernardo de Cienfuegos y la botánica de su época', *Estudios sobre la ciencia española del siglo XVII* pp. 323–335.

Baird, H. S. (1995), Document image analysis, *in* L. O'Gorman and R. Kasturi, eds, 'The skew angle of printed documents', IEEE Computer Society Press, pp. 204–208.

Bautier, R.-H. (1984), 'Normes Internationales pour l'édition des documents médievaux', *Folia Caesaraugustana* **1**, 13–64.

Bazzi, I., Schwartz, R. and Makhoul, J. (1999), 'An Omnifont Open-Vocabulary OCR System for English and Arabic', *IEEE T. Pattern Anal.* **21**(6), 495–504.

Blanco Castro, E., Morales Valverde, R. and Sánchez Moreno, P. M. (1994), 'Bernardo Cienfuegos y su aportación a la botánica en el siglo XVII', *Asclepio, revista de Historia de la Medicina y de la Ciencia* **XLVI**(1), 37–124.

Bosch, V., Toselli, A. H. and Vidal, E. (2014), Semiautomatic text baseline detection in large historical handwritten documents, *in* 'Proc. Intl. Conf. on Frontiers in Handwriting Recognition (ICFHR)', pp. 690–695.

Causer, T., Tonra, J. and Wallace, V. (2012), 'Transcription maximized; expense minimized? crowdsourcing and editing the collected works of Jeremy Bentham', *Literary and Linguistic Computing* .

Causer, T. and Wallace, V. (2012), 'Building a volunteer community: results and findings from Transcribe Bentham', *Digital Humanities Quarterly* .

Cavanilles y Centi, A. J. and Lagasca y Segura, M. (2000), *Dos noticias históricas del inmortal botánico y sacerdote hispano-valentino don Antonio José Cavanilles [...]*, Biblioteca Virtual Miguel de Cervantes. Recovered from: `http://www.cervantesvirtual.com/nd/ark:/59851/bmcpk0b6`.

Dimauro, G., Impedovo, S., Modugno, R. and Pirlo, G. (2002), A new database for research on bank-check processing, *in* 'Proc. Intl. Conf. on Frontiers in Handwriting Recognition (ICFHR)', pp. 524–528.

Jelinek, F. (1998), *Statistical Methods for Speech Recognition*, MIT Press.

Leiva, L. A., Romero, V., Toselli, A. H. and Vidal, E. (2011), Evaluating an interactive-predictive paradigm on handwriting transcription: A case study and lessons learned, *in* 'Proc. Annual IEEE Intl. Computer Software and Applications Conference (COMPSAC)', pp. 610–617.

Malleron, V. and Eglin, V. (2011), A mixed approach for handwritten documents structural analysis, *in* 'Proc. Intl. Conf. on Document Analysis and Recognition (ICDAR)', pp. 269–273.

Ortí Cárcel, M. M. (1997), *Vocabulaire international de la diplomatique*, Vol. 28, Universitat de València.

Pardo-de Santanya, M., Tardío, J. and Morales, R. (2014), Pioneers in Spanish Ethnobiology, *in* I. Svangerg and L. Luczaj, eds, 'Pioneers in European Ethnobiology', Uppsala Universitet.

Petrucci, A. (1985), *Breve storia della scrittura latina*, Bagatto libri.

Ramel, J.-Y., Leriche, S., Demonet, M. and Busson, S. (2007), 'User-driven page layout analysis of historical printed books', *IJDAR* **9**(2–4), 243–261.

Romero, V., Fornés, A., Serrano, N., Sánchez, J. A., Toselli, A. H., Frinken, V., Vidal, E. and Lladós, J. (2013), 'The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition', *Pattern Recogn.* **46**, 1658–1669.

Romero, V., Leiva, L. A., Toselli, A. H. and Vidal, E. (2009), Interactive multimodal transcription of text images using a web-based demo system, *in* 'Proc. Intelligent User Interfaces (IUI)', pp. 477–478.

Romero, V., Toselli, A. H. and Vidal, E. (2012), *Multimodal Interactive Handwritten Text Transcription*, Series in Machine Perception and Artificial Intelligence (MPAI), World Scientific Publishing. `http://www.worldscientific.com/worldscibooks/10.1142/8394`.

