The ESPOSALLES Database: An Ancient Marriage License Corpus for Off-line Handwriting Recognition

Verónica Romero[a,*], Alicia Fornés[b], Nicolás Serrano[a], Joan Andreu Sánchez[a], Alejandro H. Toselli[a], Volkmar Frinken[b], Enrique Vidal[a], Josep Lladós[b]

[a] *ITI-Universitat Politècnica de València, València, Spain*
[b] *CVC-Universitat Autònoma de Barcelona, Barcelona, Spain*

Abstract

Historical records of daily activities provide intriguing insights into the life of our ancestors, useful for demography studies and genealogical research. Automatic processing of historical documents, however, has mostly been focused on single works of literature and less on social records, which tend to have a distinct layout, structure, and vocabulary. Such information is usually collected by expert demographers that devote a lot of time to manually transcribe them. To support research in automatic handwriting recognition for historical documents containing social records, this paper presents a new database compiled from a marriage license books collection. Marriage license books are documents that were used for centuries by ecclesiastical institutions to register marriage licenses. Books from this collection are handwritten and span nearly half a millennium until the beginning of the 20th century. In addition, a study is presented about the capability of state-of-the-art handwritten text recognition systems, when applied to the presented database. Baseline results are reported for reference in future studies.

Keywords: Handwritten Text Recognition, Marriage Register Books, Hidden Markov Models, BLSTM, Neural Networks.

*Tel: +34 96 387 7253, Fax: +34 96 387 7239. e-mail: vromero@dsic.upv.es

1. Introduction

In the last years, large amounts of handwritten ancient documents residing in libraries, museums, and archives have been digitalized and made available to the general public. Many of these ancient documents have an outstanding cultural value in subjects as diverse as literature, botanic, mathematics, medicine, or religion, to name a few. However, there are still large collections of a different type of historic documents, containing records of quotidian activities. A vast majority of such records have not been digitalized and contain only limited information when considered individually, but provide an intriguing look into the historic life when considered as a complete collection and in the context of their time. Examples of these kind of documents are birth, marriage, and death records[1], military draft records[2], court records[3,4], medical forms, border crossing records, municipal census records, and property registers[5].

Each of these collections in itself is a valuable source of information in historical research, family history and genealogical research. Linking several collections together, large historical social networks with information about relationships (ancestors, couples, neighbors, etc.) and their related information (dates, places, occupation, medical and physical condition, literacy, etc.) can be constructed. Among the most popular historical social networks are the FamilySearch's genealogical database[6], the Mormon Pioneer Overland Travel database[7], the Mormon Migration database[8], and the Historic Journals website[9].

Although the structure of such information uses to be stable across cities and countries, manual inference of historical social networks requires significant time and effort spent for gathering and cross-referencing many different data sources. Therefore, some research has been focused on automatic link-

[1] http://arxiu.historic.arquebisbattarragona.cat/
[2] http://www.ancestry.com/
[3] http://stlcourtrecords.wustl.edu/index.php
[4] http://www.mcu.es/archivos/MC/ACV/index.html
[5] http://pares.mcu.es/Catastro/
[6] http://familysearch.org
[7] http://www.mormontrail.lds.org
[8] http://lib.byu.edu/mormonmigration
[9] http://journals.byu.edu

age and knowledge discovery [1, 2, 3], based on already transcribed historical documents. However, research concerning automatic transcription has largely focused on single volumes containing relevant masterpieces, while a lack of approaches for the transcription of large document collections can be observed. Instead, these transcriptions are usually carried out by expert paleographers[10]. In recent years, crowd-sourcing techniques are also gaining popularity where large amounts of volunteers help in creating a transcription[11]. citeSaund2009. For example, in [4] an annotation platform for archived handwritten documents presented, where the required annotations are produced both automatically and collectively with the help of the readers. Another example is the DEBORA project [5], which aims at developing a remote and collaborative platform for accessing digitized Renaissance books.

Yet, even with help of crowd-sourcing tools, the manual transcription and annotation of large amount of documents still requires a lot of effort. Available OCR technologies are not applicable to historic documents, due to character segmentation ambiguities and the cursive writing style. Therefore, segmentation-free continuous text recognition of entire lines or sentences is required. This technology is generally referred to as *"off-line Handwritten Text Recognition"* (HTR) [6]. Several approaches have been proposed in the literature for HTR that resemble the noisy channel approach currently used in Automatic Speech Recognition. Consequently, HTR systems are based on hidden Markov models (HMM) [6, 7], recurrent neural networks [8], or hybrid systems using HMM and neural networks [9]. These systems have proven to be useful in a restricted setting for simple tasks such as reading postal addressed or bank check legal amounts. In the context of historic documents, however, their performance decreases dramatically, since paper degradation including show-through and bleed-trough, and a lack of a standard notation between different time periods renders the task quite challenging.

In this scenario, the key to a reliable recognition is contextual information, which is specially useful for structured entries encountered in social and demographic documents. Thus, automatic systems could benefit from context information not only in the handwriting recognition step, but also in the posterior semantic information extraction and knowledge discovery. HTR technology relies on statistical learning methods which generally require sig-

[10] `http://admyte.com/e_historia.htm`

[11] `http://www.ucl.ac.uk/transcribe-bentham/`

nificant amounts of transcribed text images. While several publicly available databases of historic handwritten documents exist, such as the George Washington dataset [10], the Parzival database [11], the Saint Gall database [12], the RODRIGO database [13], or the GERMANA database [14], none of these contain structured texts to investigate such *context-benefit* systems

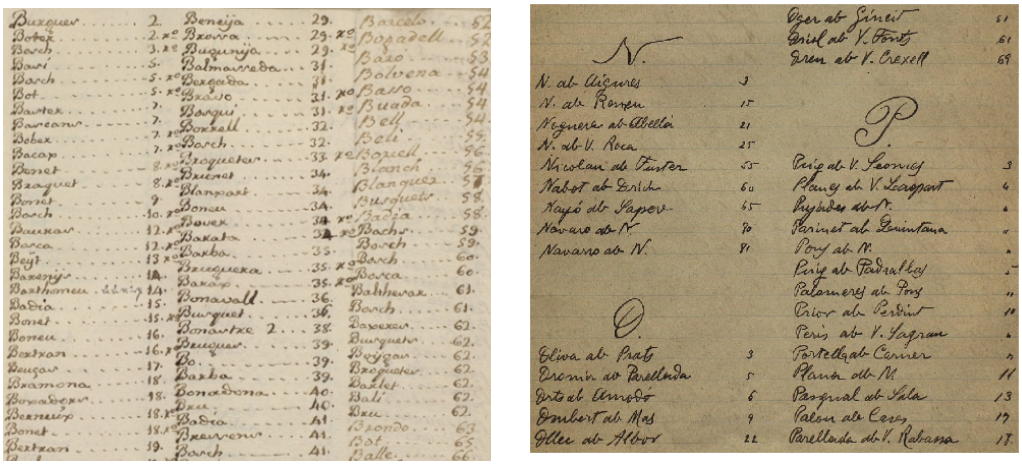
1.1. Contributions

To change this situation, the first contribution of this paper is a publicly available database, compiled from a collection of Spanish marriage license books. The collection consists of handwritten books, spanning several centuries, used to register marriages licenses in the corresponding parishes. Selected pages from different centuries are shown in Fig. 1. These demographic documents have already been proven to be useful for genealogical research and population investigation, which renders their complete transcription an interesting and relevant problem [15]. The contained data can be useful for research into population estimates, marriage dynamics, cycles, and indirect estimations for fertility, migration and survival, as well as socio-economic studies related to social homogamy, intra-generational social mobility, and inter-generational transmission and sibling differentials in social and occupational positions. Each book contains a list of individual marriage license records, analogous to an accounting book. Also, most of the books have an index (see Fig. 2) at the beginning that was used to locate each license in the book.

