Edit Transducers for Spelling Variation in Old Spanish

Jordi Porta, José-Luis Sancho, Javier Gómez

Departamento de Tecnología y Sistemas
Centro de Estudios de la Real Academia Española
c/ Serrano 187-189, Madrid 28002. Spain

{porta,sancho,javierg}@rae.es

ABSTRACT
A system for the analysis of Old Spanish word forms using weighted finite-state transducers is presented. The system uses previously existing resources such as a modern lexicon, a phonological transcriber and a set of rules implementing the evolution of Spanish from the Middle Ages. The results obtained in all datasets show significant improvements, both in accuracy and in the trade-off between precision and recall, with respect to the baseline and the Levenshtein edit distance. A qualitative error analysis suggests several potential ways to improve the performance of the system.

KEYWORDS: Old Spanish, Finite-State Transducers, Spelling Variation, Historical Linguistics.

Jordi Porta, José-Luis Sancho and Javier Gómez 2013. Edit transducers for spelling variation in Old Spanish. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013.* NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87: 70–79.

1 Introduction

When processing historical language variants, the most visible problem is the lack of a standardised orthography. The spelling of texts written in different periods of time varies because spelling conventions change over time and official orthographies, when exist, are periodically subjected to reforms. In addition, texts written during the same period of time also have been found to have variation in spelling. And, as if that were not enough, this variation can be found within the same text and even within works by the same author.

In this work, we address the problem of assigning modern citation forms (or lemmas) and word classes to historical word forms using a system for the treatment of diachronic variation found in Old Spanish. As it has been pointed out in Borin and Forsberg (2008), the assignment of word forms to citation forms is seen as a morphological analysis differing from part-of-speech tagging. While the former provides all the plausible analyses for a given word, the later assigns to words the most probable morphological analysis given the context.

The system is modular and has been implemented with weighted finite-state transducers and, in its current state, contains the devices for dealing with the phonological evolution and orthographic variation of Old Spanish. It uses also a Contemporary Spanish lexicon providing the analysis. Several experiments with different datasets have been conducted in order to assess the validity of the approach and results have been compared with those obtained using only the lexicon and in combination with the Levenshtein distance.

2 A Very Brief History of the Spanish Language

Traditionally, different periods of time are distinguished in the history of a language. In the case of the Spanish language these are: Pre-literary Romance, Medieval Spanish, Golden Age, etc. However, neither the periodisation nor the so-called 'language stages' are justified in the light of internal factors. Linguistic change may take centuries to complete and can take place in different communities at different times. The progressive abandonment of Latin in favour of the Romance languages took place during the Middle Ages, but Latin was still used as a lingua franca and in literate culture for long time. A more complete, detailed and adequate history of the Spanish language can be found in Penny (2002).

This paper focuses on historical words found in texts dated in Medieval and Golden Ages. One of the most important morphosyntactic changes of Romance from Latin is the major reduction in the nominal case system. At the syntactic level, the functions of declension were taken on by a system of prepositions and other particles. Another change is the development of synthetic future and conditional paradigms from analytic constructions in which mesoclisis becomes proclisis. According to Lloyd (1987), the development of the Old Spanish phonological system from Late Latin to Contemporary Spanish is best described as the interplay of many relaxing and simplifying processes: some vowel distinctions were dropped resulting in the current five vowel system while other vowels simply disappeared as a result of syncope, elision or apocope (notably final *-e*), to name just a few. On the consonant wagon, the relaxing trend is responsible for consonant cluster simplification, debuccalisation (consonant resulting in a vowel), palatalisation, lost of initial *f-* or various kinds of neutralisation (notably *r*/*l*). The so-called lenition, really a cascaded set of changes from geminate simplification, plosive spirantisation and voicing, also progressed to easier articulation. The most outstanding change is, by far, the devoicing of the three sibilant series that finally led to the current system. The orthographic system reflected the sound changes with a certain lag, making writing practices confusing. According to Morreale

(1978), the configuration of 'medieval spelling' is the result of the interplay of palaeographic usages (the letter shapes), graphical usages (the letter identification) and phonetic values. Experts have identified three principles governing the correspondence between phonological units and the spelling representing them: pronunciation, etymology and usage (Pombo, 2012). Although Spanish favours phoneticism from as early as 13th century, texts show alternating spelling trends rooting on dialectal, cultural (i.e. etymological trend in Late Middle Ages) or stylistic (i.e. *variatio*) reasons. As a result, we are faced with what Pombo (2012) calls *heterographs* (several spellings for one sound) and *heterphones* (several sounds for one spelling).

