Extending the tool, or how to annotate historical language varieties

Cristina Sánchez-Marco
Universitat Pompeu Fabra
Barcelona, Spain
cristina.sanchezm@upf.edu

Gemma Boleda, Lluís Padró
Universitat Politècnica de Catalunya
Barcelona, Spain
{gboleda,padro}@lsi.upc.edu

Abstract

We present a general and simple method to adapt an existing NLP tool in order to enable it to deal with historical varieties of languages. This approach consists basically in expanding the dictionary with the old word variants and in retraining the tagger with a small training corpus. We implement this approach for Old Spanish.

The results of a thorough evaluation over the extended tool show that using this method an almost state-of-the-art performance is obtained, adequate to carry out quantitative studies in the humanities: 94.5% accuracy for the main part of speech and 92.6% for lemma. To our knowledge, this is the first time that such a strategy is adopted to annotate historical language varieties and we believe that it could be used as well to deal with other non-standard varieties of languages.

1 Introduction

In the last few years, there has been a growing interest in all disciplines of the humanities to study historical varieties of languages using quantitative methods (Sagi et al., 2009; Lüdeling et al., to appear). Large corpora are necessary to conduct this type of studies, so as to smooth the great data sparseness problem affecting non-standard varieties of languages, and thus guarantee the validity of the generalizations based on these data.

Historical language varieties bear similarities to standard varieties, but they also exhibit remarkable differences in a number of respects, such as their morphology, syntax, and semantics. In addition, as orthographic rules were not established until later centuries, a great amount of graphemic variation is found in historical texts, such that one word can be written using many different graphemic variants. This variation increases considerably the number of different words and therefore the lexicon of the corresponding language variety. For instance, searching for the infinitival verb form *haber* 'have' in a historical corpus for Spanish can be a difficult task if there are, say, 5 variants of the same word (*auer, aver, hauer, haver, haber*) and the corpus does not contain any other linguistic information, such as lemma and part of speech (PoS).

In this paper we propose a strategy to automatically enrich texts from historical language varieties with linguistic information, namely to expand a preexisting NLP tool for standard varieties of a language. To our knowledge, it is the first time that such an approach is proposed and evaluated. In particular, we describe the method followed to extend a library (FreeLing[1]) for the linguistic analysis of Standard Spanish to enable it to deal with Old Spanish[2].

This general approach has four main advantages over the state-of-the-art strategies (described in section 2). First, the resulting tool can be reused (with the consequent saving of resources). Second, the tool can be further improved by other researchers. Third, it is the tool that is adapted, instead of forc-

[1] http://nlp.lsi.upc.edu/freeling. The tool for Old Spanish is available in the development version 3.0-devel, accessible via SVN.

[2] As it is considered by most scholars, we consider Old Spanish the period from the 12th to the 16th century.

Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 1–9,
Portland, OR, USA, 24 June 2011. ©2011 Association for Computational Linguistics

ing standardisation on the original texts (see section 2). Also, the strategy can be used to extend other existing tools.

The specific case study in this paper presents additional advantages. On the one hand, FreeLing is an open source resource that is well documented and actively maintained. In addition, due to the modularity of this tool, it is relatively easy to adapt. On the other hand, the result of the extension is a tool for Old Spanish across different centuries, that is to say, the tool can be used to accurately tag not only Spanish from a particular century but also to tag the language over a long period of time (from the 12th to the 16th century). The resulting tool achieves almost state-of-the-art performance for PoS-taggers: a tagging accuracy of 94.5% on the part of speech, 92.6% on lemmas, and 89.9% on the complete morphological tag including detailed information such as gender or number for nouns and tense and person for verbs.

Plan of the paper. In Section 2 we review the state of the art. In Sections 3 through 5 we describe FreeLing and the data and methodology used for its adaptation to Old Spanish. Then the results of the evaluation and error analysis are presented (Sections 6 and 7). We conclude with some discussion and suggestions for future work (Section 8).

2 Related work

Up to now, three main approaches have been followed to automatically enrich historical corpora with linguistic information: (i) automatic tagging using existing tools followed by human correction, (ii) standardisation of the words followed by automatic tagging with existing tools, and (ii) re-training of a tagger on historical texts.

