Alfaya Lamas, Elena and Garrote Espantoso, Menchu. 2021. The representativeness threshold for the *CETA* subcorpus of the *Coruña Corpus*. *Revista de lenguas para fines específicos 27.2*, pp. 125-139 · http://dx.doi.org/XXXXXXXXXXXXX

Paper received: 07 June 2021
Paper received in revised form and accepted for publication: 23 September 2021

The representativeness threshold for the *CETA* subcorpus of the *Coruña Corpus*

Elena Alfaya Lamas[1] · https://orcid.org/0000-0001-6628-6257

Universidade da Coruña

C/ Dr. Vázquez Cabrera, s/n · Campus de Ferrol, 15403 · Ferrol, A Coruña

Menchu Garrote Espantoso · https://orcid.org/0000-0001-7918-2780

Universidade da Coruña

Campus de Fábrica de Armas · Campus Tecnológico · Av. de Carlos, III, nº 21 · 45004 · Toledo · Universidad de Castilla-La Mancha

ABSTRACT

The concept of representativeness is the main distinguishing characteristic of specialised corpora in comparison to other sets of texts. The *Coruña Corpus of English Scientific Writing* currently comprises four published subcorpora (astronomy, life sciences, history, and philosophy) plus three others under compilation (physics, chemistry and linguistics). In this paper we aim to assess the lexical density of the text samples in *CETA*, the *Corpus of English Texts on Astronomy*, by means of the *ReCor* tool, *a posteriori*. The study is motivated by the following question: does quantitative representativeness analysis using *ReCor* provide, in the form of a cross-check, further validation of previous research on the representativeness of *CETA*? Previous work (Crespo and Moskowich, 2010) has indicated that the *CETA* corpus is well designed and valid for the purposes for which it was intended. We will here suggest metrics to measure these findings. The most important contribution of this study is to offer quantitative data collection results using the *ReCor* tool, which allows data triangulation and consequently ensures overall data quality. Results show that data analysis with the *ReCor* tool supports previous findings, and thus we are able to verify that *CETA* is indeed representative of the language of its time and register.

Keywords: Representativeness, ReCor, specialized Corpus, Zipf's Law, N-gram, Coruña Corpus, CETA, Astronomy.

1. Introduction

This study assesses the impact of a quantitative analysis from the field of bibliometrics and documentation, using the *ReCor* tool. It addresses previous studies from the fields of philology and linguistics that have already shown the validity of *CETA*. It is the aim of this paper to focus on the representativeness threshold of *CETA* by examining and assessing its lexical density with the *ReCor* tool, *a posteriori*. The *CETA* is a well known corpus and many studies (Crespo and Moskowich, 2012) from the fields of philology and linguistics have shown that it is well designed and valid for the purposes for it was intended, including representativeness. However, thus far no technical contribution on representativeness from the field of documentation and bibliometrics has been offered. Through the analysis presented in this paper, we aim to assess *CETA* text samples with a bibliometric method for the validation of representativeness *a posteriori*. In this respect, the study involves a comprehensive basis for data and information collected through an independent quantitative

[1] **Corresponding author** · Email: elena.alfaya@udc.es

method, one which will ensure the validity of the overall results on *CETA* representativeness, in that total data triangulation is extremely useful for the validation of evidence and support for exhaustive data quality.

Therefore, the following research question arises: Does quantitative representativeness analysis using the *ReCor* tool offer an additional form of cross-checking the validation provided by previous research on the representativeness of *CETA*? Since the corpus has been created according to defined parameters of text collection and based on predefined qualitative and quantitative criteria, we can hypothesise that quantitative lexical density analysis from the field of bibliometrics and documentation should validate previous research on *CETA* representativeness from the field of philology and linguistics.

To verify our hypothesis and answer the research question posed, we will follow a measurement method, the information we obtain being numerical in nature. The research tool used, the *ReCor* program[2], is a Java application that processes data and linguistic information and is intended to help researchers determine the representativeness of a corpus.

The *Coruña Corpus* (henceforth *CC*) consists of a group of subcorpora, including *CETA*. The compilation process of *CETA* is finished, and this subcorpus is available in a printed version and on CD (Moskowich and Crespo, 2012). A free version of the second edition is also available (Moskowich, 2011). Both editions include the *Coruña Corpus Tool* (*CCT*), a bespoke application to help researchers to explore and use the corpus (Parapar, 2007). The *CCT* has been developed by the Information Retrieval Lab - IRILaB[3] - of the University of Coruña in collaboration with the MuStE[4] research group.

All the subcorpora in the *CC* are composed of scientific texts, and thus are useful for scientific writing analysis. We hope that the results of this study will be of interest not only to all those who currently use the *CC* in their research, but also to others studying corpora representativeness. We also hope that, if our hypothesis is verified, this study may have practical and theoretical implications in terms of the representativeness of *CETA*. The results of this study may be applicable to the rest of the *CC* subcorpora, since the qualitative and quantitative criteria used for all subcorpora are the same.

Currently, work in the humanities and the social sciences are gradually adapting to the use of new technologies, as is also the case in other scientific fields.

