Linguistically-enhanced search over an open diachronic corpus

Rafael C. Carrasco, Isabel Martínez-Sempere, Enrique Mollá-Gandía, Felipe Sánchez-Martínez, Gustavo Candela Romero, and M.Pilar Escobar Esteban

Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071, Alacant, Spain

Abstract. The BVC section of the IMPACT-es diachronic corpus of historical Spanish compiles 86 books —containing approximately 2 million words. About 27% of the words —providing a representative coverage of the most frequent word forms— have been annotated with their lemma, part of speech, and modern equivalent following the Text Encoding Initiative guidelines. We describe how this type of annotation can be exploited to provide linguistically-enhanced search over historical documents. The advanced search supports queries whose search terms can be a combination of surface forms, lemmata, parts of speech and modern forms of historical variants.

1 Introduction

Diachronic corpora are a valuable source of information to understand the historical evolution of languages. This paper describes a web-based search tool built upon the Apache Lucene[1] platform. Currently, the tool supports advanced search over the BVC section of the IMPACT-es diachronic corpus of historical Spanish [4] distributed by the Impact Centre of Competence in Digitisation.[2]

The corpus contains 86 Spanish texts provided by the Biblioteca Virtual Miguel de Cervantes,[3] printed between 1482 and 1647; it covers a representative variety of authors and genres (such as prose, theatre, and verse). This corpus is one of the few collections of historical Spanish distributed under an open license.

The original spelling (even if clearly unintentional) has been preserved in order to achieve a highly accurate transcription. The metadata added to Spanish words are its lemma (in modern form), its part of speech and its modern equivalent. The words originating from other languages (less than 0.1%, and principally Latin) are labelled solely with their language. The morphological categories which have been considered are abbreviation, adjective, adverb, conjunction, determiner, interjection, noun, proper noun, numeral, preposition, pronoun, relative pronoun, and verb. The annotation process was assisted by the CoBaLT tool [1], which supports complex annotations.

[1] http://lucene.apache.org/

[2] http://www.digitisation.eu/data/browse/corpus/impact-es

[3] http://www.cervantesvirtual.com

2 The Query Language and Interface

The interface with the search engine is available at `http://bvmcresearch.cervantesvirtual.com/diasearch` where multiple query terms can be specified. As in [3], every term can be preceded by a prefix:

- If no prefix is added, the term denotes a diachronic form (verbatim text).
- The prefix `modern#` denotes a modern form.
- The prefix `lemma#` is followed by a lemma.
- The prefix `pos#` denotes a part-of-speech tag.

Multiterm queries can include different prefixes and use the rich query language[4] provided by Lucene, the open source information retrieval Java library. Words or text segments matching the query are highlighted and presented in their context (snippet).

The index is based on Lucene's synonym list where a filter is applied to expand every input token. The index contains then, for every word form, all possible modern forms, lemmata, and parts of speech; For example, the word form *celebrada* generates 5 entries (from two analysis: lemma#celebrar, pos#verb, modern#celebrada and lemma#celebrado, pos#adj, modern#celebrada) while the word form *yerro* generates 7 entries (lemma#yerro, pos#n, modern#yerro; lemma#hierro, pos#n, modern#hierro; lemma#errar, pos#verb, modern#yerro).

A diachronic form can be assigned more than one modern equivalent (e.g., the historical spelling *fijo* is compatible with the adjective *fijo* and the noun *hijo*). The diachronic form can be also compatible with multiple lemmas and parts of speech. Since the historical spelling was less constrained than modern orthography, old texts usually show a higher rate of homography.

For optimal retrieval, what is considered a word has been carefully defined: words may contain characters in the Unicode category letters separated by non-breaking symbols such as the dash, the ampersand and plus signs, dots, etc.

The tool uses the `doBVMCDiaSearch` method defined in the BVMCSearch[5] public service, which provides standard JSON output (JavaScript Object Notation)[6]. The interface is implemented in AJAX and is based on Simple Object Access Protocol (SOAP)[7], Web Services Definition Language (WSDL)[8] and the XML Schema Definition language (XSD). The tool allows for the pagination of results, navigation through the result pages, the highlighting of matches, etc. while maintaining the compatibility with the most important browsers and devices.

[4] `http://lucene.apache.org/core/3_6_0/queryparsersyntax.html`

[5] `http://app.cervantesvirtual.com/cervantesvirtual-web-services/BVMCSearchWSService?wsdl`

[6] JSON `http://www.json.org` defines a language to store and exchange textual information which is smaller, faster and easier to parse than XML.