The access to the semantic content of these documents requires efficient solutions, not only for handwritten recognition, but also for areas such document image analysis, keyword spotting, and automatic extraction of semantic information. The special features of the database presented here allows to carry out research all of these areas. The structured layout and the proximity of adjacent lines provide interesting challenges for layout analysis and line segmentation techniques. With isolated text lines, normalization and binarization, followed by handwriting recognition can be investigated. It is interesting to note that despite the structure of the text, automatic transcription of these documents is quite difficult due to the distinct vocabulary, which is composed mainly of proper names. In Section 3, detailed information about the word frequencies are given. Finally, for any further use of the data, the semantic information contained in the entries could be extracted as a set of keywords and relationships between them. However, this requires

special attention as well, even with a perfect transcription, since the structure of the text is not fixed.

Figure 1: Examples of marriage records from different centuries.



(a) (b)

Figure 2: Examples of indexes from different volumes. a) indexes with only the husband's surname. b) indexes with the husband's and the wife's surnames.

The first version of the database, presented here, consists of a relatively

small part of the whole collection outlined above. It is split into two parts, the records and the indexes. More specifically, the current database version encompasses 173 page images containing 1 747 marriage licenses and 29 images of indexes, containing 1 563 index entries. Overall, the contributed database contains more than 7 000 lines and more than 65 000 running words. All these images are annotated with the corresponding paleographic transcriptions along with other relevant ground-truth information.

Along with the database, we introduce research topics related for this sort of documents. Thus, as a second contribution of this paper, we carry out a study concerning the capability of applying current state-of-the-art handwriting recognition methods to the presented database. For this task, we apply and evaluate two HTR approaches for all textual elements found in the books and present baseline results for future comparison. This is unlike existing work on historical social documents, such as the recognition of French census data [16] or Brazilian death certificates [17], where only some parts of the documents are considered.

The rest of the paper is structured as follows. Section 2 describes the Marriage License Books collection and their main difficulties. The first version of the compiled database is presented in Section 3. Section 4 contains general description of the state-of-the-art systems used in the experiments and the experimental evaluation is given in Section 5. Finally, Section 6 summarizes the work presented and shows directions for future research.

2. The Marriage License Books collection

The Marriage License Books collection (called *Llibres d'Esposalles*), conserved at the Archives of the Cathedral of Barcelona, is composed of 291 books with information of approximately 600 000 marriage licenses given in 250 parishes between 1451 and 1905. Fig. 1 shows three pages from different centuries. One can clearly see the continuity of the layout during the centuries and the significant differences in the handwriting styles.

Each page is divided horizontally into three blocks, the *husband surname's block* (left), the *main block* (middle), and the *fee block* (right). Vertically, the page is divided into individual license records. The marriage license (see Fig. 3) contains information about the husband's and wife's names, the husband's occupation, the husband's and wife's former marital status, and the socioeconomic position given by the amount of the fee. In some cases, even

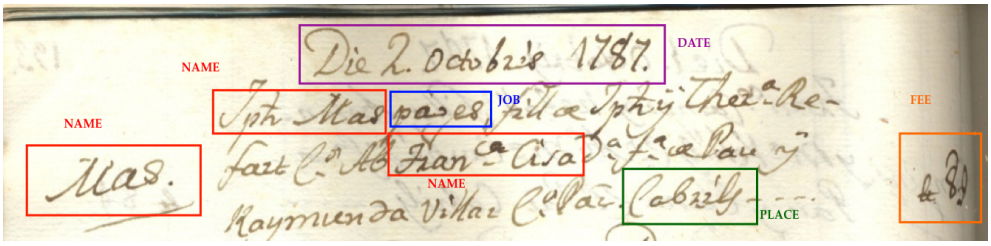


Figure 3: Marriage record where the family names, the place, date and fee are highlighted.

further information is given, such as the fathers' names and occupations, information about a deceased parent, place of residence, or geographical origin.

In addition to marriage licenses, most books include an index with the husband's surname(s) and the page number where the license entry can be found (see Fig.2.a)). In some cases, the wife's surname is also included (see Fig. 2.b)), separated by the word "ab" ("with" in old Catalan). In case one surname is unknown, it is substituted by "N.". In addition, a "V." preceding a wife's surname identifying her as a widow and that she was using her former husband's surname. In this list, indexes are sorted alphabetically according to the first letter of the man's surname. Entries starting with the same letter, however, are not alphabetically sorted. Page numbers tend to be in increasing order for entries starting with the same letter, although not always.

2.1. Difficulties of the marriage license books

The recognition of the handwritten historical marriage license books have the following difficulties:

- Degradation. The typical paper degradation problems encountered in old documents, such as significant background variation, uneven illumination, show-through, bleed-through, smear or dark spots, require specialized image-cleaning and enhancement algorithms [18].
- Handwriting styles. Long-running handwritten documents require robust recognition systems that deal with the high variability of script and writer styles within the different time periods (see Fig. 1).

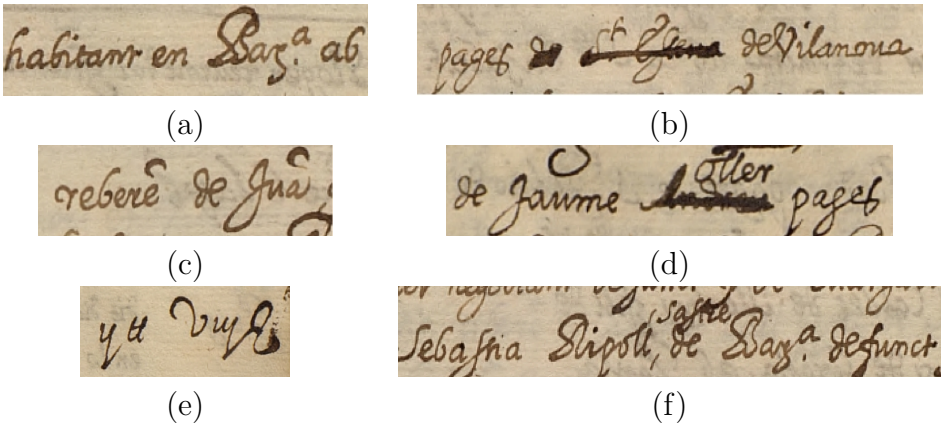
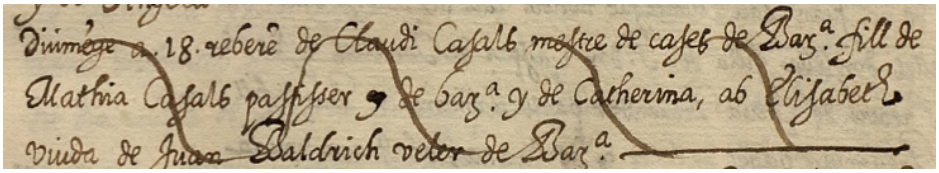


Figure 4: Examples of challenges in the notation. a) Abbreviation of *Barcelona* using *Bar.*^{a}, b) removed words, c) abbreviation of *reberem* and *Juan* using the symbol $\sim$, d) removed word and insertion of the upper word *Oller*, e) Roman numerical notation, corresponding to "*ii ll viii s*", f) the word *sastre* is written between two text lines.

