Computer-assisted Transcription of a Historical Botanical Specimen Book: Organization and Process Overview

Vicent Bosch* viboscam@upv.es
Isabel Bordes-Cabrera† isabel.bordes@bne.es
Paloma Cuenca Muñoz‡ palomacm@ghis.ucm.es
Celio Hernández-Tornero* cehertor@upvnet.upv.es
Luis A. Leiva* luileito@dsic.upv.es
Moisés Pastor* mpastorg@dsic.upv.es
Verónica Romero* vromero@dsic.upv.es
Alejandro H. Toselli* ahector@dsic.upv.es
Enrique Vidal* evidal@dsic.upv.es

ABSTRACT

We describe a protocol designed for computer-assisted transcribing a XVII century botanical specimen book, based on Handwritten Text Recognition (HTR) technology. Here we focus on the organization and coordination aspects of this protocol and outline related technical issues. Using the proposed protocol, full ground truth data has been produced for the first book chapter and high-quality transcripts are being cost-effectively obtained for the rest of the approximately 1 000 pages of the book. The process encompasses two main, computer-assisted steps; namely, image layout analysis and transcription. Layout analysis is based on a semi-supervised incremental approach and transcription makes use of an interactive-predictive HTR prototype known as CATTI. Currently, the first step of this procedure has been completed for the full book and the second step is close to be finished. Ultimately, all the data produced will be made publicly available for research and development.

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

Document Image Processing, Document Layout Analysis, Handwritten Text Recognition

*Universitat Politècnica de València; 46022 Valencia, Spain
†Biblioteca Nacional de España; 28071 Madrid, Spain
‡Universidad Complutense de Madrid; 28040 Madrid, Spain

DATeCH 2014, May 19-20 2014, Madrid, Spain ACM 978-1-4503-2588-2/14/05. http://dx.doi.org/10.1145/2595188.2595204

1. INTRODUCTION

This work is the result of a collaboration effort between the Biblioteca Nacional de España[1] (BNE), the Universidad Complutense de Madrid[2] (UCM), and the Universitat Politècnica de València[3] (UPV). The first tome of a seven volumes manuscript entitled "Historia de las Plantas" —PLANTAS for short— was selected for transcription. PLANTAS is a XVII century handwritten botanical specimen book worthy of transcription due to the knowledge it provides concerning medical plants. Furthermore, the book is written in diverse languages, being ancient Spanish the main one. At the time when this paper is being written, the whole transcription of the book is close to be finished.

Among the agreed goals of this collaboration is a longitudinal user study of CATTI, a computer-assisted transcription prototype [12] based on interactive-predictive Handwritten Text Recognition (HTR) technologies which are being developped in the European project tranScriptorium[4]. CATTI is the main tool used by transcribers provided by UCM in their transcription work. In addition, several book chapters have been transcribed in a pure manual way for comparison purposes. Eventually, the transcripts and other data gathered in this process will be exploited for developping novel interactive HTR and layout analysis approaches.

In this paper we describe the protocol we are following to digitize PLANTAS, from ground truth production (already completed for the first chapter) to final transcription (almost completed for the full book at the time of writting this paper). We believe this protocol can serve for others to replicate this work, or as an inspiration to prepare a similar setup for other transcription tasks. The resulting database will be publicly available, including page images, layout, and transcribed text. Moreover, a comprehensive report will be published, including detailed analysis of the results and conclusions of the whole computer-assisted transcription process carried out.

2. MANUSCRIPT OVERVIEW

PLANTAS was compiled by Bernardo de Cienfuegos, one of the most outstanding Spanish botanists in the XVII century. Today, the manuscript has become a source of valu-

[1] http://www.bne.es
[2] http://www.ucm.es
[3] http://www.upv.es
[4] http://transcriptorium.eu/

able information for research, specially for those interested in reviewing the botanical knowledge of that historical period. A digital reproduction of PLANTAS can be found at the Biblioteca Digital Hispánica[5] . The seven volumes of "Historia de las plantas" were written using a quill-pen (Figure 1). These books are currently available at the BNE, and are being considered for automatic indexation by means of keyword spotting techniques developed by PRHLT [13].