The first approach has been adopted in projects such as the *Penn Historical Corpora*[3] , *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor, 2007), and the *Corpus of Early English Correspondence* or *CEEEC* (Raumolin-Brunberg and Nevalainen, 2007). The second strategy, namely, to standardize the corpora prior to their annotation with NLP tools, has also been followed by other scholars (Rayson et al., 2007; Ernst-Gerlach and Fuhr, 2007; Baron and Rayson, 2008). In this approach, graphemic variants in Old English and German texts are identified and subsequently mapped onto their modern equivalents (i.e., the standardized forms). This approach is adequate for tasks such as information retrieval (Ernst-Gerlach and Fuhr, 2007), but not quite so for quantitative research for historical variants. For example, there are many words in historical varieties of languages for which a corresponding standard variant does not exist (e.g., *maguer* ‘although’ in Old Spanish). As reported in Rayson et al. (2007) the PoS tagging accuracy obtained with this method in texts from the Early Modern English period is around 85%.

Recently there have been some experiments with morphosyntactic tagging of historical data by training a model on old texts (Rögnvaldsson and Helgadóttir, 2008; Dipper, 2010). For example, Rögnvaldsson and Helgadóttir (2008) use this approach to tag Old Norse texts (sagas from the 13th and 14th century) yielding 92.7% accuracy on the tag, almost 3 points higher than that obtained in our case.

Our approach is similar in spirit to the latter, as we also train a tagger using an annotated historical corpus. However, it differs in that we consistently extend the whole resource (not only the tagger, but also the dictionary and other modules such as the tokenization). Thus, we build a complete set of tools to handle Old Spanish. Also, our work covers a larger time span, and it is able to tag texts from a wide variety of genres (hence the difference in accuracy with respect to Rögnvaldsson and Helgadóttir (2008)).

As noted in the Introduction, in comparison to state-of-the-art approaches the strategy proposed in this paper requires fewer resources, it is easily portable and reusable for other corpora and languages and yields a satisfactory accuracy.

3 The analyzer

FreeLing is a developer-oriented library providing a number of language analysis services, such as morphosyntactic tagging, sense annotation or dependency parsing (Padró et al., 2010). As mentioned in the Introduction, this tool, being open source, actively developed and maintained, and highly modular, is particularly well suited for our purposes. In addition, it provides an application programming

[3] `http://www.ling.upenn.edu/hist-corpora.`

interface (API) which allows the desired language analyses to be integrated into a more complex processing. In its current version (2.2), this resource provides services (to different extents) for the following languages: English, Spanish, Portuguese, Italian, Galician, Catalan, Asturian, and Welsh. In this paper we have focused on the adaptation of the resources for morphosyntactic tagging, but the syntactic and semantic modules can also be customized. The FreeLing processing pipeline for morphosyntactic tagging is illustrated in Figure 1. As shown in the figure, a set of texts is submitted to the analyzer, which processes and enriches the texts with linguistic information using the different modules: tokenization, dictionary, affixation, probability assignment and unknown-word guesser[4], and PoS tagger.

The tagset used by this tool is based on the EAGLES standard[5]. The first letter of each tag indicates the morphological class of the word. The remaining letters (up to 6) specify more fine-grained morphosyntactic and semantic information, such as the gender and number of nouns or the tense, mode and type (main or auxiliary) of verbs.

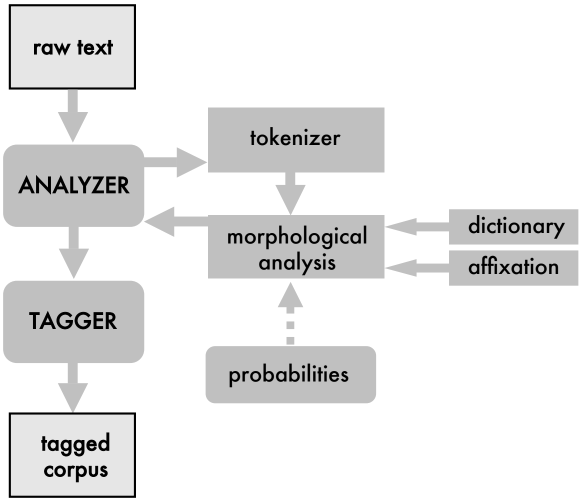

Figure 1: Processing pipeline in FreeLing.

