The RODRIGO database

N. Serrano, F. Castro, A. Juan

DSIC/ITI, Universitat Politècnica de València
Camí de Vera, s/n, 46022 València, SPAIN
{nserrano,francas,ajuan}@iti.upv.es

Abstract

Annotation of digitized pages from historical document collections is very important to research on automatic extraction of text blocks, lines, and handwriting recognition. We have recently introduced a new handwritten text database, GERMANA, which is based on a Spanish manuscript from 1891. To our knowledge, GERMANA is the first publicly available database mostly written in Spanish and comparable in size to standard databases. In this paper, we present another handwritten text database, RODRIGO, completely written in Spanish and comparable in size to GERMANA. However, RODRIGO comes from a much older manuscript, from 1545, where the typical difficult characteristics of historical documents are more evident. In particular, the writing style, which has clear Gothic influences, is significantly more complex than that of GERMANA. We also provide baseline results of handwriting recognition for reference in future studies, using standard techniques and tools for preprocessing, feature extraction, HMM-based image modelling, and language modelling.

1. Introduction

There are huge historical document collections residing in libraries, museums and archives that are currently being digitized for preservation purposes and to make them available worldwide through large, on-line digital libraries. The main objective, however, is not to simply provide access to raw images of digitized documents, but to annotate them with their real informative content and, in particular, with text transcriptions. However, automatic extraction of text blocks, lines, and handwriting recognition are still open research problems (Ramos et al., 2010; Likforman-Sulem et al., 2007; Bertolami and Bunke, 2008).

In (Pérez et al., 2009), we presented a new handwritten text database, GERMANA, to facilitate empirical comparison of different approaches to automatic extraction of text blocks, lines, and handwriting recognition. GERMANA is the result of digitizing and annotating a 764-page Spanish manuscript from 1891, in which most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines. To our knowledge, it is the first publicly available database for handwriting research, mostly written in Spanish and comparable in size to standard databases such as IAM (Marti and Bunke, 2002; Su et al., 2007). In this paper, we present another handwritten text database, which will be referred to as RODRIGO. In this case, we have selected a manuscript much older than that of GERMANA, from 1545, which is publicly available in digitized form, at 300dpi in true colors, from the Spanish "Ministerio de Cultura" web site (Rod, 2006). The original manuscript is a 853-page bound volume, entitled *"Historia de España del arçobispo Don Rodrigo"*, and completely written in old Castilian (Spanish) by a single author. We carefully annotated all text blocks, lines and transcriptions, resulting in approximately $20K$ lines and $231K$ running words from a lexicon of $17K$ words, that is, very similar to GERMANA in size. The main purpose of this work is to let this annotation known to researchers and to provide an adequate reference for future studies. The interested reader can download it from (Rod, 2010).

As GERMANA, RODRIGO is not a particularly difficult task for text and block line detection since most pages only contain a single text block of nearly calligraphed handwriting on well-separated lines. It is also a single-author manuscript on a limited-domain task and, easier than GERMANA, it is only written in Spanish. Nevertheless, RODRIGO comes from a much older manuscript, and thus the typical difficult characteristics of historical documents are more evident. In particular, the writing style, which has clear Gothic influences, is significantly more complex than that of GERMANA.

In what follows, we first describe the manuscript and the database in Sections 2 and 3, respectively. Then, in Section 4, some preliminary results are reported using a standard, HMM-based recognizer. Finally, conclusions and future work are discussed in Section 5.

Work supported by the EC (FSE), the Spanish Government (MEC, MICINN, "Plan E", under grants MIPRCV "Consolider Ingenio 2010" CSD2007-00018, iTransDoc TIN2006-15694-CO2-01, MITTRAL TIN2009-14633-C03-01 and FPU AP2007-02867) and the Generalitat Valenciana (grant Prometeo/2009/014).

