https://doi.org/10.1142/9789811203244_0002

Chapter 2

IAM-HistDB
A Dataset of Handwritten Historical Documents

Andreas Fischer

2.1 Introduction

The goal of automated reading in historical manuscripts is to receive a scanned page of the manuscript as input and to produce a machine-readable transcription of all text elements of the page. Ideally, the transcription is aligned with the scanned page, such that it can be overlaid with the image similar to printed documents that have been processed by Optical Character Recognition (OCR).

In order to develop and test systems for automated reading, the data has to be prepared accordingly. Annotations are needed for layout analysis (where are text elements on the page?) as well as handwriting recognition (what is the machine-readable transcription of those elements?). Such annotations are called *ground truth* of the pattern recognition problem and have a two-fold purpose. On the one hand, they serve as learning samples for machine learning algorithms and, on the other hand, they allow to evaluate the performance of the algorithms.

In this chapter, we describe the annotated database of historical documents that was created at the beginning of the HisDoc project at IAM institute of the University of Bern, Switzerland. The IAM-HistDB was one of the first comprehensive research datasets at the time and has been made publicly available to the research community.[1] It currently has over thousand registered users and has become a well-established benchmark in the field, allowing to test novel algorithms and compare them with previous work.

[1] http://www.fki.inf.unibe.ch/databases/iam-historical-document-database

The remainder of this chapter is structured as follows. Section 2.2 provides pointers to related research datasets, Section 2.3 presents the contents of the IAM-HistDB in greater detail, and Section 2.4 discusses the semi-automatic method used to create the ground truth. Finally, we draw some conclusions in Section 2.5.

2.2 Related Work

In contrast to printed documents, the ground truth for handwriting recognition typically does not provide annotations at character level, i.e. it does not assign a bounding box and a transcription to each individual character. For touching and overlapping characters, especially in the case of cursive handwriting styles, it is difficult to decide where a character starts and ends. Even the segmentation into individual words may be challenging when the white space between words is small, irregular, or completely omitted.

Instead, handwriting recognition systems often operate at line level, i.e. the layout analysis module attempts to find text lines within the scanned page and the recognition module attempts to find the correct sequence of characters and words within the text line images. Accordingly, a research dataset for handwriting recognition should contain the location of the text lines on the scanned page image, together with their machine-readable transcription. Unfortunately, existing electronic editions provided by researchers from the humanities usually do not contain such annotations. They may provide a transcription at page level but lack the information where the individual text lines are located on the page image.

For modern handwriting, subjects may be asked to copy a given text with their own handwriting into prepared forms that facilitate the creation of large research datasets, as for example in the IAMDB [Marti and Bunke (2002)] that contains modern handwriting samples from over 600 writers.

Creating ground truth for historical handwriting is more challenging, as the annotations have to be added to existing manuscripts. It can be done either completely manually or computer-assisted if a handwriting recognition system is available. In either case, it requires a considerable amount of time for manual interaction, especially for ancient scripts and languages that necessitate expert knowledge from the human user.

One of the first datasets that was shared with the research community was the George Washington database containing 20 annotated pages of scanned letters written by George Washington and his associates during wartime. After it has been made available by Rath et al. [Rath and Manmatha (2007)], the dataset has been used by a number of research groups for handwriting recognition and keyword spotting. The IAM-HistDB also contains these 20 images with new annotations at line level (see Section 2.3.3).

Several interactive annotation tools have been developed that led to the creation of more datasets. Early systems include DEBORA for annotating Renaissance documents [Bourgeois and Emptoz (2007)] and DMOS for annotating old civil status registers and military forms [Coüasnon *et al.* (2007)]. Several systems have been proposed for annotating old Spanish manuscripts, including STATE [Gordo *et al.* (2008)], CATTI [Romero *et al.* (2007)], and GIDOC [Serrano *et al.* (2010)], which was used to create the publicly available GERMANA[2] database [Pérez *et al.* (2009)].

More recently, the tranScriptorium[3] project (2013–2015) and the READ[4] project (2016–2019) with its Transkribus platform had a fundamental impact on the collection and annotation of new datasets. They have not only developed interactive annotation tools and made research datasets publicly available, but also created an active community around the Transkribus platform, which is expected to continue sharing and annotating collections of historical documents.