Although there might be a tendency to think that we need to collect very large quantities of those elements that we want to observe, one of the fundamental pillars of a corpus is that it is representative of the reality that it aims to reflect, and hence can be used with certain guarantees of success (Torruella and Llisterri, 1999: 1). For this reason, the operation of a natural language must be shown on a small scale as long as linguistic diversity is taken into account.

Neutrality is another important element that a corpus must comply with. However, we cannot forget that when we are compiling a corpus, we are selecting texts, which itself implies a process of exclusion; and in this sense, the corpus will no longer be as faithful as possible to reality. This is the reason why there is a tendency to think

2 2.0 Version. Designed by Gloria Corpas Pastor and Miriam Seghiri.

3 https://www.dc.fi.udc.es/ir/research.html

4 https://www.udc.es/grupos/muste/

that the more texts we integrate into a corpus, the greater the likelihood of ensuring the presence of all aspects of the language. Yet the very essence of a corpus is to be selective. It is not possible or profitable to collect everything that was written or spoken in a particular language. It is here that we find the fundamental basics of corpora, since a well-selected and representative corpus is preferable to an exhaustive one.

The main documentary source for this paper has been the *CC* itself. Also, studies published by the members of the MuStE research group and other additional studies have been central to our theoretical framework. Electronic resources have also been used, which have required the establishing of some premises to meet specific quality criteria, thus obtaining the most reliable and relevant information for the study.

CETA files are compiled in XML format, following the conventions of the *Text Encoding Initiative* (*TEI*). The Unicode standard has also been used to represent symbols and old characters in an attempt by the compilers to be true to the original texts.

In addition, in an attempt to represent only an author's prose, the members of the MuStE research group that developed *CETA* decided not to include poems, quotes, additions by editors, and numbers or symbols that do not carry a syntactic function in the sentence. Also, they introduced modifications such as the elimination of unnecessary blank spaces and the correction of obvious spelling errors. Hence, some editorial marks written in square brackets have been added to include information on the omitted parts or to disambiguate formulas or numbers that could be indexed as an English word.

2. Literature review

This section seeks to define key concepts, and to describe the methodology and theoretical approach; it will delimit, connect and clarify the relations between key content and offer a literature review that illustrates why the research problem under study here exists. Information on *CETA* and the *ReCor*[5] program is also presented.

A corpus is a representative collection of texts used for linguistic analysis. Sinclair (1991: 171) provides a precise definition of the term, identifying it as "a collection of naturally-occurring language text, chosen to characterize a state or variety of a language." Torruella and Llisterri (1999: 17) synthesize John Sinclair's position regarding the desirability of working with full texts, thus avoiding the "inconveniences of the validation of the samples" (idem). These authors refer to corpora thus:

> Es un conjunto homogéneo de muestras de lengua de cualquier tipo (orales, escritos, literarios, coloquiales, etc.) los cuales se toman como modelo de un estado o nivel de lengua predeterminado. El conjunto de enunciados incluidos en un corpus, una vez analizados, debe permitir mejorar el conocimiento de las estructuras lingüísticas de la lengua que representan (ibíd.: 8).

In addition, they offer a definition of subcorpora:

> La elección estática de textos, derivada de un corpus normalmente más general y complejo, el cual está dividido en grupos de muestras textuales más específicas; pero también puede ser una selección dinámica de textos de un corpus en crecimiento: un número determinado de textos destinados a aumentar algún apartado de un corpus general.

[5] For this study, the University of Málaga signed a License Agreement for the use of the *ReCor* application.

Despite the fundamental premises of quality and representativeness that a corpus must fulfil so as to distinguish it from other Types of textual collections, there seems to be no overall consensus as to exactly what this means.

Representativeness is a fundamental characteristic of a corpus, and indeed constitutes the central axis of its validity. As we have seen, there are many theories and a wide range of definitions of exactly what a corpus is, although most of them agree on the aspect of representativeness. Francis (1982: 17) notes that a "collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis" is a corpus. Biber, Conrad and Reppen, (1998c: 246) observe that "A corpus is not simply a collection of texts. Rather a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent."

When we use the term "representativeness" in this paper, we refer only to its quantitative aspect, that is, from the orientation of the field of documentation and bibliometrics, since the qualitative and quantitative aspect as understood within the field of philology and linguistics has already been dealt with *a priori* and explained by the compilers. To reach some conclusions on the adequacy of lexical density for the quantitative representativeness of *CETA*, we will use the algorithm of the *ReCor* application.

Seghiri (2014: 87) notes that there is a large body of work on the issue of quantity as a criterion for achieving representativeness, and that formulas are available to calculate the minimum number of words and documents necessary so that a specialized corpus can be considered representative. Many of these formulas are based on Zipf's Law, based on the idea that all texts contain a number of words that are repeated. The most frequently used words will be ranked first, and those used less frequently will follow in descending order. Zipf's Law states that there is an inverse relationship between the frequency of a word's occurrence and its rank, that is, frequency decreases when rank increases, being inversely proportional to its number on the list. Zipf's First Law states that r*f= c, that is, rank by frequency is a constant for any text.

