Improving POS Tagging in Old Spanish Using TEITOK

Maarten Janssen
CELGA-ILTEC
maarten@iltec.pt

Josep Ausensi
Universitat Pompeu Fabra
Department of Translation
and Language Sciences
josep.ausensi@upf.edu

Josep M. Fontana
Universitat Pompeu Fabra
Department of Translation
and Language Sciences
josepm.fontana@upf.edu

Abstract

In this paper, we describe how the TEITOK corpus tools helped to create a diachronic corpus for Old Spanish that contains both paleographic and linguistic information, which is easy to use for non-specialists, and in which it is easy to perform manual improvements to automatically assigned POS tags and lemmas.

1 Introduction

Although the availability of computational resources for the study of language change has experienced a considerable growth in the last decade, scholars still face considerable challenges when trying to conduct research in certain areas such as syntactic change. This is true even in the case of languages for which there already exist large corpora that are freely accessible on the internet.

One of such cases is Spanish. Despite the size and quality of the textual resources available through online corpora such as CORDE[1] or the Corpus del Español[2], researchers interested in the evolution of the Spanish language cannot conduct the type of studies that have been conducted, for instance, on the evolution of the English language due to the fact that the diachronic corpora available for Spanish are scarcely annotated with the relevant linguistic information and and the range of query options is not sufficiently broad.

This presentation reports work in progress within a project that seeks to redress this situation for Spanish. Our goal is to develop resources to study the evolution of Spanish in at least the same depth as it is now possible for English. These resources have to satisfy the following requirements: (i) the texts should also contain paleographic information, (ii) they should be enriched with linguistic information (initially POS tagging and eventually also syntactic annotation), (iii) the corpus should be easy to use by non-experts in NLP, and (iv) after the initial development stage, the corpus should also be easily maintainable and improvable by non-experts in NLP. The last requirement was especially relevant in our context because the development of corpora can be a very long term process and the financial resources to hire collaborators with the necessary technical skills are not constant and are heavily dependent on grants and projects which can be difficult to obtain for corpora that have already been financed through previous grants.

Specifically, we will discuss how the TEITOK interface helped in reaching these requirements for a diachronic corpus of Spanish (OLDES). A large portion of our corpus came from the electronic texts compiled, transcribed and edited by the Hispanic Seminary of Medieval Studies (HSMS)[3]. This is a large collection of critical editions of original medieval manuscripts which comprise a wide variety of genres and extend from the 12th to the 16th centuries. The HSMS texts were turned into a linguistic corpus enriched with POS tags and lemmas in the context of the dissertation work conducted by Sánchez-Marco (Sánchez-Marco et al., 2012). The initial version of this corpus was created in a traditional verticalized set-up using the Corpus Workbench (Evert and Hardy, 2015), henceforth CWB, and was tagged using a custom built version of Freeling (Padró et al., 2010) for Old Spanish. See Sánchez et al., (2010; 2011; 2012) for a more detailed description of the corpus as well as of the problems encountered in the initial stages of development.

[1] http://corpus.rae.es/cordenet.html
[2] http://www.corpusdelespanol.org/hist-gen/

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

2 TEITOK

The version of the corpus described here was created in the TEITOK platform (Janssen, 2015). TEITOK is an online corpus management platform in which a corpus consists of a collection of XML files in the TEI/XML format. Tokens are annotated inline, where token-based information such as POS and lemmas is modeled as attributes over those tokens. For searching purposes, an indexed version of the corpus in CWB is created automatically from the collection of XML files. With its CWB search option, TEITOK is comparable to systems like CQPWeb (Hardie, 2012), Korp (Ahlberg et al., 2013), or Bwananet (Vivaldi, 2009), with the difference that in TEITOK, the search engine additionally facilitates access to the underlying XML documents, along the lines of TXM (Heiden, 2010).

TEITOK has several attributes that make it able to respond to the four requirements mentioned in the introduction.

(i) The files of a TEITOK corpus are encoded in TEI/XML, a format that has been used extensively for encoding paleographic information. In the TEITOK interface, this information is not just present in the source code, but is graphically rendered, meaning that a TEITOK document looks like a pleasant-to-read paleographic manuscript.

(ii) TEITOK has inline nodes for tokens, as in e.g. the XML version of the BNC (BNC, 2007), which can be adorned with any type of linguistic information that is traditionally encoded in a verticalized text format, such as POS tags, lemmas, dependencies relations, etc. Furthermore, it makes a distinction between orthographic words and grammatical words, where a single orthographic word can contain multiple grammatical words (and vice-versa). This allows us to keep contractions such as *del* ('of the'), while also having the option of specifying the two grammatical words that form it: *de* ('of') and *el* ('the').

