

# HISTORICAL LIFE COURSE STUDIES

VOLUME 2  
2015



## MISSION STATEMENT

# HISTORICAL LIFE COURSE STUDIES

*Historical Life Course Studies* is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

### Methodological Articles

This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

### Research articles

This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

*Historical Life Course Studies* is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, <http://www.esf.org>), the Scientific Research Network of Historical Demography (FWO Flanders, <http://www.historicaldemography.be>) and the International Institute of Social History Amsterdam (IISH, <http://socialhistory.org/>). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at <http://www.ehps-net.eu/journal>.

**Editors: Koen Matthijs & Paul Puschmann**  
**Family and Population Studies**  
KU Leuven, Belgium  
[hislives@kuleuven.be](mailto:hislives@kuleuven.be)

**The European Science Foundation (ESF)** provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.



**The European Historical Population Samples Network (EHPS-net)** brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.  
Visit: <http://www.ehps-net.eu>.



HISTORICAL LIFE COURSE STUDIES  
VOLUME 2 (2015), 1-19, published 29-01-2015

# A Tale of Two Transcriptions

## Machine-Assisted Transcription of Historical Sources

Gunnar Thorvaldsen, Norwegian Historical Data Centre, University of Tromsø

Joana Maria Pujadas-Mora, Centre for Demographic Studies, Autonomous University of Barcelona

Trygve Andersen, Norwegian Historical Data Centre, University of Tromsø

Line Eikvil, Norwegian Computing Center, Oslo

Josep Lladós, Computer Vision Centre, Autonomous University of Barcelona

Alícia Fornés, Computer Vision Centre, Autonomous University of Barcelona

Anna Cabré, Centre for Demographic Studies, Autonomous University of Barcelona

### ABSTRACT

This article explains how two projects implement semi-automated transcription routines: for census sheets in Norway and marriage protocols from Barcelona. The Spanish system was created to transcribe the marriage license books from 1451 to 1905 for the Barcelona area; one of the world's longest series of preserved vital records. Thus, in the Project "Five Centuries of Marriages" (5CofM) at the Autonomous University of Barcelona's Center for Demographic Studies, the Barcelona Historical Marriage Database has been built. More than 600,000 records were transcribed by 150 transcribers working online. The Norwegian material is cross-sectional as it is the 1891 census, recorded on one sheet per person. This format and the underlining of keywords for several variables made it more feasible to semi-automate data entry than when many persons are listed on the same page. While Optical Character Recognition (OCR) for printed text is scientifically mature, computer vision research is now focused on more difficult problems such as handwriting recognition. In the marriage project, document analysis methods have been proposed to automatically recognize the marriage licenses. Fully automatic recognition is still a challenge, but some promising results have been obtained. In Spain, Norway and elsewhere the source material is available as scanned pictures on the Internet, opening up the possibility for further international cooperation concerning automating the transcription of historic source materials. Like what is being done in projects to digitize printed materials, the optimal solution is likely to be a combination of manual transcription and machine-assisted recognition also for hand-written sources.

**Keywords:** Nominative sources, census, vital records, computer vision, optical character recognition & word spotting.

e-ISSN: 2352-6343

PID article: <http://hdl.handle.net/10622/23526343-2015-0001?locatt=view:master>

The article can be downloaded from [here](#).

© 2015, Gunnar Thorvaldsen, Joana Maria Pujadas-Mora, Trygve Andersen, Line Eikvil, Josep Lladós, Alícia Fornés, Anna Cabré

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>

## 1 INTRODUCTION

The building of historical demographic databases covering lengthy periods requires the transcription of large amounts of nominative data, even more extensive than when covering the same territory with cross-sectional samples. This is both because it is more data-demanding to follow groups of persons in several sources or in the same source over time and because it is harder to use sampling methods in longitudinal research than in cross-sectional. Much transcription work has lately been done with low cost country labor, thus using traditional letter by letter transcription techniques offline, most notably for the long series of US and British censuses from 1850 to 1940 and 1911, respectively.<sup>1</sup> This article describes two alternative transcription systems which do not rely totally on “brute force”, but rather software based on Internet and computer vision routines. One reason for this is that the transcription of sources in small languages like Catalan and Norwegian is more difficult to outsource than materials in English. Another is that small series of sources may require the same amount of source-specific transcriber training as a uniform census with tens of millions of records. In the field of transcription of historical documents, it is also worth mentioning the *Family History Library*<sup>2</sup>, one of the major initiatives undertaken by the Mormon community to build a historical network worldwide. For years they have scanned millions of documents and entered cross-linked genealogical information. This process is done manually by thousands of volunteers using a crowdsourcing platform.

Crowdsourcing (Estelles-Arolas and González-Ladrón-de-Guevara 2012) consists in obtaining services by the collaborative contributions from a large number of people, usually online users. The philosophy consists in splitting a tedious and time-intensive task into smaller tasks that can be done in parallel by the community, so that each user performs a small portion of the greater task. Afterwards the partial results are combined in a complete solution. A pioneering and successful example of this paradigm has been Wikipedia, an open online encyclopedia where anyone can add articles, respecting the rules of conduct. For example, the Spanish version of Wikipedia currently has a data input rate of 300 items per day.

The two transcription systems explained here had different starting points, but are now coming together. The Spanish system was created to transcribe the marriage licenses books (Llibres d'Esposalles) covering the period 1451 to 1905 for the Barcelona area, likely one of the world's longest series of preserved vital records. Thus, in the Project “Five Centuries of Marriages” (5CofM) at the Autonomous University of Barcelona and Centre for Demographic Studies, directed by Professor Anna Cabré, the Barcelona Historical Marriage Database has been built. The more than 600,000 records have been transcribed by 150 transcribers working online<sup>3</sup>. The Norwegian material is cross-sectional and different as it is the 1891 census, originally recorded on one sheet per person. This format and the underlining of keywords for several variables made it easier to automate some of the data entry than when information about many persons is listed on the same page. Optical Character Recognition (OCR) for printed text is the most popular document analysis problem and is nowadays scientifically mature. Research is now focused on more difficult problems such as handwriting recognition. In the marriage project some document analysis methods have been proposed to automatically recognize the marriage licenses. Fully automatic recognition is still a challenge, but some promising results have been obtained. In both Spain and Norway the source material is made available as scanned pictures on the Internet together with unique page references. Now the two methods are being brought together since IT- specialists in the Computer Vision Centre at the Autonomous University of Barcelona cooperate with the Centre for Demographic Studies to semi-automate the transcription, while the Norwegian software is being ported from a stand-alone Windows platform to availability via the web. Like what is being done in projects to digitize printed materials, the optimal solution is likely to be a combination of manual transcription and machine-assisted recognition, also for hand-written sources.

1 <http://www.ancestry.com> and <http://www.ancestry.co.uk/>. All Internet references were checked 15 November 2014.

2 <https://familysearch.org/>

3 Not all the 150 were transcribing at the same time. The transcription lasted 2 years. <http://dag.cvc.uab.es/5cofm-ground-truth>

## 2 DIGITIZING A SINGLE SHEET CENSUS – 1891 FOR NORWAY

While the 1801 census was nominative, the decadal censuses from 1815 to 1855 were only statistical. Since 1865, however, nominative censuses were taken, mostly at ten-year intervals (Solli & Thorvaldsen 2012). For their second nominative census in 1875, the Statistical Bureau had decided to use separate forms for each domicile. This made it easier to aggregate information on the domicile and household level since the households were clearly separated. Still, it was deemed necessary to encode the information by filling in individual sheets for each person, enumerated after the questionnaires for each domicile had been returned to Oslo. This cumbersome copying of data, with differently colored cards for different types of persons, delayed the computing and publication of aggregates significantly, inspiring further reforms. Thus, the work could be rationalized further, in parallel with what Hollerith achieved with the machinery he invented for the US 1890 census (Anderson 1988, pp.106-107; US Census Bureau).

