ENGLISH STUDIES
https://doi.org/10.1080/0013838X.2020.1798635

Routledge
Taylor & Francis Group

Astronomy, Philosophy, Life Sciences and History Texts: Setting the Scene for the Study of Modern Scientific Writing

Begoña Crespo and Isabel Moskowich

Department of Letters (English Studies), University of A Coruña, A Coruña, Spain

ABSTRACT

The aim of this paper is to offer a description of four of the existing subcorpora of the Coruña Corpus of English Scientific Writing. Both the principles of compilation and the sociolinguistic variables considered during the process of text selection will be described. The editorial practice underlying the computerisation of texts, as well as several pilot studies, will also be discussed.

ARTICLE HISTORY
Received 18 October 2019
Accepted 12 June 2020

KEYWORDS
Corpus linguistics; corpus compilation; balance; representativity; scientific discourse; late Modern English

1. Introduction

The *Coruña Corpus of English Scientific Writing* (henceforth *CC*) is a purpose-built electronic corpus conceived of as a resource for the study of scientific writing in English, focussing on the two centuries prior to the language becoming the *lingua franca* of science. The project began in 2003 when some members of the MuStE group were awarded funding from the University of A Coruña to explore the historical background of English as the language of science. We soon realised that the compilation of a corpus of scientific texts from the eighteenth and nineteenth centuries would fill a gap in the field of English historical linguistics. Such a corpus, we saw, would complement the "Scientific-thought styles" project under development by Prof Taavitsainen and her colleagues at the University of Helsinki (<https://www.helsinki.fi/en/researchgroups/varieng/scientific-thought-styles-the-evolution-of-english-medical-writing>).

Initially, the Helsinki project was intended to cover the Middle Ages and the early Modern period, focusing on medical texts. The *CC*, as we will see, contains a wide range of scientific texts from fields other than medicine, and thus it was seen to be complementary. Over time the project in Finland expanded to include the eighteenth century, but was still limited to medicine. So, the *CC* continues to complement it, containing as it does subcorpora that embrace other fields, including Astronomy, Life Sciences, Philosophy, History, Languages and Chemistry.

The *CC* is founded on solid grounds: socio-external considerations, theoretical principles of Corpus Linguistics, and the technical experience gained through years of work can be observed in the final output, a carefully planned, well-structured and consistent corpus that has already provided data for a large number studies on morphology,

CONTACT Begoña Crespo bcrespo@udc.es

semantics and syntax, as well as on discursive and pragmatic issues (see Section 5, below, on exploiting the corpus), all of which confirms its value as a resource for research.

From the seventeenth century onwards, and with the increase of literacy, different types of readership in English emerged in response to a variety of discursive patterns. One of the main consequences of the shift from scholastic paradigms in the Middle Ages to modern science during the seventeenth century[1] was a dramatic change in the way knowledge and technical advances were conveyed. The *CC* tries to reflect this, and includes samples of printed texts from a variety of domains in which language and discourse were used by scientists as a means of negotiating knowledge. Such text extracts are of potential interest not only to linguists but also to historians of science, as shown, for example, by the close contact between the *CC* project and researchers at the Instituto Interuniversitario de Historia de la Medicina López Piñero (<https://www.uv.es/uvweb/instituto-universitario-historia-medicina-ciencia-lopez-pinero/es/instituto-historia-medicina-ciencia-lopez-pinero-1285893059754.html>). The *CC* is able to provide researchers with significant new tools towards a better understanding of science, language and society during the period that it covers. It allows for the study of material from a diachronic perspective, yet due to the inclusion of different genres is also useful for other approaches, such as comparative studies or other kinds of synchronic analyses within a broad diachronic framework.

Since every field of scientific endeavour has its own writing traditions and restrictions, each of the subcorpora contains samples of texts from different scientific disciplines. Such an overlapping of disciplines constitutes a fundamental difficulty in the selection of representative samples of scientific language, especially in that we are not dealing here with present-day science. Instead of designing our own taxonomy of disciplines as a starting point for compilation, we turned to an existing classification founded on rationalist premises, that of UNESCO.[2] The first subcorpus compiled was *CETA*, *Corpus of English Texts on Astronomy*,[3] and the second was *CEPhiT, Corpus of English Philosophy Texts*.[4] The third, the *Corpus of History English Texts* (*CHET*)[5] has been recently released. At the present time, a fourth is nearing completion: the *Corpus of English Life Sciences Texts* (CELiST). Plans for the inclusion of further subcorpora are currently being made, including texts on Chemistry (*Corpus of English Chemistry Text*, *CECheT*) and on Languages (*Corpus of English Texts on Languages*, *CETeL*). The characteristics of the subcorpora will be presented in what follows.

2. Compiling the *CC* Subcorpora[6]: Principles and Parameters

For the compilation of any corpus, clear principles must be established and followed. These have to do with key issues in corpus linguistics methodology, such as balance, representativeness, sampling and other external criteria such as time-span, register and the selection of disciplines.

[1]Valle; Beal.
[2]UNESCO.
[3]Moskowich *et al*. (2012)
[4]Moskowich *et al*. (2016)
[5]Moskowich *et al*. (2019)
[6]Information on CELiST corresponds to a beta version, since it has not yet been released. Hence, some slight changes in word counts may arise after the final revision of samples.

2.1. Balance and Representativeness

The text samples in the *CC* have been carefully selected and put together (rather than being "arbitrarily cut-out smaller text chunks", a criticism levelled at some corpora in the *Lampeter Corpus* manual[7]) in order to represent English science writing during a specific time-frame as a tool for research. We are conscious of the fact that historical corpora of this kind are by necessity limited to written material, and that this is an unavoidable restriction; however, we believe that such a corpus is indeed useful for scholarly research, and for this reason we have endeavoured to establish firm principles, as described below.

One of the first decisions in the compilation process had to do with number of words per sample. We compiled two ten thousand-words text files per decade, so that each of the centuries represented for a discipline contains approximately two hundred thousand words. Some pilot studies have shown that 1000-word samples, as proposed by Biber,[8] are not always sufficient for the study of variation within the scientific register, this mainly because the scientific register was not as standardised at the time as it is nowadays. Thus, the word counts for the four subcorpora are those set out in Table 1:

We have also borne in mind the principles of representativeness and balance,[9] which are perhaps most highly valued by specialists in the field of corpus linguistics. It is the reading about the history of each of the disciplines compiled that has helped us decide on which authors to include. This means that we have selected not only well-known writers but also less famous ones that we have happened to know thanks to the work of their contemporaries. In addition, it was our conscious decision to include only edited and printed texts in prose, these corresponding to first editions originally written in English, and to avoid translations from other languages. On many occasions, however, the availability of first editions is not straightforward. When problems arose here, and bearing in mind that it is generally argued that language change can be observed within a thirty-year period,[10] texts published within a thirty-year time-span from the date of first publication date were selected. In recent years an increasing number of copyright-free images of texts have become available, and this has made access to texts for compilers easier. In all cases, information about the physical location of the texts, from which we have extracted our samples, is provided in each metadata file (University libraries, collections, electronic repositories, etc.).

In order to arrive at an accurate representation of the stylistic and pragmatic devices used in Late Modern English, we have collected extracts from different parts of the works sampled, so that introductions, central chapters and conclusions are more or less equally represented. Prefaces or dedications, which are not scientific in themselves, have been excluded. In fact, prefaces to scientific works written by women have been compiled in another corpus (PreWoS) which can be found on CQPweb (<https://cqpweb.lancs.ac.uk/>).