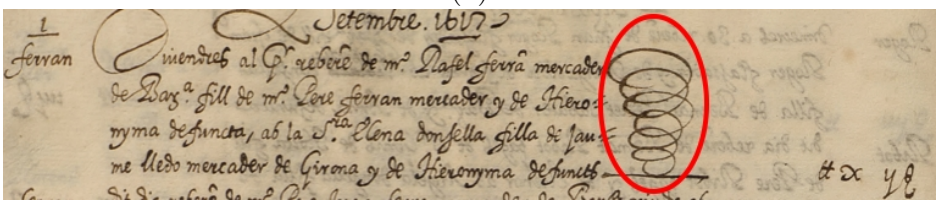
- Layout and segmentation. Documents contain voided entries, lines at the end of the marriage license, drawings (see Fig. 5), and words between text lines (see Fig. 4.d,f). In addition, the entries in the index are not aligned horizontally, as depicted in Fig. 2.a). Consequently, different columns from the same page may have a different number of lines.

- Notation and Lexicon. Documents do not follow a strict standard notation, containing a variety of special symbols and other recognition challenges (see Fig. 4). In addition, the dictionary of names and surnames is open, and new words and abbreviations are constantly appearing.

- Syntactical structure. Although the overall structure of the documents is quite stable, the syntactical structure of the marriage records is not completely known. In fact, the structure is not constant, and many variations can be found, even in the same book and writer.



(a)



(b)

Figure 5: Examples of graphical elements in the marriage records. a) A record that has been voided. b) A record with a special drawing symbol and a final horizontally line.

3. The ESPOSALLES database

In this section we present the first version of the ESPOSALLES database. It is a freely available handwritten text database[12] compiled from the Marriage License Books collection introduced in the previous section. The aim of this database is to facilitate empirical comparison of different approaches to off-line handwriting recognition applied to ancient social documents. This database can be also used to study approaches to automatic extraction of semantic content in order to generate genealogical trees, study the evolution of a family name, and the evolution of population size, among others.

The database consists of marriage license manuscripts digitalized by experts at 300dpi in true colors and saved in TIFF format. It is divided into two parts. The first one, called LICENSES, is compiled from one of the marriage license books. The second one, INDEX, is composed of pages from two of the indexes that most of the marriage license books have at the beginning. All text blocks, lines, and transcriptions have been manually annotated, resulting in a dataset that can be used to train and test handwriting recognition systems, providing a well-defined task for future studies.

The following sections describe the main characteristics of the two parts

[12] `www.cvc.uab.es/5cofm/groundtruth`

as well as the provided ground truth.

3.1. LICENSES

The first part, LICENSES, has been compiled from a single book of the marriage license books collection. The book was written between 1617 and 1619 by a single writer in old Catalan. It contains 1 747 licenses on 173 pages. Fig. 6 shows two pages of the LICENSES volume. Further characteristic details of LICENSES that can be clearly appreciated in Fig. 6 are:

- The first row of each page contains information about the month and the year and a last row with the sum of the marriage fees.
- Blank spaces between words are often omitted, usually when the word "de" is followed by a proper name; e.g., the words "de Pere" are written as "dePere".
- When the text of the last license line does not arrive to the end of the line the author introduced a straight line.
- The day of the license appears on the left block only if it is different to the previous license day.

The LICENSES database is endowed with two different types of annotations. Firstly, a layout analysis of each page has been done to indicate blocks, lines, and individual licenses. Secondly, the manuscript is completely transcribed by an expert paleographer. Details about the annotation process are explained in the following subsections.

3.1.1. Page layout structure

Document structuring and layout analysis is the first processing phase applied to each page image in order to decompose it into component regions and to understand their functional role and relationships. The layout analysis process is therefore performed in two different steps [19, 20]: the first one, called *page segmentation*, segments the document page image into homogeneous regions such as text blocks and text lines. In the second step, known as *logical layout*, a the extracted regions are classified wrt. their categories. Also relationships between different regions are established, in order to group, e.g., different text lines into registers.

For this database, the page layout involves the following structural components:

Figure 6: Pages 18 and 19 of LICENSES.

1. The bounding box of the main text block.
2. The coordinates of each text line inside the main text block.
3. The individual licenses together with their associated set of text lines.

A detection procedure for both the main text block and the corresponding text lines was conducted interactively in two phases using the GIDOC prototype [21]. Firstly, a preliminary detection was performed by a fully automatic process using standard preprocessing techniques based on horizontal and vertical projection profiles and the run-length smoothing algorithm (RLSA) [22]. Finally, the detected locations for each block and lines were verified by a human expert and corrected if necessary.

The location coordinates of a text line images reflect only their respective baselines instead of a pixel-accurate ground-truth needed to evaluate text line separation methods. Yet, this is enough for a rough detection and a basic extraction of the text lines, which, in turn, is sufficient for standard HTR recognizers. Each detected line was enclosed into a rectangle composed by 50 pixels above the baseline and 25 pixels under the baseline.

After separating the individual lines, the logical layout analysis then combined consecutive text lines into complete license records. A single record is

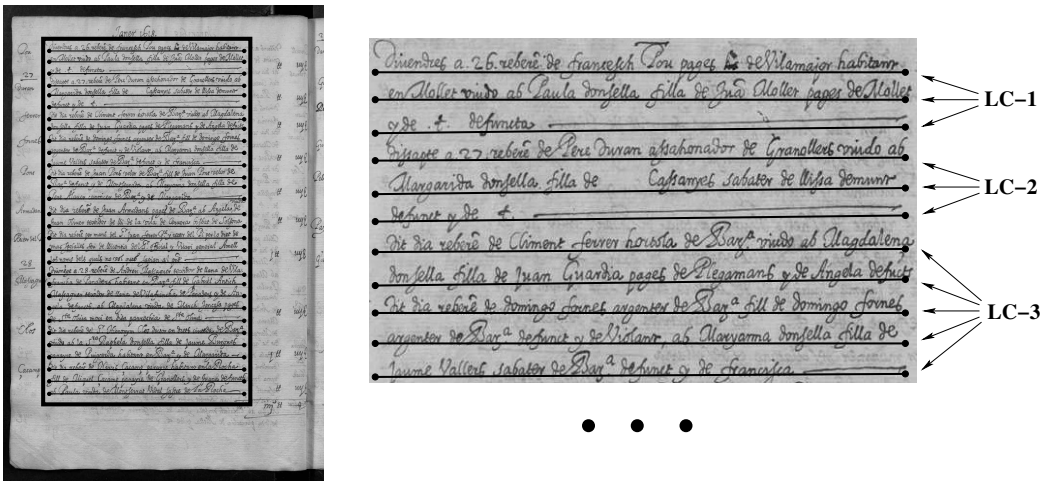


Figure 7: Visual example of annotated layout elements: main text block rectangle region, defined by its left-upper- and right-lower-corner coordinates (left), and the baseline coordinates of each text line along with the license label (LC-#) to which it belongs to.

composed of usually 3-5 lines. Finally, the license layout detection process was carried out in two steps using the GIDOC prototype. First, with help of the fee block an automatic process inferred a initial assignment of lines into licenses. This was possible since the positions of each handwritten fee is always on the height of the last line of its corresponding license. Afterwards, the assignments were manually verified and corrected. The time it takes a human expert to correct the detection and assignment errors is around 10 minutes per page.