2.3 The IAM-HistDB

The IAM-HistDB contains three databases with different handwriting styles and languages, namely the Saint Gall database, the Parzival database, and the George Washington database. Together, they provide annotations for over hundred scanned page images to train and evaluate automatic reading systems.

2.3.1 *Saint Gall Database*

The Saint Gall database is based on a medieval Latin manuscript from the 9th century written in Carolingian script, which contains the hagiography

[2] `https://www.prhlt.upv.es/wp/resource/the-germana-corpus`
[3] `http://transcriptorium.eu`
[4] `https://read.transkribus.eu`

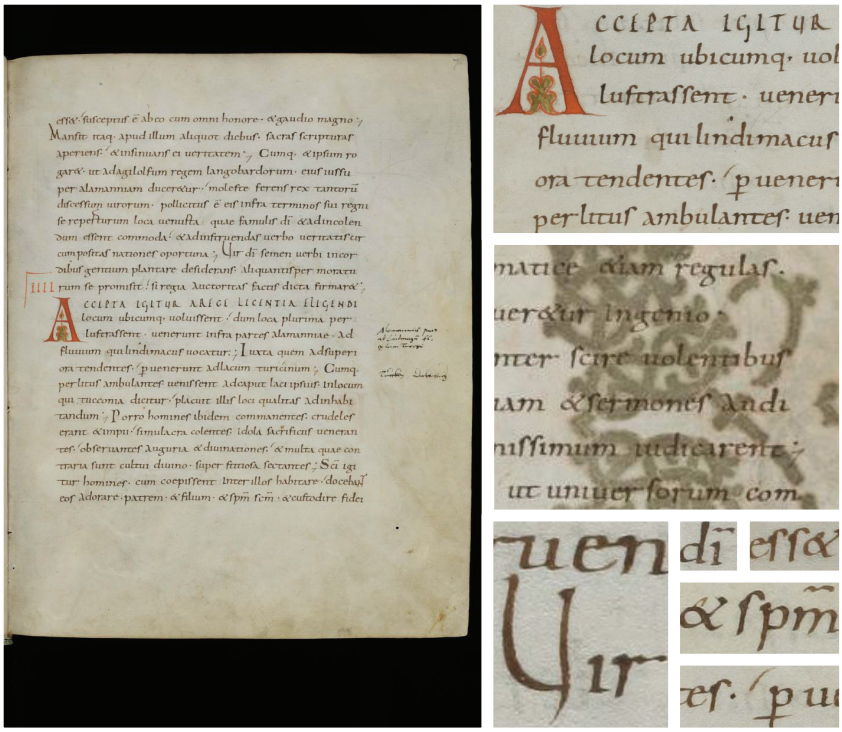

Fig. 2.1: Saint Gall database.

Table 2.1: Saint Gall database statistics.

Saint Gall database	
Century	9th
Language	Latin
Script	Carolingian
Original manuscript	Abbey Library of Saint Gall, Cod. Sang. 562
Physical format	24.0 x 30.0 cm, ink on parchment
Digital format	3328 x 4993 px, color JPEG
Writers	1
Pages	60
Text Lines	1410
Word Instances	11597
Word Classes	4890
Characters	49

Vita sancti Galli by Walafrid Strabo. The original is held by the Abbey Library of Saint Gall, Switzerland (Cod. Sang. 562), its digital images are available on the e-codices website,[5] a virtual library from the Medieval Institute of the University of Fribourg, Switzerland, and a text edition of the manuscript can be found on the monumenta website[6] based on the Patrologia Latina edition.[7] Figure 2.1 provides visual samples and Table 2.1 summarizes the main characteristics of the database.

Annotations include the text line locations on the page image in form of closed polygons and their corresponding transcription. Besides a diplomatic transcription, i.e. the exact sequence of characters and words visible in the handwriting, the database also provides an aligned version of the text edition from the monumenta website, which often deviates from the handwritten text for better readability, especially with a view to abbreviations, capitalization, and punctuation.

The Saint Gall database has a regular page layout, which contains a single column of 24 straight text lines with ample spacing between the lines. The text foreground is clearly distinguishable from the parchment background. For automated reading, only a relatively small number of 49 characters has to be modeled. Nevertheless, the database poses several typical challenges, including marginal notes, colored initial letters, holes in the parchment, ink bleed-through, lack of spacing between words, frequent word breaks at the line end, and abbreviated words (see Figure 2.1 for visual samples).