Following the studies of Moyotl-Hernández and Macías-Pérez (2016: 162), the majority of words that are most frequently used coincide with those that are shorter and easily remembered:

> Las *palabras funcionales* – también llamadas palabras vacías o ***stop words*** –, tales como artículos, pronombres, preposiciones y conjunciones son las más frecuentes en el texto, mientras que las menos frecuentes son palabras que reflejan el estilo y riqueza del vocabulario del autor. Por lo tanto, las palabras que aparecen en la zona media de la transición entre las de alta y baja frecuencia son las que representan al documento.

In this way, we can link Zipf's Law[6] with the studies by Booth[7] (1967) and Goffman (1971)[8], since these explain that there is a point at which stop words are no longer frequent (Booth, 1967), that is, the transition point, and it is here that we find the most significant terms. Goffman introduced the idea that the most significant words in a text are grouped into an intermediate area between high frequency and low frequency words: the transition point (*idem*). According to Sidorov (2013: 70), a n-gram can represent the sequences of

6 George Kingsley Zipf observed that the rank-frequency distribution is an inverse relation. His law states that frequency of any word in a given sample is inversely proportional to its rank.

7 Andrew Donald Booth was a scientist who worked with William Goffman's transition point technique.

8 William Goffman was a pioneer of information science. His contributions continue to be applicable nowadays, having stood the test of time.

words, or other elements such as numbers, as they appear in the texts. n-grams are collected from a text in a corpus. Thus, if we analyse a text with 1 gram, we obtain individual information about the word; if we analyse a text with 2 grams or more, we see relationships among words.

CETA texts are marked up with XML. Among the advantages of XML (ibid .: 12), the following stand out: it is an open technology, and is independent of the operating system; it is an international standard based on the Unicode character encoding system; it is a simple technology to use and implement; it has great power in the construction of marking vocabulary applicable to any type of document; it allows the reuse of existing texts and data in other documents for the preparation of new documents; and it provides powerful mechanisms for the search and retrieval of information.

The *CC* is a project of the MuStE research group in the University of A Coruña (Spain). This group includes linguists and documentalists from the areas of English Philology and Information and Documentation.

The corpus has been compiled for research mainly at the linguistic, historical and documentary levels. Research from these fields can help to shed light on the general characteristics of scientific English, as well as its evolution, from the first writing in the vernacular immediately following the Scientific Revolution up to the late nineteenth century.

Not much attention has generally been paid to scientific language before the 1990s, this mainly because it was not considered an object of study in itself, but rather vehicle for transmitting knowledge to which lexical, syntactic and discursive uniformity was attributed. From the end of the 19th century onwards, the growing interest in English for specific purposes runs parallel to a similar interest in its historical description, evolution and peculiarities (Crespo and Moskowich, 2010: 159).

Within the principles of the compilation of the *CC*, several aspects and parameters have been exhaustively considered to ensure accuracy: design criteria, time period (1700-1900), type of text/genre, discipline, sex and age of author, and number of words per sample, so as to achieve representativeness and balance. The compilers follow Biber, Conrad and Reppen (1998b: 4) who note that a corpus must be "a large [...] collection of natural texts". The corpus offers researchers the possibility of studying the evolution of scientific English. *CETA* compiles two sample texts containing around 10,000 words per decade. Compilers have been very rigorous in using first editions, not using more than one text by the same author, this to avoid the proliferation of stylistic idiosyncrasies (Crespo and Moskowich, 2010: 155), and not using translations. In addition, efforts have been made to maintain a proportional balance, not only in terms of the same number of words per discipline, but also the same number of disciplines per field (Exact Sciences and Humanities); only English-speaking authors who write in English have been considered.

In addition, each text is inevitably related to a particular social or extra-linguistic context that allows for socio-linguistic studies to be carried out using the corpus itself as a tool. To this end, a metadata file containing information about each author, including biographical and bibliographic material, is included with each text. The texts also include coded information on spelling, paragraphs, page numbers, abbreviations and notes, as well as sources of information. It is the intention of the *CC* to compile a more or less equivalent number of texts and words for each separate scientific field.

As a subcorpus of the *CC, CETA* contains samples of English scientific writing on astronomy, published between 1700 and 1900, by forty-two different authors, these including just two women. There is a balance between the number of words per century: 208,083 words for the eighteenth century, and 202,533 for the

nineteenth. For the compilation of the texts, compilers accessed different libraries and edited text files published in facsimile editions. The proportion of writers educated in England (41%) far exceeds those educated in Scotland (14%) or Ireland (5%). Authors who were educated and learnt to write in North America occupy the second position in this ranking (26%), these mainly in the nineteenth century. The remaining 14% correspond to the category "Unknown", and this should be interpreted as the percentage of authors for whom we must ignore the unknown provenance of their scientific writing habits. As for textual categories, *CETA* includes lectures, dialogues, academic treatises, textbooks, letters, essays, articles and other, the latter category including those samples that do not fit within any of the seven main groups.

ReCor is the application used for this study. Its algorithm analyses lexical density processing data and linguistic information, and thus determines the representativeness of corpora. It aims to establish the minimum size that a corpus must have to be considered representative, *a posteriori*, that is, once the corpus has been compiled. This program is based on a proprietary algorithm called N-Cor, and on Zipf's Law (Corpas, 2010). In this way, *ReCor* is intended to provide an effective solution, one that does not depend on the language itself or on textual typology.