Fig. 7 shows a visual example of a page with all considered layout elements. The bounding box arounf the main text block, defined by its left-upper- and right-lower-corner coordinates, as well as the baseline coordinates of all text lines along with the corresponding license identification (LC-#).

3.1.2. Transcription

The main block of the whole manuscript was transcribed line by line by an expert paleographer. This transcription was carried out in order to obtain a character accurate transcription. That is, the words are transcribed in the same way as they appear on the text, without correcting orthographic mistakes or writing out abbreviations. The time needed by the palaeographer to fully transcribe each license is around 3.5 minutes. In the Appendix, the rules followed in the transcription process are described in detail.

Table 1: Basic statistics of LICENSES text transcriptions.

Number of:	Total
Pages	173
Licenses	1 747
Lines	5 447
Running words	60 777
Lexicon size	3 465
Running characters	328 229
Character set size	85

The complete annotation of LICENSES is freely available. It contains around $60k$ running words in over $5k$ lines, split up into nearly $2k$ licenses. Table 1 summarizes the basic statistics of the LICENSES text transcriptions.

Fig. 8 shows the frequency of each different word as a function of its rank, i.e., words are sorted in decreasing order of number of samples per word. We can see that these counts approximately follow the Zipf's law [23], which states that given a corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. The word with greatest rank is *de*, with 11 366 occurrences. It is followed by several words that appear at least once in almost all licenses, such as *dia*, *doncella*, *filla* or *reberem*. Therefore, these words have a similar number of occurrences, around 1 500, which explains the plateau that can be seen at the beginning of the curve.

In Fig. 9 we can see a detailed view of the information given in Fig. 8 for low frequency words. It shows the number of word classes as a function of word frequency. From the graph we can see that about half of the different words of the corpus appear only once.

3.1.3. Partitions

The LICENSES part of the database is divided into 7 consecutive blocks of 25 pages each ($1-25, 26-50, \ldots, 150-173$), aimed at performing cross–validation experiments. Table 2 contains some basic statistics of the different partitions defined. The number of running words for each partition that do not appear in the other six partitions is shown in the out-of-vocabulary' (OOV) row.

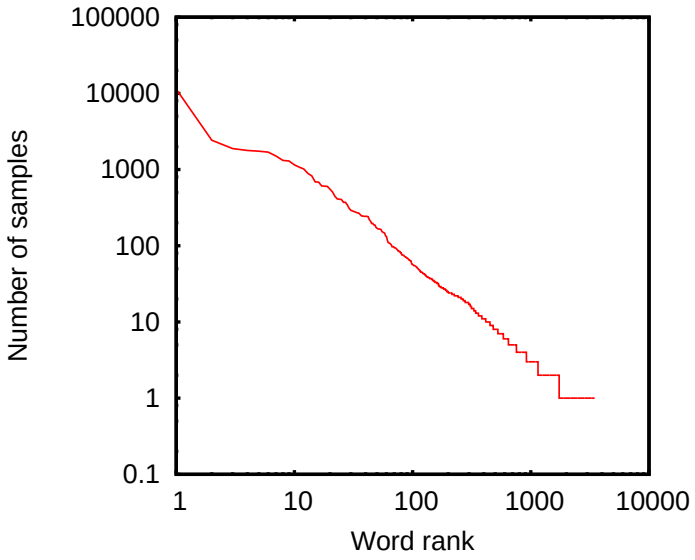


Figure 8: LICENSES Zipf's graph. Different words are sorted in decreasing rank order (horizontal axis); that is in decreasing frequency of occurrence in the corpus. The vertical axis shows the frequency (number of samples) of each different word.

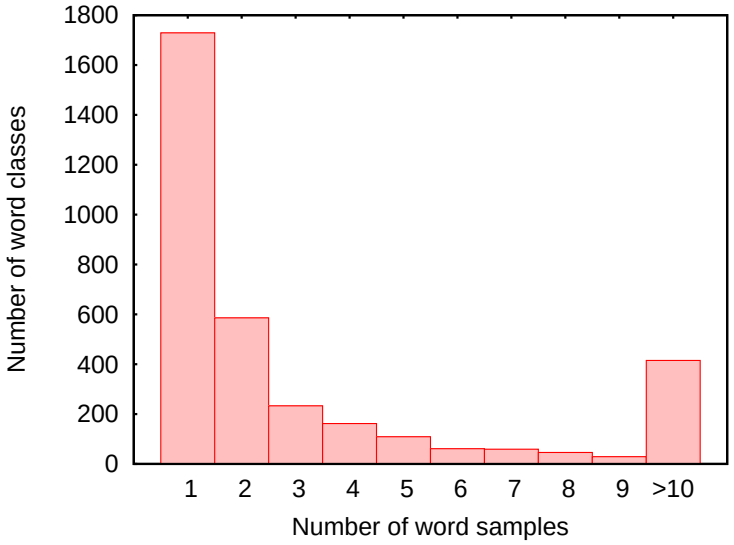


Figure 9: LICENSES word frequency histogram. The number of words classes is represented as a function of word frequency.

Table 2: Basic statistics of the different partitions for the database LICENSES

Number of:	P0	P1	P2	P3	P4	P5	P6
Pages	25	25	25	25	25	25	23
Licenses	256	246	246	249	243	255	252
Lines	827	779	786	768	771	773	743
Run. words	8 893	8 595	8 802	8 506	8 572	8 799	8 610
OOV	426	374	368	340	329	373	317
Lexicon	1 119	1 096	1 106	1 036	1 046	1 078	1 011
Characters	48 464	46 459	47 902	45 728	46 135	47 529	46 012

3.2. INDEX

The second part of the database, INDEX, is compiled from the indexes at the beginning of two volumes from the collection of marriage license books, written between 1491 and 1495 by the same writer. Fig. 10 shows a page example from the index of each volume. Note that the INDEX and LICENSES part have been deliberately chosen to be from different periods, mainly for two reasons. First, we wanted to present different writing styles. Secondly, the indexes that correspond to the LICENSES (i.e. marriage records between 1617-1619) only contain the husband's surname. More interesting data can be found in books from a different time with an enlarged lexicon where both the husband's surname and the wife's surname are given.

The INDEX part is composed of 29 text pages. Each page is divided horizontally into two columns and each column, in turn, is divided into lines, one per each marriage license, as can be seen in Fig 10. Large, single, calligraphic characters that appear in the columns represent a letter change.

As in the LICENSES part, two different annotations have been carried out. On the one hand, the different columns of each page have been detected and subsequently divided into lines. On the other hand a character-accurate transcription of the whole INDEX part was done.