One solution to rationalize census aggregation had been tried for the first time in Paris in 1817, again in France in the 1870s and was copied in Norway in 1891, a census postponed six years in order to be synchronized with the international rounds of censuses. Rather than listing all persons sequentially by family and household, the information about each person was filled directly onto a separate form by the census taker in the field, cf figure 1.<sup>4</sup> By numbering the forms sequentially for each domicile the census analysts would still know which persons belonged to the same family and household. Special single sheets were filled in with information on the household level. During aggregation, the individual forms were put into different stacks for the relevant combinations of variable values and then counted. Still the task was time consuming, single sheets could easily be misplaced and many summing errors were made when combining partial results. Individual sheets were also used in Mecklenburg, Germany for the 1891 and 1900 censuses. Here they have been preserved like in Norway, while the French forms were copied onto household schedules and supposedly destroyed (Statistics Norway 1895; INSEE no date; Haug 1979).

When the Church of Latter Day Saints microfilmed most of the Norwegian church records and censuses until 1900, they deemed it too cumbersome and cost-inefficient to include the more than two million sheets filled in 1891. Therefore, this census has been accessible only to visitors in the National Archives in Oslo, although they have been photocopied and transcribed for a few selected municipalities with altogether 120,000 inhabitants. Even if a small and non-random sample, this has been used to study, e.g., consanguineous marriages, because the forms contain a question to all married women to report whether they were married to a third cousin or a closer relative (Jåstad & Thorvaldsen 2012). Otherwise the 1891 census contains the ordinary census variables, except that the forms for the two northernmost regions ask about ethnicity and language due to the concentration of Sami and Finnish ethnic groups there. Employing up-to-date scanning equipment with automated page feeding, the whole census is now being scanned by the National Archives, and significant parts are already available as graphics files and can be browsed via the Internet.<sup>5</sup> It will be an important element in the Historical Population Register which is currently being built for Norway (Thorvaldsen 2011).

### 2.1 THE LAY-OUT OF THE FORMS

The most peculiar internal characteristic of the 1891 census sheets may be that eight demographic questions about each person are not filled in verbatim, but rather by underlining keywords for some variables. These are: sex, family position, marital status, whether spouses are related as third cousins or closer, whether born in a rural or urban municipality, whether partly or wholly provided for, if the person was mentally disturbed, deaf and dumb or blind and when the illness was contracted. In the northern regions, in addition, the two questions about ethnicity and language should be indicated by underlining keywords. In a report from the Norwegian Computing Center (Eikvil, Holden & Bævre 2010) several possibilities for automating or semi-automating the transcription of Norwegian nominative sources were evaluated, cf the appendix. Treatment of the 1891 sheets with their underlined fields is likely the easiest to implement.

This was the starting point for two rounds of software construction performed by the SAMBA - Statistical Analysis, Image Analysis and Pattern Recognition department of the Norwegian Computing Center with their comprehensive theoretical and practical knowledge in the fields of statistics, image

4 <http://www.rhd.uit.no/census/ft1891.html>

5 [http://digitalarkivet.arkivverket.no/finn\\_kilde?s=&fra=1891&til=1891&kt%5B%5D=FOLK](http://digitalarkivet.arkivverket.no/finn_kilde?s=&fra=1891&til=1891&kt%5B%5D=FOLK)



analysis and pattern recognition, cf <http://nr.no>. The software was constructed with open source tools in C/C++ under Linux, and using off-the-shelf image analysis and recognition libraries. The purpose of the first version was to read each of the images and analyze the eight questions that can be underlined in the whole country. The input is scanned images of the census in low resolution (approximately 1370x2048) jpg format, converted to an internal format for further analyses. The pictures are analyzed as RGB-pictures where only the red layer is used.

Figure 1 Individual 1891 census form with the four boxes inserted to delimit the picture before analysis.

Folketælling for Kongeriget Norge 1ste Januar 1891.

Grua Herred. Schema 2. Personseddel No. 1.

Tællingskreds No. 5. Husliste No. 1.

1. Fuldt Navn *Hans Kristian Mari*
2. Kjøen<sup>1)</sup>: Mandkjøn, Kvindekjøen.
3. Stilling til Familiens Hovedperson<sup>1)</sup>: Selv Hovedperson, Hustru, Søn, Datter, Tjenestetyende, Logerende hørende til Familien, enslig Logerende, Besøgende o. s. v.
4. Ægteskabelig Stilling<sup>1)</sup> (for Personer over 15 Aar): Ugift, Gift, Enkemand, Enke, Separeret efter Bevilling, Lovlig fraskilt.
5. For gifte Kvinder: Er De og Deres Mand indbyrdes beslegtet, som Næstsøskendebrødre (Tremenninger) eller nærmere? Ja, Nei<sup>1)</sup>.
6. Fødselsaar: 18. 42. For Børn under 2 Aar: Fødselsmaaned.....
7. Fødested: Herred, Sogn eller By..... Grua.....  
For de i Udlandet fødte: (Landets eller Stedets Navn).....
8. Hvilken Stats Undersaat (for dem, der ikke ere norske Undersaatter).....
9. Trossamfund (for dem, der ikke tilhøre den norske Statskirke).....
10. For de i selvstændig eller i underordnet Stilling Erhvervende: Erhvervsgren (Hovednæring, Bierhverv) og Stilling i samme..... Gaard..... Arbejdes..... Arbejdes.....
11. For de af Andre helt<sup>1)</sup> eller delvis<sup>1)</sup> Emsørgede:  
Forsørgerens Livstilling.....
12. Sindssvag, Døvstum eller Blind<sup>1)</sup>.
13. Er Sindssygdommen, Døvtumheden eller Blindheden medfødt (hvormed ligestilles, at den er kommen tilsyne i de første Barneaar), eller er den fremtraadt senere<sup>1)</sup>?
14. For de kun midlertidigt Tilstedeværende:  
sædvanligt Bosted.....
15. For de midlertidigt Fraværende:  
antageligt Opholdssted.....

<sup>1)</sup> De for hvert Tilfælde passende Ord understreges.

Figure 2 *Two instances of the keyword for male gender to be compared, with and without underlining.*



The method for detecting underlined data has four steps. First, the picture is delimited by identifying the four corners of the area with information on each sheet, as indicated by the four boxes inserted in figure 1. Second, the specific fields containing the information to be analyzed are delimited, so that the software knows the exact location of the sheet's constituent parts. In a third step, the image part with potential underlining is compared to a prerecorded image of the area without underlining, cf figure 2. The algorithm will then store the value of the underlined keyword and a score for the likelihood that the keyword was really underlined. If more than one underlined value is flagged, the one with the highest score is chosen. The fourth step in the program is a logical analysis of the consistency of the values of the eight fields brought together. For instance, if male gender is underlined in question two, this excludes the values "Hustru" (wife) or "Datter" (daughter) as family position in the next question.

The results were stored in a flat file with a special character delimiting the field values. When running the program on a selection of 420 images containing 1640 underlined data values, nearly all (99.3%) of the underlined data was found, and on these sheets 96.7% of all of the underlined data was identified correctly. 3 fields were incorrectly detected as underlined and 11 of the 1640 underlined fields were not detected, so that the underlining on only 14 out of 420 images was transcribed incorrectly.

A second and more advanced version of the software is being tested now, where the results are not as reliable as from the first version, simply due to the more complicated contents of the analyzed fields. The upgraded software attempts both to transcribe the fields with numbers automatically and to copy the alphabetic strings containing names, occupations, birthplaces, etc. to separate image files. This segmentation (cf paragraph 4 below) involves the detection of all the remaining fields with procedures analogous to those used for the underlined ones. Next the automatic transcription of digits with a Convolutional Neural Nets software library is being tested (Sermanet, Kavukcuoglu & LeCun 2009). Current results indicate that numbers consisting of single digits can be read with accuracies of around 85-90%. Errors in these cases are typically related to problems in the extraction of the number fields due to noisy documents, corrections made in the documents (e.g. by crossing out numbers) or where noise or corrections lead the recognizer to believe that there is more than one digit when there is not. Currently no attempt is performed to detect these cases prior to the recognition. However, some cases of the first class of problems should be possible to detect in the registration phase. Furthermore, along with each recognized number, a confidence score is included. This is currently not exploited, but could be used to identify probable errors and mark these for manual inspection.

When numbers consist of two or more digits the accuracy of the automated transcription may fall towards 70 %, which may be too low for practical purposes. These problems are related to certain difficulties the recognizer has in detecting the true number of digits, and the fact that currently less digit samples from two-digit numbers have been collected and used in the training. The solution to this will be to implement software for training, employing the database of the 120,000 forms from the 1891 census, which have been transcribed manually, as training material.

