2009 10th International Conference on Document Analysis and Recognition

The GERMANA database[*]

D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. Ramos Terrades, A. Juan
DSIC/ITI, Universitat Politècnica de València
Camí de Vera, s/n, 46022 València, SPAIN
{dperez,lionel,nserrano,francas,oriolrt,ajuan}@iti.upv.es

Abstract

A new handwritten text database, GERMANA, is presented to facilitate empirical comparison of different approaches to text line extraction and off-line handwriting recognition. GERMANA is the result of digitising 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. Due to its sequential book structure, it is also well-suited for realistic assessment of interactive handwriting recognition systems. To provide baseline results for reference in future studies, empirical results are also reported, using standard techniques and tools for preprocessing, feature extraction, HMM-based image modelling, and language modelling.*

keywords: *handwriting recognition, datasets, corpus, linguistic knowledge, historical documents*

1 Introduction

There are huge historical document collections residing in libraries, museums and archives that are currently being digitised for preservation purposes and to make them available worldwide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 digitised documents, but to annotate them with their real informative content and, in particular, with text transcriptions. Unfortunately, extraction of text lines and handwriting recognition are still open research problems [5, 4].

In this paper, we present a handwritten text database, GERMANA, to facilitate empirical comparison of different approaches to text line extraction and off-line handwriting recognition. GERMANA is the result of digitising and annotating a 764-page Spanish manuscript entitled *"Noticias y documentos relativos a Doña Germana de Foix, última Reina de Aragón"* and written in 1891 by Vicent Salvador, the Cruïlles' marquis. It has approximately $21K$ text lines manually marked and transcribed by palaeography experts.

GERMANA is not a particularly difficult task for several reasons. First, it is a single-author book on a limited-domain topic: the life of *Germana de Foix* (1488-1538), niece of King Louis XII of France and second wife of Ferdinand the Catholic of Aragon. Also, the original manuscript was well-preserved and most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines. Moreover, the manuscript comprises about $217K$ running words from a vocabulary of $30K$ words which, apparently, is a reasonable amount of data for single-author handwriting and language modelling.

It goes without saying that text line extraction and off-line handwriting recognition on GERMANA is not, by contrast, particularly easy.

[*]Work supported by the EC (FEDER/FSE) and the Spanish MCE/MICINN under the MIPRCV "Consolider Ingenio 2010" programme (CSD2007-00018), the iTransDoc project (TIN2006-15694-CO2-01), the Juan de la Cierva programme, and the FPU scholarship AP2007-02867.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.10

IEEE computer society

GERMANA has typical characteristics of historical documents that make things difficult: spots, writing from the verso appearing on the recto, unusual characters and words, etc. Also, the manuscript includes many notes and appended documents that are written in languages different from Spanish, namely Catalan, French and Latin.

All in all, we think that GERMANA entails an appropriate trade-offbetween task complexity and amount of data. 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 [6, 7]. Due to its sequential book structure, it is also well-suited for realistic assessment of *interactive* handwriting recognition systems [8]. Moreover, it can be used as well to test approaches for language identification and adaption from single-author handwriting.

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 recogniser. Finally, conclusions and future work are discussed in Section 5.