3.2.1. Page layout structure

The layout analysis of INDEX is simpler than that of LICENSES, since only physical analysis of columns and lines is necessary. For each page, the GIDOC [21] prototype was used for text block layout analysis and line segmentation. Firstly, a portion of the image including all text lines to be considered was manually marked. In this selection the text columns were detected by means of simple methods based on projection profiles. Next,

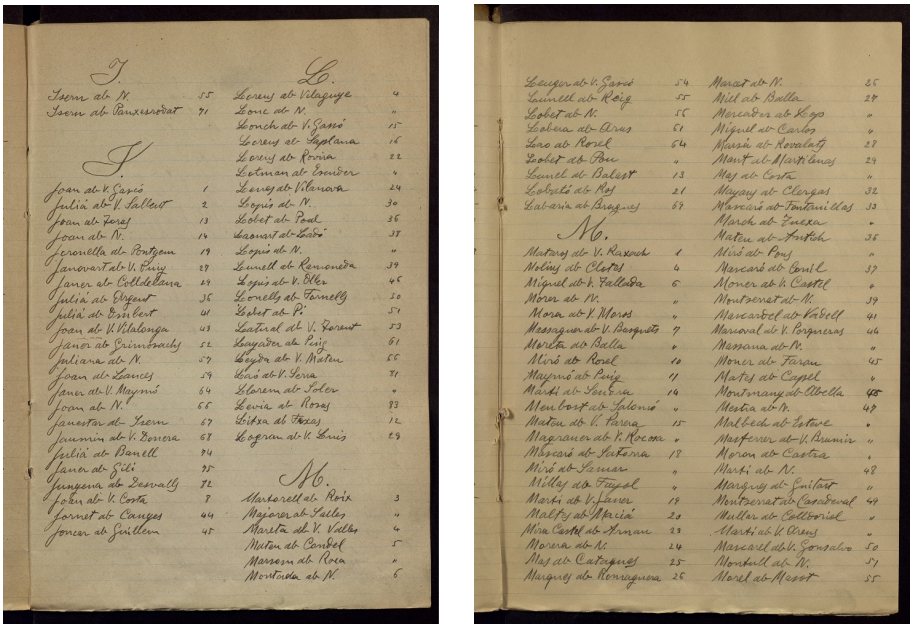


Figure 10: Pages of the INDEX database.

text lines for each text block were detected by algorithms also based on projection-profiling. Finally, all the detected blocks and lines were revised manually to correct possible errors. The time needed by a human expert to carry out this revision is around 5 minutes per page.

3.2.2. Transcription

The whole INDEX was transcribed line by line by an expert paleographer. Following the same rules than in the LICENSES partition, the words are transcribed exactly in the same way as they appear on the text and punctuation signs were copied as they appear on the manuscript. The result is a dataset of $1.5k$ text line images, containing over $6.5k$ running words from a lexicon of $1.7k$ different words. The basic statistics of the INDEX database are summarized in Table 3.

Fig. 11 shows the amount of samples available for each different word, sorted in the horizontal axis according to their rank. The graph deviates considerably from the typical Zipf's law expected for a natural language corpus [23], which clearly reflects the special nature of the INDEX database.

Fig. 12 shows the number of words classes per frequency. From the 1 725 different words, 1 202 appear only once. The majority of the running words

Table 3: Basic statistics of INDEX text transcriptions.

Number of:	Total
Pages	29
Lines	1 563
Running words	6 534
Lexicon size	1 725
Running characters	30 809
Character set size	68

of the corpus is accumulated by "*ab*", "*″*", "*V.*" and "*N.*" with 1 563, 404, 258 and 237 occurrences respectively.

3.2.3. Partitions

Four different partitions have been defined for cross-validation testing. Detailed information of the different partitions is shown in Table 4. The number of running words of each partition that do not appear in the other three partitions is shown in the OOV row (out of vocabulary).

Table 4: Basic statistics of the different partitions for the INDEX database.

Number of:	P0	P1	P2	P3
Text lines	390	391	391	391
Words	1 629	1 640	1 632	1 633
Characters	7 629	7 817	7 554	7 809
OOV	326	346	298	350

4. HTR Systems overview

In this work, we provide baseline results for reference in future studies using standard techniques and tools for HTR. Most specifically, we used two systems, the first one is based on hidden Markov models (HHM) [7] while the second one is based on artificial neural networks (ANN) [8]. Both of them use N-grams for language modeling.

Both HTR systems used in this paper follow the classical architecture composed of three main modules: document image preprocessing, line image feature extraction and model training/decoding. The following subsections describe the main modules.

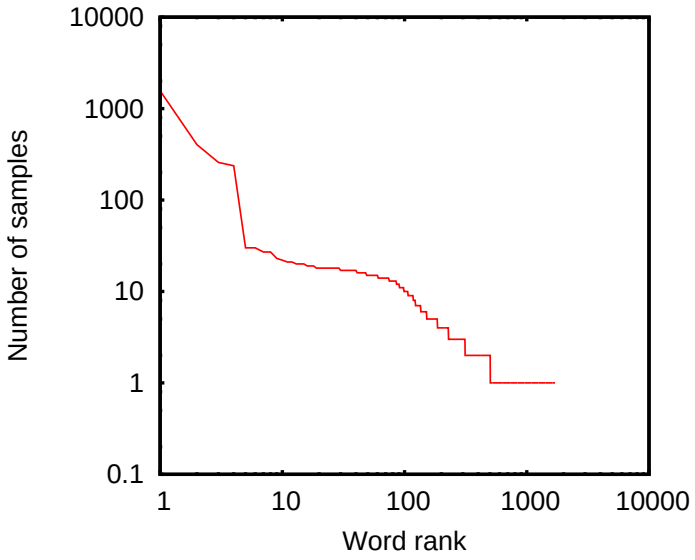


Figure 11: The INDEX Zipf's graph. This graph deviates considerably form the expected shape for a natural language data set.

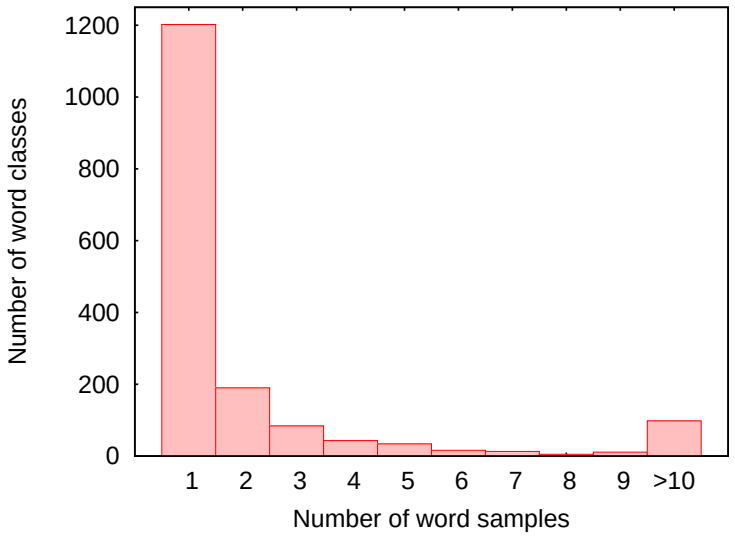


Figure 12: INDEX word frequency histogram. The number of word classes is represented as a function of word frequency.

4.1. Preprocessing

In the preprocessing module, the pages are first divided into line images as explained in the previous section. Given that it is quite common for hand-

written documents to suffer from degradation problems that decrease the legibility of the documents, appropriate filtering methods have been applied to remove noise, improve the quality of the image and to make the documents more legible. Within our framework, noise is considered as anything that is irrelevant for the text recognition. In this work, background removal and noise reduction is performed by applying a 2-dimensional median filter [24] on the entire line image and subtracting the result from the original image. To increase the foreground/background image contrast, a grey-level normalization is applied. Afterwards, the skew of each line is corrected using horizontal projections and the slant has been corrected using a method based on vertical projection profiles. Finally, the size is normalized separately for each line. A more detailed description of this preprocessing can be found in [7] and [25].

