Annotation and Representation of a Diachronic Corpus of Spanish

[1]Cristina Sánchez-Marco, [2]Gemma Boleda, [1]Josep Maria Fontana, [1,3]Judith Domingo

[1]Universitat Pompeu Fabra, [2]Universitat Politècnica de Catalunya, [3]Barcelona Media Centre d'Innovació
Roc Boronat 138 08018 Barcelona, C/ Jordi Girona 1-3 08034 Barcelona, Av. Diagonal 177 08018 Barcelona
cristina.sanchezm@upf.edu, gboleda@lsi.upc.es, josepm.fontana@upf.edu, judith.domingo@barcelonamedia.org

Abstract

In this article we describe two different strategies for the automatic tagging of a Spanish diachronic corpus involving the adaptation of existing NLP tools developed for modern Spanish. In the initial approach we follow a state-of-the-art strategy, which consists on standardizing the spelling and the lexicon. This approach boosts POS-tagging accuracy to 90, which represents a raw improvement of over 20% with respect to the results obtained without any pre-processing. In order to enable non-expert users in NLP to use this new resource, the corpus has been integrated into IAC (*Corpora Interface Access*). We discuss the shortcomings of the initial approach and propose a new one, which does not consist in adapting the source texts to the tagger, but rather in modifying the tagger for the direct treatment of the old variants.This second strategy addresses some important shortcomings in the previous approach and is likely to be useful not only in the creation of diachronic linguistic resources but also for the treatment of dialectal or non-standard variants of synchronic languages as well.

1. Introduction

The increasing availability of computational resources is opening new avenues for the study of linguistic change that not that long ago would have been unthinkable. The use of quantitative data allows linguists to track specific changes in the evolution of a language as well as to identify and describe general trends of change that would otherwise be very hard to trace accurately. Thus resources such as corpora and NLP tools are clearly becoming an indispensable tool enabling us to access diachronic data in an easier, faster and more efficient way than it was possible for traditional linguists. Some examples of the kinds of research results made possible by incorporating currently available NLP resources and techniques to the study of the evolution of a language can be seen for instance in (Han and Kroch, 2000), a study of the rise of *do*-Support in English using data from the *Penn-Helsinki Parsed Corpus of Middle English*, or (Sagi et al., 2009), which traces the semantic change of *dog*, *deer* and *do* by comparing the density of semantic vector clusters using a corpus derived from the Helsinki corpus. For other languages such as Spanish, however, the on-line resources available to the research community are rather limited. Thus, despite the quantity and quality of the documents included in electronic corpora such as *CORDE*[1] or *Corpus del Español*[2], researchers interested in the evolution of the Spanish language cannot conduct the type of studies conducted on the evolution of the English language due to the fact that the diachronic corpora available for this language are scarcely annotated with linguistic information and and the range of query options is not suffiently broad.

In order to overcome these limitations, we have embarked in a project that seeks to build resources to enable researchers to study the evolution of Spanish in the same depth as it is now possible for English. This resource has to satisfy three requirements: (i) the texts should be enriched with linguistic information, (ii) they should also contain paleographic information and (iii) the corpus should be easy to use by non-experts in NLP. To do this we are proceeding to annotate existing diachronic texts with morphosyntactic information and integrating the resulting corpus in a flexible search interface. In this paper we provide an overview of the architecture and design decisions we have made to annotate and represent this corpus.

2. Data and challenges

In the initial stages of this project, we have worked with the electronic texts compiled, transcribed and edited by the Hispanic Seminary of Medieval Studies (*HSMS*).[3] 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 words.

Working with the type of diachronic documents published by the *HSMS* poses some difficulties that are not usually encountered when working with traditional synchronic corpora. In the first place, these documents are characterized by a considerable variation in the spelling of words. Several different spellings of the same word can be found not only in texts from the same period but even within the same text (Sánchez-Prieto, 2005). An additional difficulty when working with high quality editions of medieval texts such as the ones produced by the *HSMS* is that these documents are enriched with a number of different symbols and special characters encoding information from the paleographic transcription of the old manuscripts.