Differences between the two centuries represented in the *CC* can be attested in terms of both the evolution of science itself and social consideration of science and scientists. In the nineteenth century, science became more specialised, individual journals for specific topics

[7]Claridge.
[8]Biber.
[9]McEnery and Wilson; Biber, Conrad, and Reppen; Meyer; Hardie and McEnery.
[10]Kytö, Rudanko, and Smitterberg, 92.

Table 1. Word counts.

Century	CETA	CEPhiT	CHET	CELisT	Words/century
18th c. words	208,079	200,022	202,342	200,220	810,663
19th c. words	201,830	201,107	202,815	200,085	805,837
Words/sub-corpus	409,909	401,129	405,157	403,965	1,616,500

were established, and the role of the scientist as understood by society was gradually shifting towards one of a professional.[11] Obviously, while the pursuit of knowledge was becoming fully committed to the empiricist method, and as such was far removed from the tenets of medieval scholasticism, it was still not as organised as it would subsequently come to be. It should also be noted here that external factors have played a key role in the evolution of science, and indeed might ultimately be responsible for the linguistic differences detected in samples from the two centuries represented.

In terms of the selection of disciplines, we have already noted that the starting point was UNESCO's[12] classification of science and technology into six fields, as set out in Table 2, which also indicates (in blue) the disciplines thus far included in the compilation of the *CC*.

Some of these disciplines have been re-allocated, since there is no exact correlation between the present-day conception of a scientific field and any such notion in the Modern period. The degree of branching and specialisation of present-day science cannot be found in eighteenth and nineteenth-century texts. Therefore, Figure 1 illustrates the distribution of disciplines proposed for the *CC* as well as the different corpora being compiled.

The procedure used in the compilation of text samples in the *CC* entails the paired selection of disciplines from the soft and the hard sciences in order to attain the kind of balance which can facilitate comparative studies.

We firmly believe in the interaction between language and society, and hence it was considered to be important to provide metadata files containing this information. These files are divided into two different sections: "about the author" provides information on the author, as well as some labels that can be used for searches in the Coruña Corpus Tool (henceforth *CCT*) including sex, age of the author when the work was published, geographical provenance, etc.[13] The other part of the file contains details about the text sampled, including genre/text type, the source of the sample, and cross-references to other texts in the *CC*. From the inception of the project, both text samples and metadata files have been encoded in XML format, following TEI guidelines.

2.2. Time-span Represented

We know that changes in scientific thought bring about changes in scientific discourse.[14] We have therefore used landmarks in scientific thought based upon extra-linguistic considerations, rather than those in language change itself, to set the time limits of our selection. The time-span chosen begins with the initial emergence of the scientific revolution,

[11]Puente-Castelo.
[12]UNESCO.
[13]Crespo and Moskowich.
[14]Moskowich, "The Golden Rule of Divine Philosophy".

Table 2. UNESCO classification of science and technology.

I. Natural Sciences.
Astronomy, bacteriology, *biochemistry, biology, botany, chemistry*, entomology, geology, geophysics, *mathematics*, meteorology, mineralogy, computing, physical geography, *physics*, *zoology* and other allied subjects.
II. Engineering and Technology.
Engineering sciences such as: *chemical*, civil, electrical and mechanical engineering and their specialised subdivisions; forest products; applied sciences such as geodesy, industrial chemistry, etc.; architecture; the science and technology of food production; specialised technologies of interdisciplinary fields, e.g. systems analysis, metallurgy, mining, textile technology and other allied subjects.
III. Medical Sciences.
Anatomy, stomatology, basic medicine, paediatrics, obstetrics, optometry, osteopathy, pharmacy, physiotherapy, public health services, technical health assistance and other allied subjects.
IV. Agricultural Sciences.
Agronomy, zootechnics, fisheries, forestry, *horticulture, veterinary medicine* and other allied subjects.
V. Social Sciences.
Anthropology (social and cultural) and ethnology, demography, geography (human, economic and social), law, *linguistics*, management, political sciences, psychology, sociology, organisation and methods, miscellaneous social sciences and interdisciplinary, methodological and historical S&T activities relating to subjects in this group. Physical anthropology, physical geography and psychophysiology are normally classified with the natural sciences.
VI. Humanities.
Arts (history of art and art criticism, excluding artistic "research"), *ancient and modern languages and literatures, philosophy* (including the history of science and technology), *prehistory and history*, together with auxiliary historical disciplines such as archaeology, numismatics, paleography, genealogy, etc., religion, other subjects and humanistic branches as well as other methodological and historical S&T activities relating to the subjects in this group.
UNESCO (1988)

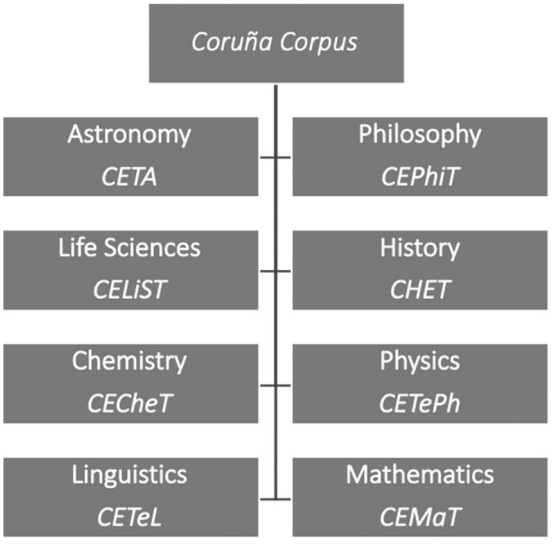


Figure 1. Disciplines and subcorpora in the *CC*.

the foundation of the Royal Society of London, and the publication by Bacon and Boyle of basic guidelines on how to present scientific works, these grounded in the ideas of clarity and simplicity of expression.[15] Empiricism promoted the development of science outside universities for the first time, a process that was probably favoured by better economic conditions and a new market for the practical application of science driven by popular

[15]Boyle.

demand. At the same time, as the dominance of religion began to decline, the importance of the observation and quantification of data as a means of reaching valid conclusions grew. These social and methodological changes resulted in the conscious creation of a new language to transmit science on the part of authors,[16] a representative sample of which we have tried to compile in the *CC*.

The earliest texts in our corpus are from 1700 (Mary Astell), 1702 (Robert Morden), 1705 (George Cheyne), and 1707 (James Douglas), a moment at which the old epistemological patterns of scholasticism were undergoing radical transformation.[17] This starting point also coincides with the new inductive method of reasoning, which in fact one of the authors included in *CEPhiT*, John Stuart Mill (1845), mentions explicitly.

The other end of the time-span, around 1900, is marked by several events which were of great importance in the history of science. These include the discovery of the electron by J. J. Thompson in 1896, the crisis in the foundations of mechanical physics, as announced by Mach, Kirchhoff and Bolzmann in the same year, Planck's introduction of quantum mechanics, at the very outset of the new century, and the publication of Einstein's paper on the Special Theory of Relativity in 1905. These developments brought with them the need to modify the discursive patterns of science, just as had happened two centuries earlier, by resorting to a simple prose with its own syntactic identity and distinctive vocabulary. Indeed, a call for a change in the discursive patterns of science was made by Thomas Huxley at the 1897 International Congress of Mathematics. From that moment, scientific discourse would change dramatically once more.

The last decade covered by the four subcorpora includes several authors,[18] and their discourse well illustrates the shifting paradigms in the expression of science that the accumulation of discoveries and the sheer weight of progress had brought about. This overwhelming manifestation of scientific facts seems, in effect, to represent the very kind objectivity that was called for by late seventeenth-century empiricists.