The first volume of PLANTAS consists of a prologue and 152 chapters which make over 1 000 pages, containing about 20 000 handwritten text lines, along with botanical drawings (Figure 1). It is worth pointing out the fact that multilingual text appears frequently in the book; namely: Latin, Greek, Arabic, Hebrew, Portuguese, Catalan, French, German, English, Flemish, Polish, Bohemian, and Hun. It is so because the author frequently identifies different plant species in different languages. This particularity makes this book a rather challenging task for HTR technology.

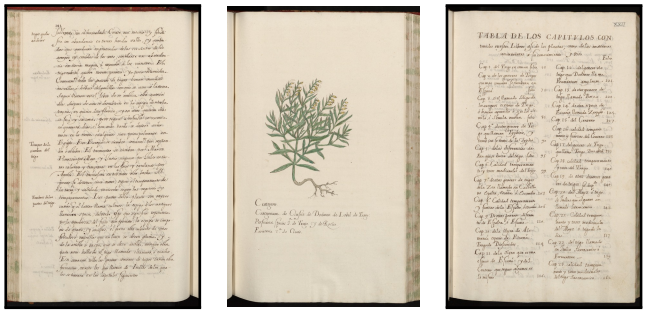


Figure 1: Examples of pages with different layouts.

2.1 Transcription Criteria

Before starting the actual transcription task, some basic criteria should be established. The most common type of transcription for historical handwritten documents used by paleographers is the so-called *diplomatic* (or *paleographic*), which tries to typographically transcribe as accurately as possible all significant characteristics of the original manuscript, including spelling, punctuation marks, abbreviations, deletions, insertions, and other text alterations.

We adopted *diplomatic* transcription rules with some minor changes, mainly to fit HTR requirements (see Section 4.2). Additionally, transcriptions can be tagged with feature-rich information. Among the transcription and tagging rules applied, the following are noteworthy (see figures 2-6):

- Transcriptions must respect as much as possible the original document. For instance, double characters should be transcribed as such. However, special characters such as large "s" or "f" must be transcribed according to currently accepted linguistic criteria. See Figure 2.
- Punctuation marks are all transcribed leaving a space after their preceding words. This is so as to reduce the total vocabulary and hence improve the handwritten text recognition performance.
- Page and line transcripts must preserve the original layout scheme as much as possible.
- Abbreviations (Figure 3) are transcribed as such together with an expansion to their full word(s).
- All foreign (non-Spanish) text is transcribed as such, and tagged with the corresponding language name; see Figure 4.
- Both the initial and final parts of hyphenated words are tagged (Figure 5).
- Crossed-out words are tagged, and also transcribed if they are readable (Figure 6).

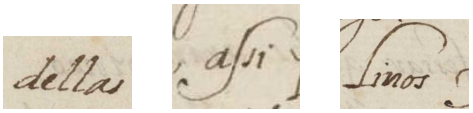


Figure 2: Examples of double-letters and large *S* and *L* letters, which are transcribed according to current transcription criteria as *dellas*, *assi* and *linos*.

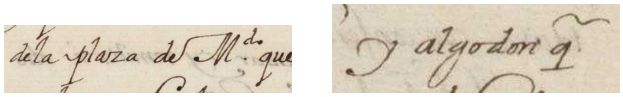


Figure 3: Handwritten pieces of text with the abbreviations M^d (Madrid) and $\tilde{q}$ (que).

Figure 4: Example of text handwritten in Latin.

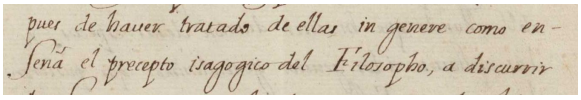


Figure 5: Example of two adjacent text lines with the hyphenated word *en-seña*.

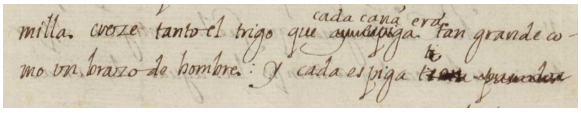


Figure 6: Example of two adjacent text lines with readable and unreadable crossed-out words.

To facilitate setting up more frequent tagging events, we define a usable set of tags based on the $ symbol. These tags prefix every transcribed word and are described in Table 1. On another line, the '#' character is used for tagging a completely illegible word. Nevertheless, All these handy tags are finally converted into TEI[6] format.