(iii) The online interface of TEITOK is designed for a broad and diverse audience, adding several features to make the corpus more easily accessible than traditional corpus interfaces: it provides an easy interface to search the corpus, in which it is possible to use the full CWB search language, but it also provides a simple form that will automatically generate a CWB search query behind the scenes. It also provides glosses for POS tags, eliminating the need to read through the tagset definitions.

(iv) Most relevantly for this paper, the same interface that is used for searching and viewing the corpus is also used to edit the corpus. This makes it easy for the administrators and authorized users to correct errors whenever they encounter them. There are also several tools available to make structural changes faster, which will be described in the next section.

Since philological information was removed in the CWB version of the corpus, we created the corpus again from the original files, this time keeping all the information provided in it. Since the two versions of the corpus were created independently, there are inevitably small differences between them: what counts as one token in one version sometimes counts as more than one in the other. This makes it close to impossible to import the tags from one version of the corpus to the other. As such, we used the Freeling parameters for Old Spanish that were developed as part of the original corpus, and applied them to the TEITOK version, resulting in a corpus that combines the linguistic and extralinguistic information in a single set of documents.

TEITOK allows for multiple orthographic realisations of the same word, which makes it possible to keep the paleographic form, and add a form in modernized orthography, making the corpus much more accessible to those not familiar with the old spelling forms. Since the lemmas provided by Freeling are in modern spelling, the modern spelling of the words was provided automatically (wherever possible), by looking up which current word corresponds to the POS tag and the modernized lemma. For instance, the word *rresçiban* was tagged as a present subjunctive (VMSP3P0) of *recibir* ('receive'). The modern Spanish lexicon for Freeling lists the form *reciban* for this, which was hence added as the modernized form.

Despite the efforts put into the initial tagging of the OLDES corpus, the level of accuracy was still not entirely satisfactory. The main objective in this stage of development was to improve the

overall quality of the tagging. For this, we decided to follow the following strategy: we set apart a selection of texts summing up to 1 million tokens, and tagged it with the Freeling tagger for Old Spanish. We then used several techniques provided by TEITOK to manually correct errors in this gold standard part of the corpus. After correcting the major errors, we trained NeoTag (Janssen, 2012) on this gold standard corpus, and applied the trained tagger to the rest of the corpus.

3 Improving POS tags

Independently of how good a POS tagger is, incorrect tags will always be created. In the case of a closed corpus like the HSMS corpus, it quickly becomes more efficient to correct errors created by the tagger than to attempt to improve the quality of the tagger. Traditionally, tagging correction has been done by hand, either in a text editor or an XML editor. Tools to facilitate tag correction are relatively new, such as ANNIS (Krause and Zeldes, 2016) or eDictor (Feliciano de Faria et al., 2010). Unlike most of these tools, TEITOK allows editing directly from the XML interface.

The TEITOK version of the HSMS texts provides a comfortable and quick way to manually correct tagging errors. The base mode of editing in TEITOK involves clicking on a word in the text. This opens up an HTML form, where any of the attributes of the word can be modified. Although this is very helpful when encountering a single error while using the corpus, it is not very efficient for large corrections. Therefore, there are three main options to speed up corrections.

The first option is the closest to the traditional way of correcting tagging errors: it is possible to get a verticalized version of a text, in which multiple tokens can be corrected at once, while still seeing the surrounding tokens. In the verticalized version the editor can correct a token in all its different layers of representation, i.e. transcription, written form, editor form, expanded form, critical form and normalized form. It is possible to see the different forms for the same token, and this renders the whole manual correcting process easier since it is possible to compare the original with the more modernized form of the same token. The verticalized version also allows the editor to correct POS tags and lemmas at the same time.

A second option is to correct errors from the text in modernized orthography. Although words still have to be corrected individually in this way, it becomes much easier to spot errors: any word that is not modernized was not recognized by the tagger, and will have an incorrect lemma, and, most likely, an incorrect POS tag as well. In many cases, if a word was recognized, but incorrectly tagged, it will have an incorrect modernized form. This makes it possible to just look for incorrect words in modern Spanish, which are much easier to spot than errors in POS or lemma. For instance, if the previous example *rresçiban* had been incorrectly modernized as *recibían* by the system, it would have been easy to recognize it by simply looking the normalized version of the text. Thus, in these cases, there is no need to check the actual POS tag (something much harder to process), because the tag can be inferred by the actual modern form.