3 Previous Related Approaches to Spelling Variation

A recent survey on the problems and the approaches to the processing of historical texts, from their acquisition to their exploitation, can be found in Piotrowski (2012). The task of analysing historical word forms using modern lexicons has been approached as an approximate string matching problem. In approximate matching an edit distance is used to measure the similarity between two strings and is defined as the minimal cost required to transform one string into another. The most commonly used distance is known as the Damerau-Levenshtein distance. The basic editing operations considered in the Levenshtein distance are the deletion, the insertion and the substitution of a letter (Levenshtein, 1966). All these operations are assigned a unit cost. Some other operations are often considered as basic edit operations, e.g. letter transposition (Damerau, 1964), which are useful correcting errors made during the fast keyboarding of texts.

In Bollmann et al. (2011), character replacement rules were derived from the alignment of the Luther's 1545 bible translation in Early New High German and a corresponding version of the bible in New High German. Normalisation for Luther texts results in 91% exact token matches (and 93% adding a word substitution list). However, performance went down when the time period was extended and the number of different authors considered was increased. For different versions of the *Interrogatio Sancti Anselmi de Passione Domini*, written between the 14th and 16th centuries, the number of exact matches at token level was of about 42%, but started in a baseline of 32%.

The approach presented here has many points in contact with Jurish (2010b). In his work, a historical word form is canonised by a modern word if they conflate through the application of a cascade of transducers implementing transliteration, phoneticization and heuristic rewrites. In order to increase precision, the context of a historical word is taken into account to disambiguate conflated analysis. For a 1.5 million word corpus of 18–19th century German, precision and recall reached at token level were of 94.3 and 99.3% respectively.

4 A Linguistic Approach Based on Transducers

Analysing a word in our model means computing the result of the composition of three transducers:

$$W \circ E \circ L \tag{1}$$

where W is the automaton representing the word, E is in general any edit transducer, and L is a lexicon relating word forms and its correspondent lemma and word class. E is currently used to model phonological and graphical variation. Note that if W is a historic word form and L is composed of modern forms, there is an implicit process of canonicalisation that leads to the modern analysis. The word analyser is implemented by merging an on-demand version of the three-way composition algorithm (Allauzen and Mohri, 2008) with the n-best strings algorithm of Mohri and Riley (2002) adapted to return only the strings with lowest cost. A very similar

solution has been also proposed in Jurish (2010a). The system has been implemented using the OpenFst library (Allauzen et al., 2007).

The *Diccionario de la lengua española de la Real Academia Española*[1] (DRAE (RAE, 2001)) is a general lexical repertoire focused on the elevated norm shared by all educated speakers all over the Spanish speaking areas. It is an ever evolving, periodically updated dictionary, originated in the early 18th century. This longevity, together with the lack of a historical dictionary that fully documents lexical and semantic evolution, account for the density of both lemmas included and grammatical and semantic distinctions made: long term unused words or meanings, words or senses specific only to some countries or areas, mildly used technical terms or even latin formulae are given an entry. On the other side, regular, on occasion very widespread words resulting from morphological processes are missing. While this impairs the reusability of the dictionary on natural language processing tasks, it is the reference dictionary of choice and as such we decided for most of the experiments to keep full contents activated, i.e. words were not selectively deactivated by chronological or geographical criteria. The lexicon derived from the 2012 amended version of DRAE has been augmented with the forms resulting from certain morphological processes: *-mente* (-ly) adverbs, superlative and augmentative or diminutive derivatives. Unlike Contemporary Spanish, Old Spanish allowed clitic affixation to any verbal form. In order to account for those complexes, we generated about 50 million entries from the combination of every verbal form with up to three enclitic pronouns.