The database has been introduced in [Fischer *et al.* (2011)] in the context of transcription alignment. The goal was to automatically align the transcription from the monumenta website with the page image.

2.3.2 *Parzival Database*

The Parzival database is based on a medieval German manuscript from the 13th century written in Gothic script, which contains the epic poem *Parzival* by Wolfram von Eschenbach. The original is held by the Abbey Library of Saint Gall, Switzerland (Cod. 857) and its digital images together with a transcription were made available on CD-ROM by the German Language Institute of the University of Bern, Switzerland.[8] Figure 2.2 provides visual samples and Table 2.2 summarizes the main characteristics of the database.

[5] `http://www.e-codices.unifr.ch`
[6] `http://www.monumenta.ch`
[7] J.-P. Migne PL114, 1852
[8] `http://www.parzival.unibe.ch/`

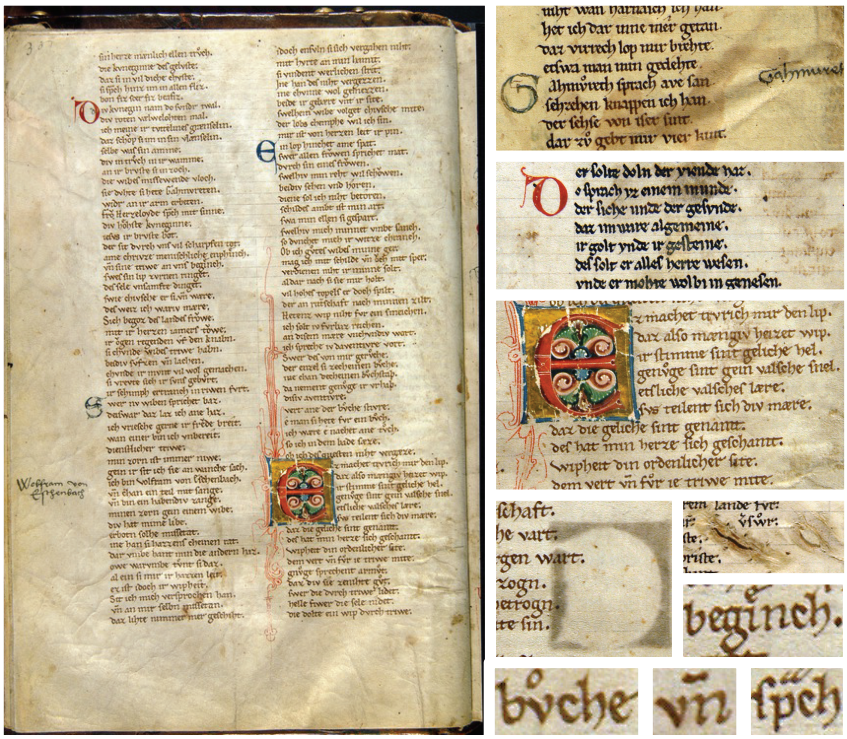

Fig. 2.2: Parzival database.

Table 2.2: Parzival database statistics.

Parzival database	
Century	13th
Language	German
Script	Gothic
Original manuscript	Abbey Library of Saint Gall, Cod. 857
Physical format	21.5 x 31.5 cm, ink on parchment
Digital format	2000 x 3008 px, color JPEG
Writers	3
Pages	47
Text Lines	4477
Word Instances	23478
Word Classes	4934
Characters	93

Annotations include the transcription for preprocessed line and word images. Preprocessing include binarization, removal of the skew, i.e. the inclination of the text line, and horizontal and vertical scaling. As it was the first database investigated during the HisDoc project, our ground truth creation procedure described in Section 2.4 was not yet in place and we have, unfortunately, not stored the location of the text lines within the page image. The diplomatic transcription includes special characters for abbreviation symbols.

The Parzival database has a two-column layout with pairwise rhyming lines. Around the main text, we find ornaments, colored initial letters, and marginal notes that were added at a later time to the manuscript. The ravages of time have affected the parchment considerably, leading to stains, wrinkles, holes, and faded ink. Some of the torn pages were repaired with visible stitches. For automated reading, a total of 93 characters have to be modeled, including a number of abbreviation symbols. Figure 2.2 provides a visual impression.