3. Methodology

According to Biber, the representativeness of a corpus is determined by "the extent to which a sample includes the full range of viability in a population" (1993: 244). This being so, we intend to measure the *CETA* subcorpus as a sample of late Modern English scientific writing, to verify its quantitative representativeness *a posteriori*, and to obtain bibliometrical information about its properties and qualities. We aim to answer the research question previously framed by analysing the *CETA* sample texts through the *ReCor* tool. Our hypothesis is that *CETA* texts and the lexical density of the corpus are quantitatively representative, from the perspective of the field of Documentation and Bibliometrics, thus confirming earlier research on *CETA* representativeness. This representativeness exists, we believe, due to the defined parameters followed *a priori* by the compilers, and shown to be so in previous studies that took philological and linguistic approaches.

Since measuring corpus representativeness is one of the most controversial aspects in the study of corpora, we will cross-check data and thus triangulate quality parameters with quantitative information obtained with the *ReCor* tool to ensure and verify the overall data quality of *CETA*'s representativeness.

In the case of specialized corpora, such as those contained in the *CC*, representativeness is key due to their small size in relation to "general" or "reference" corpora. In the absence of any consolidated theory about specialized corpora, there is no consensus as to the minimum number of documents or words that a particular corpus should have for it to be considered "valid" or representative of the sample it aims to represent.

As Miriam Seghiri (2015: 126) says, there have been several attempts to set a minimum size for specialized corpora; such proposals are either based on Zipf's Law or establish a minimum sample size *a priori*, that is, prior to compiling the corpus (Seghiri, 2011: 25). *ReCor* software establishes the size needed for a corpus to be representative *a posteriori*, that is, once the corpus has been compiled. We have made use of *ReCor* to measure *CETA* representativeness.

The methodology involves the following key components: background research, as summarized in the previous section; data collection from our primary source, *CETA*; data analysis and triangulation.

Our primary source, texts of the *CETA* subcorpus, in XML, presented metadata and tagging labels necessary for the correct identification of each text, but these interfered in some way with our methodology. Thus we removed labels and additional information that did not correspond to the original text, and conducted our study with the plain text documents.

For the data analysis, the texts needed to be in .txt format when using the "Selection of CORPUS files" option in *ReCor*, since XML is not a valid option here. Figure 1 below shows *ReCor* interface.

Figure 1. ReCor Interface (version 2.1).

We saved the XML files and dragged them to a browser, Google Chrome[9], to remove the tags manually. Then, we copied and pasted the text from the browser into Notepad[10] to save the text in .txt format. This was repeated for the 42 texts that comprise *CETA*. As the samples in the *CC* conform to TEI standards, they all contain a header preceding the body of the text. We removed this information from each file, in order to keep only the original text so that the program only processes the data required for the analysis.

On running the *ReCor* software, results are expressed in two ways: the generation of graphical representations, and the creation of output files in .txt with exportable statistical data.

We first plotted a graph to indicate the minimum size of *CETA* corpus to be considered representative, that is, the point where the corpus increases in size but not in new words or Types. These data are also offered in tables, in order to see the specific base data plotted in the graphs. Types, Tokens, Ty/To and total words with one occurrence and two occurrences are shown. These data will allow us to check that there is a point from which the compilers do not need to add more new texts since even if more sample texts are added, almost no new vocabulary would emerge. Also, we provide tables showing the frequency of word occurrence. These allow us to verify whether *CETA* complies with Zipf's Law.

9 1.0 version

10 42 version

4. Data analysis and discussion

This section describes the findings of our quantitative study after examination of a set of data from the *CETA* corpus, with the *ReCor* tool, *a posteriori* and from a bibliometric perspective. Our research question seeks to improve our knowledge on the representativeness threshold of *CETA*: Does quantitative representativeness analysis by means of the *ReCor* tool offer a cross-check, providing additional validation of previous research on the representativeness of CETA? We now provide metrics to assess the impact of a quantitative analysis of *CETA* lexical density and representativeness.

Graphic Study A (Figure II) is plotted to determine whether the corpus is quantitatively representative *a posteriori*. As Seghiri (2011: 25-26) suggests, the proportion Types/Tokens stops growing exponentially after analysing a certain number of texts. The line generated by Graphic Study A descends as we move along the horizontal axis representing the total number of documents that the corpus contains. The same happens when representativeness is calculated following the lexical density from sequences of words or n-grams, in Graphic study B, where we see that the line of the Types / Tokens ratio decreases as more Tokens are reached in the body of the texts.