4 The Data

4.1 Old Spanish Corpus

In order to adapt the tool, we have worked with the electronic texts compiled, transcribed and edited by the Hispanic Seminary of Medieval Studies (*HSMS*).[6] We will refer to the set of texts used to adapt the tool as *Old Spanish Corpus*. These texts, all critical editions of the original manuscripts, comprise a variety of genres (fiction and non-fiction) from the 12th until the 16th century and consist of more than 20 million tokens and 470 thousand types. The original texts in these compilations render the copy very closely (diplomatic transcriptions) and contain annotations encoding paleographic information, for instance about the physical characteristics of the manuscript or marks and notes by different scribes. These annotations were removed, and the original transcription of the words has been mantained preserving the similarity to the original copies.

As is the case for most languages keeping data from historical varieties, the number and type or genre of texts which have been preserved for each century varies. From this perspective, the Old Spanish Corpus used to extend the tool is representative of the language, since it covers the language of the Middle Age period, containing samples of most genres and centuries from the 12th century up to the 16th century. As shown in the first row of Table 1, the corpus contains a much lower number of tokens for the 12th century compared to the remaining centuries, as only one document from this century is included in the corpus. The 13th to 15th centuries are fairly well represented, while comparably less tokens are available for the 16th century, due to the design of the *HSMS* collections. To get an impression on the types of texts covered in the Old Spanish Corpus, the documents have been classified according to their genre or topic in *CORDE*[7]. 8 types of genres or topics have been considered: fiction (including

[4] This module has two functions: first, it assigns an *a priori* probability to each analysis of each word. Second, if a word has no analysis (none of the previously applied modules succeeded to analyze it), a statistical guesser is used to find out the most likely PoS tags, based on the word ending.

[5] Expert Advisory Group on Language Engineering Standards (`http://www.ilc.cnr.it/EAGLES96/home.html`).

[6] Corfis et al. (1997), Herrera and de Fauve (1997), Kasten et al. (1997), Nitti and Kasten (1997), O'Neill (1999).

[7] *CORDE* is a reference corpus of diachronic Spanish containing texts from the 8th century up to 1975 (`http://www.rae.es`).

novels and also other narrative books), law, didactics (treatises, sapiential literature), history (chronicles, letters and other historical documentation), society (hunting, fashion), poetry, science (medicine, astrology, astronomy), and religion (Bible). Figure 2 illustrates the distribution of texts according to their genre or topic in each century. The width and height of rows represent the proportion of texts of each genre-topic for each century. Each box corresponds to a particular type of text. On the x-axis the centuries are represented, from the 13th to the 16th century.[8] As can be seen from the size of the corresponding boxes, there is a higher number of fiction books in the later centuries. In contrast, the proportion of law and religion books decreases in time. All in all, the corpus contains a fair variety of genres and topics present in Old Spanish literature, so the language used in these types of documents is represented in the expanded tool as well.

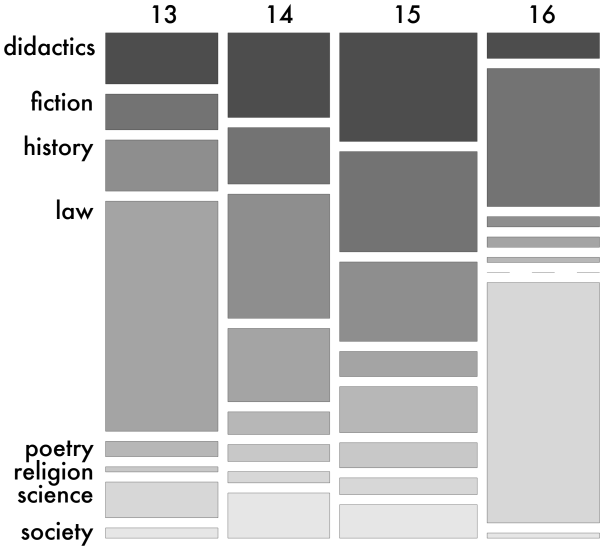

Figure 2: Distribution of genres in the Old Spanish Corpus from 13th to 16th century.