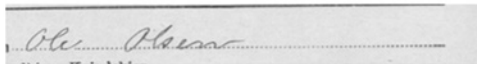
The text fields will be even more difficult to handle automatically, but per se the detection and singling out as independent pictures of the name, birthplace and occupation fields work satisfactorily. Also, it is of practical use that empty fields are identified and need not be dealt with during the manual transcription phases. We have tested algorithms for the transcription of names, occupations and birthplaces as three different processes. The name fields from several persons are displayed on the screen and as soon as the transcriber starts to write, likely names are suggested from a list prioritized by name frequencies computed in the complete database of the 1900 census (cf figure 3). Since gender has already been recognized automatically, only men's or only women's names will be displayed as a proxy for the ground truth explained in paragraph 4 below. The correct name can then be chosen by pressing a function key. For all three fields it would be rational to split the contents into separate words and group similar images, but this has not yet been implemented. Trained transcribers need much retraining before this use of pick-lists speed up transcription rates noticeably. Untrained transcribers

work faster, but tend to make more errors using pick lists than when transcribing names verbatim. Thus, we have mixed experiences with the names technique described above.

Figure 3 *MS Access application for transcribing first and last names in the 1891 census. Pick-lists with four first and four last names are displayed based on gender and name frequencies in the next transcribed census (from 1900). The 'o' key has been pressed and 'Ole' can be chosen by pressing F1 since this is the most frequent male name. The last name is carried over from the previous record. Gender (kjønn) can be overruled manually if necessary.*

Bilde

Status



Kjønn

Fornavn	Tom	Etternavn	Tom
<input type="text" value="o"/>		Olsen	<input type="text" value="Ingen (F9)"/>
<input type="button" value="Ole (F1)"/>		<input type="button" value="Torstensen"/>	
<input type="button" value="Olaf (F2)"/>		<input type="button" value="Torsen (F6)"/>	
<input type="button" value="Oskar (F3)"/>		<input type="button" value="Torstensdatter"/>	
<input type="button" value="Olav (F4)"/>		<input type="button" value="Torsdatter (F12)"/>	

The birthplaces, especially in the countryside, are place names with a high frequency, being mainly the census location itself and neighbouring municipalities. Whole groups of these can be selected before they are transcribed en masse. This could have been done with specially designed software, but it is simple, fast and intuitive to drag the birthplaces that belong together into the same folder in Windows. Once all birthplace images containing handwritten versions of e.g. "Bergen" have been collected in the folder called Bergen, it is straightforward to transfer the folder name to all the census records where these images belong – as is specified in the image file names. The occupations, however, are detailed and often referring to combinations of work roles, thus listing many different strings with low frequencies. Here the software could provide additional help if similar graphical images were grouped together, so that en masse transcription of several identical occupations could be achieved more efficiently. Such graphical grouping would enhance transcription rates also for other text strings found on the census sheets. In the 1950 census manuscripts, numeric occupation codes were written with red pencil, and current work shows that these can be read automatically with a high degree of certainty.

We expect that the graphical software being developed for the Catalan marriage books and the Norwegian 1891 census will have significant carry-over effect on the transcription of other nominative source materials. Both our own experiments with the household schedules, listing up to 20 persons in the 1950 census, and the transcription software catered for by FamilySearch in its worldwide voluntary data entry projects show that it is possible to identify the field structure in more complex materials than the 1891 census sheets. This is especially relevant for the 1920 census which also used the one sheet per person system, although with more variables and a double-sided layout. Whether it will be deemed realistic or not to outsource transcription of the large collections of vital records and census materials in Spain, Norway and elsewhere to low cost countries, we therefore expect to make rapid progress in the field of verbatim digitization of these sources during the next few years. The construction of a "reCAPTCHA" module to let genealogists and other researchers perform small voluntary data entry jobs in return for our services, is being planned in the National Archives in Oslo.



### 3 THE MARRIAGE LICENSES BOOKS OF THE CATHEDRAL OF BARCELONA (*LLIBRES D'ESPOSALLES*)

On September 27, 1409, Pope Benedict XIII<sup>6</sup> (Pedro Martínez de Luna, 1328–1432) visited Barcelona and granted the new Cathedral for a tax on marriage licenses (esposalles) to be raised on every union celebrated in the Diocese. This tax was maintained until the third decade of the 20th century. Between 1451 and 1905, a centralized register, called *Llibres d'Esposalles*, recorded all the marriages and the fees charged according to social status. This exceptional documentary treasure is conserved at the Barcelona Cathedral archives. It comprises 291 books with information on approximately 610,000 marriages celebrated in 250 parishes, ranging from the most urban core of the city to the most rural villages in the periphery of the Diocese. Their impeccable conservation for over four and a half centuries is like a miracle in a region where parish archives have suffered massive destruction through several episodes, mainly during the last 200 years.

Figure 4 Examples of pages with different marriage licenses.



These marriage licenses have been used to construct the Barcelona Historical Marriage Database, one of the main objectives of the Five Centuries of Marriages. The marriage licenses were written in Catalan until 1860; later they were recorded in Spanish. Names and family names of grooms (from 1876 two family names<sup>7</sup>), names of brides (the inclusion of the bride in the licenses started in 1485) and their marital status (mostly for the brides and for the widowed grooms) are the common items of each record during five centuries. The brides' surnames were not registered until 1643 since women were not esteemed as individuals in their own right: single brides were related to their fathers, widows to their late husbands. The information on the groom's first surname<sup>8</sup> was used to construct alphabetical indexes kept at the end of each volume. For some periods the bride's surnames were included next to the groom's.

<sup>6</sup> Benedict XIII was considered as an antipope during the Western Schism (1378 – 1418).

<sup>7</sup> The first family name (or when it is the only one) is normally a patrilineal surname. The identification of people with two family names is the consequence of the enforcing of the Civil Registration Law (1871). The first family name continued being transmitted by the father and the second one by the mother (de Salazar & Mayoralgo 1991; Salinero 2010).

<sup>8</sup> From 1481 – 1593 the indexes were kept *a posteriori* registering the first surnames of the groom and the bride. From 1593 – 1866 the indexes were built at the same time of the licenses and it had just collected the groom's surnames. From 1867 – 1905 the groom's and bride's surname were recorded originally.

Parents' information (names and whether they were alive or not at the time of their children's wedding) was registered quite often except for the period 1643 – 1750. That means there was a decrease in the quality of the licenses during that specific period, reinforced by the loss of other key variables, e. g. the previous origin or residence of the bride and groom. From 1715 onwards geographical location was recorded about the parish where the marriage was celebrated.

The amount of the tax per couple depended on their social status or occupation<sup>9</sup> - recorded for men throughout the period. The taxes mirror the whole social structure, from the nobility with the highest sum to those declared poor and exempt from tax. They were fixed in a seven-tiered scale from 1575 until 1649: 1) Nobility, 2) Military Citizens, 3) Honoured Citizens (those who could hold public office), 4) Merchants, Lawyers, Physicians, 5) Master of Guilds, 6) Farmers and small artisans and 7) The poor. From 1649 to 1857 an eight-tiered scale was found due to the inclusion of a level reserved for the merchants. After 1857 a seven-tiered scale re-emerged. The main drawback of the source from a demographic viewpoint is the missing registration of the partners' ages.

### 3.1 WEB-BASED CROWDSOURCING PLATFORM

In the 5CofM project a web-based crowdsourcing platform has been developed. This software platform has integrated the needs of the two research fields, thus combining social sciences and computer sciences. The goal is to provide a fully transcribed database for demographic research and a ground truth for document image analysis research. The first aim addressed demographic research, allowing for the transcription of the complete series of marriage license books during two years by a community of about 150 transcribers. The main advantages of this platform are not only the collaborative transcription paradigm, but also the centralized management of the tasks and monitoring of the work, assisting error detection and correction. Additionally, the platform includes a second crowdsourcing space for image annotation, generating a second modality of data. This is needed for “ground truth” generation in order to evaluate the performance of automatic image recognition software. In computer vision the term ground truth refers to gathering expected data for the recognition algorithms. In our case, expected data means the graphical position of words (delimited by bounding boxes) in the document images with the corresponding transcription. We refer to this second modality of output as the labeling mode. The inclusion of software for handwriting recognition adds a complementary tool for the manual transcription task, so the time needed for transcribing big sources can be reduced and some transcription errors can be avoided.

