Using the MGGI Methodology for Category-based Language Modeling in Handwritten Marriage Licenses Books

Verónica Romero*, Alicia Fornes†, Enrique Vidal* and Joan Andreu Sánchez*
*PRHLT Research Center, Universitat Politècnica de València, Spain
{vromero,evidal,jandreu}@prhlt.upv.es
†Computer Vision Center, Universitat Autónoma de Barcelona, Spain
afornes@cvc.uab.es

***Abstract*—Handwritten marriage licenses books have been used for centuries by ecclesiastical and secular institutions to register marriages. The information contained in these historical documents is useful for demography studies and genealogical research, among others. Despite the generally simple structure of the text in these documents, automatic transcription and semantic information extraction is difficult due to the distinct and evolutionary vocabulary, which is composed mainly of proper names that change along the time. In previous works we studied the use of category-based language models to both improve the automatic transcription accuracy and make easier the extraction of semantic information. Here we analyze the main causes of the semantic errors observed in previous results and apply a Grammatical Inference technique known as MGGI to improve the semantic accuracy of the language model obtained. Using this language model, full handwritten text recognition experiments have been carried out, with results supporting the interest of the proposed approach.**

***Keywords*-Handwritten Text Recognition, Information extraction, Language modeling, MGGI, Categories-based language model.**

I. Introduction

Historical records of daily activities provide intriguing insights into the life of our ancestors, useful for demography studies and genealogical research [1]. Handwritten marriage licenses books [2] are one of these records that have been used for centuries by ecclesiastical and secular institutions to register marriages. The information contained in these historical documents is very interesting for migratory studies, population research and genealogical investigation. Therefore, one of the goals of this kind of documents, rather than to transcribe perfectly the documents, is to extract the relevant information to allow the users to make use of it through semantic searches. Note that, if the perfect transcript is obtained, then identifying the relevant semantic information would be much easier, but it is not mandatory to obtain the perfect transcript.

For typical handwritten text images of historical documents, currently available text image recognition technologies are not suitable. Traditional Optical Character Recognition (OCR) is simply not usable since the linguistic components like characters, words or sentences can not be isolated automatically. Therefore holistic approaches that do not need prior segmentation are needed [3]. Thus, HTR of historical documents is currently based on techniques that have been used in Automatic Speech Recognition. In this way, Hidden Markov Models (HMM) [4] or hybrid HMM and Artificial Neural Networks (ANN) [5] are used for representing optical models, and n-gram models for language modeling.

The language model plays a fundamental role in the HTR process, by restricting significantly the search space. For tasks with a vocabulary of medium size current HTR state-of-the-art prototypes provide word error levels that roughly range from 10 to 40% [4], [2]. Although the training of the optical models is still an incipient research field, significant improvements can be obtained by using better language models. For example, in [6], given the regular structure of marriage licenses documents, the use of a category-based language model [7] for both better representing the regularities in marriage license books and for obtaining the relevant semantic information of each record was studied. In contrast to this, there are works where the semantic interpretation of the recognition output is carried out in a second step using natural language understanding techniques [8].

In this paper, we follow the first approach and analyze the main semantic errors occurred using category-based language models and apply a Grammatical Inference technique known as "Morphic Generator Grammatical Inference" (MGGI) [9] to improve the semantic accuracy of the language model obtained. In MGGI, a-priory knowledge is used to label the words of the training strings in such a way that a simple bigram can be trained from the transformed strings. The knowledge used allows the MGGI to produce a language model which captures dependencies of the language underlying in the handwritten records considered.