From the above description, we might also infer that social and political changes had a profound impact on the development of science and on the language of science. This will be addressed in the next section.

3. Extralinguistic Information: Communicative Format, Sex, Geographical Origin, Age

3.1. From Text Types to Genres and Communicative Formats

When we began the compilation of the *CC*, we used the label "text-type" to categorise our samples, since this tended to be the term adopted elsewhere. Obviously, our preference was to resort to established categories rather than to create our own, and for this reason we adopted those of Görlach's textual typology. As he states,[19] "proper definitions, and investigations including diachronic developments and diatopic contrasts seem to be indispensable before, for instance, corpus linguistics can claim to make reliable statements

[16]Halliday; Swales.
[17]Taavitsainen and Pahta.
[18]Alice Cooke; Montagu Burrows; Percival Lowell; Alpheus Packard.
[19]Görlach, 1.

based on a representative text selection". Hence, it seems appropriate to set out a working picture of the existing functional text categories in eighteenth and nineteenth-century scientific texts.

Although in general the terms genre and text-type have been identified with function and form, the terminological confusion regarding genre, text-type and textual category has led us to look for a solution which partially involves these concepts but which also clarifies the concept of communicative format. In previous research[20] we argued that:

> texts are produced with a clear function, in that the main aim of human language is to achieve some kind of response on the part of the receiver. However, depending on the kind of response the sender/addressor envisages, that is, the function of the text, form will vary. Hence, there is no absolute independence of form and function, and texts adopt forms depending on the function they perform (telegram, advertisement, treatise …). This mutual dependence means that form and function can be seen as a whole, one which ultimately cannot be wholly divided. For this reason we believe that the symbiosis between the form and function of a given communicative act deserves a new name: communicative format, the term we will use henceforth.

Although in many ways we prefer the term communicative format over genre or text-type, we have not modified our initial classification based on Görlach[21] since all the categories that he proposes were already in use during the late Modern period. He mentions, among others, article, essay, lecture, treatise, dialogue, textbook and letter, the definitions of which roughly coincide with those found in the *OED*. As such, we also use the *OED* for the definition of other possible formats that appear in new samples to be compiled. In addition to this initial classification, other parameters are taken into account, such as an author's explicit mention of the genre to which a work belongs, either in the preface or in any other kind of introductory material. The arrangement of the content of a work, or the very organisation of the text itself, has also been seen as yielding possible clues for the classification of samples. During the process of compilation of a particular subcorpus we have always endeavoured to avoid subjectivity as far as possible, we thus have ascribed samples to categories by applying a systematic methodology comprising five steps:

1. Use of existing classifications: Görlach's textual typology (2004) is used as the initial source for establishing taxonomies.
2. Use of *OED* definitions: The previous typology is complemented with those corresponding definitions recorded in the *Oxford English Dictionary*, looking specifically at the descriptions for the eighteenth and nineteenth centuries.
3. Titles of works: In many cases the title is self-explanatory, and in such cases is also considered for sample classification.
4. Prefatory material/audience: The aim here is to find any stated opinions by the author about his/her work. Authors may also make reference to their readership, thus providing information as to the function of the text.
5. The sample itself: Books published in the eighteenth century often contain a variety of works in different formats. Since we do not compile whole texts, it is the specific sample that needs to be assessed before ascribing any particular communicative

[20]Moskowich and Crespo, 309.
[21]Görlach.

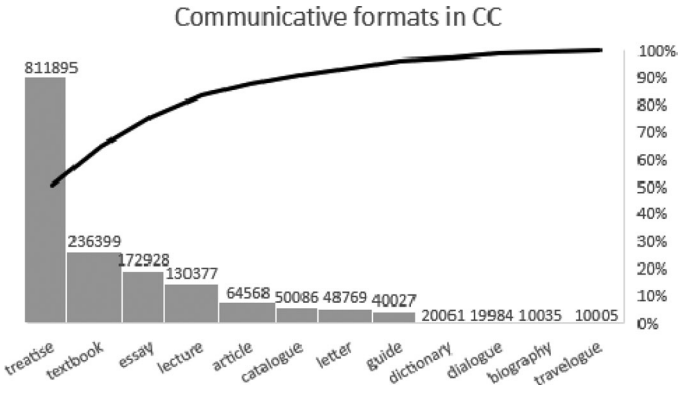


Figure 2. Communicative formats in *CC*.

format. In this way, we overcome the philologist's dilemma[22] and ensure accuracy in our classification.[23]

Once these steps have been applied in the classification process, using the same labels to distinguish formats in all subcorpora, we are able to see that their distribution is not homogeneous across disciplines, with format choice appearing to hinge on subject matter. The Pareto chart (Figure 2) shows the communicative format distribution in the *CC* as a whole. As can be observed, modern authors writing about History, Life Sciences, Astronomy and Philosophy, seem by large to prefer Treatise, followed by Textbook, Essay, Lecture and Article. These are mainly generic formats that can be adapted to any discipline. In the group of less common formats we find that most of them are discipline-specific, as we will see in the detailed analysis of formats per subcorpus.

Moessner[24] claims that a further basic consideration for text categorisation is the reader's perspective, that is, "which features make a reader interpret a text as a prototypical novel, short story, parody, etc.?" Once more, writer-reader interaction plays an interesting role in textual classification. In fact, communicative format division must, to some extent, reflect extralinguistic factors such as subject matter, purpose and discourse situation,[25] including the addressees' preferences or needs.

The different categories all seem to mirror the social reality of a world in which knowledge was no longer exclusive to universities or other cloistered institutions (where the taxonomies of lecture, treatise and textbook/handbook would fit perfectly), but was also in demand outside such institutions, as mentioned above. The vernacularisation of science and technology also brought about its popularisation, and new ways of communication had to be found. Books, especially treatises and textbooks, were especially popular in the transmission of scholarly information in the seventeenth and eighteenth centuries, but the journal article also assumed a notable role in the nineteenth century.[26] Letters, dialogues and other forms were used, although not all

[22]Rissanen.
[23]Moskowich and Crespo, 310.
[24]Moessner, 132.
[25]Rissanen.
[26]Allen, Qin and Lancaster; Crespo.

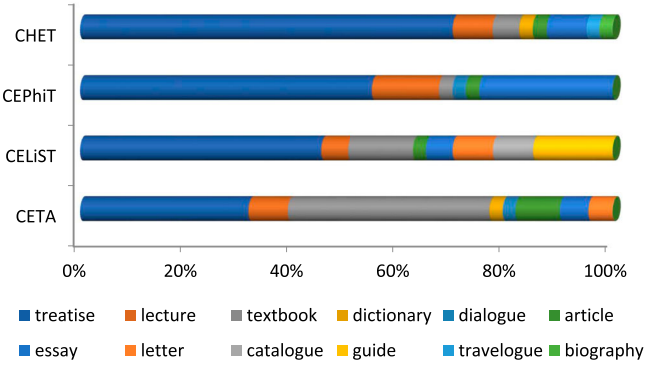


Figure 3. Distribution of formats across the corpus.

disciplines were equally represented here, since not all were equally popular in nature. In general, we can say that the distribution seen in Figure 3 broadly reflects production at the time[27]; also in the graph the number of words compiled in each subcorpus for each genre is represented:

In certain domains, writing seems to rely on just a few types of texts, whereas for others a wider range can be found. Subject matter here can be claimed to be the determining element for such choices. The first of the compiled subcorpora, CETA, contains eight different communicative formats: Treatise, Lecture, Textbook, Dialogue, Others (Dictionary), Article, Essay and Letter. Textbook (a total of 156,057 words) is the most abundant format, followed by Treatise (128,857). Article (34,014), Lecture (30,054) and Essay (22,259) are next most abundant. There are just two samples from Letters (18,717) and one example of Dictionary (10,044) and of Dialogue (9907).