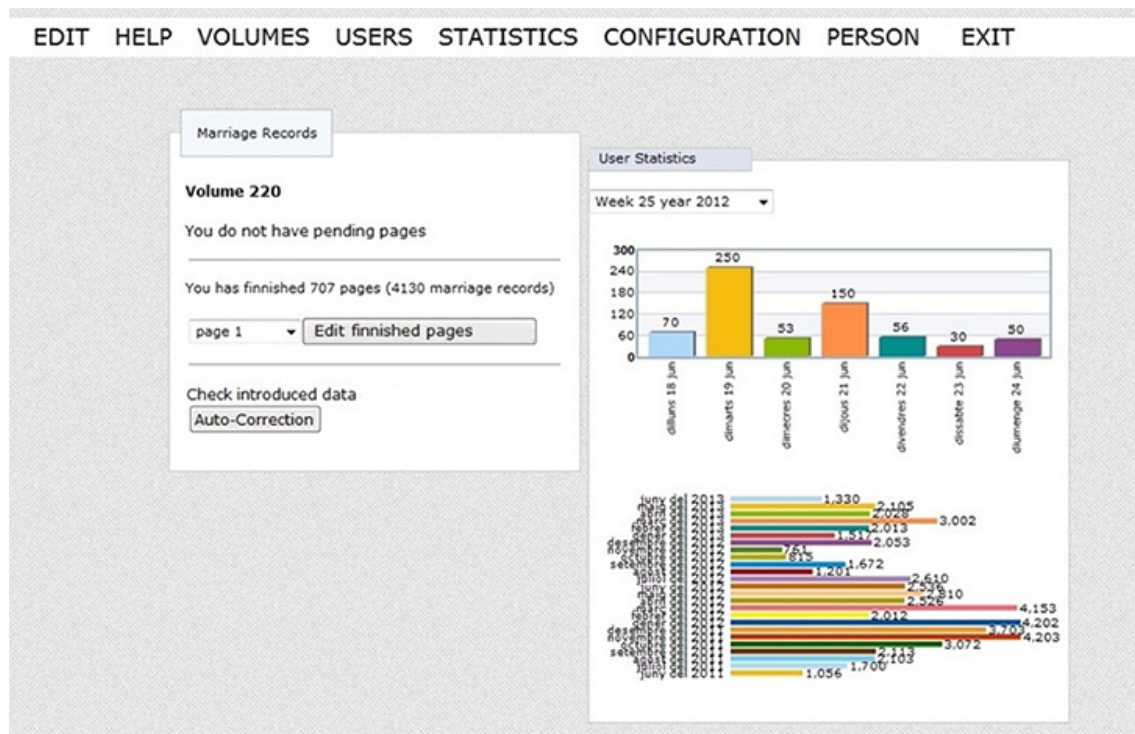
In the next subsections we describe the three abovementioned modules of the platform: users and tasks management, transcription and image annotation.

#### 3.1.1 ADMINISTRATION MODULE

As described above, a crowdsourcing platform splits the work into small tasks distributed among a community of users. In our case, the transcription is split according to groups of page images. The administration module allows us first to define the groups of images, to register users and assigns tasks to users. Additionally, the administrator can keep track of the progress of the tasks and get statistics on the number of completed pages, the profile of the work activity per user (when they work, the intensity of each work slot, etc.). Statistics are important to make decisions on new assignments of tasks, depending on user productivity or to reassign tasks from inactive users. The statistics are also shown to the users after they log into the system (see figure 5).

<sup>9</sup> The social status or the occupation was always recorded for the grooms, rarely for the brides and for parents during the period 1560 – 1663.

Figure 5 Administration module.



*Explanation: Figure 5 illustrates for a given user the work that he/she has done during a specific week of the year. From this starting page, the user can see his/her remaining work, start transcribing new pages or edit previously transcribed ones.*

### 3.1.2 TRANSCRIPTION MODULE

In the transcription interface the information is displayed in two panels (see figure 6). In the top panel the page image is presented. This is an image browser panel, so the user can zoom in, zoom out, scroll down and scroll up to easily see the register that is currently transcribed. The user can browse pages to see previous or next ones in the book without losing the data already filled in. In the bottom panel the system displays two forms with the relevant information corresponding to a wedding. The blue and pink forms correspond to the groom and the bride, respectively.

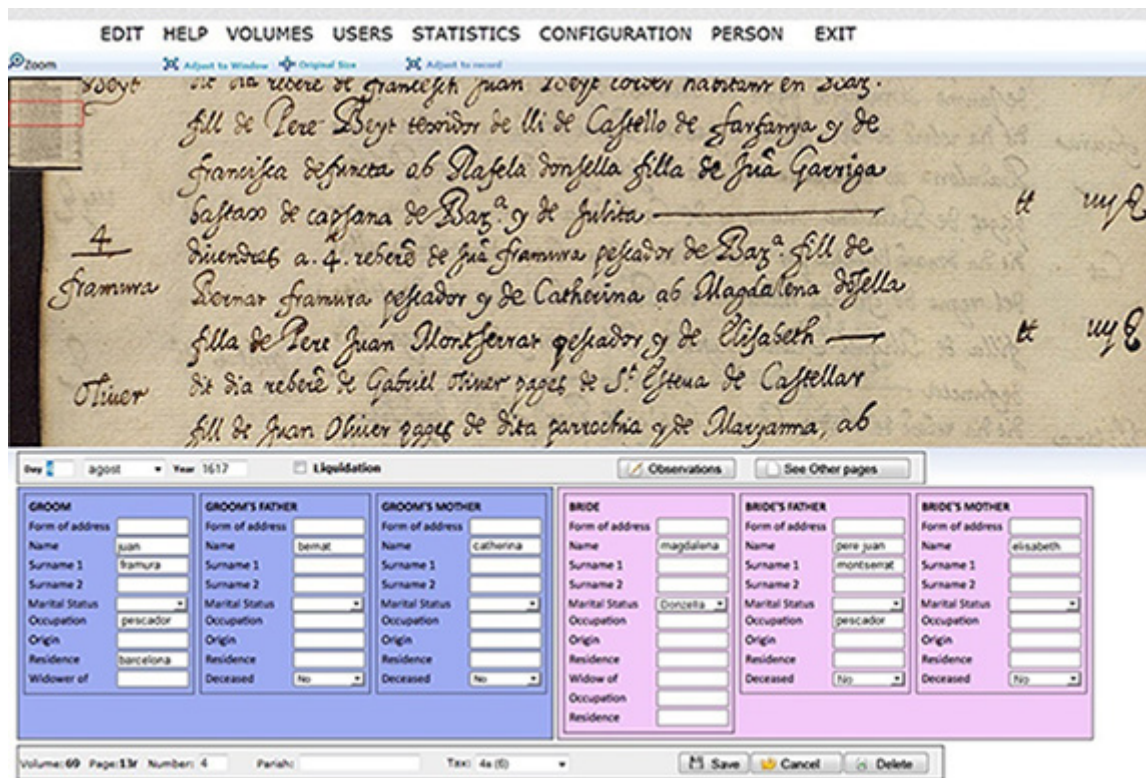
Figure 6 shows an example of a contextual information form. In the bottom part, the different fields for the husband and the wife are presented to be filled by the user. In the upper part, the user sees the image containing the records of the marriages. For example, in the marriage license shown, the name of the groom is "Juan" and the name of the bride is "Magdalena". The transcription has been done literally. Subsequently, a standardization of the data set has been carried out, including occupations, geographical locations, first names and surnames<sup>10</sup>. In the bottom part of the form, the volume number and page number are shown. The pages are automatically numbered so the user does not need to take care of this.

Some additional functions are provided. The user may write comments concerning the marriage register, e.g. if the handwriting is not readable, some information is missing or crossed-out. Once all the forms that are contained in a page are completed, the system labels that page as finished and shows the next document page to the user. An interesting functionality of the application is called "auto-correction" which is used for detecting spelling mistakes. With this feature the user can see how many times a specific word, e.g. a name, has been transcribed. Thus, the user can check for possible spelling errors since words that appear only once are more likely to contain spelling mistakes.