And finally, multiple tokens throughout the corpus can be corrected in batch mode using CWB queries. CWB can be used to search for very specific words that are frequently tagged incorrectly, and all words in the resulting KWIC list are clickable to correct any errors they contain. It is also possible to correct all matching results in one go, either by changing the lemma for all of them to a specific value, or by going through all the matches in a verticalized format. Thus, TEITOK can use the output of a CWB search to edit the underlying XML files. This provides a reliable and fast way to quickly correct the errors previously spotted on the verticalized view; sometimes an error spotted while correcting a text on the verticalized view is indicative of a more general problem that applies to the whole corpus. This renders the whole correction process faster since, by spotting a generalized error on the verticalized view, the editor can simply correct all the incorrectly tagged tokens of the whole corpus via the interface.

An example is given in figure 1, where a relatively simple query is used to identify all words starting with *rr-*, which is no longer used in current Spanish orthography. We then asked the system to edit the normalized form for all of those, where the normalized form was furthermore pre-treated automatically by replacing all double *rr* for a single *r*. This allows editing all such words in one go, independently of which XML file they appear in.

This general procedure can be enhanced via simple strategies to identify specific incorrectly tagged tokens. For instance, a recurrent incorrectly tagged token is the word *vienes*, as it is of-

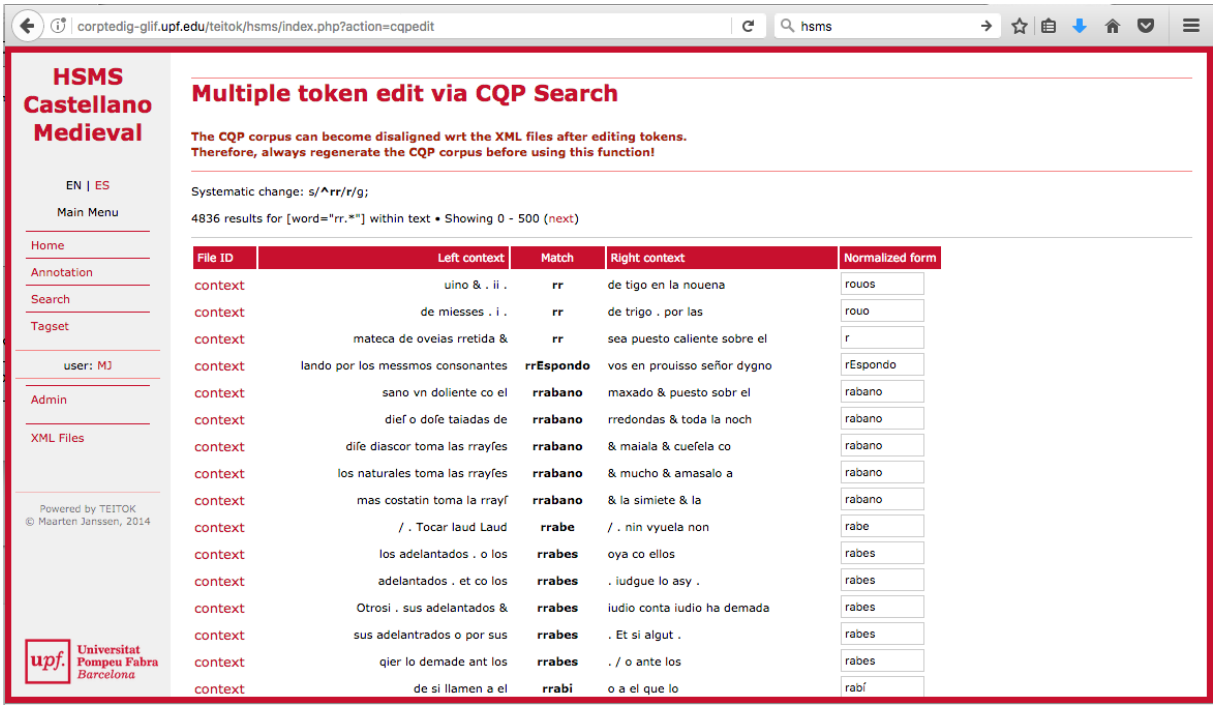

Figure 1: Multi-Editing in TEITOK

ten tagged as a verb ('you come') even though it is actually related to the modern noun spelled *bienes* ('goods'). By searching for all occurrences of the word *vienes* it is possible to correct all incorrectly marked ones in one go. It is even possible to search specifically only for occurrences of *vienes* that follow a determiner, by using a complex CQP query over multiple tokens in which the word *vienes* is marked as the target word (using the CQP operator @).