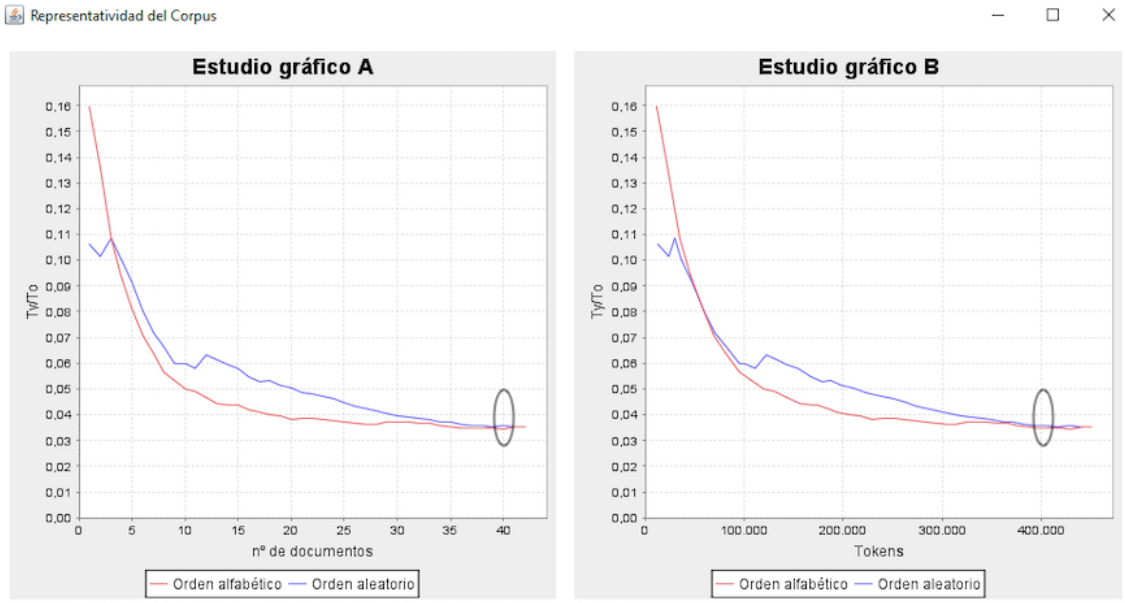

Figure 2. Graphic Studies A and B. CETA Representativeness.

These graphs show a red line, depicting the documents included alphabetically, and a blue line, depicting the documents randomly chosen by the algorithm. These lines run together as they approach zero, which indicates the minimum size required for the corpus to be considered representative. Oval marks have been added in Figure 2 to indicate the crossing points of the two lines in both cases. Graphic Study A shows that *CETA* is representative at 35 documents; in the case of Graphic Study B, it shows that a minimum of 386,000 words are needed. Therefore, we can conclude that the *Corpus of English Texts on Astronomy* is representative from 35 documents and 386,000 words thereon.

The next step is to analyse the results extracted from the corpus in the .txt files using *ReCor*. The following data tables display detailed information in a grid format of fields and records, the source again being *CETA*. Table 1 below presents quantitative data in five fields: Types, Tokens, Types/Tokens (Ty/To) or TTR, Type-

Token ratio, words with one occurrence (V1), and words with two occurrences (V2). If we compare the figures of the first two fields – Types and Tokens – we can clearly see that there is a saturation point after which there are hardly any new Types (new words) occurring in the corpus and only Tokens (repeated words) continue to be present. The first field, Types, shows the point where the increase of new words ceases at the bottom. The Figures show that Types increase is minimal: 14,210; 14,322; 14,574; 14,835; 15,595; 15,946. That is, we reach a point where the corpus increases in size but not in new words or Types. At this point the corpus is already representative. Therefore, from this saturation point the specialized terminology of Astronomy has been covered. The third field, TTR, expresses the *CETA* documents' lexical richness in terms of variety of vocabulary. We can see that *CETA* reaches quantitative representativeness with 1 gram as of 35 documents and 386,255 Tokens, as depicted in Table 1 below. Since the *CETA* corpus has 42 documents it goes past the minimum for representativeness.

Types	**Tokens**	**Ty/To**	**V1**	**V2**
1,843	11,545	0.1596362	963	286
3,131	23,049.0	0.13584103	1,629	506
3,793	34,867.0	0.10878481	1,860	595
4,357	46,459	0.09378161	2,141	651
4,704	58,022	0.081072696	2,264	676
4,926	69,754	0.070619605	2,302	696
5,215	81,993	0.06360299	2,393	756
5,361	94,668	0.056629483	2,431	763
5,710	10,6931	0.053398922	2,609	807
5,973	11,9240	0.05009225	2,704	819
6,382	13,0779	0.048799884	2,892	872
6,688	14,3439	0.04662609	2,977	950
6,965	15,6874	0.044398688	3,021	1,004
7,378	16,9073	0.043637954	3,229	1,038
7,604	17,4149	0.04366376	3,339	1,080
7,787	18,6615	0.04172762	3,397	1,133
7,957	19,3325	0.041158672	3,449	1,162
8,180	20,5189	0.039865684	3,511	1,207
8,552	21,7282	0.03935899	3,643	1,289
8,733	22,9441	0.038062073	3,704	1,309
9,322	23,9902	0.038857535	3,972	1,393
9,720	250,685	0.03877376	4,148	1,457
9,959	261,001	0.03815694	4,241	1,475
10,229	271,439	0.03768434	4,324	1,514
10,553	281,837	0.03744363	4,473	1,556
10,757	292,721	0.0367483	4,527	1,573
10,999	303,460	0.036245305	4,602	1,599
11,402	314,029	0.036308747	4,788	1,653
11,997	324,244	0.036999915	5,057	1,799
12,342	333,344	0.037024815	5,186	1,839
12,716	343,755	0.036991462	5,348	1,907