<sup>10</sup> Occupations have been coded into HISCO and geographical locations were codified according to the current Spanish ZIP. To standardize the names and family names, all different forms of a same name or surname have been grouped into their standardization form according to current Catalan grammar rules. Moreover, accents were removed as well as articles, prepositions and conjunctions in order to keep only the root of each name to facilitate the nominal linkage process.



Figure 6 Transcription module. The top panel displays the current page. The bottom panels display the two forms corresponding to the groom and the bride.



The platform also contains harmonization and record linkage functionalities. Thus, an automatic process in batch mode searches for potential links between individuals after transcription. For each couple, and starting from the parents' information, the system searches in registers corresponding to marriages celebrated some years before. Secondly, the system searches for brothers and sisters in registers having corresponding parents' information. In this search an implicit harmonization process exists so that inexact matches between names are accepted; i.e., name instances with spelling variations are considered to be similar. Finally, an expert user validates the automatically generated links between individuals.

### 3.1.3 IMAGE LABELING AND GROUND TRUTH MODULE

The second modality of the platform for entering data is the "ground truth" necessary for the validation of handwriting recognition processes. Handwriting recognition is a discipline of pattern recognition and computer vision that consists of developing computer algorithms for converting raw images of handwritten text into editable text files. Generally speaking, such algorithms can be divided into two major steps: segmentation and recognition. Segmentation consists of cropping the original image into subimages corresponding to text components, basically lines and words. Segmentation prepares the recognition step which consists of extracting relevant visual features from each word image and decoding it according to pre-learned models of character shapes, given as output consisting of the corresponding ascii string.

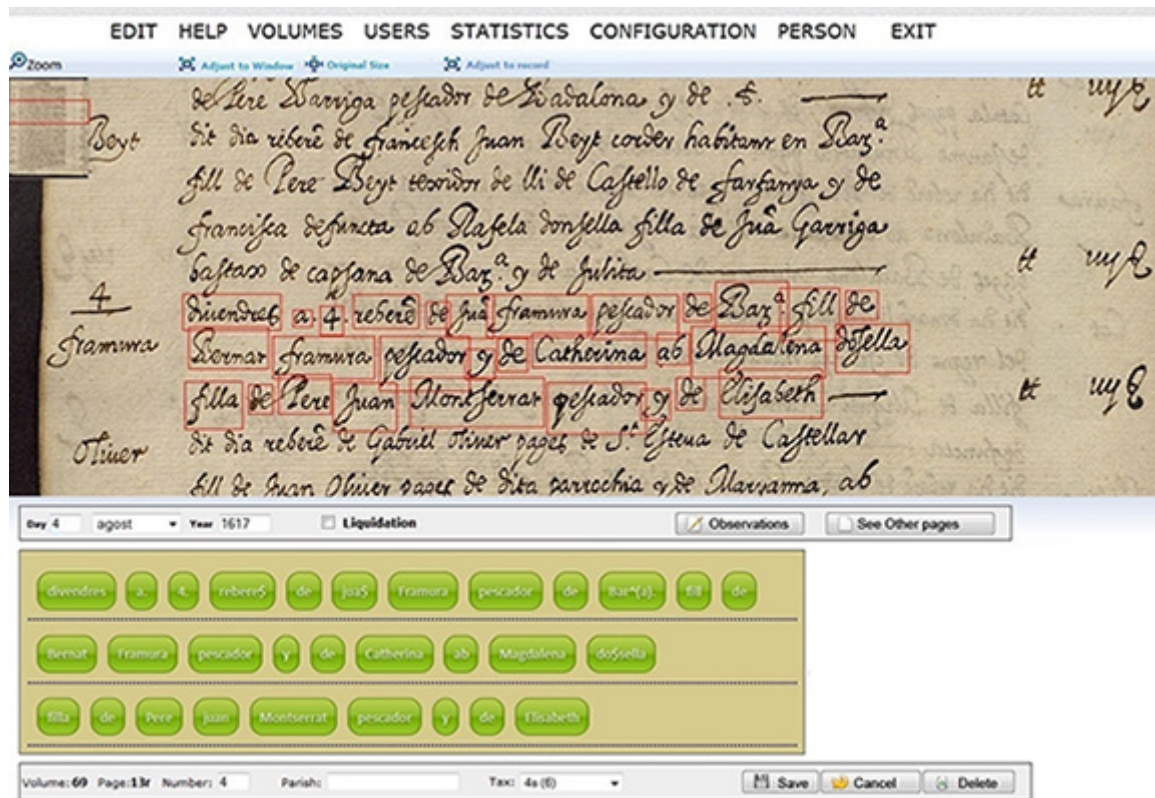
The algorithms for handwriting recognition require training and validation data. This is called the "ground truth" and consists of images labeled with the geometric position of individual lines and words and their corresponding transcription. The former is used for the validation of the segmentation algorithms, and the latter for the recognition ones.

The crowdsourcing platform is used to generate the ground truth. First, it has an interface to manually segment and transcribe line by line. Second, it contains a simple graphical interface that allows the user to draw bounding boxes for each word and store the exact position. For each word, additional metadata is recorded. This metadata corresponds to the semantic type of word (groom's name and

surname, bride's name and surname, occupation etc.) and its corresponding transcription. It is important to notice that the recorded transcription does not necessarily coincide with the one written by the users in the transcription module. In the ground truth module the transcription must be literal, without correcting spelling errors or variations in prefixes and suffixes. Handwriting recognition software generates the output as it appears in the input, so the benchmarking data must be literal too.

Figure 7 shows a snapshot of this module. In the upper part the document image is shown with the bounding boxes for the words. In the central part, the system shows the information corresponding to the people appearing in the register, as it was entered in the transcription module. In the bottom part, the words contained in each text line are shown. The user clicks every word in the text line and draws the bounding box that contains it - the red rectangle. If the word is related to some field in the form, the user validates the correspondence. Contrary, if the word does not correspond to any field in the form, it is automatically labeled as "other" in the ground truth. For example, the first word of the first line "divendres", is attributed to the category "other", while the sixth word of the first line, "jua\$", is attributed to the "groom name" category.

Figure 7 Image "ground truth" module.





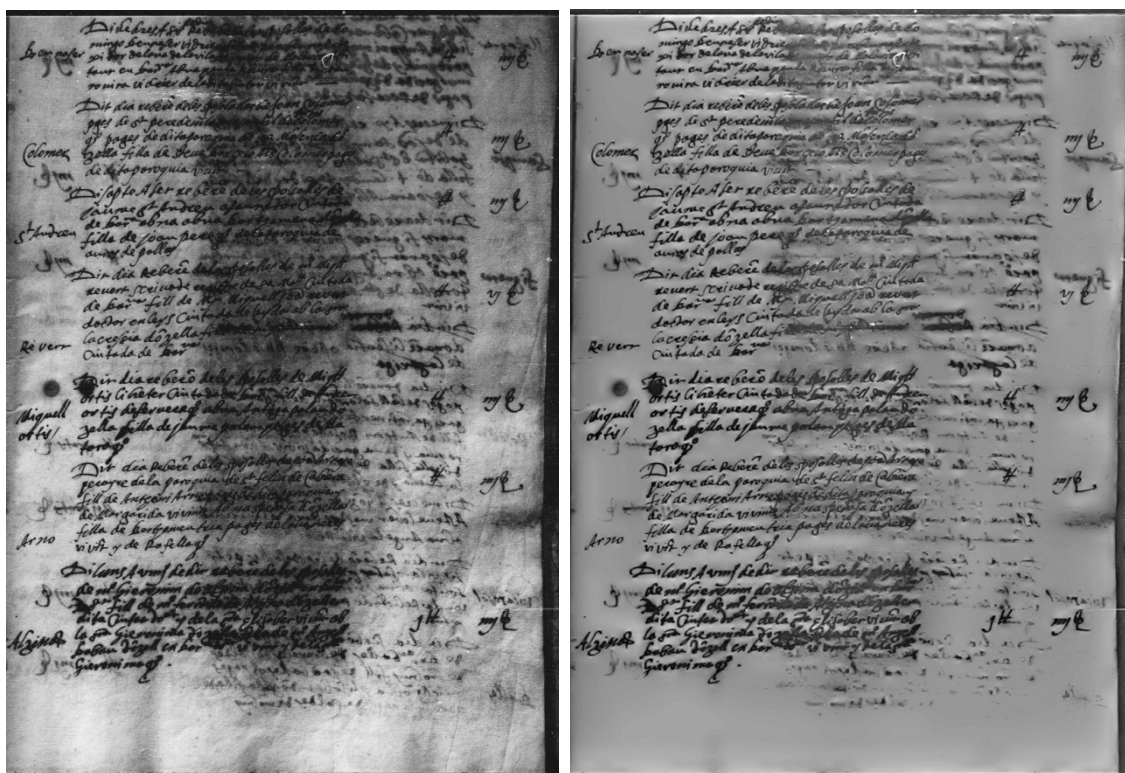
## 4 AUTOMATIC PROCESSING

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. Next, we will describe the main tasks that have been addressed concerning the marriage licenses.