Preserving this kind of information is vital for research in historical linguistics since in many cases these symbols provide clues that might prove to be very important to determine the relevance of certain linguistic data (e.g. margin annotations, scribal or editorial deletions and changes, revi-

[1] `http://www.rae.es`

[2] `http://www.corpusdelespanol.org`

[3] See Corfis et al. (1997), Herrera and de Fauve (1997), Kasten et al. (1997), Nitti and Kasten (1997), O'Neill (1999), Sánchez et al. (2003)

sions introduced in the manuscript by different scribes or in different periods, etc.; see Fontana (1993)). As we will see, these inherent difficulties determined the strategies adopted in the different stages of this project.

3. State of the art

Up to now, two different approaches have been adopted in order to enrich historical corpora with morphosyntactic information: manual annotation or automatic tagging followed by human correction. The former approach has been adopted in the annotation of large diachronic corpora such as the *Corpus do Portugués*[4] (Davies, 2002) and the *Tycho Brahe* (Britto et al., 2002). The latter has been adopted in projects such as the *Penn Historical Corpora*[5], *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor, 2007; Kroch and Taylor, 1995), the *Corpus of Early English Correspondence CEEEC* (Raumolin-Brunberg and Nevalainen, 2007).

A variant of the second strategy has recently been used in a number of projects (Rayson et al., 2007; Ernst-Gerlach and Fuhr, 2007; Baron and Rayson, 2008), namely, to standardize the corpora prior to their annotation with NLP tools. In these projects, spelling variants in historical English and German texts were identified and subsequently mapped onto their modern equivalents (i.e. the standardized or modernized forms). This is the approach we initially followed to annotate our corpus, as explained next.

4. Initial approach

The initial approach involved creating a standardized version of the source texts, and subsequently tagging the resulting normalized corpus with a modern Spanish tagger (FreeLing[6]).

In order to generate a standardized version of the texts two types of rules were applied: (i) regular rules, which map syllabic bigrams or trigrams independently of the word, and (ii) lexical rules, which map old forms onto their standard equivalents in those cases in which no regularities in the spelling are found. The tagger then assigns one morphological tag and one lemma to each form. This strategy boosts the accuracy of the tagging to 90%. This is around 15% more than the accuracy obtained in the tagging without preprocessing.

4.1. Variants and mapping rules

Spelling variation in historical Spanish texts is very noticeable throughout the medieval period. As mentioned in Section 2., several variants of the same word can be found not only in texts created within the same general historical period but also within the same text even when this has been transcribed by a single scribe. A certain normalization in the spelling of the words can be seen in the texts produced by public notaries during the kingdom of Fernando III (1217-1252) and Alfonso X El Sabio (1252-1284), who followed the uses of the Castilian chancellery. However, it is not until the 15th century that a certain unification in the spelling uses is observed. In the normalization of the spelling the most noteworthy date is 1517, when the *Reglas de Ortographía* written by Nebrija were published (Nebrija, 1517). Two centuries later, after the establishment of the Real Academia de la Lengua Española,[7] a new era of normative grammars and spelling rules began, with the subsequent decrease of spelling variants in texts. Ironically, spelling variation is starting to become prevalent again in mobile phone text messages, chats, blogs, etc., so we believe some of the techniques proposed here to process historical texts might prove useful to handle variation in these kinds of contemporary texts.

Different factors could play a role in the type of spelling variation present in the Old Spanish texts. Among them, the influence of Latin and paleographical and typographical factors stand out. As it is well-known, Latin was the most prestigious language throughout the Middle Ages in Western Europe. In the absence of clear spelling rules for the emerging Romance varieties, Latin was the only available model for scribes. Paleographical factors, such as the available space on the folio or the typography being used, could also influence the choice of one spelling variant (Torrens, 2002; Sánchez-Prieto, 2005).