In the case of CEPhiT, the texts selected can be group into a small number of genres, fewer than in other disciplines such as Astronomy, and even History or Life Sciences. Our samples of Philosophy texts are limited to six types, which in descending order are: Treatise (219,762), Essay (100,326), Lecture (50,133), Textbook (10,065), Dialogue (10,077) and Article (10,053). This may reflect the fact that in the nineteenth century Philosophy had come to be considered as simply another field of knowledge, and thus merited being known and circulated within different educational and cultural strata in society, rather than restricted to a select few.

CELiST contains eight different formats: Treatise (180,176), Guide (60,068), Textbook (50,074), Letter (30,052), Catalogue (30,045), Essay (20,020) and Lecture (20,099) and Article (9771). General formats (Treatise, Textbook) alternate with discipline-specific ones (Guide and Catalogue) which emerge due to the status and evolution of the discipline during the period.

CHET also contains eight different formats that do not fully coincide with the ones found in the other subcorpora. Treatise (282,907 words) is the most frequent format here, followed by Essay (30,323) and Lecture (30,091). There are only two samples of Textbook (20,203) and only one of Article (10,730), Dictionary (10,017), Travelogue (10,005) and Biography (10,035).

[27]Görlach, 1.

In sum, the category Treatise seems to be the favoured one for Late Modern authors of Philosophy, History and Life Sciences, as seen in Figure 3; other genres, such as Textbook, while very popular among astronomers in CETA[28] and the authors in CELiST, is found only twice in CHET and just once in CEPhiT. Essay is the next most favoured, indicating a firm preference for more formal formats. In CEPhiT, besides texts with an identifiable informative function (the most common one), there are also texts of an instructive or even entertaining nature, represented by the formats Lecture, Dialogue and Article. Catalogue and Guide are restricted to Life Sciences, whereas Travelogue is unique to History.

In general, the ascription of a sample to one or another format may be debatable, in that there are no clear-cut boundaries here or any unequivocal defining features. As Fowler[29] notes, genres may be considered as family members that "are related in various ways without necessarily having any single feature in common by all". The distribution is not identical in the two centuries compiled, and the following graph illustrates these differences, reflecting how external realities influenced text production in the field. Nineteenth-century authors resorted to a wider variety of genres than those in the preceding century, as Figure 4 illustrates:

The evolution of writing paradigms can be observed in the presence of the genre Article: just one example (in History) in the eighteenth century, compared to seven in the nineteenth century across all disciplines, four of these in Astronomy, one in Philosophy, one in Life Sciences, and one in History. This reflects the explosion in the nineteenth century of the journal article, noted above. As science became institutionalised, with specialised societies and journals established, articles evolved as a common vehicle for communication in the academic and scientific world. Travelogue disappears in the nineteenth century, whereas Lecture makes its presence felt very strongly in all disciplines in the same period.

As we have described, the taxonomy applied to samples does not rely on linguistic features exclusively; on the contrary, we have also used epistemological and social features. Thus, the corpus contains texts broadly representing the three epistemological levels of writing identifiable today: highest (typical of research articles and abstracts), high (abstracts in abstracting journals and informative scientific articles); medium (specialised non-academic articles).[30] It has been argued[31] that the *CC* is concerned with paragenres, that is, genres belonging to a single professional community[32] rather than to genres themselves. In the compilation process we have found cross-disciplinary formats (Treatise, Textbook, Essay, Article, Letter, Dialogue, Lecture) as well as discipline-specific ones (Guide, Catalogue, Biography).

Another sociolinguistic variable which has been considered is that of sex of the author, and this will now be discussed.

3.2. Sex

Sex is one of the extralinguistic variables relating to authors that we have decided to include in the *CC* as part of the metadata information. We use the term sex, rather

[28]Moskowich, "The Golden Rule of Divine Philosophy"; Moskowich "CETA as a Tool for the Study of Modern Astronomy in English" and "'A Smooth Homogeneous Globe' in CETA: Compiling Late Modern Astronomy Texts in English".
[29]Fowler, 41
[30]Fortanet *et al*..
[31]Moskowich, "The Golden Rule of Divine Philosophy".
[32]Monzó Nebot, 141.

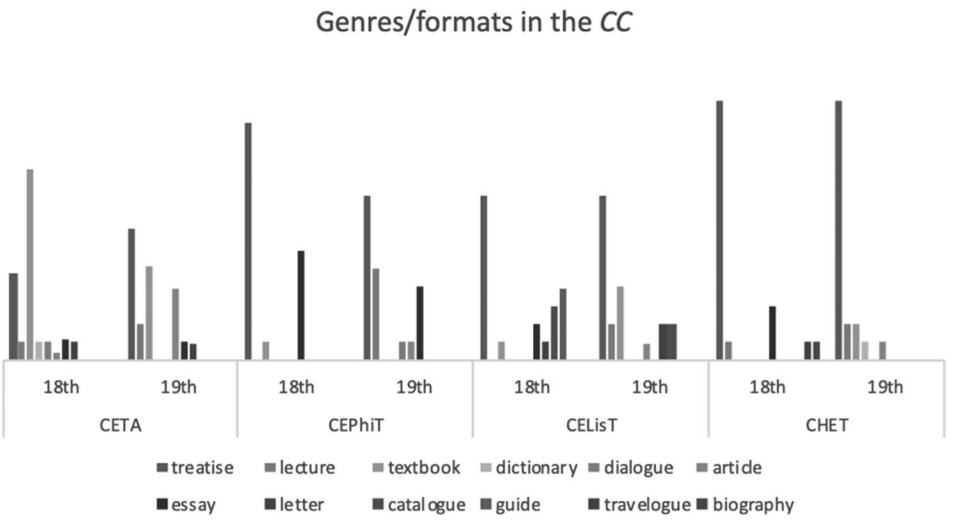


Figure 4. Words per genre in *CC* texts.

than gender, in that we are simply concerned with the biological sex of the author, not the social construct that gender represents, since we do not have access to any such information. From the two possible categories, female authors represent an evident minority.

Only a 13% of the total words compiled were written by women (212,679 words vs 1,402,262 words written by men, that is, 87%). In the Late Modern period it was not common for female activity to be made public or visible in certain social fields, and science was one of those traditionally defined as masculine. This means that many outstanding female scientists were never publicly recognised. It is difficult to trace their lives and careers because many took their husband's surname when married and some used a masculine pseudonym to ensure that their work was taken seriously.[33] Excluded from the official conduct of science, women who wanted to acquire an education had to do so by reading, by listening to other women and, occasionally, by listening to men in places other than institutions of "official knowledge", from which women were typically barred. Such social conditions constituted almost insurmountable boundaries for female authors. Thus, whereas women participated intensively in science, they often did so as mere assistants.

Overall, the increasingly social dimension of science with the expansion of literacy gradually began to afford women opportunities to create and publish their own work. This explains why the number of words produced by women authors in the nineteenth century section of our subcorpora is double that of the preceding century (see Figure 5 above).

At first, women of a high rank were able to take part in so-called "scientific circles". Everything scientific, from meetings to debates, came into fashion in the last quarter of the seventeenth century for those moving in the highest strata of society, and thus a few women were able to participate in such events.

[33]Herrero López, 75.

