BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 8 (2023) 1–10

BRILL OPEN

brill.com/rdj

The BiTe_Corpus as a Tool for Data Exploration in the History of Hispanic Linguistics

Elena Battaner Moro | ORCID: 0000-0002-5521-6445
Arts & Humanities Department, Rey Juan Carlos University, Madrid, Spain
elena.battaner@urjc.es

Cristina V. Herranz-Llácer | ORCID: 0000-0003-2406-1951
Arts & Humanities Department, Rey Juan Carlos University, Madrid, Spain
cristina.herranz@urjc.es

Ana Segovia Gordillo | ORCID: 0000-0002-0880-6581
Corresponding author
Arts & Humanities Department, Rey Juan Carlos University, Madrid, Spain
ana.segovia@urjc.es

Received 21 July 2022 | Accepted 8 June 2023 |
Published online 6 July 2023

Abstract

Digital humanities make it possible to approach traditional academic topics (such as Hispanic linguistic historiography) from a novel and revelatory perspective. Using tools such as *Voyant Tools* or *Gephi*, one can study both primary and secondary sources from the history of linguistics and extract various theoretical and historical values subject to historiographical and meta-historiographical reflection. To that end, combining corpora that facilitate analysis using digital tools is necessary. This article explains the steps that have been followed to create a corpus based on more than 3,000 abstracts gathered in the *Bibliografía Temática de Historiografía Lingüística Española: fuentes secundarias* [*BiTe*]. In doing so, researchers have access to a set of data which can be analysed in order to continue advancing in the history of Hispanic linguistics, thus offering a working methodology that can be exported to other traditions and currents in the field of linguistics.

Published with license by Koninklijke Brill NV | DOI: 10.1163/24523666-BJA10030
 | ISSN: (print) 2452-3666 (online)

Keywords

digital humanities – corpus linguistics – linguistic historiography – history of Hispanic linguistics – *Bibliografía Temática de Historiografía Lingüística Española*

– Related data set "BiTe_Corpus" with DOI www.doi.org/10.5281/zenodo.6828326 in repository "Zenodo"

1 Introduction

Hispanic linguistic historiography[1] is a living field of study which has been generating interesting reflections on the history of linguistic ideas since the end of the 20th century. The interest that this area of research has received is reflected in the number of published studies, the research projects that have been completed or are underway, the holding of conferences or the creation of associations (such as the *Sociedad Española de Historiografía Lingüística* [Spanish Society of Linguistic Historiography], which has just celebrated its 25th anniversary).

Within this framework, the *Bibliografía Temática de Historiografía Lingüística Española: fuentes secundarias – BiTe* – [*Thematic Bibliography of Spanish Linguistic Historiography. Secondary sources*] was published in 2008 (Esparza et al., 2008). It gathers secondary sources from the history of Spanish linguistics (published until 2007) in an organised manner (from the origins of the Spanish linguistic tradition until the 20th century). In *BiTe*, each of the compiled bibliographical references is accompanied by an abstract, among other information. The summaries, that can be consulted in *BiTe,* reproduce the abstract written by the author or a citation of a relevant paragraph on

1 "Hispanic Linguistics" usually refers not so much to the languages of the Iberian Peninsula as to the linguistics of the Spanish-speaking countries. "Hispanic Linguistics Historiography" is used to research the linguistic, grammatical, lexicographical, etc. traditions developed since the 15th century in such Spanish-speaking countries. This includes, to a large extent, many countries in South America and the Philippines area. The Hispanic tradition also includes all publications written in Spanish not only about other peninsular languages but also about those that existed or exist in those territories (see, for example, the studies on Missionary Linguistics). In the *Bibliografía Temática de Historiografía Lingüística Hispánica* (the work from which the abstracts have been extracted) there are different chapters concerning linguistics of other peninsular languages and all those languages that were studied in the Hispanic linguistic tradition.

the project's objective, method, or conclusions (Esparza Torres, 2006). Since its publication, *BiTe* has become an essential reference book for any scholar who wants to study the history of linguistics in works in which the Spanish or Castilian language is the subject language or meta-language.

As a new way to approach the history of linguistics, the objective of the corpus we have built is to offer a means of approaching the study of linguistic historiography through specific digital tools to extract various theoretical and historical values on which one can reflect from historiographical and meta-historiographical perspectives. Specifically, this article falls within the field of Digital Humanities, understood as an emerging interdisciplinary area in which the humanistic disciplines merge with digital technologies.