4.2. Feature extraction

Each preprocessed line image is represented as a sequence of feature vectors using a sliding window with a width of W pixels that sweeps across the text line from left to right in steps of T pixels. At each of the M positions of the sliding window, N features are extracted. In this paper we tested two different sets of features:

- **PRHLT features**
  These features are introduced in [7]. The sliding window is vertically divided into into R cells, where $R = 20$ and T is then the height of the line image divided by R. We used standard values for both parameters, R and $W = 5T$. In each cell, three features are calculated: the normalized gray level, the horizontal and vertical components of the grey level gradient. The feature vector is constructed for each window by stacking the three values of each cell ($N = 3 \cdot 20$). Hence, at the end of this process, a sequence of 60-dimensional feature vectors is obtained.

- **IAM features**
  These features correspond to the ones proposed in [6]. From each window ($W = 1$ and $T = 1$), $N = 9$ geometric features are extracted, three global and six local ones. The global features are the 0^{th}, 1^{st} and 2^{nd} moment of the distribution of black pixels within the window. The local features were the position of the top-most and that of the bottom-most black pixel, the inclination of the top and bottom contour of the word at the actual window position, the number of vertical

black/white transitions, and the average gray scale value between the top-most and bottom-most black pixel.

4.3. Modelling and decoding

Given a handwritten line image represented by a feature vector sequence, $\mathbf{x} = x_1\ x_2\ \ldots\ x_m$, the HTR problem can be formulated as the problem of finding a most likely word sequence, $\hat{\mathbf{w}} = \hat{w}_1\ \hat{w}_2\ \ldots\ \hat{w}_l$, i.e., $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \Pr(\mathbf{w} \mid \mathbf{x})$. In this paper we tested two different HTR technologies, one based on HMM and the other one based on ANN. In both cases, a bi-gram language model is used to model the concatenation of words in $\mathbf{w}$.

- **HMM-based handwriting recognition**
  The HMMs approach follows the classical inverted channel scheme [26], as presented in [7]. Using the Bayes' rule, the word sequence posterior probability, $\Pr(\mathbf{w} \mid \mathbf{x})$, is decomposed into a conditional likelihood, $\Pr(\mathbf{x} \mid \mathbf{w})$, and a language model prior, $\Pr(\mathbf{w})$ and the optimal $\hat{\mathbf{w}}$ is obtained using the Viterbi algorithm. Conditional likelihoods are modeled by means of morphological character models, concatenated into word models according to the given lexicon. Each character model is a continuous density left-to-right HMM, which models the horizontal succession of feature vectors representing this character. A Gaussian mixture model governs the emission of feature vectors in each HMM state. In the experiments, standard values of 6 HMM states per character and 64 Gaussian densities per state have been chose. These values have been proven to work well in previous handwriting recognition experiments.

- **ANN-based handwriting recognition**
  The artificial neural networks approach that was selected for the baseline experiments is the one proposed by Alex Graves et al. [8], which showed excellent results in the ICDAR handwriting recognition competition [27]. It is based on the bidirectional long short-term memory blocks (BLSTM), a recurrent neural network architecture which sophisticated memory cells for contextual information. The input features are processed by the neural network and a token passing algorithm is used to generate the output word sequence according to given lexicon and the word sequence probabilities provided by the language model.
  The experiments with the ANNs were conducted using BLSTM neural

networks with 100 LSTM memory blocks in each of the forward and backward layers. Each memory block contained one memory cell as well as an input gate, an output gate and a forget gate to control the flow of information. The tanh activation function was used for the block input and output squashing functions, while the gate activation function was the logistic sigmoid. A learning rate of $1.0 \cdot 10^{-4}$, and a momentum of 0.9 was used. These parameter settings have been found to work very well in previous handwriting recognition experiments. Similarly to the HMM approach, these are standard settings and have not been adapted to this database.

The HMMs and ANNs are trained on images of unsegmented, continuously handwritten text, transcribed into character-accurate word sequences. Finally, the concatenation of words into text lines or license records are modeled using word bi-grams, with Kneser-Ney back-off smoothing [28], estimated from the training transcriptions of the text images.

5. Baseline experiments

In order to asses the state-of-the-art HTR technology for transcribing marriage licenses books, we performed different experiments with both datasets described in Sec. 3, LICENSES and INDEX. The empirical results here reported aim to serve as reference performance benchmarks for future research.

5.1. Experimental setup and assessment measures

Experiments were carried out to test both HTR systems described in Sec. 4 (HMM and ANN), each with the two sets of features described in the same section (PRHLT and IAM). Furthermore, two slightly different recognition tasks were performed with the LICENSES data set. In the first one, called *line level*, each text line image was independently recognized. In the second, called *license level*, the feature sequences extracted from all the line images belonging to each license record are concatenated. Hence, in the license level recognition more context information is available.

In addition, two experimental conditions were considered: *Open* and *Closed Vocabulary* (OV and CV). In the OV setting, only the words seen in the training transcriptions were included in the recognition lexicon. In

CV, on the other hand, all the words which appear in the test set but were not seen in the training transcriptions (i.e., the OOV words) were added to the lexicon. Except for this difference, the language models in both cases was the same; i.e., bi-grams estimated only from the training transcriptions. Also, the morphological character models were obviously trained using only training images and transcriptions.

These two lexicon setting represent extreme cases with respect to real use of HTR systems. Clearly, only the words which appear in the lexicon of a (word-based) HTR system can ever have the chance to be output as recognition hypotheses. Correspondingly, an OV system will commit at least one error for every OOV word instance which appears in test images. Therefore, in practice, the training lexicon is often extended with missing common words obtained from some adequate vocabulary of the task considered. In this sense, the CV setting represents the best which could be done in practice, while OV corresponds to the worst situation.

CV testing simplifies reproducibility of experiments and allows better interpretations of empirical results. This kind of testing is a time-honored common practice in the field of Automatic Speech Recognition.

The quality of the automatic transcriptions obtained with the different HTR systems used in the experiments is measured by means of the word error rate (WER). It is defined as the minimum number of words that need to be substituted, deleted, or inserted to match the recognition output with the corresponding reference ground truth, divided by the total number of words in the reference transcriptions.

5.2. LICENSES results

In a set of experiments we evaluated the performance of both HTR systems (HMM and ANN) with both types of features (PRHLT and IAM) in closed and an open vocabulary settings and assuming recognition tasks both at the line and the license levels. The seven different partitions described in Section 3 are used in these experiments for cross-validation. That is, we carry out seven rounds, with each of the partitions used once as validation data and the remaining 6 partitions used as training data. The results can be seen in Table 5.

Using the PRHLT features, in the line recognition problem, the WER was 12.0% and 16.1% in the closed and open vocabulary settings, respectively. In the whole license recognition case, the WER decreased as expected, given

Table 5: Transcription Word Error Rate (WER) in LICENSES using HMM and ANN-based systems with two different kinds of features. All results are percentages.

