Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 13-19 recibido 12-11-2018 revisado 07-01-2019 aceptado 05-02-2019

The Coruña Corpus Tool: Ten Years On

El Coruña Corpus Tool: diez años después

Anabella Barsaglini-Castro, Daniel Valcarce
University of A Coruña
{anabella.barsaglini.castro,daniel.valcarce}@udc.es

Abstract: In this paper we provide a brief introduction to a new version of the Coruña Corpus Tool. Currently available for Windows, macOS and Linux, the Coruña Corpus Tool is a corpus management tool that facilitates the retrieval of information from an indexed textual repository. Although it works like most concordance programs, its distinguishing feature is that it allows users to search for old or non-standard characters and tags in texts and metadata files, as well as to extract and export specific data for the purposes of research. With a new set of advanced search features and other recent improvements, researchers now have access to functionalities that significantly enhance the previous user experience.
Keywords: corpus management, information retrieval, software tools, concordance, Coruña Corpus

Resumen: En este artículo presentamos una breve introducción a una nueva versión del Coruña Corpus Tool. Actualmente disponible para Windows, macOS y Linux, el Coruña Corpus Tool es una herramienta de gestión de corpus que facilita la recuperación de información desde un repositorio textual indexado. Aunque funciona como la mayoría de los programas de concordancia, su característica distintiva es que permite a los usuarios buscar caracteres y etiquetas antiguos o no estándar en archivos de texto y metadatos, así como extraer y exportar datos específicos con fines de investigación. Con un nuevo conjunto de funciones de búsqueda avanzada y otras mejoras recientes, los investigadores ahora tienen acceso a funcionalidades que mejoran significativamente la experiencia previa del usuario.
Palabras clave: gestión de corpus, recuperación de información, herramientas informáticas, concordancia, Coruña Corpus

1 Introduction

Created as a beta version in 2007 for the Research Group for Multidimensional Corpus-based Studies in English (MuStE[1]), and released in 2012 with the publication of the *Corpus of English Texts on Astronomy*, *CETA* (Moskowich et al., 2012), the Coruña Corpus Tool (CCT) continues to be an inseparable companion to the *Coruña Corpus of English Scientific Writing* (*CC*). From its inception in 2003, the *CC* has grown and evolved. The aim of this ongoing project is the compilation of samples of scientific texts from the late Modern English period into a specialised corpus with common principles (Crespo and Moskowich, 2010; Moskowich, 2016), making possible diachronic and synchronic studies at most linguistic levels.

In parallel with the compilation of the corpus itself, we have been developing the CCT, an information retrieval platform. This tool is designed to manage, gather and query the corpus using modern information retrieval techniques. In light of the publication of the *Corpus of Historical English Texts*, *CHET* (Moskowich et al., 2019), the CCT returns with a series of improvements in efficiency and effectiveness that will immediately be noted by users. This paper presents the main features of the CCT and its recent improvements by focusing on the basic operation of the software. Our aim is, therefore, to introduce the software to researchers interested in the analysis of discourse and language, in the belief that the tool will help them to extract specific data for their research. In order

[1] https://www.udc.es/grupos/muste

ISSN 1135-5948. DOI 10.26342/2020-64-1

to provide as much detail as possible, Section 2 deals with the general characteristics of the software, and Section 3 focuses on in improvements both as a manager for corpus compilation and as an information retrieval tool for end-users. Finally, Section 4 analyses the contributions of the CCT in contrast to similar tools.

2 About the Coruña Corpus Tool

Created as a corpus management tool to facilitate the gathering of data, the CCT has been developed by the IRLab[2] group in collaboration with the MuStE research group of the University of A Coruña, Spain (Parapar and Moskowich, 2007). Its main purpose is to retrieve information from a set of compiled documents that the *Coruña Corpus* (*CC*) comprises, in order to help linguists to extract specific data for their research. To this end, all the texts in the *CC* are compiled, marked-up, encoded and stored as XML files according to the Text Encoding Initiative P5 standard (see The TEI Consortium (2019) for the specification). In combination with the textual documents, MuStE also compiles metadata files that offer extra information about the authors and their texts.

The CCT is a multi-platform desktop application (it can be executed in Linux, macOS and Windows) written in Java. The tool parses XML files following the TEI standard and extracts the tagged fields that we wish to index, such as information about authors, date, scientific field (or the subcorpus to which a particular sample belongs), content, and document identifier. The XML files are validated using a Document Type Definition (DTD), a file that defines a set of syntax rules for XML documents. At this stage during the compilation, the tool also shows any errors found in the XML files to help coders to deal with any issues of syntax that might arise.