II. Task description

In this paper we used a book from a collection of Spanish marriage license books conserved at the Archives of the Cathedral of Barcelona and described in [2]. Fig. 1 shows a page of marriage licenses from the book used in this paper. Each marriage license typically contains information about the marriage day, groom's and bride's names, the groom's occupation, the groom's and bride's former marital status,

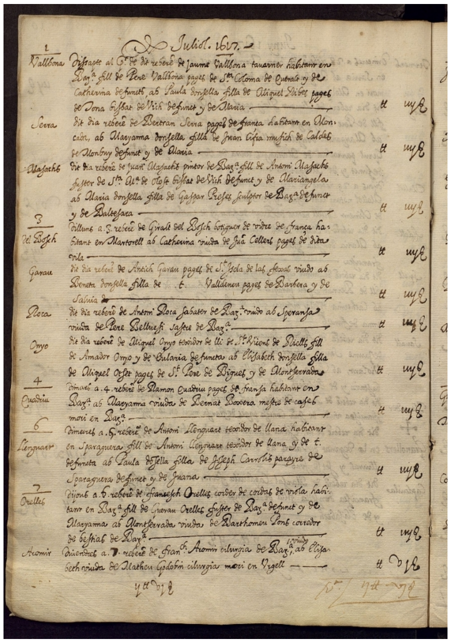


Figure 1. Example of a marriage license page.

and the socio-economic position given by the amount of the fee. This information is not written randomly but the opposite. The groom's information is written first and then the bride's information. Inside the groom's information, the given name and surnames are written first, then the birth town and then the occupation. Then the groom's father information is in a similar order, and then the bride's information. In some cases, additional information is given as well as information about a deceased parent. This structure suggests that the vocabulary changes along the license: the first part is related to the groom, with names related to men and occupations, whereas, the last part is the bride's part. Fig. 2 shows an example of an isolated marriage license.

A problem when transcribing handwritten marriage license books by means of HTR methods is that the classical n-gram language models can be very inaccurate due to the special vocabulary of the task, which is composed mainly of proper names (given names, surnames, town names, etc.). A classical n-gram language model can have difficulties to predict the probability of a word if the word to be predicted is a proper name. For example, consider a license that starts with the following sentence referred to the groom:

```
Dit dia rebere$ de Raphel Joani texidor de lli
de Vilassar ...
```

The translation of this sentence is

```
That day we received from Raphel Joani linen weaver
from Vilassar ...
```

Note that is quite difficult to predict the word `Raphel` from the previous words since any (groom's) given name can appear in this position. Something similar occurs for other words, like `Joani`, (groom's surname) `linen` or `Vilassar` (groom's town). However, if the groom's given name is categorized, the number of contexts in the n-gram model is reduced and, therefore, is easy to predict the correct word. This is the idea described in the following section.

III. Category-based HTR

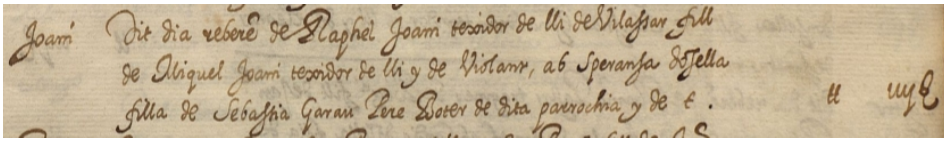
As shown in [6], the use of a category language model in the handwritten text recognition process can benefit both, the handwritten accuracy and the semantic information extraction process. This improvement is due to two main reasons. Firstly, given that category-based language models share statistics between words of the same category, category-based models are able to generalize to word patterns never encountered in the training corpus. Secondly, grouping words into categories can reduce the number of contexts in an n-gram model, and thereby reduce the training set sparseness problem.

In this paper, following the ideas presented in [6], some categories have been defined taking into account the semantic information included in the licenses: groom's (*Gr*) given name and surname, bride's (*Br*) given name and surname, parents' (*Fa* and *Mo*) given names and surnames, occupations (*Oc*), place of residence (*Resi*), geographical origin, etc. Then, a category-based language model was generated and integrated into the handwritten text recognition process. In the next text, the annotated license corresponding to the image in Figure 2 is shown. Each semantic label (marked into brackets) is immediately after the relevant word:

```
Dit dia rebere$ de Raphel[GrName] Joani[GrSurname]
texidor_de_lli[GrOc] de Vilassar[GrResi] fill
de Miquel[GrFaName] Joani[GrFaSurname]
texidor_de_lli[GrFaOc] y de Violant[GrMoName],
ab Sperensa[BrName] do$sella filla de
Sebastia_Garau[BrFaName] Pere[BrFaSurname]
Boter[BrFaOc] de dita_parrochia[BrFaResi] y
de t.[BrMoName]
```