Other general problems that can be corrected automatically include examples such as the following: the form *a* is incorrectly tagged as a preposition ('to') when it relates to the modern form spelled *ha* ('he has'); the form *él* ('he') is tagged as a pronoun when it relates to the determiner *el*, or *partida* is tagged as a noun ('departure') when it relates to the participle ('departed'). All these generalized problems can be easily corrected taking advantage of the CWB interface and looking for specific combinations of the forms and specific lemmas or POS tags. For instance, searching for occurrences of *a* marked as a preposition, that are followed by a participle, gives only occurrences that should have been normalized as *ha* from the verb *haber*, hence making it possible to change all of them in batch mode. Searching for *él* followed by a noun returns instances of *él* that should have been tagged as a determiner, or searching for the lemma *ser* (i.e. be, in any form) followed by *partida* marked as a noun will return occurrences of *partida* that should have been tagged as a participle.

Since these different methods to correct errors in tags, lemmas, and normalized forms are easy to apply and do not require specific knowledge of the computational system, or imply that the corpus has to be rebuilt by a computational linguist, TEITOK allows all administrators of the corpus to correct errors over time - either by simply correcting individual errors, or by correcting multiple instances of an error throughout the corpus in batch mode as described in the previous paragraph. This means that the process of ironing out remaining errors is put back in the hands of the historical linguists, instead of requiring the technical support of external collaborators.

4 Conclusion

In this article, we hope to have shown how the TEITOK framework makes working with annotated historical corpora much easier: not only does it allow one to keep all paleographic information with the corpus, but it also makes it possible for linguists to correct annotation errors in an easy way, without the need to have detailed knowledge of the computational processes behind it. This re-

sult is a historical corpus that is useful not only for corpus linguists or syntacticians, but also, for instance, for historical linguists or philologists, and which can be improved over time, given that it is possible to correct errors whenever they are encountered. This is especially relevant in the context of historical corpora, since there are so many different sources of possible errors.

References

Malin Ahlberg, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal, and Jonatan Uppström. 2013. Korp and karp – a bestiary of language resources: the research infrastructure of språkbanken.

BNC. 2007. British national corpus, version 3 BNC XML edition.

Ivy A. Corfis, John O'Neill, and Jr. Theodore S. Beardsley. 1997. *Early Celestina Electronic Texts and Concordances*. Madison.

Stefan Evert and Andrew Hardy. 2015. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *10th International Conference on Open Repositories (OR2015)*, June.

Pablo Picasso Feliciano de Faria, Fabio Natanael Kepler, and Maria Clara Paixão de Sousa. 2010. An integrated tool for annotating historical corpora. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 217–221, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew Hardie. 2012. Cqpweb combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3).

Serge Heiden. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, and Yasunari Harada, editors, *24th Pacific Asia Conference on Language, Information and Computation*, pages 389–398, Sendai, Japan. Institute for Digital Enhancement of Cognitive Development, Waseda University.

María Teresa Herrera and María Estela González de Fauve. 1997. *Textos y Concordancias Electrónicos del Corpus Médico Español*. Madison.

Maarten Janssen. 2012. NeoTag: a POS tagger for grammatical neologism detection. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2118–2124.

Maarten Janssen. 2015. Multi-level manuscript transcription: TEITOK. In *Congresso de Humanidades Digitais em Portugal, Lisboa, October 8-9, 2015*.

Lloyd Kasten, John Nitti, and Wilhelmina Jonxis-Henkemans. 1997. *The Electronic Texts and Concordances of the Prose Works of Alfonso X, El Sabio*. Madison.

Thomas Krause and Amir Zeldes. 2016. Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118.

John Nitti and Lloyd Kasten. 1997. *The Electronic Texts and Concordances of Medieval Navarro-Aragonese Manuscripts*. Madison.

John O'Neill. 1999. *Electronic Texts and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings*. Madison.

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'10)*, La Valletta, Malta, May.

Cristina Sánchez-Marco, Gemma Boleda, and Lluís Padró. 2010. Annotation and representation of a diachronic corpus of spanish. In *Proceedings of the Language Resources and Evaluation Conference*, Malta, May. Association for Computational Linguistics.

Cristina Sánchez-Marco, Gemma Boleda, and Lluís Padró. 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. Association for Computational Linguistics.

Cristina Sánchez-Marco, J.M. Fontana, and J. Domingo. 2012. Anotación automática de textos diacrónicos del español. In *Actas del VIII Congreso Internacional de Historia de la Lengua Española*, Universidad de Santiago de Compostela.

Jorge Vivaldi. 2009. Corpus and exploitation tool: Iulact and bwananet. In *I International Conference on Corpus Linguistics (CICL 2009), A survey on corpus-based research, Universidad de Murcia*, pages 224–239.