Types	Tokens	Ty/To	V1	V2
13,091	354,934	0.036882915	5,558	1,958
13,360	365,231	0.03657959	5,630	1,978
13,532	375,605	0.036027208	5,682	2,002
13,674	386255	0.035401482	5,734	2,023
13,855	396,399.	0.034952156	5,788	2,049
14,21	407066	0.034908343	5,928	2,119
14,322	412,394	0.034728926	5,983	2,129
14,574	419,020	0.034781154	6,093	2,149
14,835	429,577	0.03453397	6,166	2,183
15,595	441,001	0.03536273	6,531	2,284
15,946	449,664	0.035462033	6,694	2,344

Table 1. Analysis. Results in alphabetical order[11]

The third field shows the Ty/To (Types/Tokens) ratio used for Graphic Studies A and B. Columns V1 and V2 sum the total words with one occurrence and words with two occurrences, respectively.

Also, a table displaying a sample word list alphabetically sorted has been generated with *ReCor* from *CETA* data; this allows us to check the frequency of occurrence of words in *CETA*. Table 2 below shows an excerpt from 1-gram words starting with letters *a, e, m*. It is a brief excerpt of 36 entries from a total of 5,280. The complete results are not presented here because this is not the main aim of the current study.

Word	Occurrences	Word	Occurrences	Word	Occurrences
Assigned	15	ebullitions	1	moo	1
Assigning	2	ec	52	moon	1854
Assigns	2	ecb	1	moonless	1
Assist	2	eccentric	12	moonlight	7
Assistance	2	eccentrical	1	moons	70
Assistant	4	eccentrically	1	mooted	1
assisted	2	eccentricities	5	mooth	2
assists	2	eccentricity	42	moral	12
associated	1	eccentricityâ	1	morality	1
association	3	eccentrics	1	moralize	1
asume	23	eccleå	1	morals	2
assumed	34	ecde	1	morbum	1
assumes	5	echar	1	morden	2
assuming	6	echidna	1	more	896
assumption	14	echidnas	1	moreover	25
assumptions	2	echo	2	morning	64
assure	2	eclipse	114	morrow	1
assured	4	eclipsed	17	mortal	2
asterisms	1	eclipses	55	mortals	2

[11] Table generated from *CETA* data with *ReCor*.

Word	Occurrences	Word	Occurrences	Word	Occurrences
asteroids	4	eclipt	1	mortify	1
astonished	1	ecliptic	463	mosaic	1
astonishing	5	ecliptical	2	moscow	1
astonishment	2	ecliptick	114	mosely	1
Astr	33	eclipå	255	most	220
Astro	2	eclyptical	1	mostly	3
astrologer	2	eclyptick	11	mote	1
astrology	3	ecm	1	motheoros	1
astromomers	1	econd	79	mother	3
Astron	1	econdaries	4	motherus	1
astronomer	39	econdary	26	motibus	1
astronomers	99	econds	41	motion	939
astronomia	1	economy	3	motions	236
astronomical	60	ecp	2	motive	13
astronomie	7	ecphantus	1	motives	4
astronomische	1	ecq	1	mottled	5
astronomy	129	ecret	5	mottling	1

Table2. CETA alphabetically ordered 1-gram words sample (excerpt of letters a, e, m)[12].

Tables 2 and 3 contain identical data. However, Table 3, unlike Table 2, is sorted by word frequency of occurrence. It displays four fields: word; frequency (f); rank (r) and r*f (which stands for Zipf's constant). The selected samples of letters *a, e* and *m*, contain both the lexeme of the words and their corresponding derivatives. For instance, the word *eclipse* and its derivatives are displayed: *eclipsed*, *eclipses*, *eclipt*, *ecliptic*, *ecliptical*, *ecliptick*, *eclipå*, *eclyptical* and *eclyptick*, with *ecliptic* and *eclipå* being the most frequently used variants. Table 3 clearly shows words empty of meaning preceding the top. It is a basic convention of corpus linguistics to ignore the 10 first words of a frequency list due to this. They are interspersed with the most frequently used words in the Astronomy scientific field, such as earth, sun and moon, as we can see in Table 3 below.

Word	Frequency (f)	Rank (r)	r * f
the	46,047	1	46,047
of	21,318	2	42,636
and	11,813	3	35,439
to	10,971	4	43,884
in	8,887	5	44,435
is	7,867	6	47,202
a	6,577	7	46,039
that	5,031	8	40,248
be	4,721	9	42,489
it	4,334	10	43,340
as	3,898	11	42,878

[12] Table generated from *CETA* data with *ReCor*.

Word	Frequency (f)	Rank (r)	r * f
which	3,824	12	45,888
from	3,689	13	47,957
by	3,684	14	51,576
at	3,300	15	49,500
or	2,964	16	47,424
earth	2,897	17	49,249
s	2,826	18	50,868
are	2,755	19	52,345
sun	2,688	20	53,760
e	2,545	21	53,445
t	2,512	22	55,264
this	2,353	23	54,119
on	2,067	24	49,608
for	2,055	25	51,375
with	2,030	26	52,780
will	1,993	27	53,811
but	1,922	28	53,816
its	1,922	29	55,738
moon	1,854	30	55,620
we	1,672	31	51,832
not	1,524	32	48,768
have	1,511	33	49,863
their	1,487	34	50,558
one	1,457	35	50,995

Table 3. Frequency ordered CETA subcorpus 1-gram words.