As shown in the example, only some words of each license had relevant semantic information, the proposed categorization involved classifying only some words in the vocabulary and not all of them. In this way, a partially categorized corpus was obtained, that is, not each word had a category associated to it. Words that had not a category could be viewed as categories that contain a single word. For instance, we can introduce the category "DIA" containing only the word "dia". On the other hand, a word may belong to several categories. For example, the word *Ferrer* (that could be translated as Smith) could belong to the categories *husband surname*, *husband profession*, *father husband surname*, *father husband profession*, *bride surname*, *father bride surname*, etc.

Formally speaking, let $\mathbf{x} = x_1 \ x_2 \ \ldots \ x_m$ be a handwritten sentence image represented by a feature vector sequence.The HTR problem is formulated as the problem of



Joani Dit dia reberé de Raphel Joani texidor de lli de Vilassar fill
de Miquel Joani texidor de lli y de Violant, ab Speransa dozella
filla de Sebastia Garau Pere Boter de dita parrochia y de t. ll iiij

Figure 2. Example of a marriage license.

finding the most likely word sequence, $\mathbf{w} = w_1\ w_2\ \ldots\ w_l$, i.e., $\mathbf{w} = \arg\max_{\mathbf{w}} \mathrm{P}(\mathbf{w} \mid \mathbf{x})$. Using the Bayes' rule we can decompose this probability into two probabilities, $\mathrm{P}(\mathbf{x} \mid \mathbf{w})$ and $\mathrm{P}(\mathbf{w})$, representing optical-lexical knowledge and syntactic knowledge, respectively:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \mathrm{P}(\mathbf{w} \mid \mathbf{x}) = \arg\max_{\mathbf{w}} \mathrm{P}(\mathbf{x} \mid \mathbf{w}) \cdot \mathrm{P}(\mathbf{w}) \tag{1}$$

$\mathrm{P}(\mathbf{x} \mid \mathbf{w})$ is typically approximated by concatenated character models, usually HMMs [10], while $\mathrm{P}(\mathbf{w})$ is approximated by a language model, in this work we use a category-based language model [6].

By further considering the sequence of semantic categories, $\mathbf{c} = c_1\ c_2\ \ldots\ c_l$, associated to the word sequence as a hidden variable in equation (1), approximating the sum by the dominating term, and following the same assumptions presented in [6], we can rewrite previous equation as:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \mathrm{P}(\mathbf{x} \mid \mathbf{w}) \sum_{\mathbf{c}} \mathrm{P}(\mathbf{c}, \mathbf{w}) \tag{2}$$

$$\approx \arg\max_{\mathbf{w}} \mathrm{P}(\mathbf{x} \mid \mathbf{w}) \max_{\mathbf{c}} \mathrm{P}(\mathbf{c}) \cdot \mathrm{P}(\mathbf{w} \mid \mathbf{c}) \tag{3}$$

$$\approx \arg\max_{\mathbf{w}} \mathrm{P}(\mathbf{x} \mid \mathbf{w}) \max_{\mathbf{c}} \prod_{i=1}^{l} \mathrm{P}(c_i \mid c_{i-1} ... c_{i-n+1}) \cdot \mathrm{P}(w_i \mid c_i) \tag{4}$$

where $\mathrm{P}(w_i \mid c_i)$ is computed from the word-category distribution and $\mathrm{P}(c_i \mid c_{i-1} c_{i-2} ... c_{i-n+1})$ is computed from an n-gram of categories.

