The Five Centuries of Marriages Project

Joana Maria Pujadas-Mora
Center for Demographic Studies – Autonomous University of Barcelona
jpujades@ced.uab.es

The aim of the talk is to present an overview of the Advanced Grant project “Five Centuries of Marriages” (5CofM) directed by Professor Anna Cabré funded by the European Research Council through the IDEAS Programme for the period 2011 – 2016. The project is developed jointly by researchers from three different host institutions: Autonomous University of Barcelona, Center for Demographic Studies and Center for Computer Vision.

The project (http://dag.cvc.uab.es/infoesposalles/) is a long-term research initiative based on the data-mining of the Llibres d’Esposalles (Marriage Licenses Books) conserved at the Archive of the Barcelona Cathedral to construct an exhaustive database named the Barcelona Historical Marriage Database (BHMD). This extraordinary data source comprises 291 books of marriage licenses, with information of approximately 600,000 unions celebrated in over 250 parishes of the Diocese between 1451 and 1905. These books compiled information on the tax posed on each couple depending on their social class. This fiscal marker, as well as the exhaustive nature of the source and the variety of types of the parishes involved, from the city centre to the most rural villages, allows for the execution of a project oriented towards the multiple aspects of demographic research, especially in the very long view. The good preservation of the source allows the research on computer-aided recognition of manuscripts to help to the compilation of the demographic information.

Project goals.

The demographic research has the challenge of working an important variety of demographic questions through a type of event, the marriage. The main goals are as follows:
- The extraction of all the substantive information on five centuries of marriages contained in the Llibres d’Esposalles through the construction of an exhaustive database to be named the Barcelona Historical Marriage Database (BHMD)
- Chronology and geography of population estimates.
- The data on marriage fees and occupations allows the description and analysis of the importance of social groups, their geographical location and the historical change concerning them.
- Migratory flows.
- Survival, thus mortality.
- The methodological research on linkage of identities and kinship.

The computer vision objectives are the following:
- To study and implement computer algorithms for handwriting recognition.
- To develop writer adaptation and linguistic modelling approaches to improve the recognition.
- To create novel multimodal interaction paradigms to assist knowledge capture from practitioners.

The building of the Barcelona Historical Marriage Database

The Barcelona Historical Marriage Database compiles the information of approximately 600,000 marriages celebrated in over 250 parishes of the Diocese between 1451 and 1905 from the 291 Books of Marriage Licenses preserved at the Archive of Barcelona Cathedral. The first preserved volume dates from 1451, but there are reasons to believe that there was a custom of recording marriages in the diocese of Barcelona in earlier times (Baucells 2002). The origin of the books can be traced to a privilege given by Pope Benedict XIII (1328–1423) to
the diocese for construction and subsequent maintenance of the Barcelona Cathedral when he visited the city in September 1409 (Carreras Candí, 1913). The tax paid by each bridal couple was based on their socio-economic status. This exceptional source is conserved at the Archives of the Barcelona Cathedral. The source covers the nearly 250 parishes of the Dioceses of Barcelona, ranging from the most urban core of the city, Barcelona, to the most rural villages in the periphery of the Diocese. The Diocese was divided in 4 Deanships; the main one was called the Oficiality of Barcelona that concentrated the 90% of the population of the Diocese.

To transcribe the marriage licences a web-based crowdsourcing platform has been developed. It also includes functionalities to create a ground-truth for document image analysis research, necessary for the image recognition. In the transcription interface the information is displayed in two panels (see Figure 1). In the top panel the source is presented. The image of the source can be zoomed in, zoomed out, scrolled down and scrolled up. The user can browse pages to see previous or next ones in the assigned images without losing the data already filled in.

**Figure 1: Web-based crowdsourcing platform 5CoFM project. Transcription module.**


The platform also contains a functionality of standardization or harmonization for names, surnames, occupations and geographical locations. This task is essential as the marriages were recorded over a period of five centuries by different scribes with different handwriting. After that, the data is transferred into a digital format compatible with spreadsheets that are also downloaded from this platform.

**Record linkage software**

The goal related to the methodological research on linkage of identities and kinship with the aim to reconstruct genealogies has been fulfilled technically with the creation of the record linkage software called *Buscadescendències*. The nominal record linkage has been done using two different measures: the Bag distance and a slightly modified Levenshtein edit distance. This record linkage has been thought to link parents and children. After an initial link there is a data cleaning in order to discard the less likely links and reduce the amount of multiple links due to the existence of over-links with multiple candidates. At the same time the system searches for brothers and sisters in registers having corresponding parents’ information.

**Demographic Research**

Some of the research on the demographic view part of the project is concerned to the study of marriage. The approach to the nuptial phenomenon has been carried out by way of two themes, the first concerned with the processes of social reproduction or the intergenerational transmission of social status or occupation during the Ancien Régime, and the second with social change during industrialisation with the backdrop of modern population growth deriving from the fact of getting married more frequently during Lent and the explosion of consanguinity. Other topic studied within the 5CoFM is the examination of the family names contained in the BHMD. This methodology is based on the premise that surnames are transmitted from fathers
to their children in a stable form, as in the Catalan case since the fourteenth century. Thus as the transmission of surnames is, almost, stable, different family lines may be identified, allowing us to establish a direct relationship between the number of existing surnames and the demographic evolution of those who bear them.

The statistical methodology used in these studies is based on descriptive and multivariable analysis with logistic regression or log linear models.

**Computer Research**

Document Image Analysis and Recognition is a research field within Computer science concerned with converting raw images of scanned documents to an editable format with the recognized contents. Prior to the recognition of text is necessary other operations. Sometimes it has to be dealt with the degradation of paper due to paper aging. Since some of the marriage license books present severe show-through, we have proposed a show-through cancellation method (see figure 2).

*Figure 2: After and before of a page of the Marriage Licenses Books. Show-through cancellation.*

In order to start with the recognition, first it is necessary to analyze the layout of the document in order to detect the text blocks and segment the text lines. In our case study the layout analysis is used for segmenting the left column corresponding to first surname, the main text block with the license proper, the right column corresponding to fees and the total amount located at the bottom of the page. Once the main text block has been detected, the next step segments the text into text lines (see Figure 3).

*Figure 3: Example of line segmentation result.*

Now, the next step is to transcribe them automatically (see Figure 4). Current state-of-the-art technologies for handwritten text line recognition are based on Hidden Markov models (HMMs) (Rabiner, 1989) and Neural Networks (NNs). In order to improve the recognition rates language models and dictionaries, in which the historians has an important role, are typically integrated into the recognition process.

**Figure 4: Examples of recognition of new words**

<table>
<thead>
<tr>
<th>Word n-gram</th>
<th>Maryanna donsella filla de dita Vila texidor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>Maryanna donsella filla de Juà Valls texidor</td>
</tr>
<tr>
<td>Hybrid Grammar</td>
<td>Maryanna donsella filla de Juà Valta texidor</td>
</tr>
</tbody>
</table>


There are also other techniques, as word-spotting, which is the task of retrieving all instances of a given word, offers a viable solution to make historical documents amenable to searching and browsing, especially when training data is difficult to obtain as it is needed for the previous techniques (see figure 4). This technique has been also used to correct the transcription in order to detect missing information. This has been really important in variables like the marital status, mainly because it was recorded as an abbreviation, or occupation.

**Figure 4: Example of word-spotting using the word Barcelona.**


To conclude, it is important to remark that handwriting recognition as a research domain of the computer vision field, hence it is not realistic to affirm that a completely automatic process is feasible. Although handwriting recognition has experienced important progress in the last years, the full transcription is still a challenge.

**References**