2 Hypothesis

The working hypothesis was to create a new object of study of linguistic historiography in the form of a corpus. This corpus would contain the abstracts of the secondary sources of this area – taken from BiTe – with the idea that it allows advancing in new ways of studying the development of the area. Scientific journal abstracts have proved to be a stimulating source of linguistic and semantic research on, for example, the characteristics of scientific writing – the scientific 'genre' –, the description of specialty language and its translation, or its impact on different metrics, among other topics. In this case, following the methodology presented in Criado-Alonso & al. (2020, 2021), this corpus has been prepared with the intention of qualitatively and quantitatively studying the specialty language of the history of Hispanic linguistics and revealing new objects of historiographic study through content analysis and data visualization. In this regard, some research results can be found in Battaner Moro et al. (2022).

3 Methodology

The corpus has been published in a document containing only text. Unlike a rich-text document, a plain text file cannot have bold text, fonts, or any other special text formatting. Since the main purpose of the creation of this corpus is the possibility of conducting content analysis, no labelling or further processing is included beyond the strictly formal one – which is explained

in the following lines – so that, if desired, the researcher can apply on such corpus/text any labelling action he/she needs or wishes. Yet the corpus as plain text can be directly used in corpus analysis programs such as Voyant Tools, AntConc, or #LancsBox, among others.

The final corpus was the result of two phases. First, the abstracts which formed part of the bibliographical records in *BiTe* have been obtained thanks to Dr. D. Miguel Ángel Esparza Torres, whom we want to thank for his collaboration, as he extracted the abstracts of the works in question (around 3200 records) from the database hosted in *FileMaker Pro*. At the start of this initial phase, the corpus comprised a total of 296,029 words and 31,469 terms (unique appearances of words) distributed throughout 2,298 abstracts. The five most common lexical units were, in order, *a, the, no, of,* and *lengua*.

Second, this initial set of abstracts was edited. This phase lasted from October 2020 to December 2021, given that it would require detailed and meticulous work. The editing criteria published by Samper Padilla (1998), traditionally used to study lexical availability, were taken as the starting point. In addition, continuing with the idea of observing the details and singularities of the corpus, there was an attempt to maintain a conservative position (as opposed to a uniforming position) (Fernández Juncal, 2013) when it came to making decisions on editing.

The corpus was first edited into a text document in Word format. Abstracts which were published in any language other than Spanish (English, German, French, Italian, or Catalan, among others) were removed. The next step was to detach all punctuation signs. Classic stopwords were also removed (prepositions, conjunctions, determiners, pronouns, and some auxiliary verbs). In this case, despite being common words, their meaning is not lexical but rather grammatical and so they are defined as *empty words* (Cuartero Sánchez, 2002). Therefore, it was decided that there should be one corpus, essentially containing nouns, adjectives, verbs, and some adverbs.

Linked to this point, Microsoft Excel was used to be more systematic and facilitate the editing process. First, the corpus, with the changes outlined above, was copied into an Excel file so that each of the abstracts was contained within one cell (see Figure 1). After that, a list of unique words (see Figure 2) was obtained using the programming language VBA (Visual Basic for Applications) and then edited according to the criteria explained below (except where the word did not require changes, in which case a dash was used to show that it had been reviewed).

	A	B	C	D
3	objetivo libro es explicar lingüística pro			
4	pretende breves páginas presentar historia			
5	BIBLIOMET dirección está cargo profesor Ig			
6	Diccionario bibliográfico metalexicografía			
7	avanzado estado elaboración banco datos CA			
8	notas eminentemente bibliográficas pretend			
9	siguiente lista aspira ser muestra comenta			
10	perspectiva entiende propósito Bibliografí			
11	biblioteca Universidad hemos hablado misma			
12	Academia Madrid 1726 1739 Academia Madrid			
13	intento catalogar bibliografía relativa le			
14	Podría considerarse bibliografía parcial c			
15	Hemos dividido trabajo tres libros Colecci			

FIGURE 1 Partial corpus in Excel

Término original	Propuesta de modificación
abierto	abierto
abiertos	abierto
abonado	abonar
abordadas	abordar
Abordamos	abordar
abordan	abordar
Abordaré	abordar
abordaremos	abordar
abre	abrir
abreviaturas	abreviatura
abrumaros	abrumar
absoluta	absoluto/a
absoluto	absoluto/a
abstracto	-
abundando	abundar

FIGURE 2 Partial list of unique words in Excel

The following are some other decisions of semantic, syntactic, and morphological nature that were taken for the final preparation of the corpus:

– Uppercase was removed where it appeared by virtue of punctuation. Uppercase was maintained only for anthroponyms (*Maturino*)_*Gilberti,* toponyms (*Honduras*, *Cuba,* etc.), and initialisms not explained in the corpus (*DUE*, *UFS*, *VSO*, etc.).
– Graphical polymorphism was maintained. For example, *Méx/jico* or Ju/oan in the case of authors such as (*J*(*o/uan*))_*Palet* or (*J*(*u/oan*))_*Corominas*.
– Diminutives were only maintained if they indicated different realities.
– Numbers (both Arabic and Roman) were only maintained where they indicated years.
– Verbs were recorded in the infinitive.
– Nouns subject to gender variability and adjectives with two endings were recorded as follows: if they only appeared in the feminine, the feminine form was maintained (*adulta; bella; divulgadora; etc.*) and, if they only appeared in the masculine, the masculine form was maintained (*acólito; africano; aldeano; etc.*). Where both forms (masculine and feminine) appeared, it was decided to combine them into a single entry and separate the alternative endings with a slash; for example, *autor/a.*
– Each lexical unit is understood to be a single word, but anthroponyms and toponyms are an exception. For the latter, underscores (_) were used to join each of the components of the name, and parenthesis was used to show the presence or absence of elements of the word.

Once all these changes had been incorporated into the list of unique words, the project proceeded as follows. The corpus in Excel format was converted to '.txt' and processed with the *BBEdit* program, a text and HTML editor which offers numerous functions for editing, searching, and manipulating the corpus and textual data. Thanks to this program, the original terms were replaced with the modified words contained in the list of unique words. All this editing work (see Figure 3) was carried out three times to clean up the corpus and ensure that the criteria explained above were followed rigorously.

This process resulted in the final corpus in '.txt' format comprising 102,613 words and 9,270 terms. Table 1 shows an example of the modifications to which the corpus was subject.

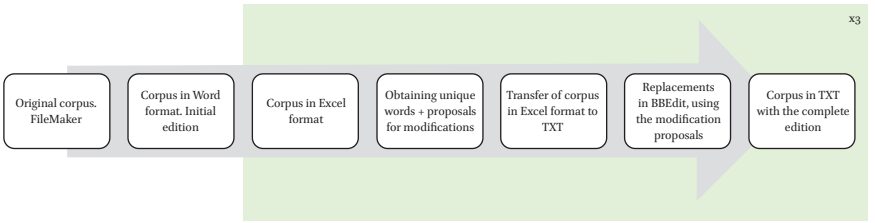

FIGURE 3 Flow chart of the editing process

TABLE 1 Original corpus vs. Edited corpus

Original abstract	Final edited abstract
130 "Aunque actualmente la historiografía sea una disciplina muy estimada e importante en la lingüística, uno podría preguntarse lo siguiente: ¿qué valor tiene el estudio de una gramática antigua hoy en día? Abordaré a continuación esta cuestión delicada del ¿por qué?, que a su vez no puede ser vista independientemente del ¿cómo?, del proceder metódico. En una segunda parte estos resultados teóricos serán aplicados a la gramática de Nebrija."	actualmente historiografía disciplina estimado/a importante lingüística preguntar siguiente valor tener estudio gramático antiguo/a hoy día abordar continuación cuestión delicado/a ser vista independientemente proceder metódico/a parte resultado teórico/a ser aplicado/a gramática (Elio)_(Antonio)_Nebrija

4 Data

- **BiTe_Corpus deposited at Zenodo** – DOI:www.doi.org/10.5281/zenodo.6828326
- **Temporal coverage**: origins of the Spanish linguistic tradition-2007

The final corpus is especially conceived for the meta-historiographical study of the history of Hispanic linguistics. In this sense, in Battaner Moro et al. (2022) different analyses are carried out using the corpus with tools such as Voyant, Excel, Tableau, or Gephi. On the one hand, with this type of corpus, purely quantitative studies can be conducted, and relevant information can be extracted. For example, the ten most frequent years in the corpus refer to the publication date of ten of the most relevant works in the history of Hispanic

linguistics. But, on the other hand, the most relevant results, so far, refer to the specific use of different terms in linguistic historiography.

If we analyse the words that usually appear together or related, for example, we can extract information about the behaviour of certain specific terms; research can be conducted regarding disciplines – grammar, orthography, lexicography, etc. –, authors – Nebrija, Menéndez Pidal, Hervás y Panduro, etc. –, or concepts – syntax, vowel, definition, etc. –, among others. Figure 4 is a repulsion graph showing the behaviour of the two apparent synonyms "Castilian" (in blue) and "Spanish" (in green), and the nodes (*i.e.*, words or terms) related to them. It can be observed that different words are associated with each of them: for example, "Castilian" usually appears with "grammar", but "Spanish" appears with "dictionary". In this way, we can see that the use made in the history of linguistics is not synonymous but is very much circumscribed to the type of work being described or studied. In this sense, we could say that for "Grammar", Nebrija's work (his *Gramática castellana*) is most influential and, for "Lexicography", the work of the Real Academica Española. Reasons for this differentiated usage, however, are complex and involve other historical, political, and sociolinguistic issues.