3. ORGANIZATIONAL ASPECTS

The adopted collaborative process of line-level transcription demands a careful human resource management and a scheduled execution plan. As we describe in Section 4.2, the transcription task is being performed by BNE and UCM, which have recruited four students specialized in paleography. Each transcriber has been assigned approximately 38

[5] http://bdh-rd.bne.es/viewer.vm?id=0000140162

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

Table 1: Set of tags and their associate events.

Description	Label-Tag
Line- end/start word hyphen	`$-prefix`, `-$suffix`
Deletion	`$/word`
Expansion	`$.word`
Catch-word	`$>word`
Underline	`$_word`
Superscript	`$^word`
Illegible text	`#word`
Foreign languages: Latin, Greek, Portuguese	`$l:word`, `$g:word`, `$p:word`

chapters of 5 pages on average, so as to balance the assigned work load. In addition, an experienced paleography expert has been assigned the task of reviewing the students' results and making decisions about the proper interpretation of transcription criteria when difficulties or novelties arise. The same expert was also in charge of checking, and amending when necessary, the layout analysis results. To coordinate the activity of the different agents involved in the project, tools such as *Google Docs* were used, along with several meetings aimed at discussing work assignments, transcription criteria, etc.

Transcribers are asked to work on their assignments individually on a line-by-line basis. Further, as one of the collaboration objectives is a longitudinal evaluation of a CATTI prototype (Section 4.3 and Figure 10), transcribers' learning skills are being assessed through two separate conditions: completely manual and predictive transcription.

For each of the manual and predictive transcription sessions, several data related with the effectiveness and efficiency of the process are collected. When the whole transcription work will be finished, these data will be compiled into adequate statistics which will allow us to analyze productivity and usability features of the underlying technology.

The final results will be available in two parts: the page images themselves and the ground truth information. For the latter we are using the PAGE XML format [7], which includes not only information about the page layout (i.e., locations of text/pictures blocks and text lines) but also page transcripts at the line level. The transcripts themselves are stored in a TEI-compliant format. Both TEI and PAGE are extensively used in the academic community.

4. TECHNICAL ASPECTS

4.1 Layout Analysis

Layout analysis aims at producing text blocks and line segmentations through a two-step top-down sequential procedure. First, text blocks are automatically detected and manually revised. Then, text lines for each of the text blocks are detected and finally extracted.

4.1.1 Text Block Detection

This step is focused on the detection and labeling of text and non-text areas on a page. Specifically the areas of interest were: *page number*, *heading*, *main text block*, *catch word* and *drawings*. Samples of verified text are shown in Figure 7a. Block detection is performed in a semi-supervised manner. First, an automatic, rough localization of text areas in the page images is performed by means of horizontal and vertical line detection methods [6, 8]. Then, the detected areas are verified, labeled and (if necessary) manually corrected by means of the GT_Tool_Page software[7].

4.1.2 Text Line Detection and Segmentation

The next step consists of detecting and segmenting the text lines that are found in each of the detected text blocks. First, the text baselines are detected using a fairly robust method based on Hidden Markov Models (HMMs) [1]. This method is enhanced by using the ground truth information of text blocks, in order to remove margins and non-text areas, the end result of which can be seen in Figure 8. Next, the text baselines are manually checked (and corrected if needed). Figure 7b shows an example of correctly detected text lines, as displayed by GT_Tool_Page.

Since initially there is no data to train the HMMs, an iterative segmentation process for baseline detection is used. It starts with a single page that is manually labeled. The data is then propagated to a new set of pages. Once the text baselines of the new set are verified, the HMMs are retrained with all the available data and then used to process the next set of pages.

Lastly, using dynamic programming, the best cutting path along each overlapping region of adjacent text lines was found, preserving as much as possible their respective ascender and descender strokes. Figure 8 shows the outcomes of the segmentation process.

4.2 Handwritten Text Recognition

A conventional HTR architecture is adopted, composed of three processes [11]: off-line preprocessing, feature extraction, and recognition.

HTR preprocessing [2] is aimed at correcting image degradations and geometry distortions: skew, slant corrections, and size normalization. Feature extraction, on the other hand, transforms a preprocessed text line image into a sequence of 60-dimensional feature vectors [11].