4.2 Gold Standard Corpus

A *Gold Standard Corpus* has been created in order to retrain the tagger and to carry out the evaluation and the error analysis. This corpus consists of 30,000 tokens which have been pre-annotated with the Standard Spanish tagger and manually corrected. Texts composing the Gold Standard Corpus have been selected from the Old Spanish Corpus so as to mirror the data in the whole corpus as far as possible. The token distribution of the Gold Standard Corpus is shown in the second row of Table 1, and the distribution of text types in the second row of Table 2.

4.3 Standard Spanish Corpus

A *Standard Spanish Corpus* has been used to establish a baseline performance for the tagger, namely, the *LexEsp* corpus (Sebastián et al., 2000), consisting of texts from 1975 to 1995 and totalling more than 5 million words. The corpus comprises a representative sample of the Spanish written variety in the 20th century (40% of the tokens in this corpus correspond to fiction, 20% science and didactics, and 40% different classes of press –sports, weekly magazines, and newspapers).

5 Method

The method proposed consists in using the existing Standard Spanish tool as a basis to create an Old Spanish processor to automatically enrich Old Spanish texts with lemma and morphosyntactic tag information. The adaptation of the existing Standard Spanish tool involves the expansion of the dictionary (section 5.1), the modification of other modules which are part of the library, such as the tokenization and the affixation modules (section 5.2), and the retraining of the tagger (section 5.3).

5.1 Dictionary expansion

Data. The Standard Spanish dictionary contains 556,210 words. This dictionary has been expanded with 32,015 new word forms, totalling more than 55,000 lemma-tag pairs, and thus increasing the number of word forms in the dictionary to 588,225. For example, the word form *y* in the expanded dictionary has 4 different lemma-tag pairs, corresponding to a coordinate conjunction, a noun, a pronoun, and an adverb, whereas in the Standard Spanish dictionary it has only 2 lemma-tag pairs, corresponding to the coordinate conjunction and noun uses. Table 3 illustrates the distribution of the categories of words which have been added to the dictionary. As could be expected from the general distribution of words across PoS categories, verbs and nouns account for more than half of the words added.

[8] The document in the 12th century data, belonging to poetry, is not represented in this graph because of its small size.

Corpus	12th c.	13th c.	14th c.	15th c.	16th c.	Total
Old Spanish	0.1	32.2	21.5	31.6	14.6	22,805,699
Gold Standard	4.5	31.3	35.1	20.5	8.6	30,000

Table 1: Size of the Old Spanish and the Gold Standard Corpus, respectively, in tokens (percentages over the *Total* column).

Corpus	Fiction	Law	Didactics	History	Society	Poetry	Science	Religion	Total
Old Spanish	22.4	21.8	18.5	17.5	6.3	6.6	3.6	3.3	22,805,699
Gold Standard	39.9	13.0	13.0	13.0	0.0	8.7	8.7	4.3	30,000

Table 2: Text type distribution in the Old Spanish and the Gold Standard Corpus, respectively, in tokens (percentages over the *Total* column).

Verbs	48.8%	Adverbs	0.4%
Nouns	20.8%	Determiners	0.3%
Adjectives	7.0%	Conjunctions	0.3%
Pronouns	0.6%	Interjections	0.2%
Prepositions	0.5%	Numbers	0.2%
		Punctuation	0.01%

Table 3: Distribution of words added to the dictionary.

Method. Two different types of mapping rules have been used in order to automatically generate the types of words to be added to the dictionary: substring rules and word rules. *Substring rules* map 54 sequences of characters from an old variant onto the corresponding standard variant. These mapping rules are based on the observed regularities in the spelling of Old Spanish texts (Sánchez-Prieto, 2005; Sánchez-Marco et al., 2010). These rules are independent of the morphophonological context, except that 18% of them are restricted to the beginning or the end of a word. Table 4 shows some examples of these rules. 81.4% of the types added to the dictionary have been generated using these rules. All words generated by this method are added to the dictionary if and only if they are contained in the corpus. This avoids the automatic generation of a very high number of variants.