The tool builds a multi-field index structure that allows searches using different criteria, both in the samples and in the metadata accompanying them. Thus, users can execute queries on the whole set of documents (an option that is shown by default); on the individual document level, which allows the selection of a single sample from the whole corpus of corpora that have been loaded; or even on a subset of samples. Search results are displayed in a table or grid, typical of concordance programs, showing the document identifier, plus the position and surrounding context of the match. Advanced search features are also available. As the texts compiled date from the eighteenth and nineteenth centuries, users can use wildcards to specify spelling variants of the same form (*e. g.,* ⟨e⟩, ⟨æ⟩, ⟨œ⟩) and regular expressions to make complex queries that match patterns such as prefixes or suffixes. Additionally, they can use phrase queries involving combinations of words with a specified number of spaces between them as a means of finding specific expressions or verbal forms. Thus, if none (0), one (1) or two (2) spaces are selected when searching for "the answer", results will include occurrences matching these consecutive terms, a term between them (*e. g.,* "the right answer"), or two elements in between them (*e. g.,* "the logical right answer"), respectively. These advanced search functionalities are implemented using Apache Lucene, an open-source information retrieval library that offers state-of-the-art search features.

The CCT is also able to generate frequency lists from the whole set of documents or a subset of the corpus or corpora (when more than one corpus is loaded). These alphabetically sorted lists also contain the number of occurrences or tokens of each term (type). Additionally, the user can use filters to select documents that satisfy the required criteria.

Finally, the tool also provides styled document rendering to view text samples and metadata files. The XML files are rendered using Cascading Style Sheets (CSS) and an integrated web browser offers a visually pleasant user experience.

The CCT is composed of two executables, called 'Manager' and 'Client'. The Manager is used by compilers to build a corpus from a set of documents and metadata files, whereas the Client (available for users in general) is intended for searching the corpus, viewing documents and generating word lists and concordances.

3 New features and improvements

Over its ten years of existence, the CCT has evolved from being an Information Retrieval platform accompanying an indexed repository of English scientific texts, to a more ma-

[2] `https://www.irlab.org`

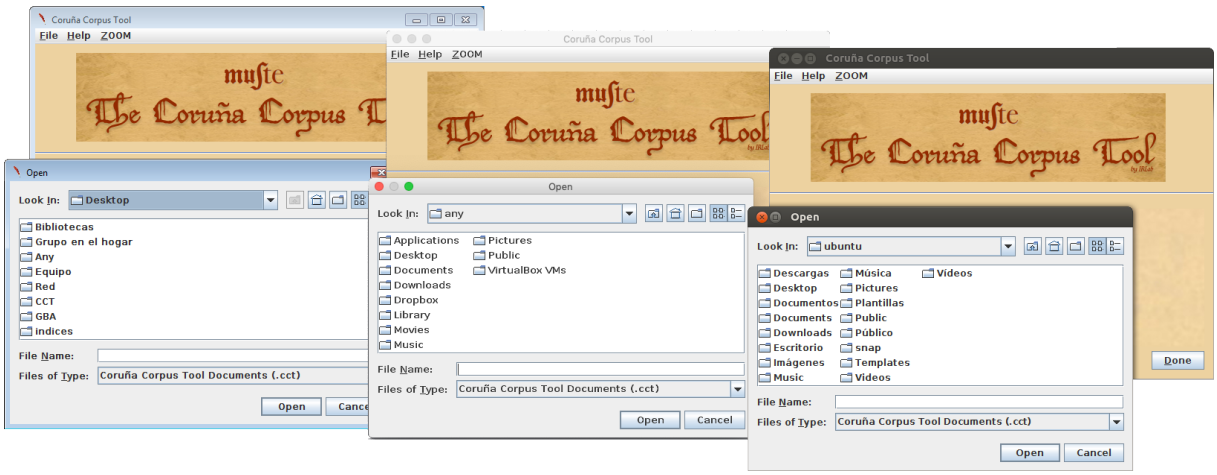

Figure 1: New "Look & Feel" on the CCT for Windows, macOS and Linux

ture kind of software that can create, manage and query the stored collections following the TEI standard (The TEI Consortium, 2019). Over the years, new functionalities have steadily been added, as well as bug fixes and improvements to usability. Nonetheless, it is important to note that compatibility is one of the fundamental principles of the development of this software. Hence, newer versions of the CCT are always backwards compatible with corpora indexed by older versions, as with the *Corpus of English Texts on Astronomy*, *CETA* (Moskowich et al., 2012) and the *Corpus of English Philosophy Texts*, *CEPhiT* (Moskowich et al., 2016). In the following subsections, we present an overview of the most notable changes.