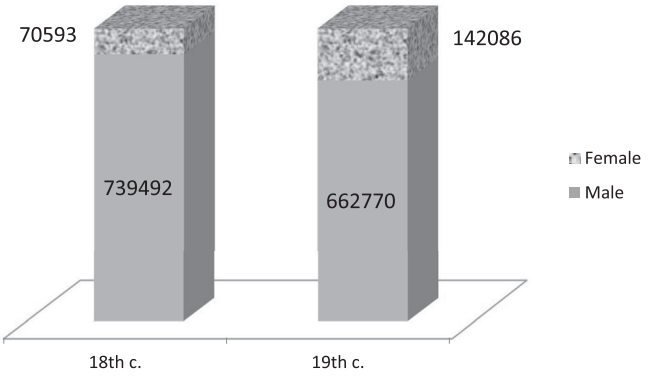


Figure 5. Distribution of sex of authors by century.

The family context and the education received by these women goes a long way to explain their ability to write on scientific matters: their fathers normally occupied positions of significant social prestige, as bankers, landowners, members of parliament or merchants with an interest in intellectual matters. Women's exclusion from scientific knowledge runs parallel to the process of the institutionalisation of science, which developed between the last part of the seventeenth century and throughout most of the eighteenth century[34] with the creation of societies and specialised associations to which women were not admitted. Nevertheless, the dissemination of science among the growing literate population also included the tentative participation of women in these matters. In fact, "from 1730 onward there was a European-wide effort led by Newtonians (…) to find a female audience for science".[35] Nevertheless, women who participated in scientific events were seen as a mere "ornament" for men in the eyes of society in general.[36] In addition, publishing was not a common female activity, as Figure 6 shows:

As a result, only two women have been included in CETA, although others are known to have existed. The two women we have chosen, Margaret Bryan for the eighteenth century and Agnes Mary Clerke for the nineteenth, signed the works they authored, and in both cases their research resulted in important advances and discoveries in the field.

Works on philosophy written by women were particularly infrequent in the Late Modern period. In fact, no women writing on philosophy in the nineteenth century have been included, and thus a total of only 30,192 words of female writing represent the whole period of compilation, the end point of which immediately predates the beginning of the suffrage movement. The three examples of female philosophy authors are Mary Astell (1700), Catharine Macaulay (1783) and Mary Wollstonecraft (1792).

There are eight female authors in the Life Sciences subcorpus: Elizabeth Blackwell (1737), the only eighteenth-century author, and seven female writers from the following century: Priscilla Wakefield (1816), Almira Phelps Lincoln (1832), Pratt (1840), Agassiz (1859), Lankester (1879) and Emily Gregory (1895).

[34]Solsona i Pairó, 86–87.
[35]Crespo, "On Writing Science in the Age of Reason".
[36]Crespo, "La intervención femenina en el desarrollo científico anglo-sajón".

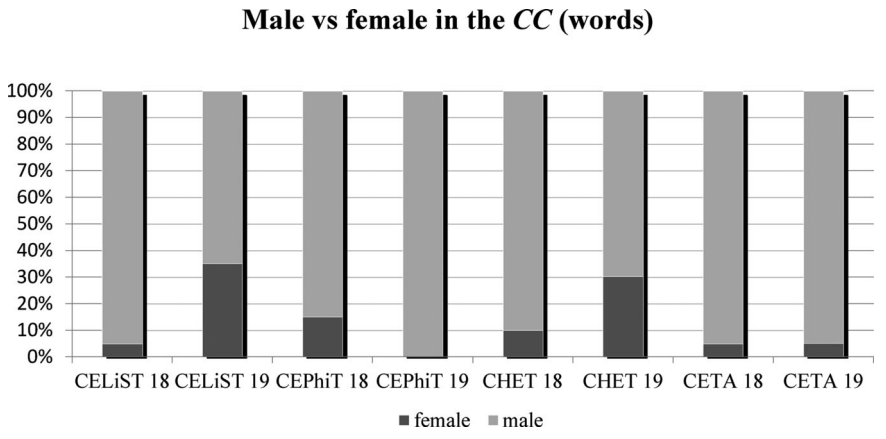


Figure 6. Male vs female writing in the *CC*.

CHET contains only two samples of eighteenth-century female writing. These women are Catherine Justice (1700) and Sarah Scott (1783). A greater number of texts from women writing about history have been collected for the nineteenth-century section of the corpus: Mercy Otis Warren (1805), Mary Callcott (1828), Lucy Aikin (1833), Elizabeth Sewell (1857), Martha Freer (1860) and Alice Cooke (1893). CHET thus reflects the scarcity of overt female activity in this field, although it is indeed higher than in other subcorpora.[37]

As for the communicative formats used, women seem to pursue the same communicative goals as men, and thus the formats they tend to adopt follow general patterns in scientific writing. Treatise (81,378) and Textbook (40,333) are the two most frequent, followed by Letters (20,043) and Guides (20,041). The remaining formats are only represented by one sample of ca. ten thousand words each. The social progress that science brought about allowed women to communicate scientific activity through the specific formats demanded by a discipline (as is the case of Guide) and in formats that conferred on women some visibility as authors (Lecture, Article), but men clearly dominated, imposing the path and the pace of change.

The following section turns to the variable of geography.

3.3. Mapping the CC: Geographical Variable

The whole *Coruña Corpus* has been designed according to a social constructionist perspective,[38] which entails that both the creation of knowledge and knowledge itself depend on context. This in turn implies that science can be seen as the interpretation of the world by a particular individual (the scientist) and not an independent entity or an absolute truth. As a result of this, language is a central element both for the interpretation of facts and for the transmission of these.

However, the linguistic performance of each speaker may vary, in that social conditioning may exert an influence on the speech or writing of an individual. Thus, the corpus has

[37]Moskowich, "The Golden Rule of Divine Philosophy"; "'A Smooth Homogeneous Globe' in CETA: Compiling Late Modern Astronomy Texts in English" and "CETA as a tool for the Study of Modern Astronomy in English"; "Philosophers and Scientists from the Modern Age".
[38]Hyland.

been compiled in such a way that it allows the researcher to analyse variation within scientific discourse, considering variables such the geographical origin of an author. The geographical distribution of authors must be taken to be the places where authors received their formal education, hence where they acquired the linguistic habits to be found in their writings. For this reason, we have used, whenever possible, texts by authors about whom we could find basic biographical information, so that this might enrich our understanding of their linguistic habits. This biographical information has been compiled in the metadata files accompanying the samples.

Following the principles of the *CC*, we have selected English-speaking authors writing in English, avoiding any sort of translation. American authors have also been included, although they are not represented in the different subcorpora in a homogeneous way, as we will show in the following pages.

The provenance of authors varies slightly depending on the discipline that is being compiled. Thus, in the corpus of astronomy texts, 50% of the authors acquired their linguistic habits somewhere in England, as opposed to 28% in North America. These are the two territories that predominate here, although Ireland is also represented (14% of authors) and Scotland (8%). Scientific knowledge was developing in the Old Continent under the influence of philosophical and cultural movements such as Empiricism. Meanwhile, North America was recovering from the effects of a very convulsive eighteenth century. Similarly, during the nineteenth century Americans were more deeply concerned with the practical application of scientific advances than with theoretical disquisitions.