2. The manuscript

As said above, the RODRIGO database corresponds to a manuscript from 1545 entitled *"Historia de España del arçobispo Don Rodrigo"*, and completely written in old Castilian (Spanish) by a single author. It is a 853-page bound volume divided into 307 chapters describing chronicles from the Spanish history. Most pages only contain a single text block of nearly calligraphed handwriting on well-separated lines. This can be seen in Fig. 1, where it is also apparent that writing style has clear Gothic influences (Millares and Ruiz, 1983).
Other characteristic details of RODRIGO that can be clearly appreciated in Fig. 1 are:

- The author tends to embellish the writing, specially in broad white spaces, resulting in the extension of some ascenders and descenders across whole words.
- Natural blank spaces between successive words are often omitted; e.g., the words "de la" are written as a single word "dela" in the third line from the bottom of page 15. Sometimes, on the contrary, artificial blank spaces are inserted within a single word; e.g., the word "llegaronse" is written as two words, "llegaron se".
- Each chapter should begin with a dropcap, but the manuscript contains no dropcaps, probably because it was never brought to an artist to do so. Instead, there is a blank area in each position where a dropcap should have been inserted and, in most cases, the corresponding letter is written in small size.
- The first words in each even page are also copied in the bottom right corner of its preceding page.

3. The database

The manuscript was carefully digitized by experts from the Spanish *Ministry of Culture*, at 300dpi in true colors, and it is publicly available at (Rod, 2006). As with historical documents in general, scanned pages have noise effects like spots, tears, ink fading and transparency of back side. Also, they show a slight warping due to book binding. Nevertheless, the manuscript can be easily read and thus we decided not to apply any preprocessing to it (apart from desaturation) for ground-truth annotation.
We followed an annotation procedure very similar to the one used for the GERMANA database (Pérez et al., 2009). First, all text blocks were annotated with minimal enclosing rectangles and, within each text block, each text line was marked by its (straight) baseline. This was done semi-automatically by means of the *GIDOC prototype* (Serrano et al., 2010). All blocks and baselines automatically detected were also manually supervised, and corrected when needed.
On the other hand, the whole manuscript was transcribed line by line, by a paleography expert, in accordance with the following transcription rules:

- Page and line breaks are copied exactly.
- Missing natural blank spaces between successive words are indicated by the symbol "◡".
- Inserted artificial blank spaces within words are indicated by the symbol "␣".
- No spelling mistakes are corrected.
- No case or accentuation change is done.
- Punctuation signs are copied as they appear.
- Word abbreviations are first copied verbatim, except for sub-indixes and super-indixes, which are written in LaTeX-like notation as `_{sub}` and `^{super}`, respectively. Then, they are followed by the corresponding word between brackets. Thus, for instance, `q^{i}er.` is transcribed as `q^{i}er[quier]`.
- The symbol "$" is appended to each line having a broken word at its end.

The total time required for a single expert to manually annotate (text blocks, baselines and transcriptions of) the whole manuscript was estimated as 500 hours; that is, approximately 35 minutes per page on average. The complete annotation of RODRIGO is publicly available, for non-commercial use, at (Rod, 2010). It comprises about $20K$ text lines and $231K$ running words from a lexicon of $17K$ words, which is comparable in size to standard databases such as IAM (Marti and Bunke, 2002; Su et al., 2007). It is worth noting that more than half of the words in the lexicon (54.4%) are singletons (*hapax legomena*), but they only account for a 4.1% of the running words. Please see Table 1 for more details.

4. Experiments

As discussed in the introduction, RODRIGO is introduced to facilitate comparison of different approaches to automatic extraction of text blocks, lines, and handwriting recognition. In this section, however, we will restrict ourselves to (automatic) transcription (handwriting recognition). More specifically, our aim is simply to provide baseline results for reference in future studies, using standard techniques and tools; i.e.,