3.1 General improvements

We have updated the CCT to work with modern versions of Java, with Java 8 being the minimum version required to run the tool. We have taken advantage of the new functionalities provided by Java 8 to improve the efficiency and security of the CCT. We also packaged all the external dependencies accompanying the older version in the same file using Maven, a Java build tool, which also offers faster compilations. In this way, users only need to interact with two applications ("manager.jar" and "client.jar" for the Manager and the Client, respectively) to access all functionalities.

We have endeavoured to solve all compatibility issues and to ensure that Windows, macOS and Linux users can all use the tool seamlessly. One of the tasks undertaken in this respect was the unification of the "Look & Feel" of the tool for the three platforms, so as to offer the same visual experience to all users (Figure 1).

3.2 User experience improvements

In this version of the CCT, we have improved the visualisation of documents (samples) by employing a modern web engine (Figure 2). As noted above, the XML files are styled using Cascading Style Sheets (CSS) and the new embedded web browser is capable of rendering the documents in high quality.

One feature that sets the CCT apart from other concordance programs is that from the first version it has always been designed with visual accessibility in mind. It includes a zoom feature to adapt the rendering size of the content. In the current revision of the tool, we have also added a more fine-grained control of the zoom level. Moreover, the state of the application is now saved to avoid losing the search results when the zoom level is modified.

The names of some labels have been updated, as has the size of the windows required to adapt better to modern screen resolutions. Nonetheless, users can resize the windows to make them larger if required. Finally, we have added a searching box to the digital manual to enable users to look for specific help.

3.3 Manager improvements

After several years using the Manager to index linguistic corpora, we found that the most common problems were related to file encoding. Therefore, to avoid issues of this kind entirely, the Manager encodes all XML