As we can observe in the table above, based on Zipf's Law, the sample tests of *CETA* contain several words that are repeated. The first words in the table are the most frequently used in the sample texts compiled. The order represents the rank (r), and the number of occurrences is expressed as frequency (f). Zipf stated that $r*f= c$, with c being a constant for any text. By comparing these data with the data obtained in table 2, we can determine that there is an inverse relationship between the frequency of occurrence and rank.

Our findings suggest, then, that *CETA* is a representative sample of a larger body of existing texts on astronomy by English-speaking authors, both American and European, in the eighteenth and nineteenth centuries. Based on the data obtained, *CETA* attains quantitative representativeness with 1 gram as of 35 documents and 386,255 Tokens. Since *CETA* consists of 42 documents and 450,000 words, it is, thus, a representative sample that covers the terminology of the specialized field that it aims to represent. We saw that Graphic Studies A and B depict the saturation point with an ellipse. In this respect, Table 1 supports this outcome, in that the data at the bottom of the first field, Types, show that the word increase is minimal, which means that at this point *CETA* increases in size but not in new words, hence *CETA* is representative of its time and field of study.

The findings of this study provide further support for previous evidence on the representativeness of *CETA*. Also, they prove our hypothesis and summarize the purpose of the study. Our findings offer an insight on impact assessment quantitative lexical analysis, using the *ReCor* tool. Does the analysis of quantitative representativeness using *ReCor* offer a cross-check to confirm and add to the validation of *CETA*'s representativeness, as found in previous research? The results shown above answer this question; this study does indeed provide further support, from the fields of bibliometrics and documentation, and compliments earlier research here from the fields of philology and linguistics. In our study representativeness was attained at 35 sample texts.

5. Conclusions

In this paper we have presented a quantitative analysis of lexical density with the aim of determining the threshold of representativeness of the *CETA* subcorpus of the *Coruña Corpus of English Scientific Writing*. The procedures used and described by the MuStE research group for the compilation of the *CC* have proved to be rigorous. According to our results, an adequate specialized corpus has been compiled. By means of the *ReCor* program algorithm, we have verified that English scientific terminology on astronomy has been successfully represented by the texts in the corpus. It can therefore be concluded that:

First, the object of the *Coruña Corpus* is to compile a more or less equal number of representative samples and words for each scientific register, in order to facilitate comparative studies of any kind, and confirming the wide range of linguistic variation in academic prose.

Second, *ReCor* has proved to be a valid application as a research tool for determining the representativeness of corpora already compiled. It shows results in graphs and allows access to the data tables, which can then be analysed. In addition, it allows us to generate lists of all the words in the sets of texts that shape the corpus which, when sorted both by frequency and alphabetically, allow secondary and parallel data checks, and then triangulation.

Third, from the quantitative numerical data obtained from *ReCor*, *CETA* is found to be representative when reaching 35 documents, with a minimum of 386,000 words. Given that *CETA* comprises 42 documents and 450,000 words, the sample texts are indeed representative, covering the terminology of the specialty field that it aims to represent.

Fourth, we have offered a technical contribution to the field of bibliometrics and documentation studies, as well as an additional method for the *a posterirori* validation of representativeness in corpora. A combination of different types of qualitative and quantitative approaches is very useful for data triangulation and cross-checking evidence and quality.

Fifth, although the concept of representativeness is imprecise, in that there is still no general agreement as to what the ideal size of a corpus should be, our quantitative analysis of the data from an *a posteriori* documental and bibliometric perspective supports the findings of previous research on the representativeness of *CETA* and proves that it is well designed and valid for the purposes for which it was intended.

To conclude, we would like to underline the fact that the findings of this study support evidence from previous studies and verify that the *CETA* subcorpus of the *Coruña Corpus* is a representative sample of texts, covering the terminology of scientific texts on astronomy written by English-speaking authors, between the eighteenth

and nineteenth centuries. That being the case, our research question and hypothesis on the representativeness of the sample texts in *CETA* are answered and verified. From the results we can add new knowledge and further evidence to existing evidence in terms of *CETA* representativeness. Future research can build on these observations on the quantitative representativeness of subcorpora, towards greater and more effective knowledge-based cross-check procedures to verify the representativeness of the *Coruña Corpus*, ensuring overall data quality and validity for the purposes for which the corpus was intended.

Declaration of conflicting interests

Las autoras han solicitado y recibido permiso para usar el *Coruña Corpus of English Scientific Writing* y la *ReCor tool*.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

About the author(s)

Elena Alfaya-Lamas obtained an MA in Germanic Philology from the University of Santiago de Compostela in 1994 and a PhD in English Historical Linguistics in 2002. From 1998 to 2000 she was a postgraduate worker and scholarship holder in the Department of Linguistics of the University of Edinburgh. In November 2001 she became an Associate Lecturer at CESUGA-University College Dubin. In October 2003 she obtained a position as an "Isidro Parga Pondal" researcher at the University of A Coruña and in October 2004 she became a Lecturer and Researcher in the area of Information and Documentation Science at the University of A Coruña.