The edit transducer plays the role of variation model. We have implemented two edit transducers for the purpose of comparison: the Levenshtein transducer and a rule-based linguistic transducer implementing the sound change in Spanish. An extensive use of rewrite rules of the kind of $\alpha \rightarrow \beta \ / \ \gamma \ \_\_\_ \ \delta$ (Chomsky and Halle, 1968) is done in the descriptions of language at different levels. Rewrite rules turn out to be efficiently compiled into finite-state transducers (Kaplan and Kay, 1994; Karttunen, 1995, 1996). The linguistic transducer implements a cascade of applications of rewrite rules organised in modules. The overall process can be summarised as follows: First, modern word forms in the lexicon are phonetically transcribed, then a set of rules expressing the phonetic and phonological change is applied and the resulting forms are transcribed back to graphemes. Finally, a set of rules for graphical variation is applied. The application of these modules can be formulated in terms of regular relations as:

$$E = (M \circ P \circ C \circ G \circ V)^{-1} \tag{2}$$

where

- M is a set of modern word forms.
- P represents a grapheme-to-phoneme transcription containing rules as the following, which map the letter c to the SAMPA (Wells, 1997) sounds $[T]$ or $[k]$ depending on the context:

$$(c \rightarrow [T] / \_\_\_ \{E, I\}) \circ (c \rightarrow [k]) \tag{3}$$

- C are the set of rules expressing phonological change. For example, the series $[ds] > [Z] > [S] > [T]$ and $[ts] > [S] > [T]$ accounts for the evolution of palatal affricates,

[1] `http://lema.rae.es/drae`

where deaffrication, devoicing (when applicable) and dentalisation are at play. Their implementation is as follows:

$$([ds] \rightarrow [Z]) \circ ([ts] \rightarrow [S]) \circ ([Z] \rightarrow [S]) \circ ([S] \rightarrow [T]) \quad (4)$$

- G translates phonological forms into surface forms. The following example maps $[k]$ to its alternating graphemic realisations:

$$([k] \rightarrow qu/\_\_\{E,I\}) \circ ([k] \rightarrow c) \quad (5)$$

- V contains graphemic equivalences, as in the following example, where $(\rightarrow)$ makes the rewriting optional and (0.2) is the weight associated to the rewrite rule:

$$(\{c,z\} \rightarrow \{s,z\}/\_\_\{E,I\}) \circ (z\ (\rightarrow)\ ç\ (0.2)/\_\_\{A,O,U\}) \quad (6)$$

It is worth noting that G differs from P^{-1} in that G implements a relaxed form of the current orthographic norms. Note also that in order to get a transducer from Old Spanish to Modern Spanish the result of the composition is inverted $(^{-1})$. These weighted rational relations are all expressed using regular expressions and context-dependent rewrite rules. Preference in rewriting is expressed using numerical values or weights in the tropical semiring (Mohri, 2009). All these rules and regular expressions are compiled into weighted finite-state transducers. For its implementation we have used the OpenGrm Thrax Grammar Compiler (Roark et al., 2012).

5 Datasets

Experiments have been conducted on different datasets of Medieval and Golden Ages that represent different gold standards. We want to note that these datasets are word lists and that we report figures on a type basis, unlike other works reporting results on running texts, i.e. on a token basis. The distribution of word classes of these datasets can be seen in Table 1.

The dataset called FL-EM basically corresponds to the lexicon found within the FreeLing[2] distribution for analysing Old Spanish. The creation of this lexicon containing variants observed in the corpus of medieval texts of the Hispanic Seminary of Medieval Studies is described in Sánchez-Marco et al. (2011). It is important to note that FreeLing bases the analysis of some words on morphological decomposition modules: verbs with enclitics, adverbs ended in *-mente*, augmentatives and diminutives, superlatives, etc. Therefore, these words are not found in the FreeLing static lexicon. Proper nouns and Roman numerals, as well as multiword and amalgamated expressions have been removed for the experiments. Neither the system of words classes nor the lemmatisation is directly comparable and some categories and lemmas of the DRAE have been changed before comparisons not to get false differences in the experiments. Notably, old lemmas from DRAE with modern counterparts were deactivaded, e.g. *fermosura*, whose modern lemma is *hermosura* (beauty).

CDH-EM and CDH-SO are the lexicons induced from the manual correction of a previous annotation of a subcorpus of the Spanish corpus *Corpus del Nuevo diccionario histórico*[3] (CDH). CDH-EM comes from a fragment of 67 661 running words from medieval texts spanning from 1064 to 1494 and CDH-SO contains 318 728 running words from Golden Ages texts (1521–1698).