		Lines		Licenses	
		Open Voc.	Closed Voc.	Open Voc.	Closed Voc.
HMM	PRHLT feat.	16.1	12.0	15.0	11.0
	IAM feat.	17.7	14.8	17.4	14.6
ANN	PRHLT feat.	15.1	12.0	15.9	13.1
	IAM feat.	12.7	9.6	12.1	9.0

that the language model in this case is more informative. In particular, the system achieved a WER of 11.0% with closed vocabulary and 15.0% with open vocabulary.

The same experiments were carried using the IAM features and the obtained error rates were slightly higher. For line recognition, a WER of 14.8% and 17.7% was obtained using a closed and an open vocabulary respectively, whereas for license recognition, the obtained WER was 14.6% using a closed vocabulary and 17.4% using an open vocabulary.

To evaluate the ANN-based approach, for each one of the 7 partitions we trained 5 randomly initialized neural networks. Note that we did not use a validation set to guide the training. Instead we fixed the number back-propagation iterations to 50.

Using the PRHLT features for single line recognition, the WER were 12% with closed vocabulary and 15% with open vocabulary. As far as the IAM features are concerned, the WER was 9.6% and 12.7% with closed and an open vocabulary, respectively. Very good performance for the whole license recognition task was achieved using the IAM features, with 12% and 9% WER for open and closed vocabularies, respectively.

When comparing the HMM and the ANN approaches, one can see that the PRHLT features perform similarly for both recognition tasks, single line and whole licenses. In contrast, the IAM features seem to perform better for the ANN-based approach, which constantly outperform the HMM-based approach. In fact, the 9% WER achieved for the closed vocabulary license recognition is the best result achieved among all the tests.

5.3. INDEX results

For the INDEX database, we conduced experiments with both HMM-based and ANN-based systems, again with both features sets, PRHLT and

Table 6: Transcription Word Error Rate (WER) for INDEX using HMM and ANN-based systems with two different kinds of features. All results are percentages.

		Open Vocabulary	Closed Vocabulary
HMM	PRHLT features	43.7	31.1
	IAM features	53.0	44.7
ANN	PRHLT features	70.4	70.1
	IAM features	63.4	59.8

IAM, and again in the open and closed vocabulary case. For cross-validation, we used the partitions defined in Section 3. The resulting error rates are given in Table 6.

Using the HMM-based approach, with the PRHLT features, a WER of 31.1% and 43.7% is obtained with open and closed vocabularies, respectively. In comparison, for IAM features, 44.7% and 53.0% WER is obtained with open and closed vocabularies, respectively.

Using the ANN approach, with the PRHLT features, the WER was about 70% both for open and closed vocabularies. The performance when using the IAM features was slightly better, with 59.8% and 63.4% WER with closed and open vocabularies, respectively (see Table 6). Clearly, these results are worse than the ones obtained using the HMM-approach. An explanation for the poor performance of the neural networks in this data set might be the few amount of training data, which is about ten times less for INDEX than for LICENSES. From this we conclude that the ANN with the standard parameter settings can not cope with only few training data as well as the HMM-based approaches can.

5.4. Discussion of results

It seems clear from the results that PRHLT features work better for HMM than the IAM features and vice-versa for the ANN recognizer. Although we have no clear explanation for this fact, it might be explained by the type of features involved in each features set. The Gaussian seem to work better with the continuous data of the PRHLT features than with the discrete data of some IAM features. However, the opposite occurs with the ANN recognizer.

From the results it is also clear that the number of available training data is a deterministic factor in the accuracy of the different classifiers. The ANN-approach obtains the best results in the LICENSES database. However, in the INDEX set, where only 29 pages are available, the HMMs approach

seems to perform better. It seems that hidden Markov models, although outperformed for large training sets, can learn more efficiently with scarce training data.

The special form of the datasets leaves room for specialized systems to further improvement of the recognition task. The INDEX dataset, e.g., is rather small and the word distribution does not follow a natural language syntax. Except for a few words and abbreviations, like "ab", which occurs in nearly every line, a large part of the words are surnames. These form a vocabulary that is extremely large in relation to the small size of the set. Therefore, under-trained HMM, ANN, and bi-gram models are to be expected, rendering the task quite difficult. Different architectures of the models, e.g. using semi-continuous HMMs, using an adequate vocabulary of names, or adding vocabularies from the same period, or name collections might decrease the WER. Furthermore, given the simple and regular syntactic structure of the lines, a fixed, prior grammar that restricts the search space in a more effective way than bi-grams seems to be promising as well. Such kind of simple language model has been recently used with good results in [29]. The obtained results suggest that, using prior knowledge, a higher recognition accuracy can be obtained.

In summary, we have presented baseline results for both LICENSES and INDEX databases with HMM and ANN-based approaches. In the LICENSES case, the IAM features with the neural networks perform slightly better. On the contrary, the PRHLT features with the HMMs recognizer work better in the INDEX database. This suggests that the recognizers could be improved by combining both sets of features and performing feature selection on the new set of features.

6. Conclusions

In this paper, a historic handwritten text database has been presented. The data is compiled from a marriage license book collection, opening the research in automatic handwriting recognition to a new field. While most of the existing historic handwriting databases focus on single literature work, the processing of records concerning everyday activities is largely unexplored. Yet, social records, especially when linking several databases, give a new perspective on family trees, persons, and events, which are valuable to researchers and the public alike.

Along with the database, we provide baseline results for state-of-the-art recognition approaches, namely hidden Markov models and recurrent neural networks. Surprisingly, this kind of structured documents turned out to be difficult to recognize. This is because the set of words is not closed and abbreviations are not standardized. Also the syntactical structure is not known and many variations appear even within the same volume. Nevertheless, the obtained results are really encouraging.

As future work, we plan to increase the presented corpus with further volumes of the same Marriage License books collection from different writers and different centuries. Furthermore, we plan to provide bounding-boxes around for each word in order to make the database suitable for research on keyword spotting [30]. Note, however, that the underlying segmentation free HTR technology tested in the experiments can already be used for some learning-based keyword spotting tasks [31, 32].

As far as recognition technologies are concerned, we plan to follow the ideas previously developed in [29] and study the integration of prior knowledge by taking advantage of the regular structure of the script.

Along a different line of research, the use of interactive systems to obtain perfect transcriptions will be explored. Heavy "post-editing" work of the output of an automatic recognition system is usually perceive as inefficient and uncomfortable and therefore hardly accepted by expert transcribers. Instead, computer assisted interactive predictive solutions such as CATTI [33] can be used. In a nutshell, the user works interactively in tight mutual cooperation with the system to obtain the final perfect transcription with minimum effort.

Appendix: ESPOSALLES transcription rules

The following rules were considered during the transcription process of the ESPOSALLES database:

- Page and line breaks were copied exactly.
- No spelling mistakes were corrected.
- Punctuation signs were copied as they appear.
- Word abbreviations with superscripts were written as ^(super). For instance, the word Bar^a was transcribed as *Bar^(a).* (see Fig. 4.a).

- Word abbreviations with deletion of some characters were written with \$. For instance, the word *Juã* was transcribed as *Jua\$* (see Fig. 4.c).
- The words written between lines are written as: *ˆˆ(words).* For example, the Fig 4.f) is transcribed as "*Sebastia Ripoll, ˆˆ(sastre) de Barˆ(a). defunct*".
- The cross-out words that can be read were transcribed between square bracket. For example, the Fig 4.b was transcribed as "*pages [de] [Sˆ(t)] [Esteva]*".
- If a complete register is cross-out, it is transcribed between double square bracket. (see Fig. 5.a).
- The cross-out words that can not be read were transcribed as */xxxxx/* (see Fig. 4.d).
- Blank areas in lines were transcribed by 5 blank spaces.
- The straight line at the end of the register was transcribed as -----.