Finally, as explained in [11], from the decoding process, we can obtain not only the best word sequence hypothesis, but also the best sequence of semantic categories used in the most probable sentence:

$$(\hat{\mathbf{c}}, \hat{\mathbf{w}}) \approx \arg\max_{\mathbf{c},\mathbf{w}} \mathrm{P}(\mathbf{x} \mid \mathbf{w}) \cdot \mathrm{P}(\mathbf{w} \mid \mathbf{c}) \cdot \mathrm{P}(\mathbf{c}) \tag{5}$$

IV. Language Modeling using MGGI

It is well known that n-gram models are just a subclass of probabilistic finite-state machines (PFSM) [12], [13]. Therefore the capabilities of n-grams to model relevant language contexts or restrictions is limited, not only with respect to more powerful syntactic models such as context-free grammars, but also even with respect to the general class of PFSMs. In fact, no n-gram can approach (word) string distributions involving the kind of long-span dependencies which are common in natural language. For instance, no n-gram (with bounded n) can approach a distribution of strings over the vocabulary $\{a, b, c, d, e\}$ such that the probability is high for the strings ab^ic or db^ie and is low or null for other strings such as ab^ie and db^ic, where i is any arbitrarily large integer. However, such a distribution can be exactly modeled by a very simple PFSM (see [13], Sec. 2.1.3).

While learning PFSMs from training strings is in general hard, there is a not-very-well-known framework which allows to learn PFSMs which can model *given*, albeit arbitrarily complex (finite-state) restrictions. This framework, known as "Morphic Generator Grammatical Inference" (MGGI), provides a methodology for using prior knowledge about the restrictions which are interesting for the task in hand, to ensure that the trained finite-state models will comply with these restrictions. For instance, in the previous example, one can rely on the known legal starting and ending words to obtain a PFSM which accurately approaches the distribution aimed at (see [13], Sec. 2.2).

MGGI was introduced in 1987 [14], within the framework of *Grammatical Inference* for Syntactic Pattern Recognition. It is based on the well known "morphism theorem of regular languages [15], which states that every regular language (generated or accepted by a finite-state machine) can be obtained by applying an appropriate word-by-word morphism to the strings of a *local language* over some suitable vocabulary. A probabilistic extension of this theorem is given in [13], where it is also shown that a probabilistic local language is exactly the same as a bigram language model.

In MGGI, a-priory knowledge is used to label the words of the training strings in such a way that a simple bigram can be trained from the transformed strings. Then an inverse transformation (the *morphism*) is applied to this bigram to obtain a PFSM which deals with the restrictions conveyed by the initial string transformation [14], [13]. A direct application of these ideas to build accurate PFSM language models for automatic speech recognition can be seen in [9].

In this work, rather than using a plain bigram trained on raw word strings, a category-based bigram is used for the application of MGGI. This way, the resulting PFSM will additionally provide the helpful generalizations entailed by word categorization. The knowledge used to define and label relevant word categories is expanded with additional word labels which allows the MGGI to produce a PFSM which captures important dependencies of the language underlying in the handwritten records considered.

To this end, we checked the most frequent errors committed by a standard category-based bigram, such as the one used in [6]. One of the most common errors, clearly due to a wrong bigram generalization, was the mis-categorization of the bride's family information as groom's information. The following example shows an example of this kind of errors, where the bride's father name has been wrongly labeled as the groom's father name and the same occurred with the surname and the profession:

```
... ab Sperensa[BrName] do$sella filla de
Sebastia_Garau[GrFaName] Pere[GrFaSurname]
Boter[GrFaOc] de dita_parrochia[BiFaResi] y
...
```

This clearly happened because the bigram *"de [GrFaName]"* had higher probability than the bigram *"de [BrFaName]"*, since groom's family information appears more often than that bride's family information. This suggests that a better generalization of the training text could be achieved by just tagging all the text tokens (categories and words) with labels that help distinguishing their relative position in the record.

In the vast majority of the records considered, the groom's and bride's information are separated by the word *"ab"* ("with" in English). Therefore, it is straightforward to label all the tokens which precede the word *"ab"* with the suffix *"G"* and those appearing after *"ab"* with *"B"* (meaning that the informations correspond to the Groom and the Bride, respectively). By applying this labeling scheme to the categorized training transcripts of the license of the Figure 2, the following training text is obtained:

```
DitG diaG rebere$G deG [GrName]G [GrSurname]G
[GrProf]G deG [GrResi]G fillG deG [GrFaName]G
[GrFaSurname]G [GrFaProf]G yG deG [GrMoName]G
,G ab [BrName]B do$sellaB fillaB deB
[BrFaName]B [BrFaSurname]B [BrFaProf]B deB
[BrFaResi]B yB deB [BrMoName]B
```

After training a category-based bigram, the inverse transformation required by MGGI (the word-by-word morphism) consists just in removing these suffixes *"G"* and *"B"*. The resulting PFSM adequately models the dependencies conveyed by the labeling adopted.