Old	Modern	Example
euo	*evo*	*nueuo* → *nuevo* ’new’
uio	*vio*	*uio* → *vio* ’saw’

Table 4: Examples of the substring rules.

The remaining 18.5% of the types incorporated into the dictionary have been created using *word rules*. These are mappings from an old variant of a word to its corresponding standard variant (created manually), to deal with the most frequent types not covered by the substring rules, such as for instance words without an accent (*consul* → *cónsul* ’consul’), or other graphemic variants (*yglesia* → *iglesia* ’church’, *catholica* → *católica* ’catholic’).

5.2 Adapting other modules

The *tokenization* of some symbols has been customized, in order to deal with the particular characteristics of the original data, for instance to account for the fact that in most cases the letter *ç* is written in the texts of the *HSMS* as *c’*, and *ñ* as *n˜* (*yac’e* ’lay’, *cin˜o* ’adhered’). Also, FreeLing analyzes forms not found in the dictionary through an *affixation* module that checks whether they are derived forms, such as adverbs ending in *-mente* or clitic pronouns (*-lo*, *-la*) attached to verbs. This module has also been adapted, incorporating Old Spanish clitics (*-gela*, *-li*) and other variants of derivation affixes (adverbs in *-mientre* or *-mjentre*).

5.3 Retraining the tagger

FreeLing includes 2 different modules able to perform PoS tagging: a hybrid tagger (*relax*), integrating statistical and hand-coded grammatical rules, and a Hidden Markov Model tagger (*hmm*), which is a classical trigram markovian tagger, based on TnT (Brants, 2000). As mentioned in Section 4, the tagger for Standard Spanish has been used to pre-annotate the Gold Standard Corpus, which has subsequently been corrected to be able to carry out the retraining. The effort of correcting the corpus is much lower compared to annotating from scratch. In this paper we present the evaluation of the performance of the extended resource using the *hmm* tagger with the probabilities generated automatically from the trigrams in the Gold Standard Corpus.

6 Evaluation

In this section we evaluate the dictionary (Section 6.1) and present the overall tagging results (Section 6.2). The resources for Standard Spanish have been used as a baseline.

6.1 Dictionary

In order to evaluate the expanded dictionary, we use three different measures: ambiguity, coverage, and accuracy and recall of automatically generated entries.

Ambiguity measures the average number of lemma-tag pairs per word form. To compute average ambiguity, each word form is assigned a score corresponding to the number of lemma-tag pairs in its dictionary entry. We have checked ambiguity in two different ways: (i) in the dictionary (type-based), (ii) in the corpus (token-based). *Coverage* measures the percentage of tokens in the corpus which are analysed by the dictionary. Uncovered or unknown words are those forms which are not included in the dictionary or analysed by the affixation module. We also evaluated the *precision* and *recall* of *automatically generated entries*, that is the percentage of correct words among those added to the dictionary by the substring rules,[9] and the percentage of the expected lemmas for those words actually added by the rules. Both measures have been obtained by checking a random sample of 512 types (corresponding to 2% of the types added with the substring rules). As only the words added to the dictionary are being evaluated, these measures have been obtained only over the Old Spanish dictionary.

The results of the evaluation are summarised in Table 5. As can be seen in this table, the Old Spanish Corpus is more ambiguous than the Standard Spanish Corpus, despite the fact that the dictionary is not (note that the 32,000 entries added are only a 5.8% increase in the Standard dictionary). The higher ambiguity in the corpus is probably due to the fact that many function words, such as the word *y* mentioned in section 5.1, have more entries in the expanded dictionary than in the Standard Spanish dictionary. The increase in ambiguity is also due to the large time span covered by the dictionary, as for instance forms that in the 13th century were lexical verbs and later changed to auxiliaries will bear both the old and the new morphosyntactic tag (*haber* changed its meaning from 'possess' or 'hold' to be the auxiliary in perfect tenses). Due to this increase in ambiguity, we can expect a higher number of errors due to ambiguity in Old Spanish than in Standard Spanish texts, as the tagger has more options to disambiguate in context and thus the overall error probability increases. As for coverage, 99.4% of the words in the Standard Spanish Corpus are covered by the Standard Spanish dictionary and affixation module. In contrast, 92.6% of the words in the Old Spanish Corpus are covered. If a word has no analysis, the probability assignment module tries to guess which are its possible PoS tags, based on the word ending. This also means that the adapted tool needs to guess the tag of a word more often, therefore increasing the number of potential errors.

