RiCL Research in Corpus Linguistics

‘A matter both of curioſity and uſefulneſs’: Compiling the *Corpus of English Texts on Language*

Leida Maria Monaco - Luis Puente-Castelo
University of Oviedo & University of A Coruña / Spain

Abstract – This paper describes the compilation of CETeL, the subcorpus on ‘Language and Linguistics’ in the *Coruña Corpus of English Scientific Writing*, and discusses the various challenges encountered during the process of selection and digitisation of material. CETeL includes forty-four samples of texts on Language, Languages, and Linguistics from the period 1700–1900, and on completion will contain around 400,000 words. The paper will examine the historical context of academic writing in that period and the way in which this context affects the process of compilation. Likewise, the criteria followed in the compilation of the *Coruña Corpus* will be discussed in order to show the extent to which these criteria have affected the compilation of CETeL, and how they contribute towards making the corpus representative of the disciplinary practices of the period. Finally, the corpus will also be described according to a series of parameters used to assure representativeness and balance, namely the date of publication of samples, their genre, and the sex and linguistic background of their authors.

Keywords – *Coruña Corpus*; corpus compilation; Late Modern English; scientific writing

1. INTRODUCTION[1]

The *Corpus of English Texts on Language* (henceforth, CETeL) is one of the many twin subcorpora of the *Coruña Corpus of English Scientific Writing*, currently under compilation at the Universidade da Coruña by the Research Group for Multidimensional Corpus-Based Studies in English (MuStE, http://www.udc.es/grupos/muste). This paper covers the process of compilation and selection of samples in CETeL, which has now been completed,[2] and discusses the challenges faced here, focusing in particular on the

[1] The research reported here has been funded by the Spanish Ministry of Economy and Competitiveness (MINECO), grant number FFI2016-75599-P. This grant is hereby gratefully acknowledged.
[2] The initial process of computerisation of CETeL is complete, and a process of revision is about to start.

CC BY

difficult task of collecting a set of samples sufficiently representative of the type of language used in writing about language between 1700 and 1900, and on how these challenges were approached.

Section 2 presents the general design of the *Coruña Corpus*, to which CETeL belongs, while Section 3 explains its general compilation criteria. The *status quo* of Language and Linguistics studies in the eighteenth and nineteenth centuries is dealt with in Section 4, and Section 5 provides an analysis of the difficulties in reconciling general criteria and disciplinary particularities during the compilation of CETeL. Finally, a thorough description of CETeL is offered in Section 6, looking at a series of parameters including the distribution of samples over time, their topics, their genres, and the sex and linguistic background of their authors, followed by brief concluding remarks in Section 7.

2. The *Coruña Corpus*

Designed to be a "purpose-built electronic corpus conceived of as a resource for the study of scientific writing in English" (cf. Moskowich 2012: 35), the *Coruña Corpus* contains samples of texts of a scientific nature from the eighteenth and nineteenth centuries, allowing research at all linguistic levels except phonology. The corpus will consist of ten subcorpora (see Figure 1), all with the same design and principles of compilation, and one for each field of knowledge or scientific discipline.

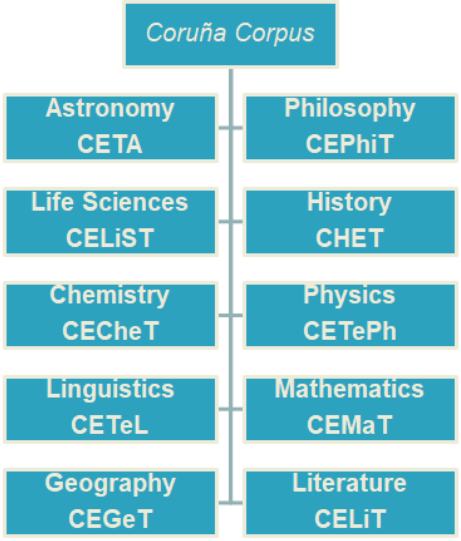

Figure 1: Current plan for subcorpora in the *Coruña Corpus*

Three of these ten subcorpora – on Astronomy (CETA), Philosophy (CEPhiT), and History (CHET) – have already been published. The former two were originally released along with collections of pilot studies (cf. Moskowich and Crespo 2012 and Moskowich *et al.* 2016, respectively).[3] An edited volume with pilot studies involving CHET is also in the final stages of preparation (cf. Moskowich *et al*. 2019).