V. Experimental Framework

To assess how using the MGGI for category-based language modeling can benefit the handwriting recognition and the semantic information extraction, different experiments were carried out. The corpus, the assessment measures and the obtained results are explained next.

A. Corpus

The experiments were performed on the publicly available ESPOSALLES[1] database [2], which was compiled from a marriage license book conserved at the Archives of the Cathedral of Barcelona. We used the version labeled with the semantic information presented in [6].

The corpus was written in old Catalan by only one person between 1617 and 1619. It is composed by 173 pages that contain 5,447 lines grouped in 1,747 licenses. The whole manuscript was transcribed line by line by an expert paleographer. The complete annotation contains around 60,000 running words from a lexicon of around 3,500 different words. The database was also labeled with the semantic information of the licenses. 40 different categories were defined by demographer experts and the relevant words in each license were manually labeled with the corresponding category as shown in section III.

Table I
Basic statistics of the database and average values from the 7 different partitions.

Number of:	Total	Average
Pages	173	24.7
Licenses	1,747	249.6
Lines	5,447	778.1
Run. words	60,777	8682.4
OOV	–	361
Lexicon	3465	1070
Semantic labels	21386	3055.1

The standard partition proposed in [2], consisting of seven consecutive blocks of 25 pages has been used in the experiments. Table I shows the average values of the statistics related with the different partitions.

B. System setup

The seven different partitions were used in these experiments for cross-validation. That is, we carried out seven rounds, with each of the partitions used once as test data and the remaining six partitions used as training data.

The pages were divided into line images as explained in [2]. Then, appropriate filtering methods were applied to remove noise, improve the quality of the image and to make the documents more legible. Afterwards, the skew and the slant of each line were corrected. Finally the size was normalized separately for each line.

Each preprocessed line image was represented as a sequence of feature vectors. In this work we used the features described in [4] based on the gray level of the image. As explained in [2], the ESPOSALLES database is provided at two different levels: line level and license level. In this work we have carried out experiments at license level. Given that the lines belonging to each license was known, feature sequences extracted from the lines could be easily merged into whole license line images.

The characters were modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. These models were estimated from training text images represented as feature vector sequences using the Baum-Welch algorithm.

A category-based bi-gram was estimated using the MGGI methodology from the training transcriptions of the text line

[1] It is publicly available at: http://www.cvc.uab.es/5cofm/groundtruth

images. The out of vocabulary (OOV) words (words of the test partition that do not appear in the training partition) belonging to a category seen in training were added as singletons to the corresponding word category distribution. For OOV words that belong to a category that has not been seen in training, we add the category in the category-based 1-gram and the word in the category distribution as singleton. The word category distributions were modeled by uni-grams. The decoding was carried out by the Viterbi algorithm [10].

C. Assessment Measures

Different evaluation measures were adopted to assess the HTR and the relevant information extraction performance. The quality of the transcription is given by the well known *Word Error Rate* (WER). It is defined as the minimum number of words that need to be substituted, deleted or inserted to convert the sentences recognized by the system into the reference transcriptions, divided by the total number of words in these transcriptions.

On the other hand, to asses the quality of the information extraction performance we have used the standard precision and recall measures. We define precision and recall in terms of the number of relevant words in the dataset and the number of relevant words retrieved by the system. Relevant words are those words that belong to one of the 40 relevant categories defined by demographer experts. For instance, in the example shown in previous sections, the relevant words are those associated to a category: *Raphel, Joani, texidor_de_lli, Vilassar, Miquel, ...* Let R be the number of relevant words contained in the document, let D be the number of relevant words that the system has detected, and let C be the number of the relevant words correctly detected by the system. Precision (π) and recall (ρ) are computed as:

$$\pi = \frac{C}{D} \qquad \rho = \frac{C}{R}$$

Finally, it must be said that we consider an error whenever the semantic category or the transcription are incorrect. This means that if a word transcription is incorrect, then we will consider it also a semantic labeling error, although its category is correct. Consequently, the computation of the semantic labeling error is pessimistic, which means that it will never be lower than WER.