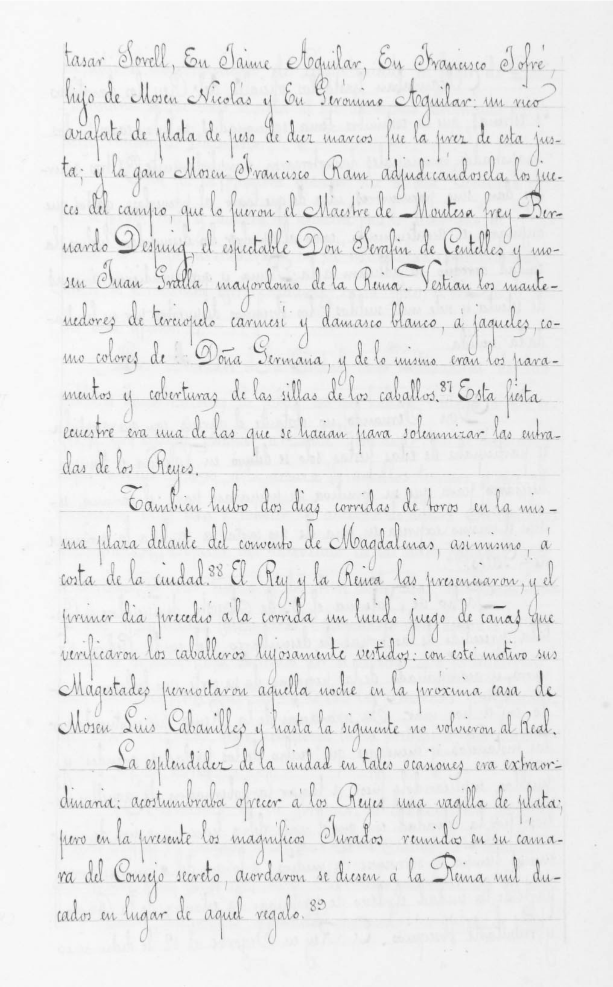
2 The manuscript

As said in the introduction, GERMANA is the result of digitising and annotating a Spanish manuscript from 1891 on the life of Germana de Foix. The original manuscript is preserved in the Nicolau Primitiu Collection at the Valencian Library [1]. It is a 764-page bound volume which, according to its index on page 728, is divided into 17 sections.

For simplicity, we will distinguish only 7 parts of the manuscript:

1. *Front matter (pp* 1–6*):* a half title, a title and a portrait of *Doña Germana de Foix*.
2. *The chapters (pp* 7–180*):* 174 pages divided into 6 chapters, each one devoted to a distinct period in the life of Germana.
3. *Notes (pp* 181–282*):* 290 numbered notes referenced in the chapters.
4. *Biography notes (pp* 283–302*)* of 8 relevant persons mentioned in the second part.
5. *Documents (pp* 303–540*):* handwritten copies of 71 historical documents related to the life of Germana.
6. *Illustrations (pp* 541–716*):* 4 documents with their own notes appended at the end.
7. *Back matter (pp* 717–764*):* various indices and images.

Most pages only contain handwritten text aligned to horizontal rules in a simple template of either 24 (pp 1–180 and 729–764) or 32 (pp 181–728) lines. As an example, the page 67 is shown in Figure 1. Note that the handwriting is easily readable and tightly aligned to horizontal rules.



tasar Sorell, En Jaime Aguilar, En Francisco Jofré, hijo de Mosen Nicolás y En Geronimo Aguilar: un rico arafate de plata de peso de diez marcos fue la prez de esta justa; y la ganó Mosen Francisco Ram, adjudicandosela los jueces del campo, que lo fueron el Maestre de Montesa frey Bernardo Despuig, el espectable Don Serafin de Centelles y mosen Juan Gralla mayordomo de la Reina. Vestian los mantenedores de terciopelo carmesí y damasco blanco, á jaqueles, como colores de Doña Germana, y de lo mismo eran los paramentos y coberturas de las sillas de los caballos.[87] Esta fiesta ecuestre era una de las que se hacian para solemnizar las entradas de los Reyes.

Tambien hubo dos dias corridas de toros en la misma plaza delante del convento de Magdalenas, asimismo, á costa de la ciudad.[88] El Rey y la Reina las presenciaron, y el primer dia precedió á la corrida un lucido juego de cañas que verificaron los caballeros lujosamente vestidos: con esté motivo sus Magestades pernoctaron aquella noche en la proxima casa de Mosen Luis Cabanilles y hasta la siguiente no volvieron al Real.

La esplendidez de la ciudad en tales ocasiones era extraordinaria: acostumbraba ofrecer á los Reyes una vagilla de plata; pero en la presente los magnificos Jurados reunidos en su camara del Consejo secreto, acordaron se diesen á la Reina mil ducados en lugar de aquel regalo.[89]

Figure 1. Page 67 of GERMANA.