Sanchez, J. A., Toselli, A. H., Romero, V. and Vidal, E. (2015), ICDAR 2015 competition HTRtS: Handwritten Text Recognition on the tranScriptorium dataset, *in* 'Document Analysis and Recognition (ICDAR), 2015 13th International Conference on', IEEE, pp. 1166–1170.

Srihari, S. N. and Keubert, E. J. (1997), Integration of Handwritten Address Interpretation Technology into the United States Postal Service Remote Computer Reader System, *in* 'Proc. Intl. Conf. on Document Analysis and Recognition (ICDAR)', Vol. 2, pp. 892–896.

Toselli, A. H., Juan, A., Keysers, D., González, J., Salvador, I., H. Ney, Vidal, E. and Casacuberta, F. (2004), 'Integrated Handwriting Recognition and Interpretation using Finite-State Models', *IJPRAI* **18**(4), 519–539.

Toselli, A. H., Romero, V., Pastor, M. and Vidal, E. (2010), 'Multimodal interactive transcription of text images', *Pattern Recogn.* **43**(5), 1814–1825.

Toselli, A. H., Vidal, E., Romero, V. and Frinken, V. (2016), 'HMM Word graph based keyword spotting in handwritten document images ', *Information Sciences* **370-371**, 497 – 518.
URL: *http://www.sciencedirect.com/science/article/pii/S0020025516305461*

Vinciarelli, A., Bengio, S. and Bunke, H. (2004), 'Off-line recognition of unconstrained handwritten texts using HMMs and statistical language models', *IEEE T. Pattern Anal.* **26**(6), 709–720.

Appendix

In this section we provide more details about the task assignments and the results of the longitudinal evaluation.

A Task Assignments

Table 7 shows statistics of the batches and their corresponding chapters. It shows in addition the transcription task assignment distribution among the participants (including the supertranscriber) for manual and interactive modalities, the latter using CATTI. As observed, the assignments are quite balanced, according to the number of pages that would be transcribed in both modalities.

It is worth mentioning that columns labeled with **#Lns(M:I)** present the total number of manually and interactively transcribed lines which were actually performed. For example, according to the first row of **Batch 6**, column **#Pgs(M:I)**, the original plan was to transcribe *all* pages (18) using the interactive modality. However, as some wordgraphs could not be generated due to image preprocessing issues, the corresponding image lines were eventually transcribed manually (in this case 17 lines).

Table 7: Book partition and corresponding transcription assignment tasks. **P01,...,P04**: transcribers, **ST**: supertranscriber, **# Chps**: number of chapters, **# Pgs (M:I)**: number of pages to be (M)anually ir (I)nteractively transcribed and **# Lns (M:I)**: number of lines (M)anually or (I)nteractively transcribed.

Batch	User	Chapters	#Pgs(M:I)	#Lns(M:I)
0	ST	Prologue	38:00	1200:000
1	P01	1	18:00	581:000
	P02	8-13	19:00	551:000
	P03	22-24	20:00	479:000
	P04	37-42	16:00	492:000
2	P01	2-3	00:19	017:593
	P02	14-19	00:18	012:549
	P03	25-29	00:18	014:399
	P04	43-45	00:22	021:583
3	P01	4-7	16:00	510:000
	P02	20-21	18:00	569:000
	P03	30-35	17:00	481:000
	P04	46-49	22:00	498:000
4	P01	50-53	00:20	015:242
	P02	63-66	00:21	017:547
	P03	36	00:11	010:298
	P04	85-87	00:20	017:476
5	P01	54-57	20:00	292:000
	P02	67-69	14:00	411:000
	P03	73b-74	00:10	008:187
	P04	88-90	22:00	523:000
6	P01	58-59	00:18	017:483
	P02	70-73	00:20	026:389
	P03	75-78	20:00	505:000
	P04	91-97	00:26	031:494
7	P01	60-62	19:00	544:000
	P02	111-112	17:00	206:000
	P03	79-80	00:20	021:427
	P04	141-146	24:00	416:000
8	P01	98-100	00:24	048:417
	P02	113-114	00:21	046:269
	P03	81,83-84	22:00	477:000
	P04	147-149	00:10	007:247
9	P01	101-102	24:00	381:000
	P02	115-118	24:00	659:000
	P03	121-124	00:16	026:230
	P04	150-151	00:16	038:121
10	P01	103-106	10:20	158:120
	P02	119-120	00:20	018:470
	P03	125-128	20:12	208:085
	P04	—	00:00	000:000
11	P01	107-110	10:12	198:296
	P02	—	00:00	000:000
	P03	129-140	38:30	678:302
	P04	—	00:00	000:000

