Procesamiento del Lenguaje Natural, nº39 (2007), pp. 289-290 recibido 05-06-2007; aceptado 22-06-2007

The Coruña Corpus Tool*

Javier Parapar
IRLab, Computer Science Dept.
University of A Coruña, Spain
Fac. Informática, Campus de Elviña
15071, A Coruña, SPAIN
javierparapar@udc.es

Isabel Moskowich-Spiegel
MUSTE, English Philology Dept.
University of A Coruña, Spain
Fac. Filología, Campus de Zapateira
15070, A Coruña, SPAIN
imoskowich@udc.es

Resumen: El Coruña Corpus de documentos científicos será usado para el estudio diacrónico del discurso científico en la mayoría de los niveles lingüísticos, contribuyendo de esta forma al estudio del desarrollo histórico del inglés. El Coruña Corpus Tool es un sistema de recuperación de información que permite compilar conocimiento sobre el corpus.
Palabras clave: Lingüística de corpus, inglés científico-técnico, recuperación de información.

Abstract: The Coruña Corpus of scientific writing will be used for the diachronic study of scientific discourse from most linguistic levels and thereby contribute to the study of the historical development of English. The Coruña Corpus Tool is an information retrieval system that allows the extraction of knowledge from the corpus.
Keywords: Corpus linguistics, English scientific writing, information retrieval.

1. Introduction

The Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing was carried out since 2003 by the Muste Group of the University of A Coruña. The corpus compilation is still in progress, at the moment we have gathered together samples of 10,000 words approximately belonging to the field of eighteenth- and nineteenth-century mathematics and astronomy.

In order to manage all the information that will be present in the corpus and to facilitate linguists the gathering of data, a corpus management tool, the Coruña Corpus Tool (CCT) has been developed in collaboration with the IRLab of the University of A Coruña. In this demo we would like to present to the natural language processing community the main characteristics of the corpus compilation process and its management tool.

* Acknowledgements: The research which is here reported on has been funded by the Xunta de Galicia through its Dirección Xeral de Investigación e Desenvolvemento, grant number PGIDIT03PXIB10402PR (supervised by Isabel Moskowich-Spiegel). This grant is hereby gratefully acknowledged. The first author also has to acknowledge the funds of the "Secretaría de Estado de Universidades e Investigación" and FEDER (MEC TIN2005-08521-C02-02) and "Xunta de Galicia"(PGIDIT06 PXIC10501PN).

2. The Coruña Corpus

The Coruña Corpus (CC) has been designed as a tool for the study of language change in English scientific writing in general, as well as within the different scientific disciplines. Its purpose is to facilitate investigation at all linguistic levels, though, in principle, phonology is not included among our intended research topics. The CC contains English scientific texts other than medical produced between 1650 and 1900. Medical texts have been disregarded since they are being compiled by Taavitsainen and Pahta and their team in Helsinki (Taavitsainen and Pahta, 1997). Our project proposes to complement other corpora pertaining to the history of what we nowadays call ESP, such as the well-known Corpus of Early English Correspondence, the Corpus of Early English Medical Writing, and the Lampeter Corpus of Early Modern English Tracts.

From the six areas into which UNESCO divides Science and Technology we are compiling samples of texts, at the moment, from: *Exact and Natural Sciences*: Mathematics, Astronomy, Physics and Natural History; *Agricultural Sciences* and *Humanities*: Philosophy and History. We intend to compile the same number of samples for each scientific field in order to facilitate comparative studies. For each discipline we have selected two texts per decade, with each sample con-

ISSN: 1135-5948

taining around 10,000 words, excluding tables, figures, formulas and graphs.

3. The Coruña Corpus Tool

In order to retrieve information from the compiled data, we decided to create a corpus management tool. This software application is currently in its testing phase. It is designed to help linguists to extract and condense valuable information for their research. The Coruña Corpus Tool (CCT) is an Information Retrieval (IR) platform (see Figure 1) where the indexed textual repository is the set of compiled documents that constitutes the CC. The texts that conform the CC we-

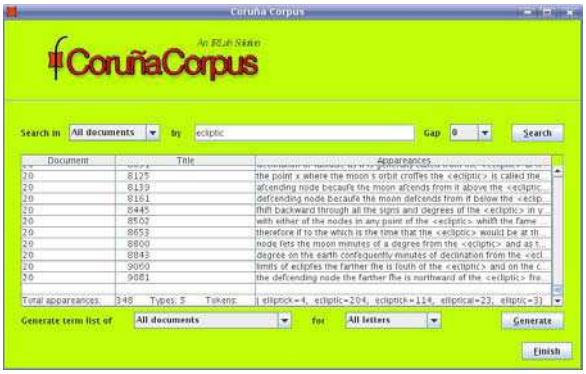

Figura 1: A snapshot of the application.

re coded and stored as XML documents. We chose to tag the information following the recommendations of the TEI (Text Encoding Initiative) (Sperberg-McQueen and Burnard, 2002) standard. Several tagged fields that we desire to index are extracted from the documents. In this sense we have to notice that we build a multi-field index to allow searches using different criteria; we store, for instance, information about authors, date, scientific field, corpus document identifier, etc.

It is fair to mention here that we used some existing open-source libraries for the system implementation. Among them we would like to mention Lucene: it is an indexing library (Apache, 2007) widely used in the development of IR applications.

3.1. Features

The system offers among others the next functionalities:

Document validation: if the document is not correctly constructed according to the DTD rules, the syntax validator will show the coders the errors present in the document so they can be fixed.

Basic term search: it can be launched over the whole set of indexed documents or at individual document level. As the result of a user query all the occurrences of a word are shown. For each one the following information is available: document identifier, word position and concordance.

Advanced search: a certain number of custom search characteristics are implemented to facilitate the extraction of research results:

- Wild card use: the inclusion of wild card characters are allowed to specify the searching of spelling variations of the same form along time.
- Regular expression searching: to allow searching using patterns, it is useful for example to search by suffixes or prefixes.
- Phrase search: combinations of words can be specified as a query indicating the gap between the words. This can be used for instance to look for expressions or verbal forms.

Term list generation: generation of lexicon lists of the whole corpus or inside each document (as chosen). An alphabetical sorted list of words with the number of appearances is generated filtered by the user criteria.

4. Conclusions

As previously explained, the CC is still a work in progress. We have a lot of text to compile and codify yet. But the CCT is designed to be scalable and adaptable to the new needs of the corpus compilation process. The CCT is currently an option to manage any TEI encoded corpus and offers the features more often demanded by linguists.

References

Apache, Foundation. 2007. *Lucene: http://lucene.apache.org/.*

Sperberg-McQueen, C. M. and L. Burnard. 2002. TEI P4: Guidelines for electronic text encoding and interchange. In *Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen.*

Taavitsainen, Irma and Päivi Pahta. 1997. Corpus of early english medical. In *ICAME '97: Proceedings of the International Computer Archive of Modern and Medieval English Conference*, pages 71–78. Kluwer Academic Print on Demand.