The manuscript is solely written in Spanish up to page 180. After this page, however, the reader can also find text in Catalan, French, Latin and, to a lesser extent, German and Italian. In the third part, there are 33 notes (mostly) written in Catalan (4, 47, 50, 73, 78, 79, 81, 82, 84, 85, 87-91, 94-96, 134, 177, 194, 205, 209, 214, 227, 229, 236, 238, 261, 266-268 and 270); 18 in French (1, 2, 15, 22, 23, 25, 29, 44-46, 71, 109, 110, 119, 155, 170, 257 and 280); and 1 in German (180). Also, there are 24 documents in the fifth part that are written in Catalan (7, 8, 27, 29, 31-33, 36-40, 44, 48-54, 59, 64, 68 and 69); 10 in Latin (2, 4-6, 12, 24, 34, 42, 43, 70); 1 in French (7); 1 in German (25); and 1 in Italian (65). Biography notes and Illustrations are primarily written in Spanish, though there is also some content in Catalan (a short excerpt of 13 lines starting at the last line on page 300; notes 39, 47 and 61 of illustration C; and note 17 of illustration D).

The interested reader is referred to [3] for a deep study of the manuscript from a historian's point of view.

3 The database

The manuscript was carefully scanned by experts from the Valencian Library at 300dpi in true colours. 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 for the purpose of annotating ground-truth.

Ground-truth annotation of GERMANA consisted of two parts. On the one hand, all text blocks were marked 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 *GNU Image Manipulation Program (GIMP)* [2] and certain GIMP *plug-ins* we developed specifically for block and line annotation of GERMANA. All blocks and baselines detected automatically were also manually supervised, and corrected when needed.

On the other hand, the whole manuscript was transcribed line by line, by palaeography experts. The transcription process did not start from scratch, but from a partial transcription produced by experts from the Valencian Library during 2002. This partial transcription covered most of the manuscript (76%), but it was not directly applicable to handwriting research, mainly because it did not include original page and line breaks. Therefore, to produce the final transcription, this partial version was first reviewed and then completed. This was done more recently, during 2007. It was done again by palaeography experts, in accordance with the following transcription rules:

- Page and line breaks are copied exactly.
- Blank space is only used to separate words.
- 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 subindices and superindices, 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, D$^{\mathtt{a}}$. is transcribed as `D^{a}.[Doña]`.

Also, to facilitate language-dependent processing of the manuscript, each transcribed line was manually labelled in accordance with its dominant language. The total time required for a single expert to manually transcribe the whole manuscript was estimated as 232 hours; that is, approximately 30 minutes per page on average.

Table 1 contains some basic statistics drawn from our GERMANA transcription. These statistics were computed after applying the following preprocessing steps:

1. Substitution of abbreviations by their corresponding words.
2. Concatenation of hyphenated words at line ends with their remainders.
3. Isolation of punctuation signs.

				Lexicon		Char
Lang.	Pages	Lines	Words	Size	Sing.	set
			(K)	(K)	(%)	
Spanish	595	16599	176.8	19.9	55.6	111
Catalan	87	2417	26.9	4.6	63.2	86
Latin	29	951	8.3	3.4	69.2	87
French	8	266	3.0	1.1	71.1	82
German	8	228	1.5	0.6	52.7	71
Italian	2	68	0.8	0.3	67.3	59
None	35	0	0.0	0.0	0.0	0
All	764	20529	217.2	27.1	57.4	115

Table 1. Basic statistics of GERMANA (Sing=Singletons, words occurring only once).

Note that the Spanish part of GERMANA comprises about $17K$ text lines and $177K$ running words from a lexicon of $20K$ words, which is comparable in size to standard databases such as IAM [6, 7]. It is also worth noting that 56% of the words only occur once (singletons). Regarding the other, non-Spanish parts, it is clear that they are not large enough to reliably estimate independent models for them (c.f. HMMs and n-gram language models). Instead, it would be very interesting to see how models trained with different data can be adapted to them. In particular, character HMMs trained with the Spanish part might be very well reused without significant changes.

The database is available at the PRHLT website (`prhlt.iti.es`) for non-commercial research. Also, an independent, printed transcription of the manuscript can be found in [3] though, as it was not intended for handwriting research, it was reformatted for better readability.