The compilation of the *Corpus of English Philosophy Texts* (*CEPhiT*) allows for the representation of social and political shifts, extralinguistic concerns that have played a part in the presence of authors from one territory or the other. The impact of the American Civil War is one such example that accounts for the absence of American authors writing on philosophy in the nineteenth century. Another historical situation can explain the low percentage of philosophy writers: during the eighteenth-century, Ireland lived through the Protestant Ascendancy which meant that the native Irish population was excluded from power and public life[39]; with England as the coloniser, it is little wonder that most scientific texts were published by English authors. However, there is an unexpected number of authors that acquired their writing habits in Scotland. The reason might lie in the important philosophical tradition to be found in that territory, often referred to as the Scottish Enlightenment.[40]

When we turn to Life Sciences texts, the situation regarding the presence of North American authors seems to vary somewhat; only 13% of the authors in the whole subcorpus are from this territory. The percentage of authors acquiring their linguistic habits in Scotland (18%) is more in line with that found in the Astronomy subcorpus than in Philosophy. It was in Europe that the systematising process in Life Sciences began, with scholars such as Carl Linnaeus. The creation of taxonomies for the natural world was reinforced by the discovery of new peoples and species, a passion for many scientific gentlemen who became naturalists during the era of colonialism. England, as one of the main colonisers, produced the highest number of authors in our samples (58%).

Once again, Europe that was producing most works in the field of History (see Figure 7), whereas North America, as a newly-born nation that had lived through a difficult eighteenth

[39]Claydon and McBride.
[40]Abbagnano.

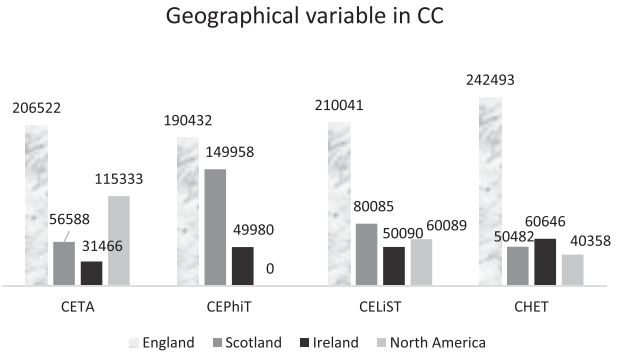


Figure 7. Geographical variable in the *CC*.

century, was arguably more concerned in the following century with the practical application of scientific advances and how to forge its own history than with the narration of past facts. In this sense, CHET, as well as CEPhiT, CELiST and CETA, are a small-scale mirror of reality.

The fourth and final sociolinguistic variable included in this paper looks at the age of authors.

3.4. Age of Authors

Age is the last variable to be explored in the four subcorpora of the *CC*. It should be noted that age here refers to how old the author was when his/her work was published. Therefore, in order to present our data, we have distributed writers according to six age groups, each of 10 years, up to the age of 75; the first group will thus encompass authors aged between 25 and 35, the second 36–45, and so on, with authors older than 75 forming the sixth and final group. Figure 8 shows the distribution of ages in the *CC*:

The second group (36–45 years old) contains the highest number of authors (43). This is not surprising, as this range of ages corresponds to the moment when authors have probably finished their training and are most productive. This is followed by the 33 authors who published their works in the following age group (46–55 years). It is worth noting that authors older than 76 do not abound, as might be expected if we bear in mind life expectancy for the 18th and 19th centuries.

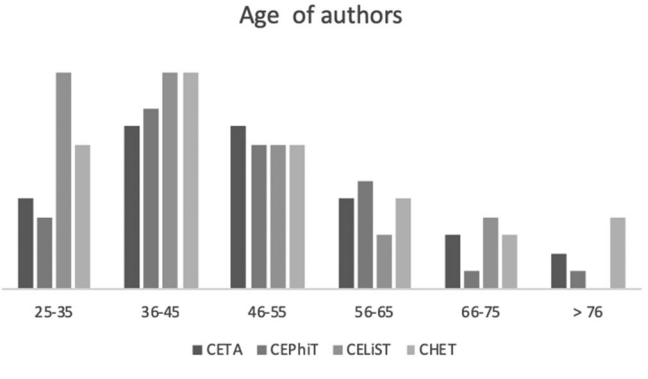


Figure 8. Age of authors in the *CC*.

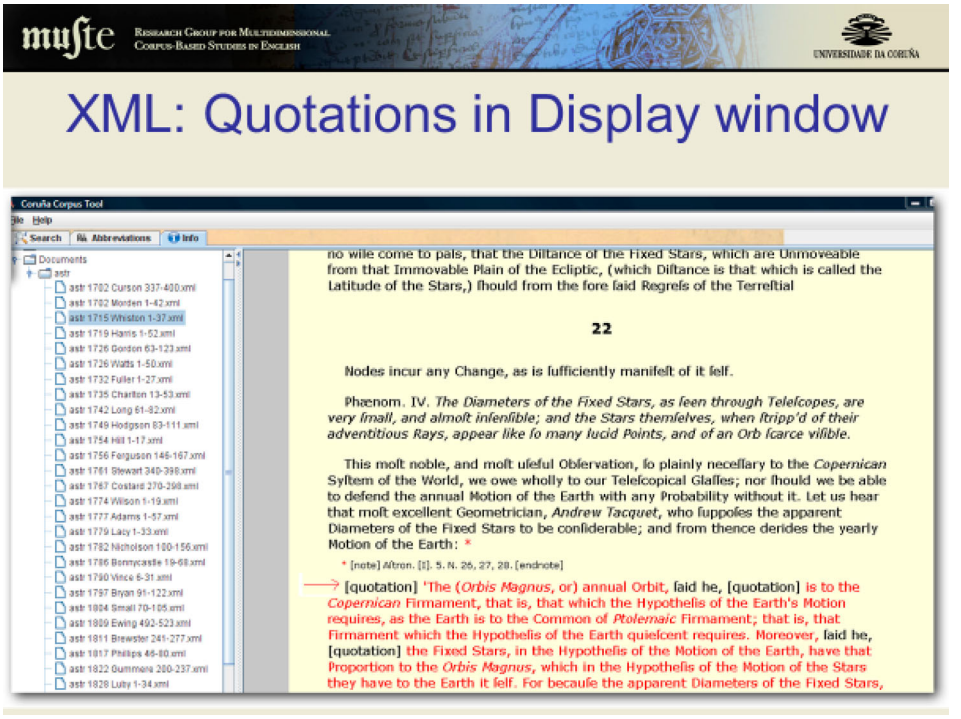


Figure 9. XML representation of elided text in red.

Turning to a breakdown of the data per discipline, we see that the youngest authors in the *CC* (aged 25–35) are those in Life Sciences (CELisT). This may be due to the attraction of all natural sciences for the younger generations, in that these were areas which were undergoing constant expansion and development, including many new discoveries, thus drawing the attention of those setting out on a life of scientific exploration. In fact, no writer older than 76 is registered in this field; on the contrary, the oldest authors in our samples are those writing History, a more classical and perhaps less immediately exciting discipline for the younger mind.

4. Digitising the Texts, Editorial Decisions, the *CCTool*

From the very beginnings of the compilation process, in 2005, texts were keyed in using XML. The XML format was preferred due to its multi-platform dimension. Since the intention of the compilers was to reach a wide community of researchers, the files containing the text samples were prepared following the guidelines provided by the Text Encoding Initiative (TEI).