The database has been introduced in [Fischer *et al.* (2009)] in the context of handwriting recognition using hidden Markov models (HMM) as well as bi-directional recurrent neural networks with long short-term memory cells (BLSTM). The goal was to transcribe the line and word images.

2.3.3 *George Washington Database*

The George Washington database is based on English letters from the 18th century written in longhand script by George Washington and his associates. The originals are held by the Library of Congress and their digital images have been made available online by the library together with a transcription.[9] Figure 2.3 provides visual samples and Table 2.3 summarizes the main characteristics of the database.

Annotations include the transcription for preprocessed line and word images. Similar to the Parzival database, preprocessing include binarization, removal of the skew, and horizontal and vertical scaling. In addition, the slant, i.e. the inclination of the characters, has been removed as well.

The George Washington database has a less regular layout than the medieval manuscripts. Apart from the main text, the letters contain page numbers, rulers, stamps, and signatures. Faded ink and stains render the task of

[9] George Washington Papers at the Library of Congress from 1741–1799, Series 2, to be found at `http://memory.loc.gov/ammem/gwhtml/gwseries2.html`

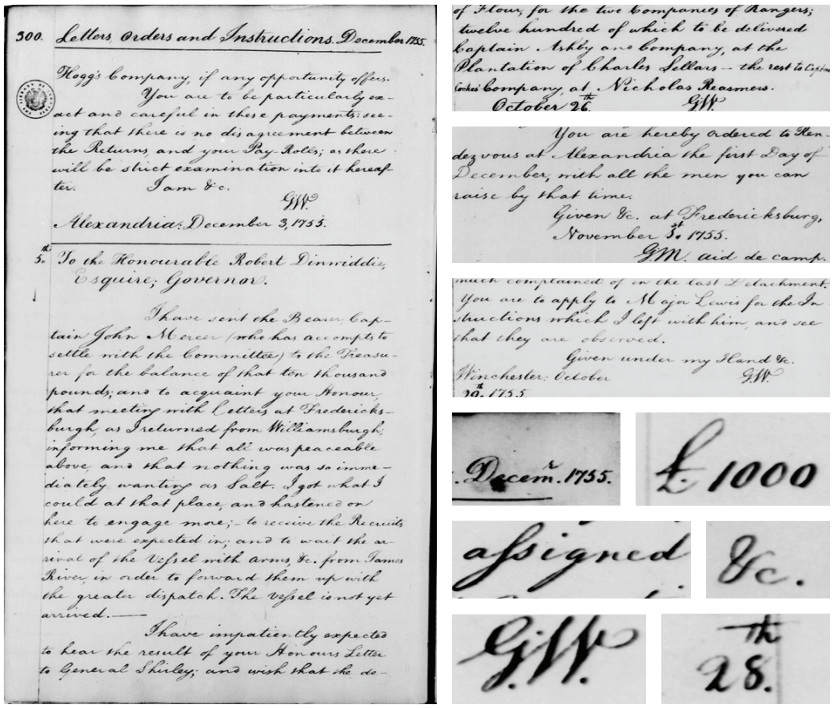

Fig. 2.3: George Washington database.

Table 2.3: George Washington database statistics.

George Washington database	
Century	18th
Language	English
Script	Longhand
Original manuscript	Library of Congress
Physical format	Ink on paper
Digital format	2034 x 3286 px, grayscale JPEG
Writers	2
Pages	20
Text Lines	656
Word Instances	4894
Word Classes	1471
Characters	82

automated reading more difficult. A total of 82 characters have to be modeled, including numbers, currency symbols, double "s" letters, "th" written on top of numbers, the frequently used "etc." symbol, and signatures (see Figure 2.3 for visual samples).

The 20 page images included in the George Washington database have been used in several studies before, e.g., in [Lavrenko *et al.* (2004)] for word recognition. Our database has been introduced in [Fischer *et al.* (2010b)] in the context of keyword spotting. The goal was to retrieve text line images that contain specific search terms provided by the user as free text.

2.4 Semi-Automatic Ground Truth Creation

In this section, we describe the semi-automatic procedure we have developed during the HisDoc project for creating ground truth annotations for automated reading. The aim of this procedure was to find a reasonable balance between the work performed by human users and tasks that can be performed automatically. Also, we wanted laypersons without special knowledge in computer science or linguistics to be able to perform all required manual interactions. The procedure is detailed in [Fischer *et al.* (2010a)]. It consists of five consecutive steps illustrated in Figure 2.4.