VI. Results

Our proposed model has been compared to a baseline system proposed in [6], which consists in a HMM-based HTR system using a category-based 2-gram language model (CB-HTR). Table II presents the experimental results in terms of WER, Precision and Recall. Although the WER remains the same because the MGGI technique is focused on the semantic labeling, the performance in information extraction significantly improves. In the first case, the mean Precision and Recall are computed for the absolute number of instances. In the second case, the mean Precision and Recall are computed by averaging the Precision and Recall for each one of the categories. As it can be observed, the absolute values are higher because there are some categories that appear in few cases, and consequence, the ability of the model to learn is lower.

Table II
Word Error Rate (WER), precision (π) and recall (ρ) obtained with the category-based HTR system (CB) and with the MGGI HTR system (MGGI). The mean is computed for the absolute number of instances (I) and for categories (C). All results are percentages.

	WER	I-π	I-ρ	C-π	C-ρ
CB	10.1	79.2	66.6	73.5	65.2
MGGI	10.1	85.3	76.2	78.3	72.2

Table III, shows the detailed results for some categories. It is worth to notice that, in both methods, π and ρ are usually high in very populated categories (e.g. Groom's and Bride's names), and tend to decrease in categories with very few instances (e.g. Bride's residence and origin). This behavior is probably due to the few training data for these low populated categories.

Concerning the comparisons, the first observation is that π and ρ are usually higher when using MGGI rather than CB. In many categories, usually in those corresponding to the bride information (e.g. Father's Bride Name and Father's Bride Surname) the improvement is over 20 points, whereas in others (e.g. Father's Groom Profession, Groom's Origin) the performance is similar. In fact, the CB methodology outperforms the MGGI only when the number of instances is lower (e.g. Bride's residence and origin). Secondly, the absolute number of repeated categories rep that are found in the same register decreases when using the MGGI technology. In this sense, whenever the same category is found several times in one register, then only the first one is taken into account.

VII. Conclusions

In this paper, we have studied the use of the MGGI methodology for category-based language modeling to relevant information extraction and for automatically transcription of a marriage license book. Given the fixed structure of the information included in the license, we have used it to label the words of the training strings. The labels are chosen in such a way that a bigram trained with the labeled strings deals with restrictions that a simple category-based language model can not. From the results we can see that using the MGGI methodology can be useful to automatically extract the relevant information, helping the user in this hard task.

Finally, given that the MGGI requires a-priory knowledge of the task that it is not always available, as future work, we intend to compute the anchor points to apply the new

Table III
DETAILED RESULTS OBTAINED WITH THE CATEGORY-BASED (CB) AND WITH THE MGGI-BASED (MGGI) SYSTEMS. R IS THE NUMBER OF RELEVANT WORDS, D IS THE NUMBER OF DETECTED WORDS, C IS THE NUMBER OF CORRECTLY DETECTED WORDS, M IS THE NUMBER OF MISSED WORDS, AND rep IS THE NUMBER OF CATEGORIES THAT THE SYSTEM HAS DETECTED MORE THAN ONCE IN THE SAME RECORD. PRECISION (π) AND RECALL (ρ) ARE IN PERCENTAGES.