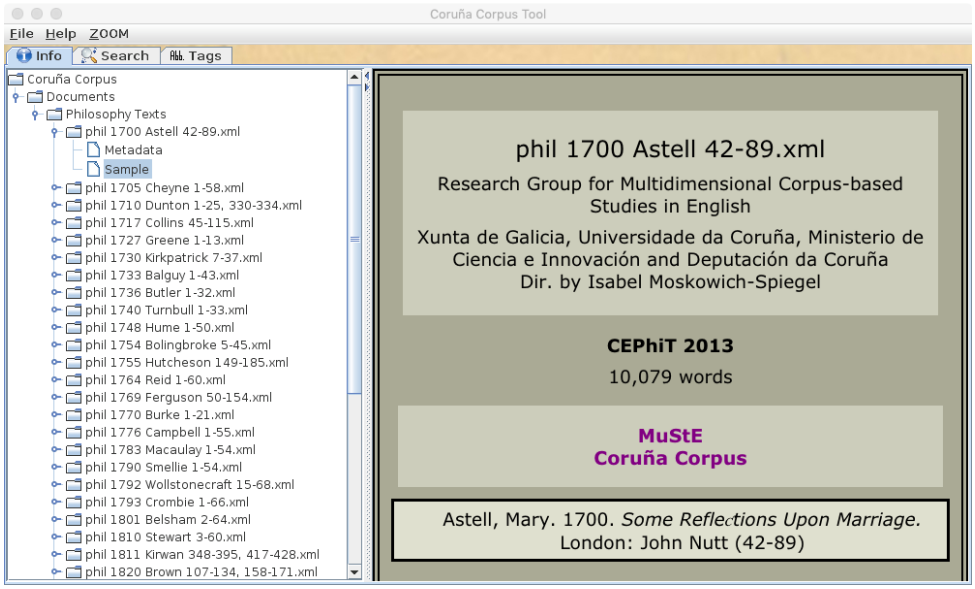

Figure 2: Visualization of XML samples

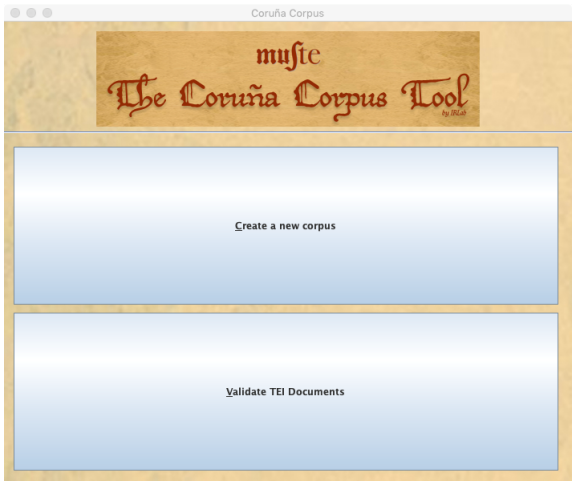

Figure 3: Manager's simplified features

files in UTF-8 format prior to indexing; UTF-8 is a Unicode encoding standard with worldwide compatibility (The Unicode Consortium, 2019). The Manager also removes the Byte Order Mark (BOM) if it is found in a file. This mark is discouraged with UTF-8 encoding (The Unicode Consortium, 2019) and can lead to compatibility problems on some platforms. Before indexing the files, during the validation process, the Manager performs these encoding tasks automatically and the compiler is informed of the result.

Additionally, the updated Manager now uses a personalised identifier for each text. When indexing, the Manager extracts the "idno" (identification number) field from the XML files —as defined by the TEI Consortium (The TEI Consortium, 2019). These identifiers or permanent codes not only facilitate the comparison between occurrences in a search, but also allow for distinguishing between one text and another regardless of the specific corpus to which each one belongs, that is, within the *CC* as a whole, and not within each index, as was the case in previous versions. In this way, documents can be referenced using unique identifiers if compilers specify the desired identifiers in each XML file.

Since previous functionalities, such as "Manage existing corpus" and "Browse existing corpus", were no longer required for the creation of new indexes, we have also simplified the options of the Manager by keeping the two main features: the creation of a new corpus and the validation of TEI documents (Figure 3).

3.4 Client improvements

With the new version of the tool, users can search several corpora at the same time. To load a new corpus when one is already loaded, users can decide to either replace or combine the current corpus with the new one. We have added warning messages to inform the user about the status of the current loaded corpora, as well as alerts that appear when the loading process is taking place. These messages prevent unintentional actions that might provoke undesirable behaviour. This new utility of the CCT allows users to make searches across multiple cor-