As for precision, the lemmas and tags which have been automatically generated using substring rules and added to the dictionary achieve 99.2%. Only 0.8% of the lemmas and tags are incorrect. These are mostly cases either of Latin words (*sedeat*) or proper nouns (*maaçe*, *lameth*), which in any case are words not easily treated with automatic rules. Also in this evaluation sample, there are some incomplete entries, lacking 1 or more lemmas and tags. Cases of entries lacking some lemma (1.4% of the evaluation sample, yielding 98.6% recall) are proper nouns (*valenc'ia, thesis*), Latin words (*mjlites, euocat*), already incomplete entries in the Standard Spanish dictionary (*escanpado* 'cleared up'), and lemma-tag pairs not generated by any of the rules (*baiassen* 'went down'). Entries lacking some tags (5.3% of the evaluation sample, yielding 94.7% recall) are mostly cases of some verbal tenses, for example words in which the tag for the future or simple past is not included (*pessara* 'he will regret', *affronto* 'he faced'). The old variant typically lacks the diacritics,

[9] The word rules and manual mappings have not been evaluated, as they have been manually created.

	Old Spanish		Standard Spanish	
	Type-based	Token-based	Type-based	Token-based
Ambiguity	1.21	1.85	1.20	1.68
Coverage		92.6%		99.4%
Precision	99.2%			
Recall	98.6% (lemmas), 95% (PoS)			

Table 5: Evaluation of the dictionary.

so the morphosyntactic tag for the accented variants is not generated.

	Lemma	PoS-1	PoS-2
C0	72.4	70.9	77.4
C1	90.7	86.0	91.0
C2	91.2	87.5	91.9
C3	92.3	89.5	93.7
C4	92.6	89.9	94.5
SS	99.1	94.0	97.6

Table 6: Accuracy obtained for lemma, PoS-1, and PoS-2 in the 5-fold cross-validation for the Old Spanish tagger on the Gold Standard Corpus (rows C0 to C4) and for Standard Spanish (row SS).

Condition	C0	C1	C2	C3
C1	l, p1, p2			
C2	l, p1, p2	p1, p2		
C3	l, p1, p2	l, p1, p2	l, p1, p2	
C4	l, p1, p2	l, p1, p2	l, p1, p2	p2

Table 7: Statistical significance in the tagging with the different conditions. If there is a statistically significant difference at a 99% confidence degree according to a χ^2 test with 1 d.f., *l* (for lemma), *p1* (for PoS-1), and *p2* (for PoS-2) are written.

6.2 Tagging

In order to evaluate the performance of the tagger, the *accuracy* in the tagging of lemmas, PoS-1 (the whole label, containing detailed morphosyntactic information; 6 characters of the tag in total), and PoS-2 (word class; 1 character in total) has been checked. In all cases, this measure has been obtained as a result of a 5-fold cross-validation. As described in Section 5, the method proposed involves (a) adapting the dictionary and other modules, (b) retraining the tagger with Old Spanish texts. To assess the relative impact of these two adaptations, we report the results of evaluating the tagging under several conditions. To assess (a), we report two scores obtained using: (C0) original tools for Standard Spanish, and (C1) the expanded dictionary and other modules combined with the Standard Spanish tagger. To assess (b), and, specifically, the impact of the size of the tagger retraining corpus, we report the results of retraining the tagger with: (C2) 10,000-token, (C3) 20,000-token, and (C4) 30,000-token subsets of the Gold Standard Corpus, always using the expanded dictionary and other modules.

The accuracy scores obtained on the Gold Standard Corpus are summarised in Table 6. This table shows that in each of the conditions, the accuracy increases. As can be seen in Table 7, most of the improvements are significant at a 99% confidence level (χ^2 test, 1 d.f.). Exceptions are the lemma when comparing C2 and C1, and the lemma and tag when comparing C4 and C3, which do not obtain a significant improvement (not even at the 95% level).