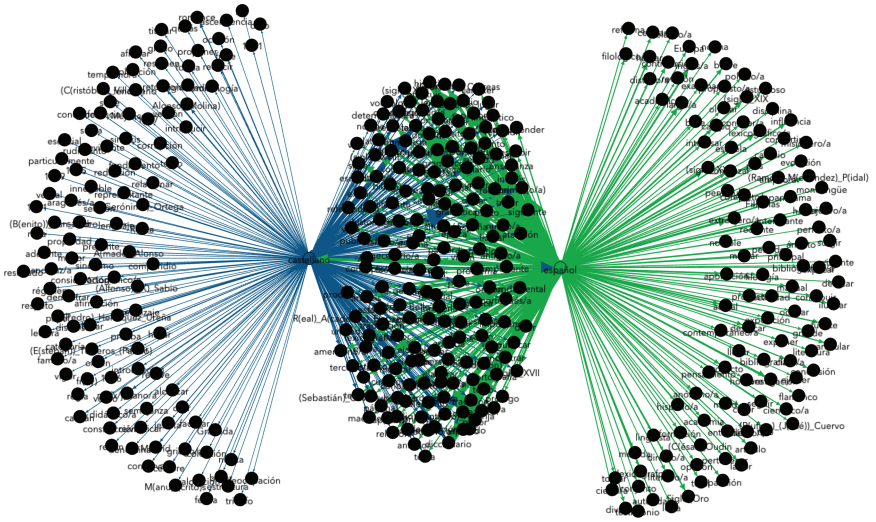

FIGURE 4 Graph of the most common collocations of the terms "castellano/a" (left) and "español/a" (right). Each node is a word that collocates with one of the two terms term (on the ends) or both (at the centre).
BATTANER MORO ET AL., 2022

5 Concluding Remarks

The edited corpus with the Spanish abstracts contained in the *Bibliografía Temática de Historiografía Lingüística Española: fuentes secundarias* presented in this article opens new lines of research in the field of Hispanic linguistic historiography. It does not only offer the possibility of analysing new objects of study (abstracts from secondary sources on Spanish linguistic historiography) but also suggests new research pathways (analysis of collocations, for example). This methodology, applied in this case to the Hispanic linguistic historiography, can consequently be exported to carry out research in any linguistic/scientific tradition, author, or discipline in any language.

One of the main objectives of Digital Humanities is not only the development or use of new methodologies or tools but the achievement of new objects of study. In this case, the corpus that we have prepared and published, understood to be one such new object of study, certainly expands the theoretical and historiographical possibilities of the digital realm but also allows for the scientific advancement of our area of research.

References

Battaner Moro, E., Herranz-Llácer, C. V., & Segovia Gordillo, A. (2022). Corpus y herramientas digitales para el estudio de la historiografía lingüística hispánica. *Boletín de la Sociedad Española de Historiografía Lingüística*, *16*, 11–39.

Criado-Alonso, Á., Battaner-Moro, E., Aleja, D., Romance, M., & Criado, R. (2020). Using complex networks to identify patterns in specialty mathematical language: a new approach. *Social Network Analysis and Mining*, *10*(1), 69. www.doi.org/10.1007/s13278-020-00684-1.

Criado-Alonso, Á., Battaner Moro, E., Aleja, D., Romance, M., & Criado, R. (2021). Enriched line graph: A new structure for searching language collocations. *chaos. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena, Vol 142*, 110509. www.doi.org/10.1016/j.chaos.2020.110509.

Cuartero Sánchez, J. M. (2002). "Significado léxico" y "significado gramatical" en las gramáticas del español moderno. *Boletín de la Sociedad Española de Historiografía Lingüística*, *3*, 43–78.

Esparza Torres, M. Á. (2006). Materiales para una historia de la lingüística española: Bibliografía Temática de Historiografía Lingüística Española. *Caminos actuales de la historiografía lingüística*, *1,* 517–528.

Esparza Torres, M. Á. (Ed.); with Battaner Moro, E., Calvo Fernández, V., Álvarez Fernández, A., and Rodríguez Barcia, S. (2008). *Bibliografía Temática de Historiografía Lingüística Española. Secondary sources*. Helmut Buske Verlag.

Fernández Juncal, C. (2013). *Léxico disponible en Cantabria. Estudio sociolingüístico*. Ediciones Universidad de Salamanca.

Samper Padilla, J. A. (1998). Criterios de edición del léxico disponible: Sugerencias. *Lingüística*, *10*, 311–333.