In addition, three other subcorpora – on Life Sciences (CELiST), Chemistry (CECheT), and Linguistics (CETeL) – are under compilation, each currently at a different level of completion, whereas the subcorpora on Mathematics, Physics, Literature, and Geography are still at very early stages of development.

2.1. Size

As noted above, all the subcorpora of the *Coruña Corpus* have the same design and structure. Each subcorpus contains a series of samples of approximately 10,000 words in length, at a rate of two samples per decade, leading to a total of 20,000 words per decade and discipline, and hence 200,000 words per century and per discipline, and 400,000 words per subcorpus. This has been done with a view to making the *Coruña Corpus* approximately 4,000,000 words long when completed, a size which, arguably, should allow enough variety of texts and genres to dilute the influence of any idiosyncrasies in particular texts and to make the corpus representative of the scientific writing of the period.

The size of the samples has also been a matter of conscious selection, and the number of approximately 10,000 words is far from arbitrary. Despite the fact that Biber (1993) has argued in favour of samples as small as 1,000 words long, the compilers of the *Coruña Corpus* took into consideration the fact that the scientific register was less standardised between 1700 and 1900 than it is today, and hence the possibility that 1,000-word samples might not provide a good representation of the register during this period.

A further problem with such small samples is that they would inevitably slow down the process of compilation due to the difficulty in attaining a corpus of an adequate overall size, given the limited number of valid texts available for inclusion. In

[3] All three are now freely available in open access form at the Universidade da Coruña Open Repository at https://ruc.udc.es/dspace/handle/2183/21846.

terms of content, samples are usually selected in such a way that they cover all parts and sections of texts (such as introductions, methods, results, discussions, or/and conclusions, but as a general rule excluding prefaces), in order to avoid accusations of arbitrariness such as those mentioned by Claridge *et al*. (1999: introduction), who see text samples as "arbitrarily cut-out smaller text chunks" and suggest that full texts should be selected instead.

It is worth noticing that 'register' is understood here in Biber and Conrad's (2009) sense, namely as a variety of language characterised by having particular communicative purposes for particular situations (i.e. scientific texts). We also understand register as a scalable concept, which can be "defined at varying levels of specificity" (cf. Gray 2011: 3), as registers can be influenced by several situational factors at once. Thus, in our study, texts on language are considered a subregister of scientific texts on account of the particular constraints of a discipline, and texts by women are considered a subregister on account of the particular constraints faced by women authors. We refer to these types of situational factors as 'parameters' (sex, linguistic background, discipline, etc.), as they are both used to assure the representativeness of the selection of texts and as possible parameters for study, and the possible values these parameters have as 'categories' (female, Irish, linguistics). By contrast, 'genre' is understood here as a recurrent formal structure adopted by a variety of language as a result of conventions on how information is organised formally in order to achieve a given purpose (i.e. a research article). Thus, it is considered as one of the above-explained parameters, accounting for the particular constraints posed by formats and formal issues.

2.2. Timespan

The *Coruña Corpus* contains samples of scientific writing from the eighteenth and nineteenth centuries, a period of profound change both in science and in the way science was written (cf. Beal 2012). This period is delimited by two important events which might be considered as chronological bookends.

The early eighteenth century marks the culmination of the process of change in science which had begun in the seventeenth century with the works by Francis Bacon and Boyle and which saw scholasticism being replaced by a new scientific paradigm

(cf. Taavitsainen and Pahta 1998: 162). This coincides with the dissemination of Newton's ideas on gravity, which revolutionised the understanding and practice of physics and would go on to influence a great deal of scientific research over the following two centuries. The first years of the twentieth century, in turn, coincide with several major scientific breakthroughs, perhaps the most important being Einstein's 1905 paper on the Special Theory of Relativity, which is still considered a foundation for research in many disciplines.