Although at first sight this variation seems arbitrary, some regularities can be found. For example, the *ñ* usually appears in old texts as *n* or *nn*, as the Latin geminated variant (*donna/dona*, 'mistress'). The characters *i*, *j*, and *y* are also found in the texts representing the same sound. The *u* represents both a consonant and a vocal sound. Thus, besides representing the vocalic sound */u/*, the letter *u* was also often used to represent a consonantic sound (possibly a voiced labiodental fricative or perhaps a voiced bilabial fricative as some authors have suggested) in words such as *cauallo* ('horse') or *ueer* ('to see').

In the initial approach, 49 mapping rules were created that automatically mapped sequences of characters in an old spelling variant onto the corresponding modern standard variant. These mapping rules were based on the observed regularities in the spelling of Old Spanish texts (Sánchez-Prieto, 2005). These rules are independent of the morphophonological context, although 18% of them are restricted to the beginning or the end of a word. Figure 1 shows some examples of these rules.

Old variant	Modern variant	Example
eio	*ejo*	*meiorar → mejorar*
oie	*oge*	*coier → coger*
euo	*evo*	*nueuo → nuevo*
uio	*vio*	*uio → vio*
sçe	*ce*	*paresçe → aparece*
uen-	*ven*	*uenir → venir*
-rt	*-rte*	*cort → corte*

Figure 1: Examples of the spelling rules.

[4] http://www.corpusdoportugues.org

[5] http://www.ling.upenn.edu/hist-corpora
The *Penn Historical Corpora* includes the *Penn-Helsinki Parsed Corpus of Middle English*, second edition (*PPCME2*), the *Penn-Helsinki Parsed Corpus of Early Modern English* (*PPCEME*), and the *Penn Parsed Corpus of Modern British English* (*PPCMBE*)

[6] http://www.lsi.upc.edu/~nlp/freeling

[7] http://www.rae.es

Additionally, a total of 7000 lexical mapping rules were created to deal with the most frequent variants for words not covered in the 49 regular mapping rules, such as those produced by joined forms (*quelos* → *que los*, *conel* → *con el*) or words without accent (*consul* → *cónsul*, *perdon* → *perdón*). These rules directly transformed old variants into the corresponding current spellings.

4.2. Architecture

The architecture adopted in the initial stage of the project presented here is illustrated in Figure 2. In this initial approach the source texts were not directly analyzed by the tagger. In the initial stage, the paleographic symbols were deleted and the variants standardized by means of the two types of rules explained in Section 4.1. Once the source texts had been modernized, they were tagged. The whole corpus was tagged following this strategy.

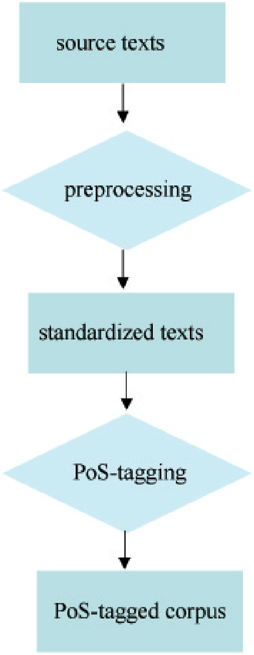


Figure 2: Architecture for the initial approach.

Figure 3 illustrates what the source text and the output of the preprocessing and tagging look like.[8] In this process, the analyzer assigns one tag and one lemma to each token in the texts. Apart from standardazing the spelling (*uenir* → *venir* 'to come'), note that the tokenization of Old Spanish is sometimes altered: The contracted form *destos* from the original transcription is converted to *de estos* 'of these'. Also note that the paleographic symbols (<, >, %) are deleted in the process, as they cannot be handled by the tagger.

4.3. Results

As shown in Table 1, this approach yields a POS-tagging accuracy of 91.5%, which is 14% more than the accuracy achieved by simply tokenizing the original text.

Input	Output
en <e>l otro q<ue> ha de uenir. % Pero destos	en en SPS00 el el DA0MS0 otro otro PI0MS000 que que PR0CN000 ha haber VAIP3S0 de de SPS00 venir venir VMN0000 . . Fp Pero pero CC de de SPS00 estos este DD0MP0

Figure 3: Excerpt of the corpus, before and after standardization and POS-tagging. In the POS-tagged version, each line contains one token, together with its lemma and POS tag.