		MGGI						CB					
CATEGORY	R	D	C	M	rep	π	ρ	D	C	M	rep	π	ρ
Groom's Name	1738	1537	1441	201	374	**93,8**	**82,9**	1433	1274	305	500	88,9	73,3
Bride's Name	1736	1570	1462	167	39	93,1	84,2	1567	1463	171	41	**93,4**	**84,3**
Father's Bride Name	1302	1141	1007	200	38	**88,3**	**77,3**	977	626	410	179	64,1	48,1
Father's Bride Surname	1300	1148	945	191	64	**82,3**	**72,7**	996	614	388	212	61,6	47,2
Father's Groom Profession	963	718	593	306	131	**82,6**	**61,6**	679	560	339	132	82,47	58,1
Groom's Origin	590	590	544	33	16	**92,2**	92,2	597	546	31	28	91,5	**92,5**
Bride's Residence	365	429	299	51	13	69,7	**81,9**	364	291	60	11	**79,9**	79,7
Bride's Origin	15	21	10	4	0	47,6	**66,7**	11	6	8	0	**54,5**	40,0

labelling required in the MGGI automatically. This can be done using combination techniques based in confusion networks, such as that presented in [16]

ACKNOWLEDGMENT

This work has been partially supported through the European Union's H2020 grant READ (Ref: 674943), the European project ERC-2010-AdG-20100407-269796, the MINECO/FEDER, UE projects TIN2015-70924-C2-1-R and TIN2015-70924-C2-2-R, and the Ramon y Cajal Fellowship RYC-2014-16831.

REFERENCES

[1] A. Esteve, C. Cortina, and A. Cabré, "Long term trends in marital age homogamy patterns: Spain, 1992-2006," *Population*, vol. 64, no. 1, pp. 173–202, 2009.

[2] V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition," *Pattern Recognition*, vol. 46, pp. 1658–1669, 2013.

[3] U.-V. Marti and H. Bunke, "Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System," *IJPRAI*, vol. 15, no. 1, pp. 65–90, 2001.

[4] A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta, "Integrated Handwriting Recognition and Interpretation using Finite-State Models," *IJPRAI*, vol. 18, no. 4, pp. 519–539, June 2004.

[5] S. España-Boquera, M. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martínez, "Improving offline handwriting text recognition with hybrid HMM/ANN models," *IEEE Trans. on PAMI*, vol. 33, no. 4, pp. 767–779, 2011.

[6] V. Romero and J. A. Sánchez, "Category-based language models for handwriting recognition of marriage license books," in *Proc. of ICDAR 2013*, 2013, pp. 788–792.

[7] T. Niesler and P. Woodland, "A variable-length category-based n-gram language model," in *Proc. of ICASSP-96*, vol. 1, may 1996, pp. 164 –167 vol. 1.

[8] C. Raymond, F. Béchet, R. D. Mori, and G. Damnati, "On the use of finite state transducers for semantic interpretation," *Speech Communication*, vol. 48, no. 3-4, pp. 288 – 304, 2006, spoken Language Understanding in Conversational Systems.

[9] E. Vidal and D. Llorens, "Using knowledge to improve n-gram language modelling through the mggi methodology," in *Proceedings of the 3rd ICGI'96*, 1996, pp. 179–190.

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

[11] F. J. Nevado, J.-A. Sánchez, and J.-M. Benedí, "Lexical decoding based on the combination of category-based stochastic models and word-category distribution models," in *IX Spanish Symposium on Pattern Recognition and Image Analysis*. Publicacions de la Universitat Jaume I, 2001, pp. 183–188.

[12] E. Vidal, F. Thollard, C. De La Higuera, F. Casacuberta, and R. C. Carrasco, "Probabilistic finite-state machines-part I," *IEEE Transactions on PAMI*, vol. 27, no. 7, pp. 1013–1025, 2005.

[13] ——, "Probabilistic finite-state machines-part II," *IEEE Transactions on PAMI*, vol. 27, no. 7, pp. 1026–1039, 2005.

[14] P. Garcia, E. Vidal, and F. Casacuberta, "Local languages, the succesor method, and a step towards a general methodology for the inference of regular grammars," *IEEE Transactions on PAMI*, no. 6, pp. 841–845, 1987.

[15] S. Eilenberg, *Automata, Languages, and Machines*, ser. Automata, Languages, and Machines. Academic Press, 1974, no. pt. 1.

[16] E. Granell and C. D. Martínez-Hinarejos, "Combining Handwriting and Speech Recognition for Transcribing Historical Handwritten Documents," in *In Proc. 13th ICDAR*, 2015, pp. 126–130.