The results indicate that both adapting the dictionary and other modules and retraining the tagger have a positive impact on the overall performance of the extended tool on Old Spanish texts. The factor that has the highest impact is the dictionary expansion (together with the adaptation of the tokenization and affixation modules), with improvements ranging from 13.6% for PoS-2 to 18.3% for lemma. However, retraining the tagger, even if it is with a small corpus, also pays off in terms of precision: With 30,000 words, the performance on PoS-identification increases from 91.0% to 94.5%. The best result with the full set of tags (PoS-1) is 89.0% and 94.5% for the main PoS.

To compare the Old Spanish and Standard Spanish taggers on the same basis, we retrained the FreeLing Standard Spanish tagger on a 30,000-token fragment of the LexEsp corpus. The results for Standard Spanish, shown in the last row of Table 6, are still significantly higher (χ^2 test, 1 d.f., 99% conf. level) than those for the Old Spanish tagger: The accuracy over PoS-2 is 97.6%, 3 points higher than the 94.5% obtained for Old Spanish. The error analysis presented below shows the causes of these errors, giving clues as to how this performance could be improved.

7 Error analysis

The analysis of errors has been conducted over the 100 most frequent errors in tagging obtained with the Old Spanish tagger under condition C4. This analysis shows that most of the errors in the tagging are due to the ambiguity in the dictionary, as could be expected given the discussion in the previous section. Specifically, 90% of the errors corresponds to words for which the correct tag is available in the dictionary, but the tagger has not selected it. More than half of these errors (57.8%) are due to types which are also ambiguous in the Standard Spanish dictionary. The most frequent errors involve (i) function words such as determiner vs. clitic readings of *la, las* 'the/it' and relative pronoun vs. subordinating conjunction readings of *que* 'that', (ii) first and third person singular of verbal forms, which are homographs in Old Spanish (*queria* 'I|he wanted', *podia* 'I|he could'). The remaining 42.2% of the errors due to ambiguity are mostly words lacking the accent in Old Spanish. These are ambiguous verbal forms of the present and simple past (*llego* 'arrive|arrived'), pronouns (*que* 'what|that'), and adverbs (*mas* 'more|but'). Other errors correspond to types which were more ambiguous in Old Spanish, such as the already mentioned ambiguity for the coordinating conjunction (*y* 'and'). The 10% errors that are not due to ambiguity correspond to words which were not added by any of the methods used to expand the dictionary, mostly proper nouns (*pierres, antolinez*), but also other words not covered by any rule (*ovo* 'had', *coita* 'wish'). This low percentage shows that the dictionary expansion is quite thorough.

8 Discussion and future work

In this paper we have presented a method to extend an existing NLP tool in order to enable it to deal with historical varieties of a language. To our knowledge, this is the first time that such an strategy is pursued to automatically enrich Spanish historical texts with linguistic information. The modules for Standard Spanish of an existing tool, especially the dictionary and affixation modules, have been adapted using evidence from a large and representative Old Spanish corpus. Also the tagger has been retrained, using a 30,000-token Gold Standard Corpus. Thus, the tool for Standard Spanish has been extended, profiting from the similarity between the historical and standard varieties of Spanish, such that constructing a resource for Old Spanish required a relatively modest effort (around 6 person-months). As a result, we have obtained a reusable tool, which can be used to tag other corpora and be maintained and improved by other scholars.

The quality of the tagging is quite good: The tagger is able to correctly identify word lemmas in 92.6% of the cases, and in 94.5% the main PoS. The performance is still below the state-of-the-art for standard varieties of languages, and below the performance on a Corpus of Standard Spanish, but it is good enough to carry out quantitative analyses of historical data. We have shown that the lower performance is due to two factors: First, the increased ambiguity in the dictionary due to the large time span considered (the tool is able to tag texts from the 12th to the 16th centuries). Second, the small size of the training corpus. It is expected that the performance could improve by using the same methods to deal with PoS-disambiguation using context information in state-of-the-art tools. For instance, adding manual rules to the hybrid tagger included in FreeLing may improve the performance. Also, a spelling corrector could help solving the 10% of the errors which are not due to ambiguity but to orthographic variation.