Accuracy	POS-tags	Lemmas
Original corpus	77.5	76.1
Standardized corpus	91.5	91.2

Table 1: Accuracy in POS-tags and lemmas, estimated on four randomly chosen texts totalling 1500 tokens.

5. IAC: Corpora Interface Access

To enable non-expert users to make use of this new resource, the corpus and its linguistic annotations have been integrated in IAC (*Corpora Interface Access*), a corpus interface created by Barcelona Media Centre d'Innovació and Universitat Pompeu Fabra.[9] IAC is a flexible and powerful tool that allows for the creation of multilingual user-friendly interfaces between a given corpus and the underlying search tool.[10]

To create an interface in IAC, the corpus must be formatted according to the IAC specification (tabular format for attributes at word level and XML format for attributes affecting groups of words and metadata). The search interface is then designed using a graphical tool (included in IAC) according to type of corpus and linguistic annotation involved in each case.

The resulting on-line search interface is adapted to the characteristics of the corpus, mainly with respect to the types of information that can be searched (for example, a lemmatized corpus will allow searches based on lemma). The interface allows for 3 types of searches: simple, expert and frequency-based.

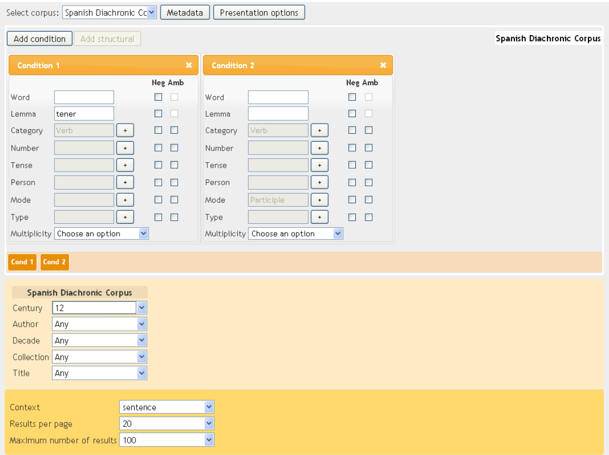
Simple searches make it possible to search a specific word (by form or lemma) and its POS (without features). Advanced search allows searching key words in context (KWIC). Moreover, each token can be restricted by form, lemma and PoS tag (divided into features, i.e. gender or number for nouns, mode for verbs, etc.). Searches can also be restricted by metadata such as author, century, etc.

[8] For further details on the standardizing procedure and the architecture, see Sánchez-Marco et al. (in press).

[9] http://www.barcelonamedia.org/

[10] IAC uses the IMS Open Corpus Workbench (CWB; http://cwb.sourceforge.net).

See Figure 4 for an example illustrating a search for *tener* (lemma) followed by a participle on texts from the 12th century. The results obtained by the user are shown in Figure 5.



Select corpus: Spanish Diachronic Cc

Condition 1

Word □ □
Lemma: tener □ □
Category: Verb □ □
Number □ □
Tense □ □
Person □ □
Mode □ □
Type □ □
Multiplicity: Choose an option

Condition 2

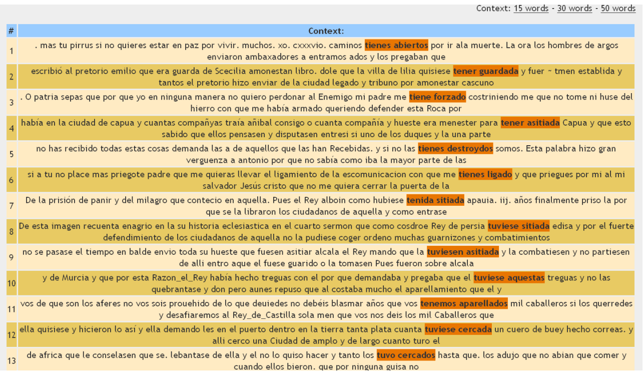
Word □ □
Lemma □ □
Category: Verb □ □
Number □ □
Tense □ □
Person □ □
Mode: Participle □ □
Type □ □
Multiplicity: Choose an option

Spanish Diachronic Corpus

Century: 12
Author: Any
Decade: Any
Collection: Any
Title: Any

Context: sentence
Results per page: 20
Maximum number of results: 100

Figure 4: Example of a search on the diachronic corpus using IAC. The user searches for a participle (*Condition 1*) followed by a verb (*Condition 2*) on texts from the 12th century (see meta-information at the bottom).