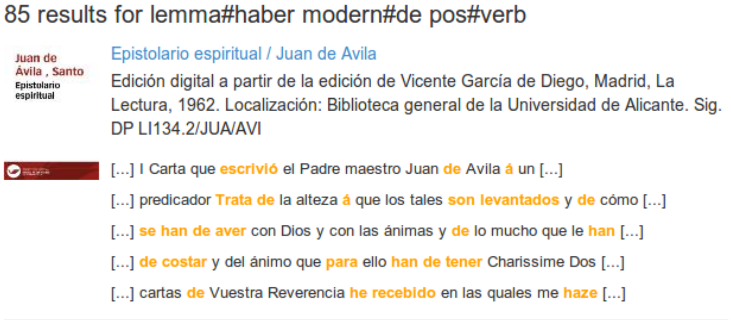
[7] urlhttp://www.w3.org/TR/2007/REC-soap12-part0-20070427/

[8] `http://www.w3.org/TR/wsdl`

Parameter	Type	Default	Example
`q`	String		pos#n
`start`	int	0	0
`maxResults`	int	10	10
`fragmentNumber`	int	0	0
`fragmentSize`	int	10	10

3 Main Features and Possible Improvements

Although only a fraction of the BVC section of the IMPACT-es corpus has been annotated, the information of the words with morphological tags can be extrapolated to non-annotated words and the coverage then grows from 27% to 92% (with a decrease in precision). Fortunately, for multiterm queries (with only a few terms) the implicit intersection often disambiguates the interpretation of the text [3]. Figure 1 shows the results retrieved when the BVC section of the IMPACT-es corpus is interrogated with the query `"lemma#haber modern#de pos#verb"`. Note that, although the word form *a* can be a form of the verb *haber* or a preposition, the first option is never followed by *de*.


85 results for lemma#haber modern#de pos#verb

Epistolario espiritual / Juan de Avila

Edición digital a partir de la edición de Vicente García de Diego, Madrid, La Lectura, 1962. Localización: Biblioteca general de la Universidad de Alicante. Sig. DP LI134.2/JUA/AVI

[...] I Carta que **escrivió** el Padre maestro Juan **de** Avila **á** un [...]

[...] predicador **Trata de** la alteza **á** que los tales **son levantados** y **de** cómo [...]

[...] **se han de aver** con Dios y con las ánimas y **de** lo mucho que le **han** [...]

[...] **de costar** y del ánimo que **para** ello **han de tener** Charissime Dos [...]

[...] cartas **de** Vuestra Reverencia **he recebido** en las quales me **haze** [...]

Fig. 1: Sample results after the query `"lemma#haber modern#de pos#verb"`.

The lexicon is the set of all unique word forms annotated in the corpus (25423 words with over half a million attestations). Almost one tenth of the word forms can be assigned more than one lemma but only 1% of the words admit more than one modernisation. This is an indication that automatic modernisation can be applied with a reasonable performance.

The ambiguity raises a sample of the corpus is analysed (rather than the lexicon): the number of words with more than one modern equivalent raises to 17.2% due to the ubiquity of the most frequent words (the so-called *stop-words*), most of which, remarkably, happen to be ambiguous. In the collection, just five word forms among the 10 most frequent words add up to 14.89% of the cases with multiple modern forms.

This ambiguity can be often resolved if the exact part of speech of the word is known. This suggest using standard part-of-speech taggers, which reach 97% precision [2]. We have found that 92% of the multi-lemma word forms are amenable to disambiguation with a standard part-of-speech tagger trained over Spanish text. Of course, there are exceptions such as the Spanish form *yerro* which can be assigned the lemma *errar* (a verb) and the lemma *hierro* (a noun), but also the lemma *yerro* (also a noun).

A small fraction (about 1%) of the analysed word forms allow for more than one modern equivalent but one half of them are disambiguated with the part-of-speech information. Therefore, a very high precision can be obtained for this type of search. Next table shows the fraction of ambiguous terms in the lexicon (25423 word forms) and in a sample of the corpus (containing 1,982,882 word forms).

Ambiguous annotations	Lexicon	Texts
Lemmata	2,518 (9,9%)	688,452 (37.5%)
Parts of speech	4,515 (17.8%)	697,539 (38.1%)
Modern forms	277 (1.1%)	314,616 (17.2%)

4 Conclusions

We have implemented an on-line service which allows to search over a diachronic corpus using a combination of query terms that may refer to historical forms, modern forms, lemmata or parts of speech. In order to increase recall we have extrapolated the annotations to all the occurrences of each word form. Although, this generalization may reduce the precision, it provides accurate results when the query contains a multiple search terms. The accuracy can be further improved with the integration of a part-of-speech tagger.

References

1. Kenter, T., Erjavec, T., Dulmin, M.Z., Fiser, D.: Lexicon construction and corpus annotation of historical language with the CoBaLT editor. In: Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 1–6. Association for Computational Linguistics, Avignon, France (April 2012)
2. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press (2001)
3. Sánchez-Martínez, F., Forcada, M.L., Carrasco, R.C.: Searching for linguistic phenomena in literary digital libraries. In: Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage. Aarhus, Denmark (September 2008)
4. Sánchez-Martínez, F., Martínez-Sempere, I., Ivars-Ribes, X., Carrasco, R.C.: An open diachronic corpus of historical Spanish. Language Resources and Evaluation (2013), dOI: 10.1007/s10579-013-9239-y