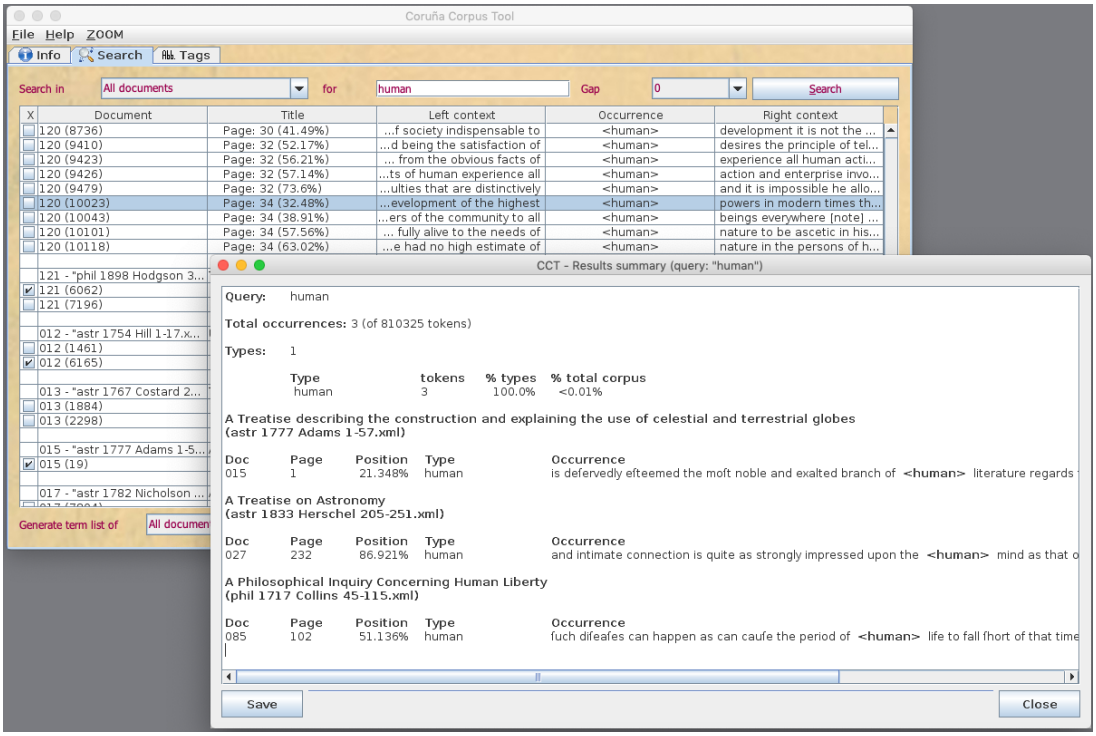

Figure 4: "Results Summary" window

pora files (or across any specified subset of documents) simultaneously. They can also generate frequency lists from documents from different corpora. To implement this feature, we had to change the internal architecture of the information retrieval engine, again using Lucene multi-search capabilities, to launch queries over multiple indexes. This feature had been repeatedly demanded by linguists because they wanted to conduct contrastive analyses between different subcorpora of the *CC* without having to close and reopen the application to load a different corpus whenever they needed to perform a search (or even to request personalised indexes), as was the case with the previous version of the tool. Thanks to this, researchers can now get results from the different corpora loaded (*i. e.,* more than one at the same time) and can easily visualise and compare them.

As Figure 4 illustrates, we have also added checkboxes to the search results display, to allow users to mark the occurrences on which they want to compute the summary statistics (*i. e.,* the "Results Summary" window). If no checkbox is checked, summary statistics are computed over all occurrences, as in the previous version. In addition, we have included an "unmark all" button (at the top of the frame) to uncheck all checkboxes of selected examples without having to scroll through the results window manually before saving results in the desired format.

Prior to this new version, when users clicked on a search result, the document was displayed in plain text only, in a new window. Now, users can also click on the "View" button to show the document with a more appealing rendering.

We have improved the save options of the CCT to export the desired search occurrences or summaries to an external file. All the files are encoded in UTF-8 to avoid formatting errors on any platform. These files can be opened with text editors as well as spreadsheets applications such as those provided in the LibreOffice or Microsoft Office suites.

4 *Related work*

To provide a brief overview of the main features of the CCT in contrast to similar software in the field, we will compare it with some well-known and web-based tools such as CQPweb (Hardie, 2012) and Wmatrix (Rayson, 2009), as well as with AntConc (Anthony, 2019).

CQPweb (Hardie, 2012) is a web-based corpus analysis system that provides a graphical user interface (GUI). Especially useful for large corpora, it is compatible with any cor-

pora and allows word-level annotation and text-level metadata. Moreover, it is also available as open-source software. Likewise, Wmatrix (Rayson, 2009) provides a user-friendly web interface that offers the analysis of word frequency lists, keywords in context (KWIC) concordances and collocations, as well as the comparison of specific domains with larger corpora.

In contrast to these web-interface tools that allow corpora storage in a web server database, AntConc (Anthony, 2019) family tools comprise a series of downloadable and executable freeware additional software such as file converters, corpus analysis toolkits (for concordancing and text analysis for different languages) and even taggers that can be directly used without any installation or Internet connection requirements. In addition to that, AntConc provides greater flexibility in the use of corpora because it is not linked to a particular corpus. However, although all these features might represent a considerable advantage for quick searches and analysis of data, the software-corpora connection should be considered. Thus, the same way that the software designed for a specific operating system proves to be more functional and stable than third-party software, the link between the CCT and its corpora provides much more accurate and reliable search results[3].

Although the CCT works like most concordance programs by allowing basic search by single terms, concordance generation (KWIC), regular expressions search, with or without term-distance specification or wildcard searches, as well as word frequency lists and searches among several corpora or within the same corpus, it also offers some special features such as the representation of original spellings (<æ>, <œ>, <ſ>), searches discriminating spelling variants (<e>, <æ>, <œ>, or <s>, <ſ>, etc.), and the possibility to select subsets of samples by using socio-external variables such as age or sex of the author and genre of the sample, among others. In the same line to AntConc, the CCT is a free and open-source software that can be downloaded and installed on any computer (Windows, macOS and Linux), by avoiding the first-time register requirements, licenses agreements and even subscription fees that web-based services might have. Thus, despite the fact of its installation requirements or even the necessity of loading a previously indexed corpora to work with, the CCT offers a series of options that the aforementioned tools lack such as the possibility of adjusting the zoom according to the users' visual needs and a more user-friendly display of the samples and metadata. Furthermore, users can select and export their searches results into a wider variety of formats (txt, docx, and xlsx, respectively) to work with the data without needing the tool nor a network connection.