B Individual Evaluation Plots

In the following figures we provide the longitudinal HTR evaluation results separately for each user. We want to stress out the fact that the evaluation was focused on measuring the system's performance, not that of the users'. The inclusion of these plots is therefore anecdotal and should be seen as supplementary data.

Figure 17: Average word error rate after transcribing each text line image. WERr: user against supertranscriber; WERs: system against supertranscriber; WERu: user against system.

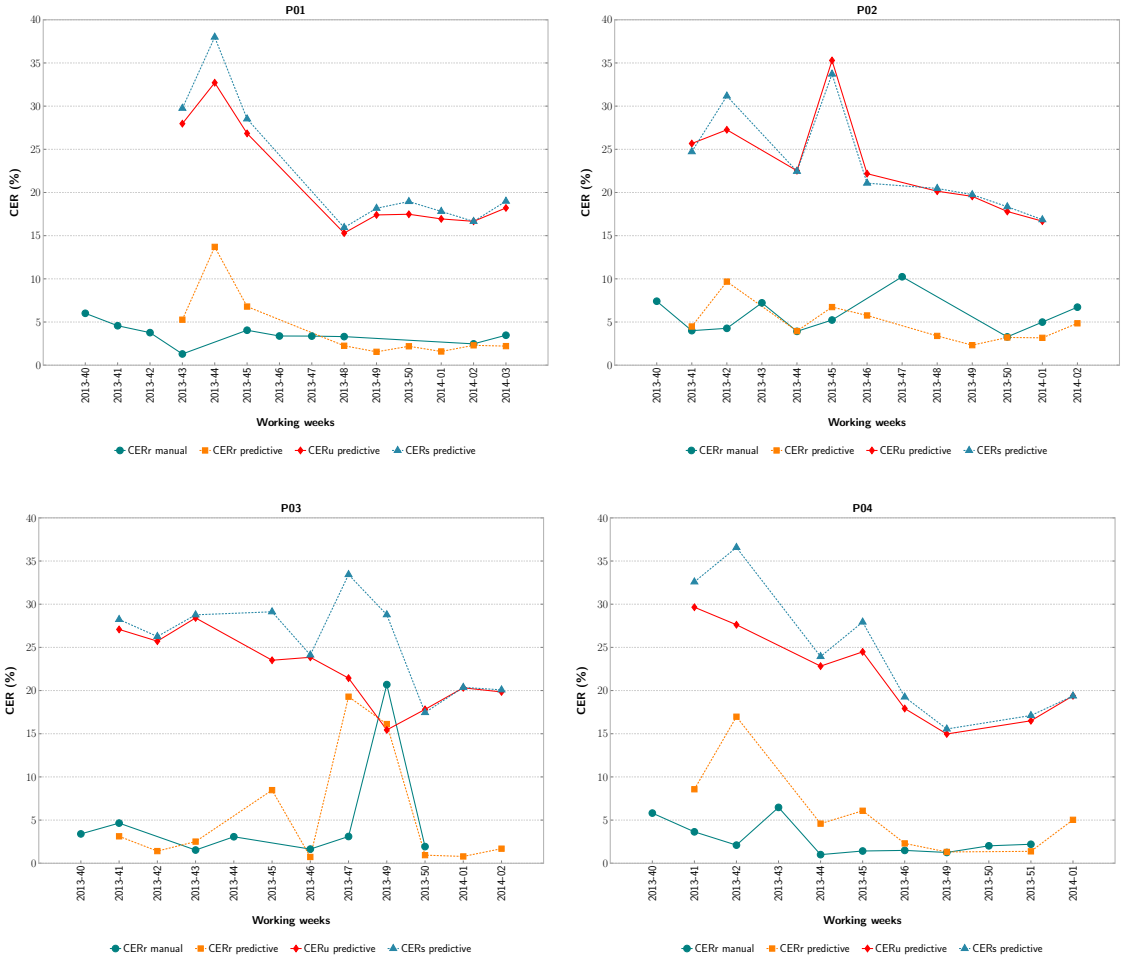


Figure 18: Average character error rate after transcribing each text line image. CERr: user against supertranscriber; CERs: system against supertranscriber; CERu: user against system.