The recognition process is based on HMMs. Characters are modeled by continuous density left-to-right HMMs, using 12 states and 32 gaussian mixture components per state. The gaussian mixture is a probabilistic approach to model the emission of feature vectors in each HMM state. The optimum number of HMM states as well as the number of Gaussian densities per state are empirically tuned.

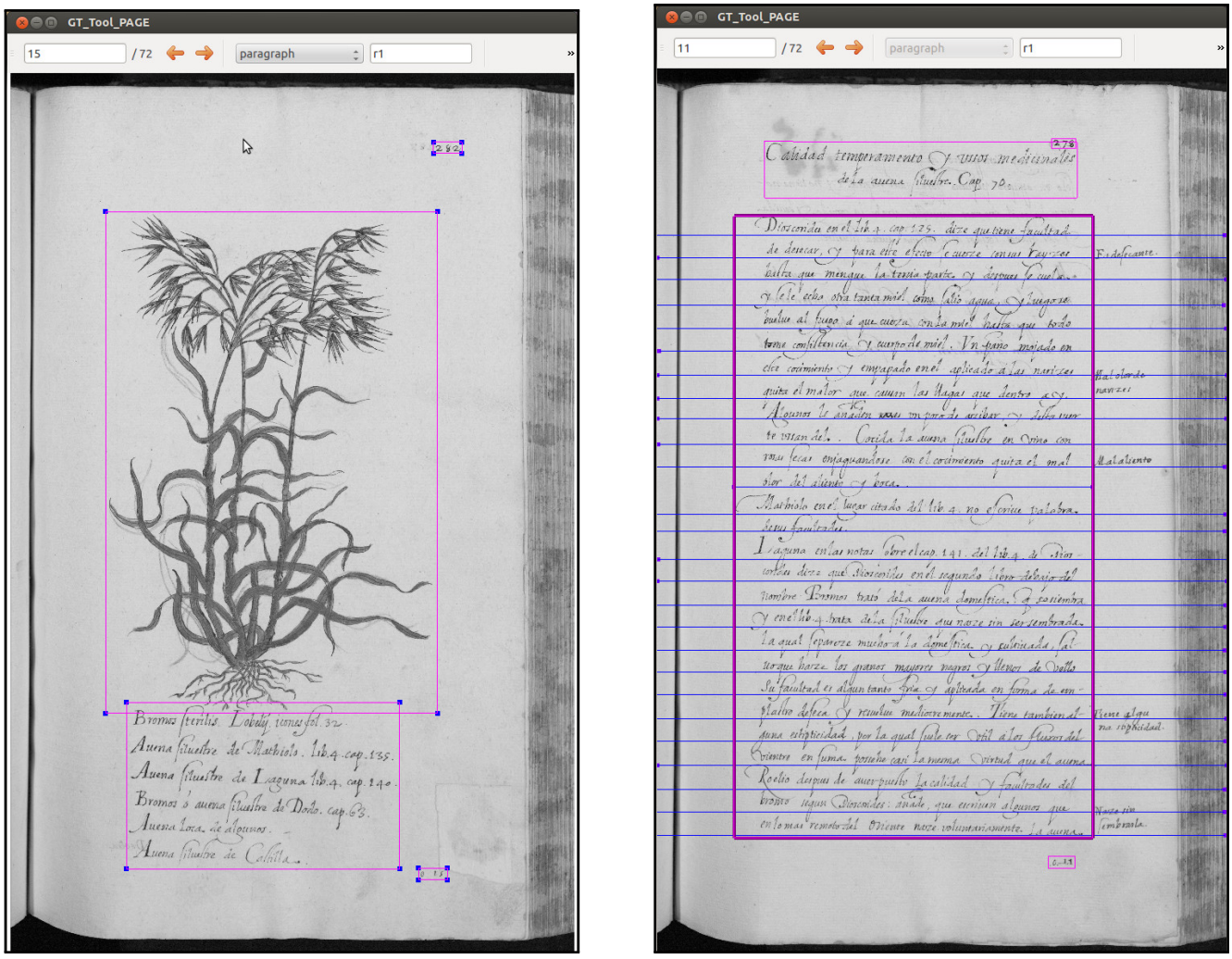
Each lexical word is then modelled by a stochastic finite-state automaton, which represents all possible concatenations of individual characters that compose a word. On the other hand, text sentences are modelled using bi-grams with Kneser-Ney back-off smoothing [4], estimated directly from the training transcriptions of the text line images.