The approach proposed could be followed to deal not only with historical varieties of languages, but also with other non-standard varieties, such as dialects or texts found in chats, blogs, or SMS texts. In the future, we will test it with so-called "Spanish 2.0".

Acknowledgments

This work is supported by the Ministry of Education and the Ministry of Science and Technology of Spain under contracts AP2006-03547 (FPU), FFI2010-15006 (OntoSem 2), TIN2009-14715-C04-04 (KNOW-II), and JCI2007-57-1479, and by the European Union via the EU PASCAL2 Network of Excellence (FP7-ICT-216886).

References

Alistair Baron and Paul Rayson. 2008. Vard 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham, UK. Aston University.

Thorsten Brants. 2000. Tnt - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.

Ivy A. Corfis, John O'Neill, and Jr. Theodore S. Beardsley. 1997. *Early Celestina Electronic Texts and Concordances*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Stefanie Dipper. 2010. Pos-tagging of historical language data: First experiments. In *emantic Approaches in Natural Language Processing: Proceedings of the Conference on Natural Language Processing (KONVENS 2010)*.

Andrea Ernst-Gerlach and Norbert Fuhr. 2007. Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL)*, Vancouver, British Columbia, Canada.

María Teresa Herrera and María Estela González de Fauve. 1997. *Concordancias Electrónicos del Corpus Médico Español*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Llyod Kasten, John Nitti, and Wilhemina Jonxis-Henkemens. 1997. *The Electronic Texts and Concordances of the Prose Works of Alfonso X, El Sabio*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Anke Lüdeling, Hagen Hirschmann, and Amir Zeldes. to appear. Variationism and underuse statistics in the analysis of the development of relative clauses in german. In Yuji Kawaguchi, Makoto Minegishi, and Wolfgang Viereck, editors, *Corpus Analysis and Diachronic Linguistics*. John Benjamins, Amsterdam.

John Nitti and Lloyd Kasten. 1997. *The Electronic Texts and Concordances of Medieval Navarro-Aragonese Manuscripts*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

John O'Neill. 1999. *Electronic Texts and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010), ELRA, La Valletta*, Malta, May 2010.

Helena Raumolin-Brunberg and Terttu Nevalainen. 2007. The York-Toronto-Helsinki Parsed Corpus of Old English Prose. In J.C. Beal, K. P. Corrigan, and H. L. Moisl, editors, *Creating and Digitizing Language Corpora. Volume 2: Diachronic Databases*, pages 148–171. Palgrave Macmillan, Hampshire.

P. Rayson, D. Archer, A. Baron, and N. Smith. 2007. Tagging historical corpora - the problem of spelling variation. In *Proceedings of Digital Historical Corpora, Dagstuhl-Seminar 06491, International Conference and Research Center for Computer Science*, Schloss Dagstuhl, Wadern, Germany, 3rd-8th December 2006.

Eiríkur Rögnvaldsson and Sigrún Helgadóttir. 2008. Morphological tagging of old norse texts and its use in studying syntactic variation and change. In *2nd Workshop on Language Technology for Cultural Heritage Data, LREC 2008 workshop*, Marrakech.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In Roberto Basili and Marco Pennacchiotti, editors, *Proceedings of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, Athens.

Cristina Sánchez-Marco, Gemma Boleda, Josep Maria Fontana, and Judith Domingo. 2010. Annotation and representation of a diachronic corpus of spanish. In *Proceedings of Language Resources and Evaluation (LREC)*, Malta, May 2010.

Pedro Sánchez-Prieto. 2005. La normalización del castellano escrito en el siglo xiii. Los caracteres de la lengua: grafías y fonemas. In Rafael Cano, editor, *Historia de la lengua española*, pages 199–213. Ariel, Barcelona.

Núria Sebastián, M. Antònia Martí, Manuel Francisco Carreiras, and Fernando Cuetos. 2000. *Léxico informatizado del español*. Edicions Universitat de Barcelona, Barcelona.

Ann Taylor. 2007. The York-Toronto-Helsinki Parsed Corpus of Old English Prose. In J.C. Beal, K. P. Corrigan, and H. L. Moisl, editors, *Creating and Digitizing Language Corpora. Volume 2: Diachronic Databases*, pages 196–227. Palgrave Macmillan, Hampshire.