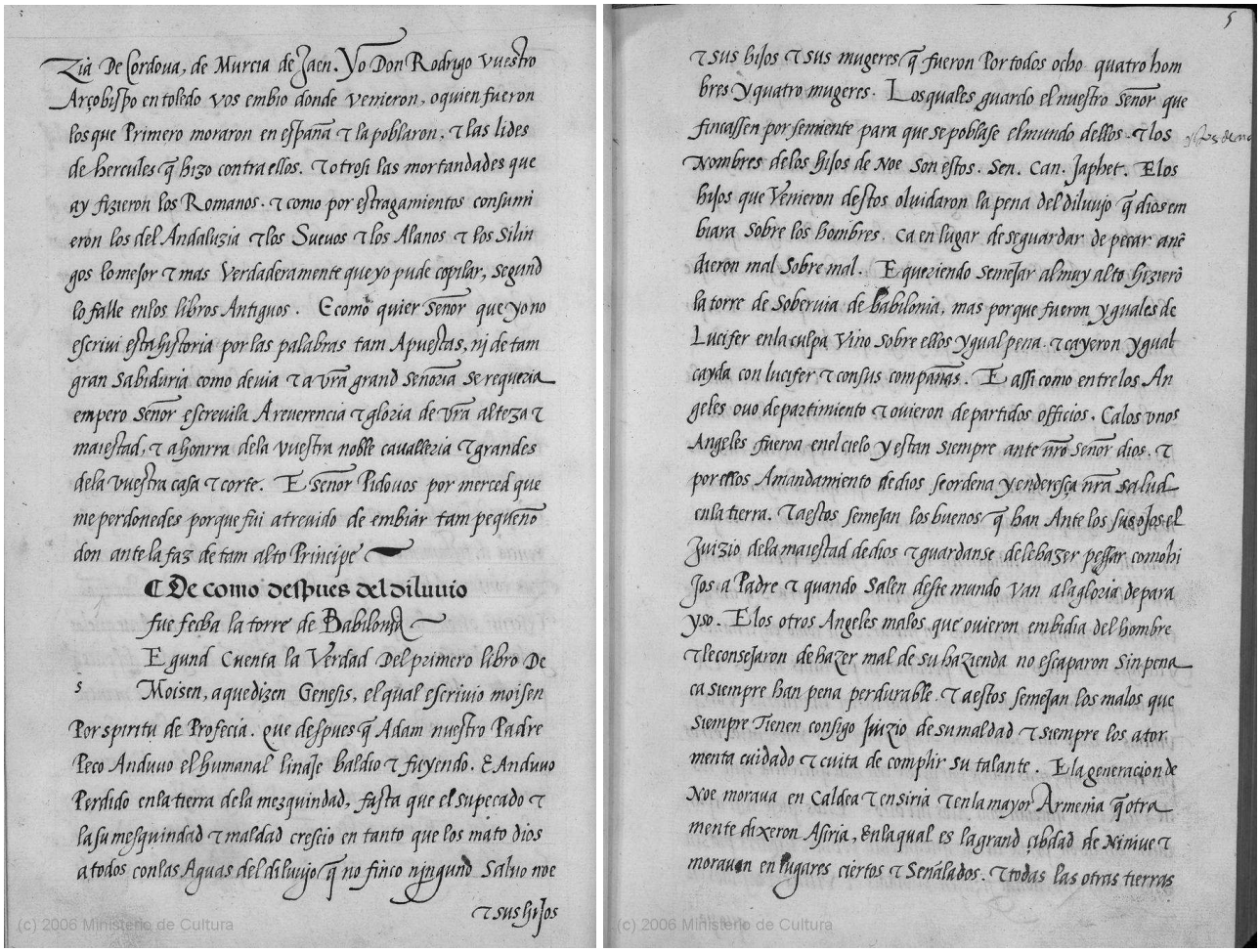

Figure 1: Pages 15 and 16 of RODRIGO.

Pages	853
Lines	20357
Running words	$232K$
Perplexity	166
Lexicon size	$17.3K$
Singletons (%)	54.4
Character set size	115

Table 1: Basic statistics of the RODRIGO text transcriptions (with isolated punctuation signs and abbreviations substituted by their corresponding words). Perplexity was computed using a bigram language model and a 100-fold cross-validation experiment. Singletons refers to words occurring exactly once.

HMM-based text image modeling and n-gram language modeling (Pérez et al., 2009).

Due to its sequential book structure, the very basic task on RODRIGO is to transcribe it line by line, from the beginning to the end. We assume that an automatic transcription system is used, and that each (automatically) transcribed line is supervised and, if necessary, amended by an expert. Clearly, after processing a block of lines or pages, all supervised transcriptions may be very well used for better (re-)training of image and language models, and thus improving system accuracy (Serrano et al., 2010).