In addition to common TEI tags, there is a list of special editorial marks used by compilers to identify particular elements, especially to indicate fragments in the samples that do not represent the original language of the author, as with quotations from other authors. The Coruña Corpus Tool (*CCT*), as a retrieval application, shows (in red) that a particular fragment does not represent the language of the author. Figure 9 illustrates

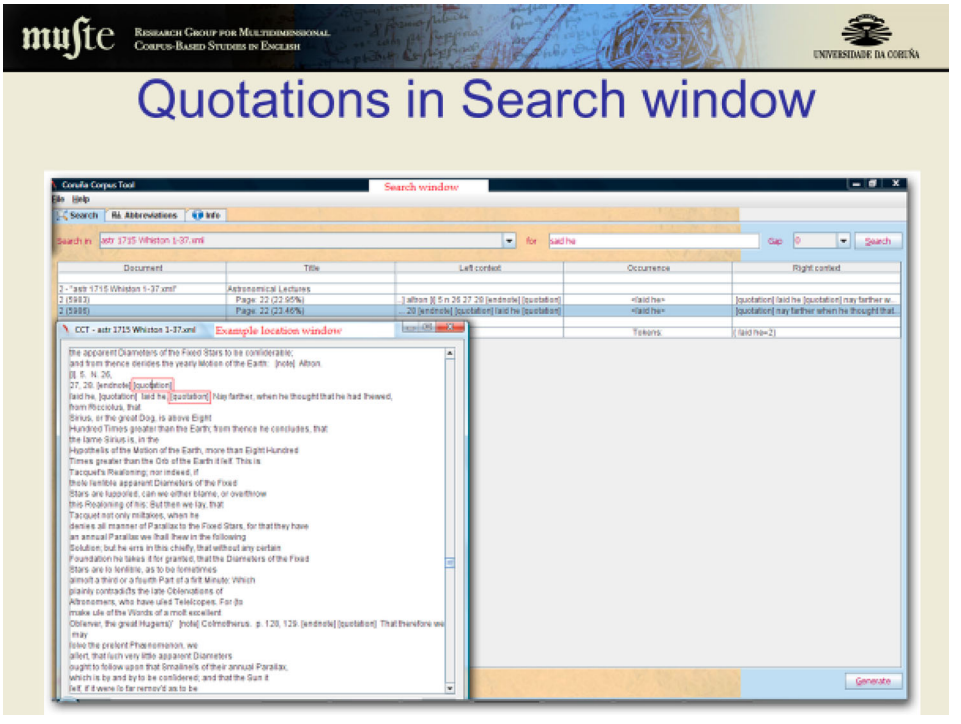


Figure 10. Sample as indexed by the *CCTool*.

this, whereas Figure 10 shows that the quoted fragment is not actually indexed in compilation and, therefore, no queries can be made on that material.

On occasion, it has been also necessary to represent mathematical or astronomical symbols and certain old characters (eighteenth and nineteenth-century spelling) and this has been done by the use of Unicode characters. As a result, the subcorpora offer extracts that come very close to the original texts.[41]

Figure 11 shows two of the peculiarities of the *CC*: symbols (in this case, astrological symbols) and special characters (here the long <s> typical of eighteenth-century printed texts) have all been retained.[42] A full account of the editorial policy[43] adopted can be found in the manual accompanying the *CCTool*.

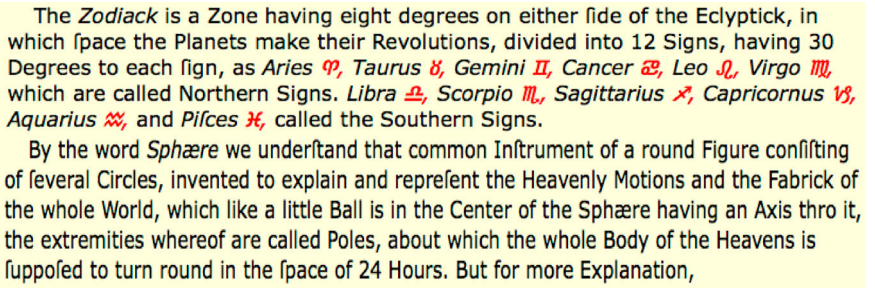
5. Further Work

The aim which motivated the creation of the different subcorpora that form the *Coruña Corpus of English Scientific Writing* was to allow the scholarly community to conduct research into the historical underpinnings of English for Specific Purposes. This interest was reinforced by a gradual increase, from the final decade of the twentieth century, in

[41]CETA has been tagged for POS (Gray and Biber), as has CHET (Degaetano, Menzel, and Teich) although different methodologies have been used in these two cases.

[42]Special characters are also used to show alternative spellings found in the texts. The CCTool has been designed to retrieve any spelling variant characteristic of this period in the English language, where homogeneity in writing tends to be lacking.

[43]Camiña and Lareo.


The *Zodiack* is a Zone having eight degrees on either ſide of the Eclyptick, in which ſpace the Planets make their Revolutions, divided into 12 Signs, having 30 Degrees to each ſign, as *Aries* ♈, *Taurus* ♉, *Gemini* ♊, *Cancer* ♋, *Leo* ♌, *Virgo* ♍, which are called Northern Signs. *Libra* ♎, *Scorpio* ♏, *Sagittarius* ♐, *Capricornus* ♑, *Aquarius* ♒, and *Piſces* ♓, called the Southern Signs.

By the word *Sphære* we underſtand that common Inſtrument of a round Figure conſiſting of ſeveral Circles, invented to explain and repreſent the Heavenly Motions and the Fabrick of the whole World, which like a little Ball is in the Center of the Sphære having an Axis thro it, the extremities whereof are called Poles, about which the whole Body of the Heavens is ſuppoſed to turn round in the ſpace of 24 Hours. But for more Explanation,

Figure 11. Representation of special characters and symbols.

the number of studies on genre conventions and special languages, as well as a similar interest in the study of eighteenth and nineteenth- century science.[44] In line with principles established in the field of corpora design, we have endeavoured to adhere to the principles of balance, representativeness, stratified sampling methods, and delimitation of the period covered, this determined by extralinguistic facts and realities. Pilot studies with compiled texts from the disciplines of Astronomy, Life Sciences, History and Philosophy have proved the *CC* to be a valuable resource for the description of the characteristics of academic writing and disciplinary conventions of scientific English in the Late Modern period.

At the time of writing, two further subcorpora are planned: CECheT, on Chemistry, and CETeL, dealing with language and linguistics. The digitisation and mark-up of texts is now under way, with the hope these new subcorpora will be as useful as the existing ones for future research into historical ESP.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Spanish Ministerio de Economia, Industria y Competitividad (MINECO) [grant number FFI2016-75599-P].

References

Abbagnano, N. *Historia de la Filosofía*. Vol. 2. Barcelona: Hora, 1982.

Allen, B., J. Qin, and F. W. Lancaster. "Persuasive Communities: A Longitudinal Analysis of References in the Philosophical Transactions of the Royal Society, 1665–1990." *Social Studies of Science* 24, no. 2 (1994): 279–310.

Beal, J. *English in Modern Times: 1700–1945*. London: Arnold, 2004.

Biber, D. "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8, no. 4 (1993): 243–257.

Biber, D., S. Conrad, and R. Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

[44]Guerrini.

Boyle, R. *A PROEMIAL ESSAY, WHEREIN, With Some Considerations Touching EXPERIMENTAL ESSAYS in General, Is Interwoven Such an Introduction to All Those Writ/ten by the Author, as is Necessary to Be perus'd for the Better Understanding of Them*. London, 1661.

Camiña, G., and I. Lareo. "Editorial Policy in the Corpus of English Philosophy Texts: Criteria, Conventions, Encoding and Other Marks." In *'The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy*, edited by I. Moskowich, G. Camiña, I. Lareo, and B. Crespo, 45–60. Amsterdam: John Benjamins, 2016.