### 4.1 DOCUMENT ENHANCEMENT

The recognition of historical documents implies dealing with the degradation of paper due to paper aging. In case of double-sided documents a common problem is the show-through effect: the back side of the document interferes with the front side because of the transparency of the document or ink bleeding. Since some of the marriage license books present severe show-through, we have proposed a show-through cancellation method (Fornés, Otazu & Lladós 2013). We hypothesize that show-through are low contrast components, while foreground components are high contrast ones. Our method has three main steps. First, we decompose the image into a Multiresolution Contrast representation, which allows us to obtain the contrast of components at different spatial scales. Second, we enhance the image by reducing the contrast of low spatial frequency components. Finally, we cancel the show-through phenomenon by thresholding these low contrast components. This decomposition is also able to enhance the image - removing shadowed areas by weighting spatial scales. As a result, the enhanced documents are more readable, and consequently, better transcribed. Similarly, the performance of handwriting recognition and word spotting algorithms are also improved. An example of a resulting enhanced document image can be seen in figure 8.

Figure 8 Original document to the left and enhanced document to the right.



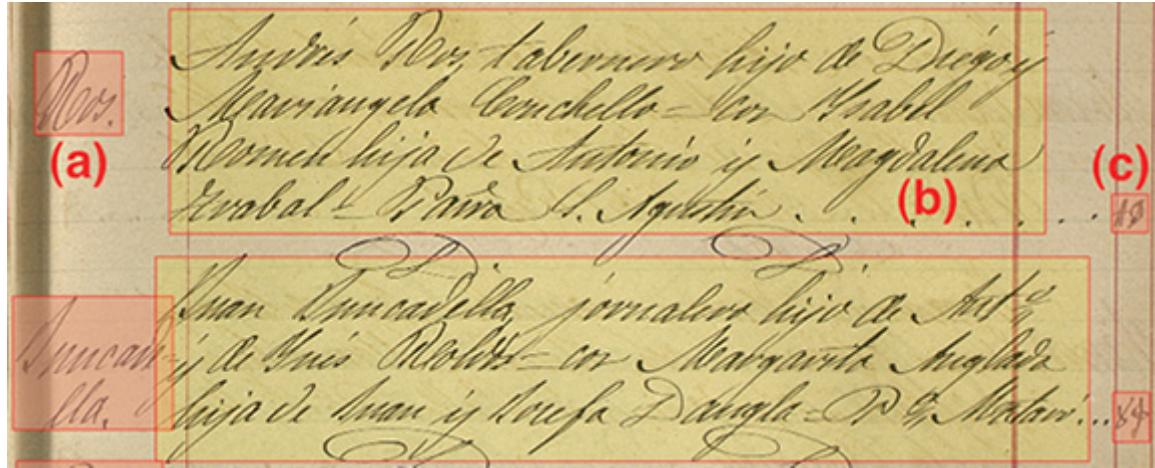
### 4.2 LAYOUT ANALYSIS AND LINE SEGMENTATION

Prior to the recognition of text, it is necessary to analyze the layout of the document in order to detect the text blocks and segment the text lines. In case of heterogeneous documents, it may be necessary to separate text from graphic elements. 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. The method proposed (Cruz & Ramos-Terrades 2012) first encodes the spatial relationships between elements using



Relative Location Features based on probability maps, and computes textural features based on Gabor filters. Then, a Conditional Random Field framework is used for learning the segmentation of the different elements from a collection of ground truth documents. An example of the resulting image is shown in figure 9.

Figure 9 Layout analysis example. a: Family name, b: Body, c: Fee



Once the main text block has been detected, the next step segments the text into text lines. Since many marriage license records contain touching lines and horizontally overlapping characters, we have proposed a graph-based approach for segmenting lines (Fernández et al. 2012). First of all, a graph is constructed using the skeleton of the background of the image. Then, a path-finding algorithm is used to find the best path which corresponds to the minimum cost path between each pair of potential starting and ending nodes. Local cost functions associated with the graph nodes are defined to find the best continuation path in terms of the potential configurations. In order to solve the problem of touching text lines, we add virtual edges between the candidate nodes that are around the involved characters. An example is shown in figure 10.

Figure 10 Example of line segmentation result.




### 4.3 HANDWRITING RECOGNITION

Once the text lines have been segmented, the next step is to transcribe them. In the case of handwritten text recognition, it is generally impossible to reliably isolate the characters, or even the words, that compose a cursive handwritten text. Therefore, the recognition of the text line is usually performed as a whole, without segmenting the image into either characters or words. Current state-of-the-art technologies for handwritten text line recognition are based on Hidden Markov models (HMMs) (Rabiner 1989) and Neural Networks (NNs) (Graves et al. 2009), which come from the field of speech recognition. First, images of handwritten text lines are normalized and encoded into sequences of local feature vectors. Then, these sequences are modeled with HMMs at the lexical level. Concerning NNs, they have been shown to obtain higher recognition rates and better robustness against noise in the data. However, the number of parameters to train in such neural networks is many orders of magnitude larger. Hence, if little training data is available, neural networks do not perform well. A comparison of HMMs and NNs for the recognition of the old marriage registers books can be found at (Fernández et al. 2013).

In order to improve the recognition rates, language models and dictionaries are typically integrated into the recognition process. Some examples of language models are n-grams, which estimate the probability of word sequences and grammar, defining the syntactically valid sequences of word categories and structure. Nominative sources, such as marriage registers, have some regularities that can be exploited in order to improve the recognition process. Thus, we have proposed a grammar that not only takes advantage of the order of appearance of the information in the marriage record, but also is able to associate a semantic category to each recognized word and fill the corresponding database. In figure 11 the system could identify that the word “Maryanna” corresponds to the name of the bride, the word “texidor” (weaver) corresponds to the occupation of the father of the bride, etc.

Finally, concerning dictionaries, the goal is to return the most similar word that exists in a predefined thesaurus. As a result, many ambiguities in the shape of the handwriting can be corrected during the recognition step. However, since the use of a dictionary containing all the existing words is computationally infeasible, we must use a limited set. Thus, any language model's performance that uses a thesaurus is upper-bounded by the impossibility of recognizing out-of-thesaurus words. Since the probability of the appearance of new words in the marriage licenses is very high, we have proposed a hybrid grammar (Cirera et al. 2013) that is able to recognize new words appearing in the following categories: names, surnames, occupations and places. In such categories the system is allowed to recognize, character by character, an unknown new word by entering into a character loop. As an illustration example, figure 11 shows that the Hybrid Grammar is able to recognize the surname “Valta”, although the thesaurus of surnames does not contain this word.

Figure 11 *Example of recognition of new words. Contrary to other methods, the hybrid grammar can correctly recognize the surname “Valta”, although this word was not in the thesaurus of surnames.*



Word n-gram	Maryanna donsella filla de dita Vila texidor
Grammar	Maryanna donsella filla de Juâ Valls texidor
Hybrid Grammar	Maryanna donsella filla de Juâ Valta texidor

### 4.4 WORD SPOTTING

In the previous section we have shown how the automatic extraction of information from the marriage records could be feasible. However, these handwriting recognition techniques require large amounts of annotated images to train the recognizer. Therefore, 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. In this scenario, the user selects either one or a few words by looking at the data set, and the system retrieves all words with a similar shape.