Context: 15 words - 30 words - 50 words

#	Context:
1	. mas tu pirrus si no quieres estar en paz por vivir. muchos. xo. cxxxvio. caminos **tienes abiertos** por ir ala muerte. La ora los hombres de argos enviaron ambaxadores a entramos ados y los pregaban que
2	escribió al pretorio emilio que era guarda de Scecilia amonestan libro. dole que la villa de lilia quisiese **tener guardada** y fuer ~ tmen establida y tantos el pretorio hizo enviar de la ciudad legado y tribuno por amonestar cascuno
3	. O patria sepas que por que yo en ninguna manera no quiero perdonar al Enemigo mi padre me **tiene forzado** costriniendo me que no tome ni huse del hierro con que me había armado queriendo defender esta Roca por
4	había en la ciudad de capua y cuantas compañyas traía añibal consigo o cuanta compañía y hueste era menester para **tener asitiada** Capua y que esto sabido que ellos pensasen y disputasen entresi si uno de los duques y la una parte
5	no has recibido todas estas cosas demanda las a de aquellos que las han Recebidas. y si no las **tienes destroydos** somos. Esta palabra hizo gran verguenza a antonio por que no sabía como iba la mayor parte de las
6	si a tu no place mas priegote padre que me quieras llevar el ligamiento de la escomunicacion con que me **tienes ligado** y que priegues por mi al mi salvador Jesús cristo que no me quiera cerrar la puerta de la
7	De la prisión de panir y del milagro que contecio en aquella. Pues el Rey alboin como hubiese **tenida sitiada** apauia. iij. años finalmente priso la por que se la libraron los ciudadanos de aquella y como entrase
8	De esta imagen recuenta enagrio en la su historia eclesiastica en el cuarto sermon que como cosdroe Rey de persia **tuviese sitiada** edisa y por el fuerte defendimiento de los ciudadanos de aquella no la pudiese coger ordeno muchas guarnizones y combatimientos
9	no se pasase el tiempo en balde envio toda su hueste que fuesen asitiar alcala el Rey mando que la **tuviesen asitiada** y la combatiesen y no partiesen de alli entro aque el fuese guarido o la tomasen Pues fueron sobre alcala
10	y de Murcia y que por esta Razon_el_Rey había hecho treguas con el por que demandaba y pregaba que el **tuviese aquestas** treguas y no las quebrantase y don pero aunes repuso que al costaba mucho el aparellamiento que el y
11	vos de que son los aferes no vos sois prouehido de lo que deuiedes no debéis blasmar años que vos **tenemos aparellados** mil caballeros si los querredes y desafiaremos al Rey_de_Castilla sola men que vos nos deis los mil Caballeros que
12	ella quisiese y hicieron lo así y ella demando les en el puerto dentro en la tierra tanta plata cuanta **tuviese cercada** un cuero de buey hecho correas. y alli cerco una Ciudad de amplo y de largo cuanto turo el
13	de africa que le conselasen que se. lebantase de ella y el no lo quiso hacer y tanto los **tuvo cercados** hasta que. los adujo que no abian que comer y cuando ellos bieron. que por ninguna guisa no

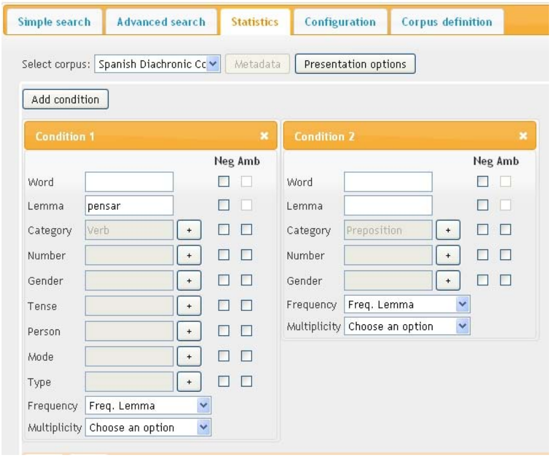
Figure 5: Results of the advanced search in Figure 5.