Step 1 – Text Selection. The first step is manual and consists in selecting the main text areas on the page image with a polygon.[10] Such areas, typically text columns, are defined as blocks of consecutive text lines. By means of this selection, the main text is separated from margin notes, ornaments, drawings, colored initial letters, and page numbers. The selection is stored in Scalable Vector Graphics (SVG) format.

Step 2 – Foreground Detection. The second step is automatic and aims to separate the text foreground from the page background. We apply a Difference of Gaussian (DoG) filter on grayscale images to locally enhance edges, i.e. we subtract two Gaussian blurs with different radii. Afterwards, a global threshold is applied for binarization. The two radii and the threshold are optimized by visual inspection of a few sample pages, in order to exclude ink bleed-through and stains as well as possible. After binarizing the whole page, the main text areas are cut out for further processing, according to the text selection from the previous step. For the IAM-HistDB, this approach

[10] We used the *Paths* tool of the GIMP software, `http://www.gimp.org/`.

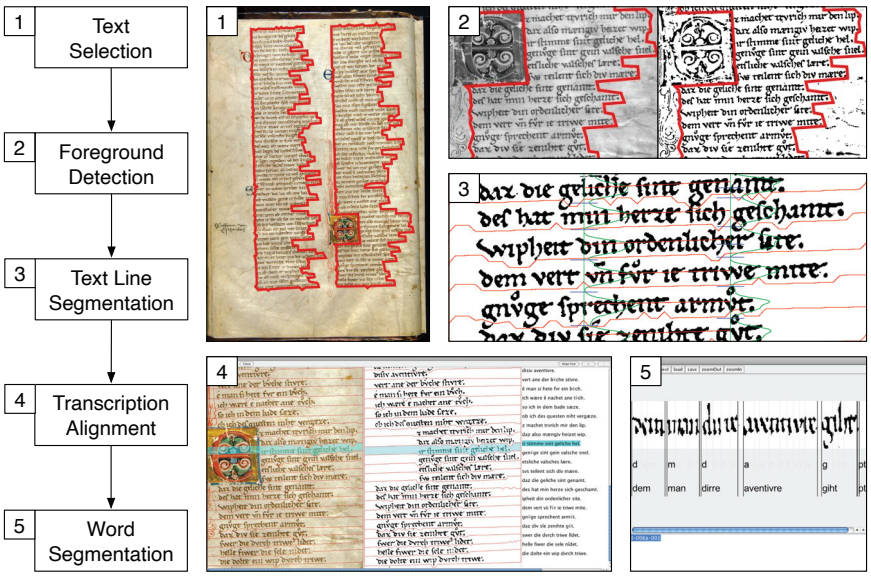

Fig. 2.4: Semi-automatic ground truth creation.

led to acceptable results in all cases, despite the fact that the three databases under consideration are quite different.

Step 3 – Text Line Segmentation. The third step consists of automatic text line segmentation, followed by manual correction. For segmenting the text lines, we use a seam carving approach to cut consecutive lines. First, start and end points of the seams are estimated from horizontal projection profiles. Afterwards, seams are computed from left to right using dynamic programming, avoiding text foreground if possible and staying close to a straight line. For more details on the seam carving algorithm, we refer to [Fischer *et al.* (2010a)]. The resulting seams are inspected and corrected by a human user in a graphical user interface. They can drag with the mouse to adjust the closest seam. Combining the text selection from the first step with the top and bottom seams from the third step, we obtain a polygon around each text line that is stored in SVG format.

Step 4 – Transcription Alignment. The fourth step aims to align the available transcription with the line images. It consists of automatic parsing of the text, followed by manual correction. Parsing is needed to clean the

text, standardize its encoding to Unicode, and to split text lines based on heuristics. Manual correction is performed in a text editor, which is integrated with the line segmentation results as illustrated in Figure 2.4. For the IAM-HistDB, the manual interaction ranged from rapid corrections of line breaks to more time-consuming corrections of abbreviations.

Step 5 – Word Segmentation. The fifth and last step consists of automatic word segmentation followed by manual correction. For this purpose a handwriting recognition system based on hidden Markov models (see Chapter 5) is applied in forced alignment mode to compute word boundaries. Afterwards, the boundaries are manually corrected in a graphical user interface. For the IAM-HistDB, the automatic alignment was near-perfect and only few manual corrections were needed.