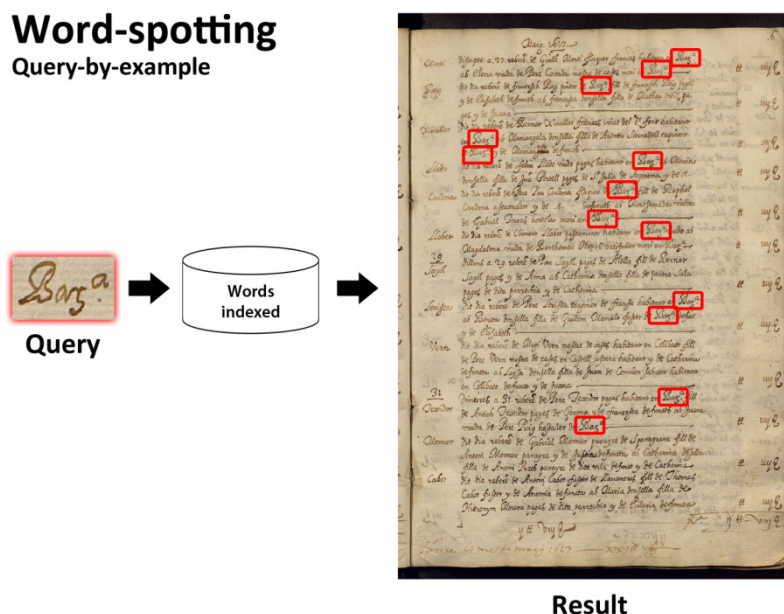


The first advantage is that word spotting can be performed on-the-fly: the user can crop a word in a new document collection, and the system will start searching for similar words without any training step. The second is that, since the query word is treated as a shape, shape-based word spotting approaches are also able to retrieve graphical elements, such as symbols, seals, or stamps.

In the case study of marriage registers, we have proposed different word representations (Lladós et al. 2012) to assess the advantages and disadvantages of statistical and structural models, see figure 12. The first approach is based on a bag-of-visual-words (BoVW) representation. First of all, we randomly select a reference set of some of the word images in order to perform a clustering of the SIFT descriptors based on gradients to build the codebook. Once we have the codebook, the word images are encoded by the BoVW model. In a last step, in order to produce more robust word descriptors, we add some coarse spatial information to the orderless BoVW model. This approach is segmentation-free, so there is no need to segment the text into either lines or words. The second approach is a pseudo-structural model based on a Loci features representation. For each pre-segmented word image, we compute a descriptor based on LOCI features, which consists of a histogram of intersections between background and foreground pixels. Afterwards, a hash-table is used for storing all these feature vectors, allowing a fast search of the query word image. The third word spotting approach consists of a structural representation based on graphs. From the skeleton of the word image, a graph is first constructed. Then, the graphs are represented using uni-dimensional structures. Finally, they are clustered in order to search and compare graphs in a fast way. From the experiments we could conclude that the statistical representations usually obtain better results, although its high memory requirements can be a problem in the case of large databases.

The performance of word spotting approaches can be improved in several ways. For example, the combination of statistical and semi-structural methods has shown good performance while maintaining the advantages of both types of methods (Almazán et al. 2012). In addition and following the idea of language models for handwriting recognition, a contextual word spotting approach (Fernández et al. 2013) can improve the initial word spotting results thanks to the knowledge of the structure of the marriage records.

Figure 12 Example of word-spotting using the word Barcelona.



## 4.5 DISCUSSION

In this section we have described the research in document image analysis conducted for an automatic transcription of manuscripts. Two major conclusions can be drawn. First, it is important to remark that handwriting recognition is 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. More intelligent systems which use linguistic and semantic context, like humans do, should be developed. A second conclusion is related to the ability of the proposed system to be transferred to other objects of study (countries, times, languages). The use of the developed tools to transcribe, for example, Hungarian or Portuguese documents would require the adaptation to different dictionaries for such languages.

## 5 CONCLUSION

In 1968 social historian Emmanuel le Roy Ladurie wrote about the future of his discipline: “the historian of tomorrow will be a programmer or he will be nothing” (le Roy Ladurie 1973, p. 14). Critics of this statement tend to interpret this too broadly; he obviously meant quantitative historians. Ladurie was right in the sense that program development would be necessary for successful work in such disciplines as historical demography, and over the years program packages have emerged that make quantitative and database research easier for us. However, IT specialists, rather than historians, constructed these program packages, which have enabled the researchers to work efficiently with their source materials and their research questions rather than spend time on the kind of third generation programming that was usual among quantitatively oriented historians a few decades ago. Now we see a continuation of this trend, most importantly, in the area of turning graphical images of source files into transcribed textual contents. Again, a distribution of efforts between IT specialists and historians is efficient in order that the software we need becomes flexible and reliable.

Longitudinal and vital event databases, even more than cross-sectional databases with the same coverage in time and space, are resource demanding with respect to transcription of the source materials. Therefore, it is crucial that alternative methods are considered for optimal cost-efficiency when deciding how to spend the available resources. In addition to own employees, paid and unpaid crowd-sourcing and use of transcribers in low-cost countries and the development of more rational transcription techniques can help increase the transcription rates. While optical character recognition is available as off-the-shelf software, machine interpretation of hand-written images is still in a developmental stage – especially with regard to the contents of unstandardized, historical sources. This article has explained how two separate projects try to implement new computer vision routines for the 1891 individual census sheets in Norway and the marriage protocols from 1451 to 1904 from Barcelona.

Both cases are more concerned with segmenting relevant parts and recognizing them by document image analysis and recognition techniques than with translating the contents to ASCII characters as is done with OCR methods for printed texts. The Norwegian software finds the eight to ten underlined information fields and automatically interprets them so that in most cases the gender, marital status, etc. of the persons are known before the more complicated fields with names, occupations and birthplaces are transcribed manually. Since these source entries exist as independent images they are transcribed separately with software that speeds up the process, for instance, because the gender is known and pick lists are constructed from name frequencies based on the 1900 census. Pick lists can make transcription more error-prone, however.

The Five Centuries of Marriages project has transcribed the marriage protocols with paid crowd-sourcing via the Internet by building a full-fledged data-entry program package containing modules for transcription, administration, document enhancement, layout analysis, handwriting recognition and word spotting. The Barcelona project brings these techniques further by isolating each word in the marriage entry and attempt to match them with a thesaurus containing most words occurring in the protocols. In both cases many records have been transcribed with manual methods, providing training materials which are valuable also when transferring similar graphics techniques to other source materials.

A large-scale project to turn Norwegian literary classics into e-books might indicate how computer vision and manual transcription can be integrated also for handwritten materials in the future (Fjellberg 2013). This concerns printed materials. So after scanning, OCR is used to create one digital text version. The other version is created by transcribers contracted to key in the same Norwegian texts via keyboards in an operation organized by commercial partners in India. The OCR and the transcribed



versions are then compared word by word, character by character by software flagging all inconsistencies before final proofreading. A similar combination of transcriptions and output from computer vision software to deal with handwritten texts after proofreading would create not only valuable digital materials for research, but also ground truthed texts that can be used to enhance the functioning of the computer vision software. It is likely that an optimal solution will emerge by combining the strength of computer vision, low-cost labor and crowd-sourcing.

We have learned a lot from the fruitful interchange of ideas between social historians and IT experts. The competence of computer scientists is needed to define models for handwriting recognition and document interpretation. Issues like readability, harmonization of variables, record linkage, etc. have to be translated into technical methods as algorithms where historians and demographers need to work with computer scientists to reach the necessary level of precision. The main benefit to the technological researchers acquired from the collaboration with social scientists may be the latter's ability to understand the requirements and needs. This expertise not only comprises the skills for example in paleography to be able to identify the writing style, but also the knowledge about the historical context associated to the document. Since major parts of the computer vision methods are language independent, we believe that international cooperation in the field of semi-automating the transcription of handwritten sources might be in the interest of several historical database projects. We aim to release all software developed according to open source principles and will welcome in kind contributions of software modules from international partners.