Another characteristic shared by the CQPweb (Hardie, 2012), Wmatrix (Rayson, 2009) and the CCT —and questioned by Anthony (2013, pp. 153)— is the inability users have "to observe the raw data directly with their own eyes", due to the fact of being indexed, and hence, requiring the tool to be visualized and make use of it. Although the CCT does not provide an entire solution to this respect, it does allow the visualization of the texts in two ways. Parallel to the default display that these tools tend to offer (i.e., in txt format), the CCT also provides a clear and digitalised version of the original samples, giving the user the opportunity of accessing and reading the texts without being affected by any filtering effect the tool might cause.

Overall, despite some limitations that could not been implemented yet, the CCT combines some of the best features most used in this field, with powerful and user-friendly functionalities that represent a before and after in its use.

5 Final remarks

In this paper, we have described the new features of the Coruña Corpus Tool. This brief overview has been intended to clearly illustrate the main characteristics and functionalities of the tool, and thus to allow users to take advantage of its current full potential. This software is extensively used by the MuStE group for the management and indexation of linguistic corpora, but also by users

[3] As such, this does not imply that the CCT is not compatible with other corpora. It is simply a factor that facilitates the analysis and accuracy of results. An example of this can be found in the frequency lists generated by those tools. Thus, the greater the software-corpus linkage is, the more accurate the results will be. Otherwise, and due to the fact that certain tools do not filter punctuation marks, the total word count is increased, forcing the researcher to perform a manual normalisation of frequencies to provide a reliable analysis.

more generally for the study of the historical development of English, for specific purposes and from different perspectives. The new features that the IRLab group has been developing in collaboration with the MuStE group at the University of A Coruña will enable linguists to easily obtain reliable data for their research and improve their user experience.

In the future, we plan to continue working on new functionalities that will improve search filters and the recovery of special characters, as well as the distinction between formulas, subscripts, and certain other elements included in scientific texts, by using specific labels. Likewise, we aim to provide a more advanced and varied display of results and to facilitate the accessibility of the tool from other devices, as well as its portability and compatibility with other corpora and/or platforms.

Acknowledgements

The research reported here has been funded by the Spanish Ministry of the Economy, Industry and Competitiveness (MINECO), grant number FFI2016-75599-P. This grant is hereby gratefully acknowledged. The second author also acknowledges the support of the Spanish Ministry of Science, Innovation and Universities, grant number FPU014/01724.

References

Anthony, L. 2013. A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2):141–161.

Anthony, L. 2019. AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from `https://www.laurenceanthony.net/software`.

Crespo, B. and I. Moskowich. 2010. CETA in the Context of the Coruña Corpus. *Digital Scholarship in the Humanities*, 25(2):153–164.

Hardie, A. 2012. CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Moskowich, I. 2016. Philosophers and scientists from the Modern Age. In I. Moskowich, G. Camiña Rioboo, I. Lareo, and B. Crespo, editors, *'The Conditioned and the Unconditioned': Late Modern English texts on philosophy*. John Benjamins, Amsterdam, chapter 1, pages 1–23.

Moskowich, I., G. Camiña Rioboo, I. Lareo, and B. Crespo. 2012. *Corpus of English Texts on Astronomy*. John Benjamins, Amsterdam.

Moskowich, I., G. Camiña Rioboo, I. Lareo, and B. Crespo. 2016. *Corpus of English Philosophy Texts*. John Benjamins, Amsterdam.

Moskowich, I., B. Crespo, L. Puente-Castelo, and L. M. Monaco. 2019. *Corpus of History English Texts*. Universidade da Coruña, A Coruña.

Parapar, J. and I. Moskowich. 2007. The Coruña Corpus Tool. *Procesamiento del Lenguaje Natural*, 39:289–290.

Rayson, P. 2009. Wmatrix: a web-based corpus processing environment. *Computing Department, Lancaster University*.

The TEI Consortium. 2019. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium, Charlottesville, VA.

The Unicode Consortium. 2019. *The Unicode Standard, Version 12.1.0*. The Unicode Consortium, Mountain View, CA.