The period in between these turning points is one of constant innovations and, at the linguistic level, broadly corresponds to what is referred to as late Modern English. Although the English language may be considered to have remained almost intact at the phonological, morphological and syntactic levels over the two hundred years prior to the twentieth century, it does, however, experience a gradual but consistent development of a distinct scientific register during that time, with a specialised terminology and a distinctive genre of its own, the research article, following Boyle's (1661) ideas on the five compulsory characteristics it should present: 'brevity', 'lack of assertiveness', 'perspicuity', 'simplicity of form', and 'objectivity' (cf. Allen *et al.* 1994; Atkinson 1996; and Gotti 1996, 2001, 2003, 2005). The end of this period of development is also marked by linguistic change, with the early 1900s witnessing several arguments in favour of a new scientific register, such as that called for by Thomas Huxley at the 1897 'International Congress of Mathematics', resulting in the consolidation of a relatively standardised scientific register as we know it today.

3. General compilation criteria in the *Coruña Corpus*

Each sample included in the *Coruña Corpus* has been selected in such a way as to create a set of samples which mirror scientific writing (and each discipline) as faithfully as possible during the period, ensuring the representativeness of the corpus.

Representativeness is assured by means of two processes:

1. The selection of suitable specific texts as examples of genuine scientific writing comprising a series of requisites to be fulfilled in order to be considered for inclusion.
2. The conformation of a balanced selection of samples, including examples of different types of scientific writing being produced during the period, with the

aim of achieving, when considered as a whole, a balanced representation of the register during the Late Modern English period.

3.1. Criteria for inclusion of particular texts

There are five main criteria that a text must satisfy to be considered a genuine manifestation of English scientific writing, and thus being eligible for inclusion in the *Coruña Corpus*.

First, only written, edited and published manifestations of scientific writing in prose are considered. Oral texts are excluded on the grounds that oral data is impossible to obtain for most of the period, although both transcriptions of lectures and scripts intended to be read aloud are eligible. Also excluded are texts in verse, since the inherent constraints in the language used in these texts imply a distorted or deliberately manipulated use of English, thus rendering such texts unrepresentative of the register.

Second, only texts written by native speakers are selected, since the use of English by non-native writers would not be representative of the English used in scientific writing during the period. Moreover, authors who completed all their training in English-speaking territories are prioritised on the assumption that these writers would be likely to present more genuine linguistic habits than those who lived and studied abroad.

In the same spirit, only texts written directly in English are selected, thus excluding translations, even where authors were the translators themselves, because interferences from the original language might have appeared in the translated text. This criterion is problematic, since a good proportion of the scientific production of the period was originally written in Latin, particularly at the start of the eighteenth century.

A further criterion is that only one work per author can be selected, thus avoiding jeopardising representativeness by over-representing the idiosyncrasies of particular authors. This limitation is applied at the corpus level rather than at the subcorpus level, so that only one work by any given author is selected for the whole of the *Coruña Corpus*.

Finally, first editions are preferred whenever possible, in order to avoid distorting the results on the diachronic axis by including samples from subsequent editions.

However, where first editions are not available, samples from editions published within a thirty-year timespan from the publication of the first edition are eligible, following Kytö *et al.*'s (2000: 92) assumption that thirty years is the minimum timespan in which language change can typically be observed.

3.2. Criteria to conform a balanced and representative set of samples

In order to achieve the desired balance and representativeness across the whole set of samples, each sample has to be selected very carefully in relation with all other samples in the subcorpus. In order to do so, each eligible sample is classified according to a series of parameters. Alongside the discipline and the period of the text, these include genre, plus the sex and linguistic background of the author.

These parameters in the classification of each individual sample are compared with information drawn from a detailed consideration of the history of the discipline over this period, taking into account its particularities and characteristic uses. In this way we can achieve as faithful a representation of the reality of the register during the period as possible. Some relevant aspects of the development of early studies in Language and Linguistics are discussed in what follows.