Acknowledgements

Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), MITTRAL (TIN2009-14633-C03-01) and KEDIHC (TIN2009-14633-C03-03) projects. This work has been partially supported by the European Research Council Advanced Grant (ERC-2010-AdG-20100407: 269796-5CofM) and the European 7th framework project (FP7-PEOPLE-2008-IAPP: 230653-ADAO). Also supported by the Generalitat Valenciana under grant Prometeo/2009/014 and FPU AP2007-02867. We would also like to thank the Center for Demographic Studies (UAB) and the Cathedral of Barcelona.

References

[1] D. J. Kennard, A. M. Kent, W. A. Barrett, Linking the past: discovering historical social networks from documents and linking to a genealogical database, in: Proc of the 2011 Workshop on Historical Document Imaging and Processing (HIP 2011), New York, USA, 2011, pp. 43–50.

[2] D. W. Embley, S. Machado, T. Packer, J. Park, A. Zitzelberger, S. W. Liddle, N. Tate, D. W. Lonsdale, Enabling search for facts and implied facts in historical documents, in: Proc. of the 2011 Workshop on Historical Document Imaging and Processing (HIP 2011), New York, USA, 2011, pp. 59–66.

[3] S. Athenikos, Wikiphilosofia and pananthropon: Extraction and visualization of facts, relations, and networks for a digital humanities knowledge portal, in: ACM Student Competition at the 20th ACM Conference Hypertext and Hypermedia (Hypertext 2009), Torino, Italy, 2009.

[4] B. Coüasnon, J. Camillerapp, I. Leplumey, Access by content to handwritten archive documents: generic document recognition method and platform for annotations, International Journal on Document Analysis and Recognition 9 (2007) 223–242.

[5] F. Le Bourgeois, H. Emptoz, Debora: Digital access to books of the renaissance, International Journal on Document Analysis and Recognition 9 (2007) 193–221.

[6] U.-V. Marti, H. Bunke, Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System, Int. Journal on Pattern Recognition and Artificial Intelligence 15 (1) (2001) 65–90.

[7] A. H. Toselli, et al., Integrated Handwriting Recognition and Interpretation using Finite-State Models, Int. Journal on Pattern Recognition and Artificial Intelligence 18 (4) (2004) 519–539.

[8] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (5) (2009) 855–868.

[9] S. España-Boquera, M. Castro-Bleda, J. Gorbe-Moya, F. Zamora-Martínez, Improving offline handwriting text recognition with hybrid hmm/ann models, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (4) (2011) 767–779.

[10] V. Lavrenko, T. Rath, R. Manmatha, Holistic word recognition for handwritten historical documents, in: Proc. of 1rst IEEE International Conference on Document Image Analysis for Libraries (DIAL 2004), Washington DC, USA, 2004, pp. 278–287.

[11] M. Wuthrich, M. Liwicki, A. Fischer, E. Indermuhle, H. Bunke, G. Viehhauser, M. Stolz, Language model integration for the recognition of handwritten medieval documents, in: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, IEEE, 2009, pp. 211–215.

[12] A. Fischer, V. Frinken, A. Fornés, H. Bunke, Transcription alignment of latin manuscripts using hidden markov models, in: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, HIP '11, New York, NY, USA, 2011, pp. 29–36.

[13] N. Serrano, F.-M. Castro, A. Juan, The RODRIGO database, in: Proceedings of the The seventh international conference on Language Resources and Evaluation (LREC 2010), 2010.

[14] D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. Ramos, A. Juan, The GERMANA database., in: Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009), Barcelona, Spain, 2009, pp. 301–305.

[15] A. Esteve, C. Cortina, A. Cabré, Long term trends in marital age homogamy patterns: Spain, 1992-2006, Population 64 (1) (2009) 173–202.

[16] C. Sibade, T. Retornaz, T. Nion, R. Lerallut, C. Kermorvant, Automatic indexing of french handwritten census registers for probate geneaology, in: Proc. of the 2011 Workshop on Historical Document Imaging and Processing (HIP 2011), New York, USA, 2011, pp. 51–58.

[17] A. B. S. Almeida, R. D. Lins, G. de F. Pereira e Silva, Thanatos: automatically retrieving information from death certificates in Brazil, in: Proc. of the 2011 Workshop on Historical Document Imaging and Processing (HIP 2011), New York, USA, 2011, pp. 146–153.

[18] F. Drida, Towards restoring historic documents degraded over time, in: Proc. of 2nd IEEE International Conference on Document Image Analysis for Libraries (DIAL 2006), Lyon, France, 2006, pp. 350–357.

[19] A. Namboodiri, A. Jain, Document structure and layout analysis, in: Digital Document Processing, 2007, pp. 29–48.

[20] K. Kise, A. Sato, M. Iwata, Segmentation of page images using the area voronoi diagram, Computer Vision and Image Understanding 70 (1998) 370–382.

[21] N. Serrano, L. Tarazón, D. Pérez, O. Ramos-Terrades, A. Juan, The GIDOC prototype, in: Proceedings of the 10th PRIS 2010, Funchal (Portugal), pp. 82–89.

[22] K. Y. Wong, F. M. Wahl, Document analysis system, IBM Journal of Research and Development 26 (1982) 647–656.

[23] C. D. Manning, H. Schütze, Foundations of statistical natural language processing, MIT Press, 1999.

[24] E. Kavallieratou, E. Stamatatos, Improving the quality of degraded document images, in: Proc. of 2nd IEEE International Conference on Document Image Analysis for Libraries (DIAL 2006), Washington DC, USA, 2006, pp. 340–349.

[25] V. Romero, M. Pastor, A. H. Toselli, E. Vidal, Criteria for handwritten off-line text size normalization, in: Proc. of the 5th Int. Conf. on Visualization, Imaging and Image (VIIP 2006), Palma de Mallorca, Spain, 2006.

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

[27] E. Grosicki, H. El Abed, ICDAR 2009 handwriting recognition competition, in: Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009), 2009, pp. 1398–1402.

[28] R. Kneser, H. Ney, Improved backing-off for m-gram language modeling, Vol. 1, Detroit, USA, 1995, pp. 181–184.

[29] V. Romero, J.-A. Sánchez, N. Serrano, E. Vidal, Handwritten text recognition for marriage register books, in: Proc. of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China, 2011, pp. 533–537.

[30] T. Rath, R. Manmatha, Word spotting for historical documents, International Journal on Document Analysis and Recognition 9 (2007) 139–152.

[31] A. Fischer, A. Keller, V. Frinken, H. Bunke, Lexicon-free handwritten word spotting using character hmms, Pattern Recognition Letters 33 (7) (2012) 934 – 942.

[32] V. Frinken, A. Fischer, R. Manmatha, H. Bunke, A novel word spotting method based on recurrent neural networks, IEEE Transactions on Pattern Analysis Machine Intellignece. 34 (2) (2012) 211–224.

[33] A. Toselli, V. Romero, M. Pastor, E. Vidal, Multimodal interactive transcription of text images, Pattern Recognition 43 (5) (2009) 1824–1825.