Her main research interests are historical linguistics, cognitive linguistics, discourse analysis, gender studies and mind-consciousness studies. She studied with the Mindfulness Association and the Kagyu lineage for years, developing competence in the range of skills necessary to teach Mindfulness, passing the Universities of Bangor, Exeter and Oxford Mindfulness-based Interventions, Teaching Assessment Criteria, MBI:TAC. She is currently co-heading the Mindfulness Association in Spain.

She teaches Informational Behaviour, Historical Archives and Records Management, Scientific Research Techniques and Digital and Information Management.

Menchu Garrote se gradúa en Información y documentación en la Facultad de Humanidades y Documentación de la Universidade da Coruña en 2019. Obtiene el Premio Extraordinario Fin de Estudios (Universidade da Coruña) y Premio Excelencia Académica de Galicia (Xunta de Galicia). Actualmente cursa el Máster Universitario en Patrimonio Histórico: Investigación y Gestión en el Campus de Toledo de la Universidad de Castilla-La Mancha. Comenzó su acercamiento a la investigación cuando obtuvo la beca de colaboración en formación complementaria en los departamentos universitarios de los centros propios de

la UDC durante el curso 2018/19. Tutorizada por la Dra. Alfaya-Lamas se adentró en la investigación sobre el Coruña Corpus diseñado por el grupo de investigación MUSTE de la UDC.

References

Biber, D. (1993). "Using Registered-diversified Corpora of General Language Studies". *Computational Linguistics*, 19 (2), 219-241.

Biber, D., Conrad, S. & Reppen, R. (1998a). Preface. In: D. BIBER, S. Conrad & R. Reppen (eds.), *Corpus Linguistics: Investigating Language Structure and Use* (pp. ix-x). Cambridge: Cambridge University Press.

Biber, D., Conrad, S. & Reppen, R. (1998b). Introduction Goals and Methods of the Corpus-based Approach. In: D. Biber, S. Conrad & R. Reppen (eds.), *Corpus Linguistics: Investigating Language Structure and Use* (pp. 1-18). Cambridge: Cambridge University Press.

Booth, A. D. (1967). "A Law of Occurrences for Words of Low Frequency". *Information and Control*, 10 (4), 386-393.

Corpas, G. y Seghiri, M. (2010). "Size Matters: A Quantitative Approach to Corpus Representativeness". In R. Rabadán, (ed.) *Lengua, traducción, recepción. En honor de Julio César Santoyo* (pp. 112-146). Secretar: Universidad de Alicante.

Crespo, B. & Moskowich-Spiegel, I. (2010). "CETA in the Context of the Coruña Corpus". *Literary and Linguistic Computing*, 25(2), 153-164.

Francis, W. N. (1982). Problems of Assembling and Computerizing Large Corpora. In S. Johansson (ed.) *et al*. *Computer Corpora in English Language Research* (pp. 7-24). Norway: Norwegian Computing Centre for the Humanities

Moskowich-Spiegel, I., Lareo, I., Camiña, G. & Crespo, B. (comps.) (2012). *Corpus of English Texts on Astronomy*. Amsterdam: John Benjamins.

Moskowich-Spiegel, I. (2011). "The Golden Rule of Divine Philosophy: Exemplified in the Coruña Corpus of English Scientific Writing". *Revista de Lenguas para Fines Específicos*, 17, 167-197.

Moskowich, I. & Crespo García, B. (eds.) (2012). *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins

Moyotl-Hernández, E. & Macías-Pérez, M. (2016). "Método para autocompletar consultas basado en cadenas de Markov y la ley de Zipf". *Research in Computing Science,* 115, 157-170.

Parapar, J. & Moskowich-Spiegel, I. (2007). "The Coruña Corpus Tool". *Revista de Procesamiento del Lenguaje Natural* 39, 289–290.

Sidorov, G. (2013). "N-gramas sintácticos no-continuos". *Polibits*, 48, 69-78.

Seghiri, M. (2011). "Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad". *Revista de Lingüística teórica y Aplicada* 49 (2), 13-30.

Seghiri, M. (2014). "Too Big or not too Big: Establishing the Minimum Size for a Legal *ad hoc* Corpus". *Hermes: Journal of Language and Communication in Business* 27 (53), 85-98.

Seghiri, M. (2015). Determinación de la representatividad cuantitativa de un corpus ad hoc bilingüe (inglés-español) de manuales de instrucciones generales de lectores electrónicos. In M. T. Sánchez (ed.), *Corpus-based Translation and Interpreting Studies: From description to application* (125- 146). Frankfurt: Frank & Timme.

Sinclair, J. (1991). Glossary. In: J. Sinclair (ed.) *Corpus, Concordance, Collocation* (pp. 169-176). Oxford: Oxford University Press.

Torruella, J. & Llisterri, J. (1999). Diseño de corpus textuales y orales. In: J. M. Blecua (ed.) *et al*. *Filología e informática. Nuevas tecnologías en los estudios filológicos* (pp. 45-77). Barcelona: Universidad Autónoma de Barcelona.