[2] `http://nlp.lsi.upc.edu/freeling/` (accessed 2013-03-03)

[3] `http://www.frl.es/Paginas/Corpusdiccionariohistorico.aspx` (accessed 2013-03-03)

Word Class	FL-EM
Adjectives	4 048
Nouns	11 257
Verbs	20 339
Prepositions	64
Determinants	172
Pronouns	292
Adverbs	254
Conjunctions	160
Interjections	117
Other	6
Total	36 709

Word Class	CDH-EM	CDH-SO	MAP-EM	MAP-SO
Adjectives	1 714	4 974	10 230	5 218
Nouns	3 505	9 855	23 776	11 533
Verbs	5 967	16 046	45 021	22 275
Prepositions	55	56	153	83
Articles	12	15	24	22
Pronouns	235	319	839	353
Adverbs	274	476	1 459	555
Conjunctions	52	52	198	103
Interjections	114	169	310	145
Other	1	6	13	10
Total	11 929	82 023	31 968	40 297

Table 1: Word class distributions in the FL-EM dataset and CDH and MAP datasets.

The last datasets, MAP-EM and MAP-SO, come from a list of historical forms that were not analysed or were incorrectly analysed by a previous system developed for annotating first the historical corpus CORDE[4] (Sánchez et al., 1999) and then the CDH in 2006–2009. The list was manually analysed with the aid of several specialists and contains valuable information about dating and the phenomena involved in the transformation of these old forms into modern ones. The original list contains 96 790 entries and has been split up into two lists corresponding to the Ages considered in this work.

6 Experiments and Analysis of Results

The starting point for assessing the validity of the system proposed on each of the different datasets is the performance of the lexicon derived from the DRAE using as edit transducer the identity. We will refer to this experiment establishing the baseline as ID. To have a clearer picture of the performance of the proposed system, we have carried out some experiments using the Levenshtein transducer with maximum distance costs of one and two. It is important to note that in each of the cases only the set of analysis with the lowest cost is returned. These experiments will be referred to as LEV. Finally, we have experimented with the proposed system, that will be referred to as LIN, with different maximum costs, yielding not major differences in the results. Consequently, maximum distance has been set to two. In order to compare the systems we have computed on the basis of the analysis confusion matrix the standard measures of precision, recall, their harmonic mean F, and accuracy. Formulas and results for the different datasets can be seen in Table 2.

As can be seen in Table 2, quantitatively, the LIN system obtains the best F result in all datasets, indicating the better trade-off between precision and recall, and the best accuracy rates in each dataset, while ID has better precision at CDH and LEV with a maximum distance of two obtains good results also in CDH but at the expense of overgeneration of analysis.

Most of the false negatives, i.e. missing analysis, returned by our system are due to diverging criteria regarding lemmatisation and/or categorisation which are, at least, arguable. Consider, for example, the possibility of attributing to *cerda* (sow) the masculine lemma *cerdo* (pig) or the feminine lemma *cerda* (bristle). It is worth noting that false negatives caused by alternative conventions are usually accompanied by false positives, and that these mismatches impair both

[4] `http://corpus.rae.es/cordenet.html` (accessed 2013-03-03)