The use of IAC also provides an easy way to obtain meaningful quantitative data from corpus searches. A search to find the prepositions subcategorized by the verb *pensar* (*to think*) is shown in Figure 6. The results are grouped in a frequency table with the lemma *pensar* followed by the different prepositions (see Figure 7). It is also possible to access the sentence examples linked to the frequencies. The results (statistics and sentences) produced by IAC (simple, advanced and statistics search) can be also downloaded for further processing.

6. Discussion

6.1. Shortcomings of the initial approach

The approach we adopted at the initial stages of this project had, however, some significant shortcomings. First, the tagging accuracy is still far from the accuracy rate achieved by state-of-the-art taggers (95% and above). Secondly, the tokens of the source texts are lost in the standardizing preprocess (recall examples *uenir* and *destos* from the previous section) and establishing a mapping between each modernized form and its corresponding variants in the original texts is not a trivial task. Thirdly, recovering the paleographic information that is lost in the process also involves considerable technical difficulties. These problems could be solved by indexing the non-standard variants from the original texts with the corresponding variants resulting from the preprocessing tasks while also keeping indexes for all the deleted paleographic symbols.



Select corpus: Spanish Diachronic Cc

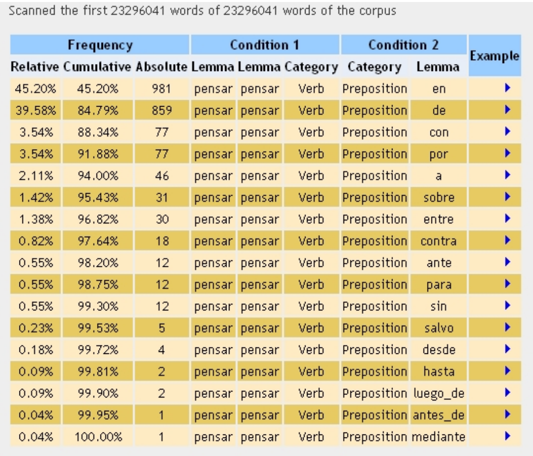
Condition 1

Word □ □
Lemma: pensar □ □
Category: Verb □ □
Number □ □
Gender □ □
Tense □ □
Person □ □
Mode □ □
Type □ □
Frequency: Freq. Lemma
Multiplicity: Choose an option

Condition 2

Word □ □
Lemma □ □
Category: Preposition □ □
Number □ □
Gender □ □
Frequency: Freq. Lemma
Multiplicity: Choose an option

Figure 6: Example of a frequency-based search on the diachronic corpus in IAC. The user searches for *pensar* (*Condition 1*) followed by a preposition (*Condition 2*) in texts of the 12th century.



Scanned the first 23296041 words of 23296041 words of the corpus

Frequency			Condition 1			Condition 2		Example
Relative	Cumulative	Absolute	Lemma	Lemma	Category	Category	Lemma	
45.20%	45.20%	981	pensar	pensar	Verb	Preposition	en	▶
39.58%	84.79%	859	pensar	pensar	Verb	Preposition	de	▶
3.54%	88.34%	77	pensar	pensar	Verb	Preposition	con	▶
3.54%	91.88%	77	pensar	pensar	Verb	Preposition	por	▶
2.11%	94.00%	46	pensar	pensar	Verb	Preposition	a	▶
1.42%	95.43%	31	pensar	pensar	Verb	Preposition	sobre	▶
1.38%	96.82%	30	pensar	pensar	Verb	Preposition	entre	▶
0.82%	97.64%	18	pensar	pensar	Verb	Preposition	contra	▶
0.55%	98.20%	12	pensar	pensar	Verb	Preposition	ante	▶
0.55%	98.75%	12	pensar	pensar	Verb	Preposition	para	▶
0.55%	99.30%	12	pensar	pensar	Verb	Preposition	sin	▶
0.23%	99.53%	5	pensar	pensar	Verb	Preposition	salvo	▶
0.18%	99.72%	4	pensar	pensar	Verb	Preposition	desde	▶
0.09%	99.81%	2	pensar	pensar	Verb	Preposition	hasta	▶
0.09%	99.90%	2	pensar	pensar	Verb	Preposition	luego_de	▶
0.04%	99.95%	1	pensar	pensar	Verb	Preposition	antes_de	▶
0.04%	100.00%	1	pensar	pensar	Verb	Preposition	mediante	▶