4 Experiments

As discussed in the introduction, GERMANA may be used either, to test text line extraction methods, or to evaluate off-line handwritten text recognition techniques. 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; that is, HMM-based text image modelling and n-gram language modelling [8].

Due to its sequential book structure, the very basic task on GERMANA 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 to (re-)train the automatic transcription system. This should help in improving the system accuracy, at least in the transcription of the first GERMANA pages. Fortunately, the first two parts of GERMANA are solely written in Spanish and thus, at least, the lack of training data is not combined with multilingual input.

Taking into account the above discussion, we decided to only try GERMANA transcription of the first two parts, up to page 180. Starting from page 3, we divided GERMANA into 9 consecutive blocks of 20 pages each ($3-22$, $23-42$, $\dots$, $163-180$). Then, from block 2 to block 9, each block was automatically transcribed by the system trained with all preceding blocks. As indicated above, we used standard techniques and tools for preprocessing, feature extraction, HMM-based image modelling, and language modelling [8]. The results are shown in Figure 2, in terms of word error rate (WER) per block.

As expected, the WER decreases as the amount of training data increases. In particular, the system achieves around 37% of WER for the last two blocks, which is not too bad for effective computer-assisted transcription. Although we think that there is room for significant improvements, it must be noted that most errors are caused by the occurrence of out-of-vocabulary (OOV) words. This can be also observed in Figure 2, where a curve is plotted showing the part of the WER due to the occurrence of such words. Note that, in relative terms, this part is of increasing importance. For instance, while OOV words account for 54% of the errors in the first transcribed block, this figure increases to 64% in the last block. Moreover, it can increase even more in the remaining parts

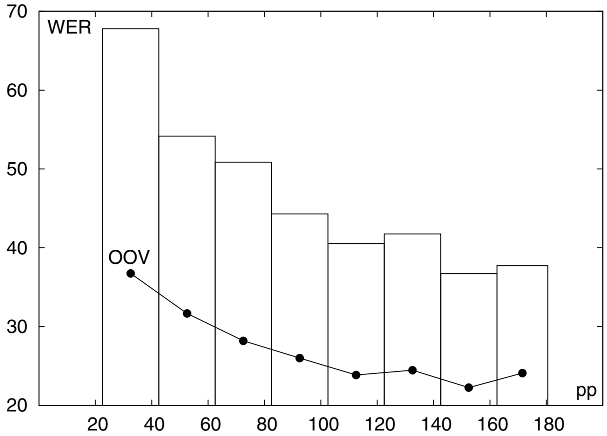


Figure 2. Transcription Word Error Rate (WER) on GERMANA as a function of the block of pages transcribed (pp). For each block, the transcription system is trained with all the pages in preceding blocks. Also shown is the part of the WER due to the occurrence of out-of-vocabulary (OOV) words.

of GERMANA due to their multilingual nature.

5 Conclusions and future work

A new handwritten text database, GERMANA, has been presented to facilitate empirical comparison of different approaches to text line extraction and off-line handwriting recognition. To our knowledge, it is the first publicly available database for handwriting research, mostly written in Spanish 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 modelling, and language modelling. Although we think that there is room for significant improvements, the word error rates obtained are already acceptable for effective computer-assisted transcription.

We are now completing the preliminary experiments reported here, that is, the complete GERMANA transcription, which involves language identification and adaptation due to the multilingual nature of GERMANA.

References

[1] Biblioteca Valenciana. http://bv.gva.es/.

[2] GNU Image Manipulation Program (GIMP). http://www.gimp.org/.

[3] E. Belenguer, editor. *Germana de Foix, última reina de Aragón.* Univ. de València, 2007.

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

[5] L. Likforman-Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. *Int. J. of Doc. Analysis and Recognition*, 9:123–138, 2007.

[6] Marti and H. Bunke. A full english sentence database for off-line handwriting recognition. In *Proc. of ICDAR 1999*, pages 705–708, 1999.

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

[8] A. H. Toselli, V.Romero, L. Rodríguez, and E. Vidal. Computer Assisted Transcription of Handwritten Text. In *Proc. of ICDAR 2007*, pages 944–948, 2007.