In summary, the principal ground truth annotations for automated reading consist of polygons around text lines, stored as SVG, together with their diplomatic transcription, stored as plain text in Unicode. These annotations allow to train and evaluate layout analysis systems, which aim to automatically extract line images, and handwriting recognition systems, which aim to transcribe the line images into machine-readable text. For recognition experiments at word-level, we also provide word images, which are annotated with their transcription.

In terms of time consumption, the first bottleneck of the proposed procedure is the manual text selection, which took about 2 minutes per column on average for the IAM-HistDB. The second bottleneck is the manual correction of the text line segmentation together with the manual correction of the transcription alignment, which together took about 3.5 minutes per column if the user was familiar with the task.

2.5 Conclusions

The creation of a comprehensive research database with a variety of different scripts and languages was an important first step in the HisDoc project, as there was almost no data available for training and evaluating handwriting recognition in the context of historical manuscripts at the time. With the Saint Gall, Parzival, and George Washington databases, we have focused on relatively regular layouts and Latin alphabets, in order to facilitate our first attempts towards automated reading.

The proposed semi-automatic procedure for creating ground truth has proven fairly efficient with a bit more than five minutes of manual interaction per page on average. Nevertheless, our experience was that the interaction quickly became repetitive and rather dull. Clearly, there is much room for improvement in that regard, e.g. designing the procedure in a way that the human user focuses on difficult cases, interacts naturally and rapidly with the proposed annotations, and sees how the system gets better at suggesting annotations over time.

References

Bourgeois, F. L. and Emptoz, H. (2007). DEBORA: Digital AccEss to BOoks of the RenAissance, *Int. Journal on Document Analysis and Recognition* **9**, pp. 193–221.

Coüasnon, B., Camillerapp, J., and Leplumey, I. (2007). Access by content to handwritten archive documents: Generic document recognition method and platform for annotations, *Int. Journal on Document Analysis and Recognition* **9**, 2, pp. 223–242.

Fischer, A., Frinken, V., Fornés, A., and Bunke, H. (2011). Transcription alignment of latin manuscripts using hidden Markov models, in *Proc. 1st Int. Workshop on Historical Document Imaging and Processing*, pp. 29–36.

Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., and Stolz, M. (2010a). Ground truth creation for handwriting recognition in historical documents, in *Proc. 9th Int. Workshop on Document Analysis Systems*, pp. 3–10.

Fischer, A., Keller, A., Frinken, V., and Bunke, H. (2010b). HMM-based word spotting in handwritten documents using subword models, in *Proc. 20th Int. Conf. on Pattern Recognition*, pp. 3416–3419.

Fischer, A., Wüthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., and Stolz, M. (2009). Automatic transcription of handwritten medieval documents, in *Proc. Int. Conf. on Virtual Systems and Multimedia*, pp. 137–142.

Gordo, A., Llorens, D., Marzal, A., Prat, F., and Vilar, J. (2008). State: A multimodal assisted text-transcription system for ancient documents, in *Proc. 8th Int. Workshop on Document Analysis Systems*, pp. 135–142.

Lavrenko, V., Rath, T. M., and Manmatha, R. (2004). Holistic word recognition for handwritten historical documents, in *Proc. Int. Workshop on Document Image Analysis for Libraries*, pp. 278–287.

Marti, U.-V. and Bunke, H. (2002). The IAM-database: an English sentence database for offline handwriting recognition, *IJDAR* **5**, 1, pp. 39–46.

Pérez, D., Tarazón, L., Serrano, N., Castro, F.-M., Ramos-Terrades, O., and Juan, A. (2009). The GERMANA database, in *Proc. 10th Int. Conf. on Document Analysis and Recognition*, pp. 301–305.

Rath, T. M. and Manmatha, R. (2007). Word spotting for historical documents, *IJDAR* **9**, 2–4, pp. 139–152.

Romero, V., Toselli, A. H., Rodríguez, L., and Vidal, E. (2007). Computer assisted transcription for ancient text images, in *Proc. 4th Int. Conf. on Image Analysis and Recognition*, pp. 1182–1193.

Serrano, N., Tarazón, L., Pérez, D., Ramos-Terrades, O., and Juan, A. (2010). The GIDOC prototype, in *Proc. 10th Int. Workshop on Pattern Recognition in Information Systems*, pp. 82–89.