Figure 19: Average number of keystrokes required to transcribe each text line image.

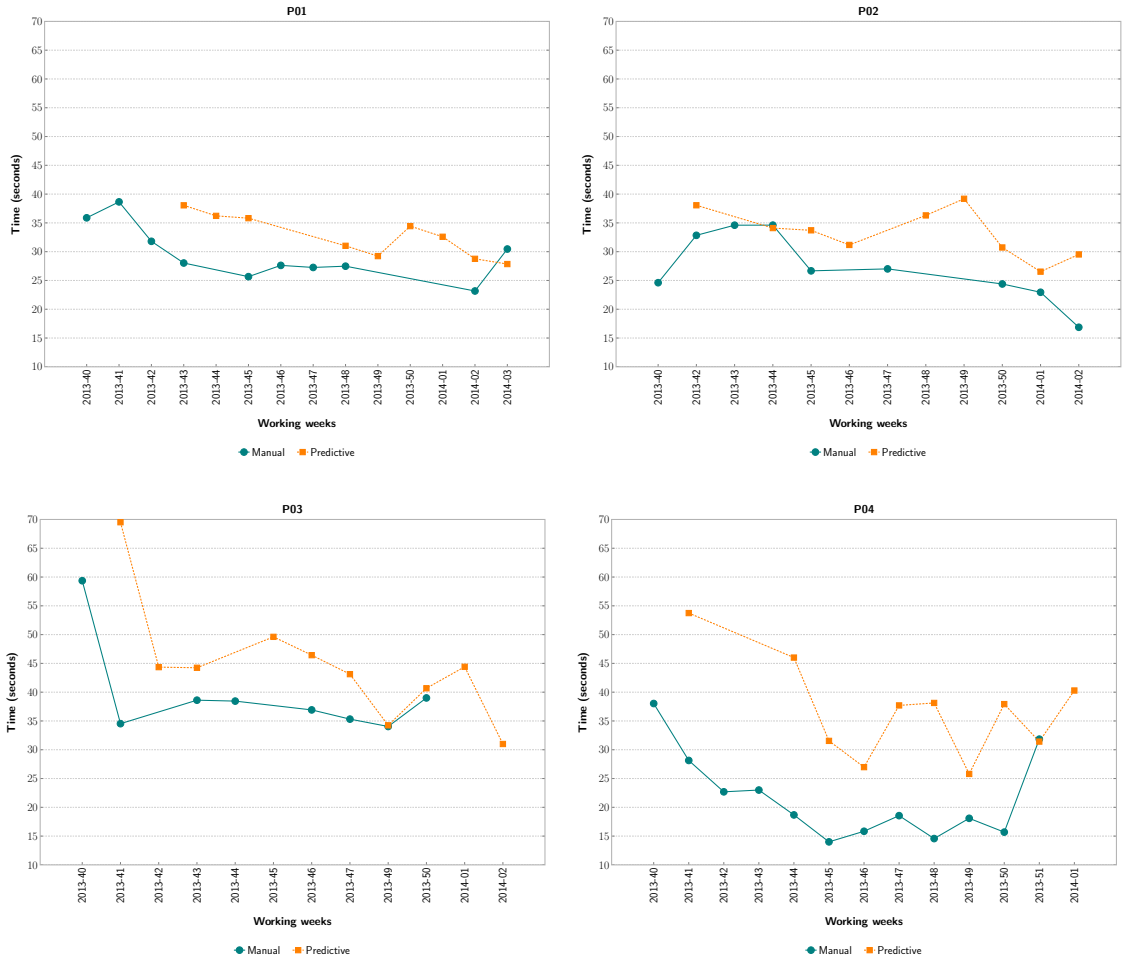


Figure 20: Average time spent transcribing each text line image.