All these finite-state models (character, word, and sentence) can be easily integrated into a single global model, on which a decoding process is efficiently performed by means of the Viterbi algorithm [3]. This can be adapted also for the search required in the CATTI prototype, as outlined in Section 4.3. In Figure 9 an example of the recognition outcome is given.

4.3 Computer Assisted Transcription

CATTI, shorthand of Computer Assisted Transcription of Text Images, is a system aimed at assisting the user in the transcription process. That is, the system eases and

[7]Publicly available.



(a) **(b)**

Figure 7: Example of text block detection (7a) and horizontal text lines detection (7b).

at the same time speeds up the task of transcribing text, by responding intelligently to keyboard and mouse interactions.

Both the system and the human transcriber cooperate together to generate the final transcript of the text images. The user validates portions of a sentence (named *prefixes*) that are then used by system to *predict* suitable text continuations (named *suffixes*). The rationale behind this approach is to combine the accuracy provided by the transcription expert with the efficiency of HTR technology.

The CATTI prototype implements ideas that have been already studied in the fields of speech recognition and machine translation [14]. Instead of giving an extensive review on the CATTI technology, the reader is redirected to previous work such as [15, 16] and the consolidated summary under a uniform formulation of multimodal interactive processing by Toselli *et al.* [12].

4.4 Graphical User Interface

The aforementioned CATTI prototype follows a client-server communication model. The web interface (the client) is responsible for rendering the application and capturing the user actions. The HTR engine (the server) loads language models of the text images and builds smart auto-completions from partially user-validated hypotheses.

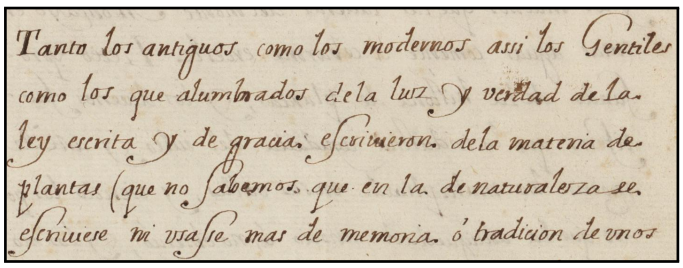
The user can perform different operations through the web interface, namely:

- *Substitute:* The user replaces the first erroneous word by the correct word.
- *Reject:* The system replaces the first erroneous word by another (possibly 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 words are concatenated to generate a correct word.
- *Split:* A word is divided into two different words.
- *Validate:* The full transcription is accepted.

Each operation is assigned a numerical code, and is sub-

(a) Original image. **(b)** Clean image. **(c)** Detected baselines. **(d)** Segmented lines.

Figure 8: From text line detection to segmentation.



llano los antiguos como los moderno assi los temple
como los que al sembrador dela los y verdad dela
ley escrito y desgracia escrivieron dela materia de
plantas que no $l:saturare que en la de naturaleza se
escrivieron viesse mas de memoria bendicion de uno

Figure 9: Example of recognized text piece image of five lines. Right panel show the output sentence hypotheses with the miss-recognized words underlined.

mitted together with the user-validated prefix to the HTR engine. Figure 10 shows some interface screenshots. The prototype has been used in the past for other transcription tasks [5, 9, 10], and the current version is available at `http://cat.iti.upv.es/iht2`.

5. REMARKS AND CONCLUSIONS

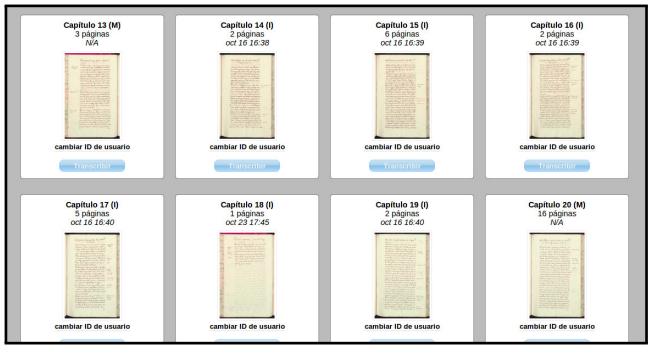
In this paper we have commented on a protocol originally designed to transcribe a botanical specimen book of the XVII century. The protocol has led us to generate ground truth data for a part of the book and to accurately transcribe most of the remaining book page images. In the near future, statistics of all the information generated throughout the process will be compiled, analyzed and published.