## ACKNOWLEDGEMENTS

This article is part of the "Norwegian Historical Population Register" project financed by the Norwegian Research Council (grant # 225950) and the Advanced Grand Project<sup>11</sup> "Five Centuries of Marriages" (2011-2016) funded by the European Research Council (# ERC 2010-AdG\_20100407) and whose principal investigator is Professor Anna Cabré. The research team of the project on marriages includes researchers from the Universitat Autònoma de Barcelona and its Centre for Demographic Studies and Computer Vision Centre.

## REFERENCES

- Almazán, J., Fernandez, D., Fornés, A., Lladós, J. & Valveny, E. (2012). A coarse-to-fine approach for handwritten word spotting in large scale historical documents collection. *Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 453-458.
- Anderson, M. (1988). *The American Census. A Social History*. New Haven: Yale University Press.
- Cirera, N., Fornés, A., Frinken, V. & Lladós, J. (2013). Hybrid grammar language model for handwritten historical documents recognition. *Pattern Recognition and Image Analysis, Lecture Notes in Computer Science*, 7887, 117-124. Berlin/Heidelberg: Springer-Verlag.  
DOI: 10.1007/978-3-642-38628-2\_13
- Cruz, F. & Ramos-Terrades, O. (2012). Document segmentation using relative location features. *21st International Conference on Pattern Recognition*, 1562–1565.
- de Salazar, J. & Mayoralgo, J.M. (1991). *Génesis y evolución histórica del apellido en España*. Madrid: Real Academia Matritense de Heráldica y Genealogía.
- Eikvil, L., Holden, L. & Bævre, K. (2010). *Automatiske metoder som hjelp til transkribering av historiske kilder*. Oslo: Norsk regnesentral (Norwegian computing center).
- Estellés-Arolas, E. & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2), 189-200.
- Fernández, D., Manmatha, R., Lladós, J. & Fornés, A. (2012). On influence of line segmentation in efficient word segmentation in old manuscripts. *Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp.759-764.

11 ERC Advanced Grants allow exceptional established research leaders of any nationality and any age to pursue ground-breaking, high-risk projects that open new directions in their respective research fields or other domains. Cf <http://erc.europa.eu/advanced-grants>

- Fernández, D., Marinai, S., Lladós, J. & Fornés, A. (2013). Contextual word spotting in historical manuscripts using Markov logic networks. *2nd International Workshop on Document Imaging and Processing (HIP)*, pp. 36-43.  
DOI: [10.1145/2501115.2501119](https://doi.org/10.1145/2501115.2501119)
- Fjellberg, A. (2013 November 22). Hamsuns digitale reise. *Morgenbladet*.
- Fornés, A., Otazu, X. & Lladós, J. (2013). Show through cancellation and image enhancement by multiresolution contrast processing. *12th International Conference on Document Analysis and Recognition (ICDAR)*, pp.200-204.  
DOI: [10.1109/ICDAR.2013.47](https://doi.org/10.1109/ICDAR.2013.47)
- Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H. & Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5), 855-868.  
DOI: [10.1109/TPAMI.2008.137](https://doi.org/10.1109/TPAMI.2008.137)
- Haug, J. (1979). Manuscript census materials in France: The use and availability of the listes nominatives. *French Historical Studies*, 11(2) Autumn, 258-274.
- INSEE. (no date). *Le recensement de la population dans l'Histoire*.
- Jåstad, H. & Thorvaldsen, G. (2012). The incidence of consanguinity in Norway in the late 19th century. In: E. Beekink & E. Walhout. (Eds.). *Frans van Poppel: a sort of farewell: liber amicorum* (pp. 58-62). Den Haag: NIDI.
- le Roy Ladurie, E. (1973). *Le territoire de l'historien*. Paris: Gallimard.
- Lladós, J., Rusiñol, M., Fornés, A., Fernández, D. & Dutta, A. (2012). On the influence of word representations for handwritten word spotting in historical documents. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 26(5), 1263002.1-1263002.25.  
DOI: [10.1142/S0218001412630025](https://doi.org/10.1142/S0218001412630025)
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77(2), 257-286.  
DOI: [10.1109/5.18626](https://doi.org/10.1109/5.18626)
- Salinero, G. (2010). Sistemas de nominación e inestabilidad antroponímica moderna 9-27, G. Salinero, I. Testón. (Eds.) *Un Juego de Engaños. Movilidad, nombres y apellidos en los siglos XV a XVIII*. Casa de Velázquez: Madrid.
- Sermanet, P., Kavukcuoglu K. & LeCun, Y. (2009). EBLearn: Open-source energy-based learning in C++, *Proc. International Conference on Tools with Artificial Intelligence, IEEE*.  
DOI: [10.1109/ICTAI.2009.28](https://doi.org/10.1109/ICTAI.2009.28)
- Solli, A. & Thorvaldsen, G. (2012). Norway: From colonial to computerized censuses. *Revista de Demografia Historica*, XXX (I), 107-136.
- Statistics Norway. (1895). *Fremgangsmåden m.v. ved den i Januar 1891 afholdte Folketælling*. [Instructions for the 1891 census.] Kristiania. Last accessed: 5/1/2015
- Thorvaldsen, G. (2011). Using NAPP Census Data to Construct the Historical Population Register for Norway. *Historical Methods*, 44(1), 37-47.  
DOI: [10.1080/01615440.2010.517470](https://doi.org/10.1080/01615440.2010.517470)
- US Census Bureau. (no date). *The Hollerith Machine*. Last accessed: 5/1/2015

## Appendix Automatic Transcription of Historical Sources

*This is an English summary (by Gunnar Thorvaldsen) of a report from the Norwegian Computing Center authored by Line Eikvil, Lars Holden and Kåre Bævre in 2010, cf [http://www.rhd.uit.no:8080/nhdc/HBR\\_notat\\_okt-2010.pdf](http://www.rhd.uit.no:8080/nhdc/HBR_notat_okt-2010.pdf)*

Throughout this report, we considered various automatic techniques that may be relevant to aid in transcription of historical sources, and identify current techniques applied to similar problems. We have looked at partitioning the questionnaire structure, color separation and recognition of handwritten text.

Automatic partitioning of Scheme structure is identified as a basic element in a system providing automated tools to aid transcription. Automatic recognition in the traditional OCR sense is most appropriate for numbers, while for continuous handwriting it is not yet possible to achieve high detection rates. Here we suggest instead the use of techniques that can support the manual process. In addition, we suggest methods that can reduce the problem by using information from other sources to limit the number of possible interpretations for a given word or a number. Sometimes this can reduce the detection problem to a verification problem, immediately making automatic techniques more appropriate even for text.

Good internet solutions with wiki functionality can facilitate the recruitment of voluntary manual transcribers to help with transcription work. How such resources can be used will depend on whether the specific archival series are open or closed to the public.

Automatic partitioning of the questionnaire structure is identified as a basic element in a system to provide automated tools to help the transcribing process. Regardless of to what degree manual or automatic transcription is used further on in the process, such partitioning will be useful because it provides an opportunity to work with the transcription of the forms, either row by row, column by column or cell by cell. Therefore, we have looked into how this can be resolved and have implemented a prototype designed for the analysis of household forms from the 1950 census.

Automatic detection in the traditional sense is most appropriate for numbers, while for continuous handwriting it is not possible to achieve high detection rates. Here we suggest instead using techniques which can be a support to the manual process, such as automatic grouping or retrieval of similar words. In addition, we suggest, where possible, to reduce the problem by limiting lists. The use of information from other sources and linking this information is a way to shorten the dictionaries, thus reducing the detection problem to a verification problem and making automatic techniques more appropriate for text.

In addition to the solutions mentioned above, interactive Internet solutions, such as the Historical Population Register Wiki, enable the recruitment of genealogists and local historians to assist in the automatic methods for transcribing of historical sources. For open archives such resources should be utilized straightforwardly for verbatim transcription. As for archives with limited access for legal reason, the partition of forms into isolated cells and anonymization will be the key to legally outsource the transcription to the public or low-cost countries. The mobilization of such transcribers can be done with forced "reCAPTCHA" transcription to get access to genealogical resources, volunteer transcription with a competitive element or with a salary.

In conclusion, the choice of the right automated techniques and tools can ease the manual labor, reducing the cost of transcription of historical sources substantially. Good tools and Internet solutions can also provide an opportunity to utilize large volunteer resources in transcription, which can reduce both costs and the time it will take to transcribe large archives.