Dataset	Edit	d	TP	FP	FN	Prec.	Rec.	F	Acc.
FL-EM	ID	–	1 340	30 386	35 369	4.22	3.65	3.92	2.00
FL-EM	LEV	1	25 987	53 340	10 722	32.76	70.79	44.79	28.86
FL-EM	LEV	2	31 396	101 010	5 323	23.71	85.50	37.13	22.80
FL-EM	LIN	2	**32 679**	**14 171**	**4 030**	**69.75**	**89.02**	**78.22**	**64.23**
CDH-EM	ID	–	8 441	**2 769**	3 488	**75.30**	70.76	72.96	57.43
CDH-EM	LEV	1	10 516	5 878	1 413	64.15	88.15	74.26	59.06
CDH-EM	LEV	2	**10 805**	8 500	**1 124**	55.97	**90.58**	69.19	52.89
CDH-EM	LIN	2	10 802	3 796	1 127	74.00	90.55	**81.44**	**68.69**
CDH-SO	ID	–	25 719	**5 012**	6 249	**83.69**	80.45	82.04	69.55
CDH-SO	LEV	1	29 714	9 869	2 254	75.07	92.95	83.06	71.02
CDH-SO	LEV	2	**30 131**	12 546	**1 837**	70.60	**94.25**	80.73	**67.69**
CDH-SO	LIN	2	29 489	7 649	2 479	79.40	92.25	**85.34**	74.44
MAP-EM	ID	–	1 681	63 720	80 342	2.57	2.05	2.28	1.15
MAP-EM	LEV	1	49 825	130 974	32 198	27.56	60.75	37.92	23.39
MAP-EM	LEV	2	**63 597**	242 393	**18 426**	20.78	**77.54**	32.78	19.60
MAP-EM	LIN	2	60 201	**30 770**	21 822	**66.18**	73.40	**69.60**	**53.37**
MAP-SO	ID	–	852	31 172	39 445	2.66	2.11	2.36	1.19
MAP-SO	LEV	1	31 616	64 596	8 681	32.86	78.46	46.32	30.14
MAP-SO	LEV	2	**35 600**	92 089	**4 694**	27.88	**88.34**	42.38	26.89
MAP-SO	LIN	2	33 367	**11 779**	6 930	**73.91**	82.80	**78.10**	**64.07**

Table 2: Results of the identity transducer (ID), the Levenshtein edit transducer (LEV) with several maximum distances (d) and of the linguistic transducer (LIN) with a cost distance of two. In the table, column TP represents true positive matches, FP represents false positives, and FN false negatives. Precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, Recall $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, $\mathrm{F} = 2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}/(\mathrm{Precision}+\mathrm{Recall})$ and Accuracy $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$.

recall and precision. Some other false negatives are old word forms with no possible matches in our lexicon. They correspond to old suppletive forms or disappeared present participles. These cases are especially frequent in CDH-EM and CDH-SO and point to future work. Other false negatives found in the FL-EM dataset correspond to full Latin forms or some errata, which can not be matched by our system. For MAP datasets, there are a number of very specific formalizations that include not only morphological changes but also their combination with changes in the sound pattern and specific graphical representation found in particular texts.

As for the false positives, i.e. analysis presumably given in excess by our system, they sometimes correspond to valid lexical or morphological analysis not present in the datasets. Our lexicon contains about 4 000 old lemma variants not amenable to deactivation. Consider *faba* (bean), having both current and old uses according to DRAE. These cause false positive analysis on their word forms. It is possible to track a certain bias in both FL-EM and, much more noticeably, CDH-EM and CDH-SO datasets: CDH texts from which the lexicon is induced were manually revised and corrected, when needed, and show some particularised lemma and/or part of speech tags (or lack thereof) unattainable by our general purpose system. The lack of selective deactivation of the irrelevant fragment of the lexicon also plays an important role. Consider how *ionico* is clearly related to *jónico* (Ionian) in a medieval setting and unrelated to *iónico* (ionic), a word introduced in recent times.

Finally, regarding the core of the analysis engine, we have identified some room for improvement in the treatment of velar and sibilant orders, regressive lateral assimilation, lenition, tonal shift and rule weighting policies.

7 Conclusions and Future Work

A modular architecture for the treatment of diachronic variation has been implemented within the framework of finite-state models. The homogeneity and soundness of the framework allow different levels of linguistic variation to be easily modelled, composed and extended using previously existing resources such as modern lexicons, morphological analysers, phonological transcribers or lexical diachronic descriptions. The results obtained in all datasets show significant improvements in accuracy and in the trade-off between precision and recall with respect to the baseline and the Levenshtein distance.

Besides several adjustments to the rules, to the weights, and to the lexicon of the current system, a more in-depth qualitative analysis of the errors suggests several potential ways to further improve the performance of the system. From a lexical point of view, precision could be improved by deactivating lexicon entries corresponding to words not belonging to the ages considered. Also recall could be improved by introducing disappeared words into the lexicon and inflecting lemmas according not only to current patterns (e.g. old *andé* together with current *anduvo* (I walked)). Some changes in the morphology and morphosyntax of Spanish are not being covered by the phonological and graphical variation model (e.g. *tornarsa*, a synthetic future with a mesoclitic *tornar+se+ha* (She will come back)). This suggests the need to add a new component for this level. Finally, moving from the analysis of types to the analysis of tokens in context using language models could lead to improvements in accuracy.