Taking into account the above discussion, we divided RODRIGO into 20 consecutive blocks of 1000 lines each ($1-1000$, $1001-2000$, $\ldots$, $19001-20357$). Then, from block 1 to 19, the system was (re-)trained using all preceding blocks, with block 2 also used for further adjustment of a few, key parameters. After each retraining, the system accuracy was measured in terms of Word Error Rate (WER) on both, the next block to supervise and the last block. The resulting curves are shown in Fig. 2. Also shown is the part of the WER due to the occurrence of out-of-vocabulary (OOV) words.

As expected, the WER on the last block decreases as the amount of training data increases. Interestingly, however, the WER on the next block curve reveals considerable fluctuations in the recognition complexity of intermediate blocks. Nevertheless, this curve also tends to decrease and, indeed, both curves converge to a WER around 36.5% for block 20. We think

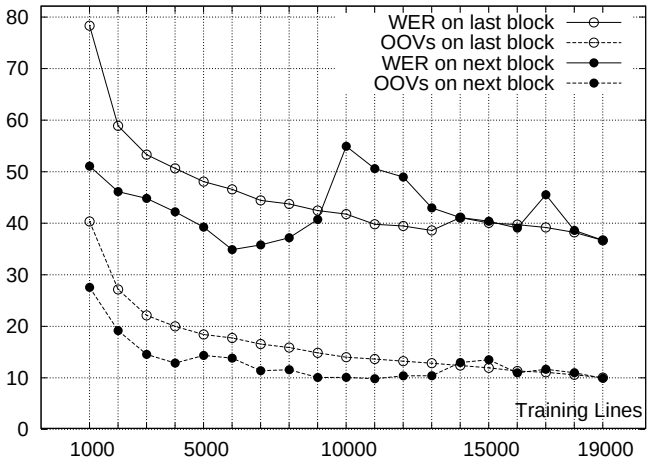

Figure 2: Transcription Word Error Rate (WER) on RODRIGO as a function of the blocks of (1000 lines) already supervised and thus available for training (Training lines). The WER is computed for both, the next block to supervise (solid line with black circles) and the last block (lines $19001 - 20357$). Also shown is the part of the WER due to the occurrence of out-of-vocabulary (OOV) words (dashed lines).

that this WER is not too bad for effective computer-assisted transcription, though for sure there is room for significant improvement. Note that, as can be observed from the OOV curves, many errors are caused by the occurrence of out-of-vocabulary words.

5. Conclusions and future work

A new handwritten text database, RODRIGO, has been presented to facilitate empirical comparison of different approaches to text line extraction and off-line handwriting recognition. RODRIGO is completely written in old Castilian (Spanish) by a single author and comparable in size to standard databases. Some preliminary empirical results have been also reported, using standard techniques and tools for preprocessing, feature extraction, HMM-based image modeling, and language modeling. Although we think that there is room for significant improvements, the word error rates obtained are already acceptable for effective computer-assisted transcription. For future work, we plan to also provide annotated data in accordance with the guidelines for Electronic Text Encoding and Interchange of the Text Encoding Initiative Consortium.

6. References

R. Bertolami and H. Bunke. 2008. Hidden Markov model-based ensemble methods for offline handwritten text line recognition. *Pattern Recognition*, 41:3452–3460.

L. Likforman-Sulem, A. Zahour, and B. Taconet. 2007. Text line segmentation of historical documents: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 9:123–138.

U. V. Marti and H. Bunke. 2002. The IAM-database: an English sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 5:39–46.

A. Millares and J. M. Ruiz. 1983. *Tratado de paleografía española*, volume 1. Espasa-Calpe, 3rd edition.

D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. Ramos, and A. Juan. 2009. The GERMANA database. In *Proccedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2010)*, pages 301–305, Barcelona (Spain).

O. Ramos, N. Serrano, and A. Juan. 2010. Interactive-predictive detection of handwritten text blocks. In *Proc. of the 17th Document Recognition and Retrieval (DRR 2010)*, San Jose, CA (USA).

2006. The RODRIGO database: digitized data. `bvpb.mcu.es`.

2010. The RODRIGO database: annotated data. `prhlt.iti.es/rodrigo.php`.

N. Serrano, A. Sanchis, and A. Juan. 2010. Balancing error and supervision effort in interactive-predictive handwritten text recognition. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI 2010)*, pages 373–376, Hong Kong (China), June.

T. Su, T. Zhang, and D. Guan. 2007. Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text. *International Journal on Document Analysis and Recognition*, 10:27–38.