Figure 7: Results of the frecuency-based search

However, spelling variation is not the only difference between old and modern Spanish. There are also relevant lexical and syntactic differences that pose a more difficult challenge for the tagging and representation of the corpus. For instance, syntactic constructions such as clitic postposition (*dixol que* 'told him that') that no longer exist in modern Spanish texts (*le dijo que*) add some further complication to the indexing scheme contemplated above. Thus, the state of the art approach we adopted as explained in Section 4. turned out to be rather impractical and seemed to create some problems whose solutions seemed to be far from trivial from a technical point of view.

6.2. Adapting the tagger

For these reasons, it seemed to us introducing the necessary modifications to the tagger in order to adapt it for the processing of diachronic varieties of Spanish could have significant advantages over a strategy involving the adaptation the source text to the tagger in its current form. The new approach essentially involves using the existing modern Spanish tagger as a basis to create an Old Spanish tagger that automatically annotates old Spanish texts with a lemma and a PoS tag. This approach avoids the standardizing preprocess, which was the source of the difficulties described above. Adapting the tagger is also a useful endeavor in itself, as the resulting tool can be re-used to enrich other historical texts. Since Freeling is an open source tool, it can be further enhanced by the research community so that its use can be extended to similar projects. Thus, much of the work involved could be easily repurposed for the processing of diachronic texts from other Romance languages such as Catalan, French or Italian.

The adaptation of the existing FreeLing Spanish library will involve the expansion of the dictionary with the addition of variants present in the source texts and the modification of some other modules which are part of the library. Currently, the modules of FreeLing which are already adapted to Old Spanish are the affixation and the corrector module. The affixation module, linked to the dictionary, has been expanded with productive suffixes present in old texts, such as Old Spanish clitics (*-gelo, -l, -li*), variants of *-mente* adverbs (in *-mientre, -mjente, -mjentre, -ment*) or superlative suffixes such as *-issimo*.[11] Asystematic orthographic variation will be handled via a corrector module. In the future the grammar will also be adapted to the type of syntactic variation present in these texts.[12]

The main advantages of this approach are that the original text does not need to be altered and that we can use FreeLing as a library (see Figure 8). Note that with the new architecture both the paleographic symbols and the plain tokens are also parsed. The corresponding indices can thus be preserved and included in the corresponding offset position in the tagged corpus. This approach is therefore more robust, as it does not rely on an independent standardizing process.

As for the representation of the corpus, the new architecture greatly facilitates the inclusion of the different annotation levels in a stand-off format. Stand-off annotation stores metadata (structural, linguistic, and paleographic information) separately from the data it describes (source text). The main advantage of this representation is again that the source text is not altered. Moreover, additional types of annotation (e.g., syntactic or semantic information) can be added to the corpus in the future [13].

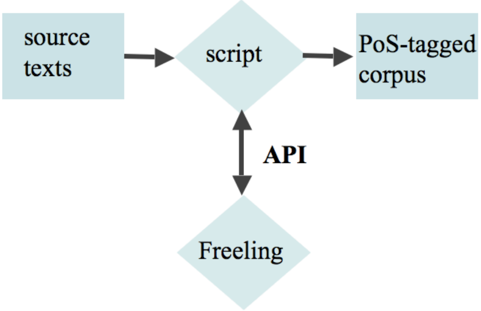


Figure 8: New architecture for the annotation of the diachronic corpus.