We plan to use the experience and the data obtained in this process to easy and speed up the transcription of other books written by Bernardo de Cienfuegos, as well as other similar historical books.

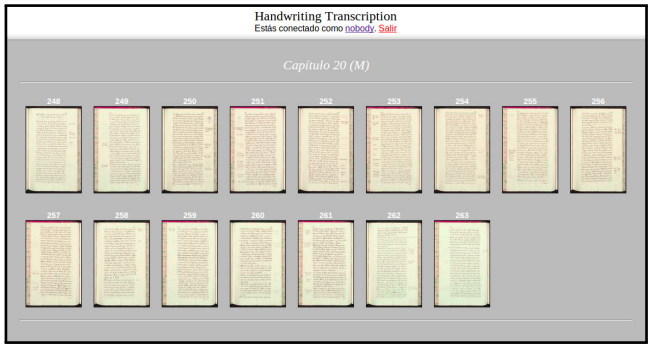
Ultimately, the resulting database will be made publicly available. We hope this will open a door to new research on handwriting transcription methods. Finally, the software we are developing in the framework of the tranScriptorium project will also be made available.

6. ACKNOWLEDGMENTS

The authors thank Jorge Martínez, who developed the GT_Tool_Page software and the trancription team composed by Andrea T. Zivny Valenzuela, David A. Rey Gómez, Roberto J. Alonso Sánchez and Lucía Sánchez Tacero. 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) and the Spanish MEC under the STraDa (TIN2012-37475-C02-01) project.



(a) Index view, aimed at selecting a group of pages.



(b) Pages view, aimed at selecting a single page.

(c) Main view. Transcription is performed inline.

Figure 10: Screenshots of the CATTI prototype.

7. REFERENCES

[1] V. Bosch, A. H. Toselli, and E. Vidal. Natural language processing framework for handwritten text line detection in legacy documents. In *Proc. LaTeCH*, 2012.

[2] F. Drira. Towards restoring historic documents degraded over time. In *Proc. DIAL*, pages 350–357, 2006.

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

[4] R. Kneser and H. Ney. Improved backing-off for n-gram language modeling. In *Proc. ICASSP*, pages 181–184, 1995.

[5] L. A. Leiva, V. Romero, A. H. Toselli, and E. Vidal. Evaluating an interactive-predictive paradigm on handwriting transcription: A case study and lessons learned. In *Proc. COMPSAC*, pages 610–617, 2011.

[6] V. Malleron and V. Eglin. A mixed approach for handwritten documents structural analysis. In *Proc. ICDAR*, pages 269–273, 2011.

[7] S. Pletschacher and A. Antonacopoulos. The PAGE (page analysis and ground-truth elements) format framework. In *Proc. ICPR*, pages 257–260, 2010.

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

[9] V. Romero, L. A. Leiva, V. Alabau, A. H. Toselli, and E. Vidal. A web-based demo to interactive multimodal transcription of historic text images. In *Proc. ECDL*, pages 459–460, 2009.

[10] V. Romero, L. A. Leiva, A. H. Toselli, and E. Vidal. Interactive multimodal transcription of text images using a web-based demo system. In *Proc. IUI*, pages 477–478, 2009.

[11] 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. *IJPRAI*, 18(4):519–539, 2004.

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

[13] A. H. Toselli, E. Vidal, V. Romero, and V. Frinken. Word-graph based keyword spotting in handwritten document images. Under review, 2013.

[14] E. Vidal, L. Rodríguez, F. Casacuberta, and I. García-Varea. Interactive pattern recognition. In *Proc. MLMI*, pages 60–71, 2007.

[15] A. Vinciarelli, S. Bengio, and H. Bunke. Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. *IEEE Trans. on PAMI*, 26(6):709–720, 2004.

[16] M. Zimmermann, J. Chappelier, and H. Bunke. Offline grammar-based recognition of handwritten sentences. *IEEE Trans. on PAMI*, 28(5):818–821, 2006.