Acknowledgments

We gratefully acknowledge the personnel of the *Centro de Estudios de la Real Academia Española* who worked during 2006–2009 in the CDH project from which we have created some of the datasets used in this work. We want also to thank the *Fundación Rafael Lapesa* for creating the opportunity to develop this work for the future annotation of old texts. Finally, we thank our colleagues Adelaida Fernández and Encarna Raigal for proofreading of the manuscript.

References

Allauzen, C. and Mohri, M. (2008). 3-way composition of weighted finite-state transducers. In *Proceedings of the 13th International Conference on Implementation and Application of Automata (CIAA–2008)*, pages 262–273, San Francisco, California, USA.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA–2007)*, pages 11–23, Praque, Czech Republic.

Bollmann, M., Petran, F., and Dipper, S. (2011). Applying rule-based normalization to different types of historical texts — An evaluation. In *Proceedings of the 5th Languange and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 339–344, Poznan, Poland.

Borin, L. and Forsberg, M. (2008). Something old, something new: A computational morphological description of Old Swedish. In *LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH–2008)*, pages 9–16, Marrakech, Morocco.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper & Row, New York.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Jurish, B. (2010a). Efficient online k-best lookup in weighted finite-state cascades. In Hanneforth, T. and Fanselow, G., editors, *Language and Logos: Studies in Theoretical and Computational Linguistics*, volume 72 of *Studia grammatica*, pages 313–327. Akademie Verlag, Berlin.

Jurish, B. (2010b). More than words: Using token context to improve canonicalization of historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–39.

Kaplan, R. M. and Kay, M. (1994). Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378.

Karttunen, L. (1995). The replace operator. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL–95)*, pages 16–23, Cambridge, Massachusetts, USA.

Karttunen, L. (1996). Directed replacement. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL–96)*, pages 108–115, Santa Cruz, California, USA.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Lloyd, P. M. (1987). *From Latin to Spanish*. American Philosophical Society, Philadelphia.

Mohri, M. (2009). Weighted automata algorithms. In Droste, M., Kuich, W., and Vogler, H., editors, *Handbook of Weighted Automata*, pages 213–254. Springer, Berlin.

Mohri, M. and Riley, M. (2002). An efficient algorithm for the *n*-best-strings problem. In *Proceedings of the International Conference on Spoken Language Processing 2002 (ICSLP–2002)*, Denver, Colorado, USA.

Morreale, M. (1978). Trascendencia de la *variatio* para el estudio de la grafía, fonética, morfología y sintaxis de un texto medieval, ejemplificada en el MS Esc. I.I.6. In *Annali della Facoltà di Lettere e Filosofia dell'Università di Padova*, volume II, pages 249–261, Florence, Italy.

Penny, R. J. (2002). *A history of the Spanish Language*. Cambridge University Press, Cambridge, second edition.

Piotrowski, M. (2012). Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.

Pombo, E. L. (2012). Variation and standardization in the history of Spanish spelling. In Baddeley, S. and Voeste, A., editors, *Orthographies in Early Modern Europe*, pages 15–62. De Gruyter Mouton, Berlin, Boston.

RAE (2001). *Diccionario de la lengua española*. Espasa, Madrid, 22th edition.

Roark, B., Sproat, R., Allauzen, C., Riley, M., Sorensen, J., and Tai, T. (2012). The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the ACL 2012 System Demonstrations*, pages 61–66, Jeju Island, Korea.

Sánchez, F., Porta, J., Sancho, J. L., Nieto, A., Ballester, A., Fernández, A., Gómez, J., Gómez, L., Raigal, E., and Ruiz, R. (1999). La anotación de los corpus *CREA* y *CORDE*. In *Proceedings of SEPLN 1999*, volume 25, pages 175–182, Lleida, Spain.

Sánchez-Marco, C., Boleda, G., and Padró, L. (2011). Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–9, Portland, OR, USA.

Wells, J. C. (1997). Sampa computer readable phonetic alphabet. In Gibbon, D., Moore, R., and Winski, R., editors, *Handbook of Standards and Resources for Spoken Language Systems*, pages 684–732. Mouton de Gruyter, Berlin and New York.