7. Conclusions and future work

In this paper we have examined two different strategies that can be considered in the development of linguistically annotated diachronic corpora. Taggers developed for current language varieties are poorly equipped to handle the wide range of spelling variants that is typically encountered in medieval texts. The first approach we considered involved the standardization of the source texts so that existing taggers could be used in the annotation of the transformed versions of the texts.

This initial approach, however, has some significant shortcomings with respect to both the annotation and the representation of the corpus. In order to overcome these problems, we have proceeded to adapt an existing open source tagger (Freeling) to handle the lexicon and grammar of the Old Spanish texts. Together with the flexible web interface provided by *IAC*, the Spanish diachronic corpus that we are in the process of creating will allow scholars to pursue new avenues of research that were previously not open. The flexible architecture we have adopted will also make it easy to continue to expand this corpus both in terms of its coverage and in terms of its usefulness by adding new texts and additional layers of linguistic annotation.

As far as we know, this is the first time that an existing tagger has been adapted to process diachronic varieties of the same language. We believe the methodology we have described could also be used to adapt similar existing taggers to process non-standard linguistic varieties, whether historical or contemporary (geographical or social dialects, cyberlanguages, etc).

Acknowledgments

This research has been partially funded by the Spanish government (projects KNOW: TIN2006-1549-C03-02, KNOW2: TIN2009-14715-C04-04, OntoSem: HUM2007-60599/FILO, programme Juan de la Cierva: JCI-2007-57-1479, FPU: AP2006-03547). We also thank three anonymous reviewers for their comments and suggestions.

[11] Note that the strategy of using spelling correction techniques for old texts has already been used with considerable success for English and German (Rayson et al., 2007; Baron and Rayson, 2008; Ernst-Gerlach and Fuhr, 2007).

[12] Note that the *relax* tagger in FreeLing is hybrid, allowing for the addition of manual rules to the basic statistical tagger.

[13] For details of the specific stand-off annotation architecture we propose, see Sánchez-Marco et al. (to appear).

8. 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.

Helena Britto, Marcelo Finger, and Charlotte Galves. 2002. Computational and linguistic aspects of the construction of the tycho brahe parsed corpus of historical portuguese.

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.

Mark Davies. 2002. Corpus del español (100 millones de palabras, 1200s-1900s).

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.

Josep Maria Fontana. 1993. *Phrase Structure and the Syntax of Clitics in the History of Spanish*. Ph.D. thesis, University of Pennsylvania.

Chung Hye Han and Anthony Kroch. 2000. The rise of do-support in english: implications for clause structure. In *Proceedings of the North East Linguistic Society (NELS 30)*, Washington D.C. Georgetown University.

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.

Anthony Kroch and Ann Taylor. 1995. *The Penn-Helsinki Parsed Corpus of Middle English*. University of Pennsylvania, Philadelphia, Department of Linguistics.

Antonio Nebrija. 1517. *Reglas de orthographía en la lengua castellana*. Ed. Antonio Quilis. Instituto Caro y Cuervo, Bogotá, 1977.

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.

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.

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.

María Nieves Sánchez, María Teresa Herrera, and María Purificación Zabía. 2003. *Textos medievales misceláneos*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Cristina Sánchez-Marco, Josep Maria Fontana, and Judith Domingo. in press. Anotación automática de textos diacrónicos del español. In *Proceedings of the VIII Congreso Internacional de Historia de la Lengua Española*, Santiago de Compostela. Universidad de Santiago de Compostela.

Cristina Sánchez-Marco, Gemma Boleda, and Josep Maria Fontana. to appear. Propuesta de codificación de la información paleográfica y lingüística para textos diacrónicos del español. Uso del estándar TEI. In *Proceedings of the Congreso Internacional Tradición e innovación: Nuevas perspectivas para la edición y el estudio de documentos antiguos*.

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.

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.

M. Jesús Torrens. 2002. *Edición y estudio lingüístico del Fuero de Alcalá (Fuero viejo)*. Fundación Colegio del Rey, Alcalá de Henares.