Claridge, C. *Life is Ruled and Governed by Opinion: The Lampeter Corpus of Early Modern English Tracts. Manual of Information. 1999/2003*. ICAME Collection, 1999.

Claydon, T., and I. McBride. *Protestantism and National Identity. Britain and Ireland, c.1650 ±c.1850*. Cambridge: Cambridge University Press, 1998.

Crespo, B. "Astronomy as Scientific Knowledge in Modern England." In *Astronomy 'Playne and Simple': The Writing of Science between 1700 and 1900*, edited by I. Moskowich, and B. Crespo, 15–34. Amsterdam: John Benjamins, 2012.

———. "La intervención femenina en el desarrollo científico anglo-sajón." *Cuadernos del IEMYR* 23 (2015): 105–120.

———. "On writing Science in the Age of Reason." *Revista Canaria de Estudies Ingleses (RCEI)* 72 (2016b): 53–78.

Crespo, B., and I. Moskowich. "CETA in the context of the Coruña Corpus." *Literary and Linguistic Computing* 25, no. 2 (2010): 153–164. doi:10.1093/llc/fqp038.

Degaetano-Ortlieb, S., K. Menzel, and E. Teich. "Typical Linguistic Patterns of English History Texts from the Eighteenth to the Nineteenth Century: An Information-theoretic Approach." In *Writing History in Late Modern English: Explorations of the Coruña Corpus*, edited by I. Moskowich, B. Crespo, L. Puente-Castelo, and L. Monaco, 58–81. Amsterdam: John Benjamins, 2019.

Fortanet Gómez, I., S. Posteguillo Gómez, J. Francisco Coll García, and J. Carlos Palmer Silveira. "Linguistic Analysis of Research Article Titles. Disciplinary Variations." In *Perspectivas Pragmáticas en Lingüística Aplicada*, edited by I. Vázquez-Orta and I. Guillén-Galve, 443–448. Barcelona: Anubar, 1998.

Fowler, A. *Kinds of Literature: An Introduction to the Theory of Genres and Modes*. Harvard: Harvard University Press, 1982.

Görlach, M. *Text Types and the History of English*. New York: Walter de Gruyter, 2004.

Gray, B., and D. Biber. "The Emergence and Evolution of the Pattern N + PREP + V-ing in Historical Scientific Texts." In *Astronomy 'Playne and Simple'. The Writing of Science Between 1700 and 1900*, edited by I. Moskowich, and B. Crespo, 181–198. Amsterdam: John Benjamins, 2012.

Guerrini, A. "The Material Turn in the History of Life Science." *Literature Compass* 13, no. 7 (2016): 469–480.

Halliday, M. A. K. "On the Language of Physical Science." In *Registers of Written English: Situational Factors and Linguistic Features*, edited by M. Ghadessy, 162–178. London: Pinter Publishers, 1988.

Hardie, A., and T. McEnery. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambrdge University Press, 2011.

Herrero López, C. "Las mujeres en la investigación científica." *Criterio* 8 (2007): 73–96.

Hyland, K. *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins, 1998.

Kyto, M., J. Rudanko, and E. Smitterberg. "Building a Bridge between the Present and the Past: A Corpus of 19th Century English." *ICAME Journal* 24 (2000): 85–97.

McEnery, T., and A. Wilson. *Corpus Linguistics, An Introduction*. Edinburgh: Edinburgh University Press, 1996.

Meyer, C. F. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press, 2002.

Moessner, L. "Genre, Text Type, Style, Register: A Terminological Maze?" *International Journal of English Studies* 5, no. 2 (2001): 131–138.

Monzó Nebot, E. *La professió del traductor jurídic i jurat: descripció sociològica del professional i anàlisi discursiva del transgènere*, Unpublished PhD, Castellón: Universitat Jaume I, 2002.

Moskowich, I. "CETA as a tool for the Study of Modern Astronomy in English." In *Astronomy 'Playne and Simple': The Writing of Science Between 1700 and 1900*, edited by I. Moskowich, and B. Crespo, 35–57. Amsterdam: John Benjamins, 2012b. http://hdl.handle.net/2183/22178.

——— "Philosophers and Scientists from the Modern Age: Compiling the Corpus of English Philosophy Texts (CEPhiT)." In *'The Conditioned and the Unconditioned: Late Modern English Texts on Philosophy*, edited by I. Moskowich, and B. Crespo, 1–23. Amsterdam: John Benjamins, 2016.

———. "'The Golden Rule of Divine Philosophy' Exemplified in the Coruña Corpus of English Scientific Writing." *Lenguas para fines específicos* 17 (2011): 167–198. http://hdl.handle.net/2183/21562.

———. "'A Smooth Homogeneous Globe' in CETA: Compiling Late Modern Astronomy Texts in English." In *Creation and Use of Historical English Corpora in Spain*, edited by N. Vázquez González, 21–35. Cambridge: Cambridge Scholars Press, 2012a.

Moskowich, I., G. Camiña Rioboó, I. Lareo, and B. Crespo. (comps). *A Corpus of English Texts on Philosophy (CEPhiT)*. Amsterdam: John Benjamins, 2016 (Also open acces at https://ruc.udc.es/dspace/handle/2183/21847).

Moskowich, I., and B. Crespo. "Classifying Communicative Formats in CHET, CECHeT and others." In *CILC2016. 8th International Conference on Corpus Linguistics*. EPiC Series in Language and Linguistics, 1, edited by A. Moreno Ortiz and C. Pérez-Hernández, 308–320, 2016. [cited November 2018]. Available from: https://easychair.org/publications/volume/CILC2016.

Moskowich, I., I. Lareo, G. Camiña Rioboó, and B. Crespo. (comps). *Corpus of English Texts on Astronomy*. Amsterdam: John Benjamins, 2012 (Also open access at https://ruc.udc.es/dspace/handle/2183/21848).

Moskowich, I., I. Lareo, P. Lojo Sandino, and E. Sánchez-Barreiro (comps.). *Corpus of English Texts*. A Coruña: University of A Coruña, 2019. http://hdl.handle.net/2183/21849.

Oxford English Dictionary online. http://www.oed.com.

Puente-Castelo, L. "Explaining the Use of If … then … structures in CEPhiT." In *'The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy*, edited by I. Moskowich, G. Camiña, I. Lareo, and B. Crespo, 167–181. Amsterdam: John Benjamins, 2016a.

Rissanen, M. "Genres, Texts and Corpora in the Study of Medieval English." In *Anglistentag 1995 Greifswald: Proceedings*, edited by J. Klein, and D. Vanderbeke, 229–242. Tübingen: Max Niemeyer Verlag, 1996.

Solsona i Pairó, N. *Mujeres científicas de todos los tiempos*. Madrid: Talasa, 1997.

Swales, J. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press, 1990.

Taavitsainen, I., and P. Pahta. "The Corpus of Early English Medical Writing: Linguistic Variation and Prescriptive Collocations in Scholastic Style." In *To Explain the Present: Studies in Changing English Language in Honour of Matti Rissanen*. (Mémoires de la Société Néophilologique de Helsinki, 52), edited by T. Nevalainen, and L. Kahlas-Tarkka, 209–225. Helsinki: Société Néophilologique, 1997.

UNESCO. 1988. *Proposed International Standard Nomenclature for Fields of Science and Technology*. UNESCO/NS/ROU/257. Paris: United Nations Educational Scientific and Cultural Organization.

Valle, E. *A Collective Intelligence: The Life Sciences in the Royal Society as a Scientific Discourse Community, 1665–1965*. Turku: University of Turku, 1999.