4. LANGUAGE AND LINGUISTICS DURING THE EIGHTEENTH AND NINETEENTH CENTURIES

Although interest in language appears in the very earliest works of Philosophy, studies on Language and Linguistics would not emerge as a distinct discipline of study until the nineteenth century, when a growing interest in biological evolution and diversification brought about an inevitable curiosity in the evolution of the world's different languages and the ontological meaning and transcendence of language as such. For a very long time, the study of language had been restricted largely to Latin, the official language for both the church and academic activity, with little attention paid to vernaculars, which were considered mere tools for communication, or in the case of poetry as an endeavour related more to entertainment than culture (cf. Bailey 1985; Beal 2004, 2012; Crespo 2004).

In the seventeenth century, however, it became apparent that English was slowly but steadily gaining popularity as an object of intellectual curiosity thanks to the

coincidence of a number of factors. The expansion of the British Empire between the late-sixteenth and early-eighteenth centuries led to a rise in the status of the English language, which was now associated with power, prestige and wealth. Using good English now became a means of social advancement, and for those who wished to have a certain status in society it became important to speak and write correctly (cf. Beal 2004; Hickey 2010; Millward and Hayes 2012). In the eighteenth century, the possibility of rising economically (and, to some extent, socially) in the spheres of trade and commerce created a "linguistically insecure middle class" (cf. Beal 2008: 22–23), whose financial success appeared to depend largely on their mastery of the linguistic register of their culturally superior clients. On the other hand, the translation of the Bible and the progressive substitution of Latin by English in academic and other official contexts, which had in fact begun far earlier (cf. Taavitsainen and Pahta 1998), created the need for wider and more conscious instruction in the vernacular, with a consequent proliferation of grammars and manuals for correct usage and pronunciation.

The preoccupation of philologists with grammar became very apparent in the eighteenth century, as can be seen in specific works by Swift (1712), Stackhouse (1731), Johnson (1747), and Fisher (1753), all of which are included in our corpus. While some were particularly concerned with the correct use of spelling and syntax, this as a reflection of a more cultivated social status through writing, others were unhappy with a number of linguistic trends of the time, most of which were considered linguistic corruptions that needed to be corrected.[4] As a result, many of the English grammars in this period can be regarded as style manuals, in that they often included extended essays on the *status quo* of the English language, along with lists of 'corrupt' terms or expressions which they advised readers to avoid. On the other hand, a simultaneous interest in the etymology and internal organisation of the vernacular awakened in philologists a renewed interest in classical languages and in the way that these were approached, which itself led to several attempts to revise and improve Greek and Latin grammars and manuals (such as Sheridan 1714, or Squire 1741, also included in CETeL).

In the nineteenth century, the German linguist and philosopher Willhelm von Humbolt observed that human language was a rule-governed system, and as such deserved to be described (cf. Schmidt 1975; Di Cesare 1990). Already by the end of the

[4] In fact, both Swift (1712: 16) and Johnson (1747: 10) were rather pessimistic about language change.

1700s, language began to be treated as an object of study of natural sciences, and languages themselves were treated as living entities and thus classified into families according to their origins, their evolution, and their behaviour (cf. Campbell 2001). Heavily influenced during this period by Darwinism, the study of languages entailed the reconstruction of their origins back to Proto-Indoeuropean, culminating by the end of the century in the work of the Neogrammarians (cf. Robins 1978, 1997). At the same time, a growing interest in Asian languages and cultures – the result of a new scientific interest in the colonies (cf. De la Cruz Cabanillas 2001; Beal 2004) – led to the extension in the scope of ancient languages under study to those outside Europe, as well as in the increasing habit of working on more than one language at a time, a practice which opened the doors to modern Comparative Linguistics.

All the trends summarised above can be found in the samples included in CETeL, and some of these trends are directly related to specific challenges faced during the process of compilation. These difficulties will be described in the next section.

5. Criteria applied: Difficulties faced during the process of compilation

As described above, the selection of samples in all subcorpora of the *Coruña Corpus* is conducted in such a way as to make the set of samples representative of the disciplinary practices of the time, and the selection of samples in CETeL is no exception. However, in this case, the process has been particularly challenging, especially for the beginning of the period, as the result of several factors.

First, as already noted, the development of Linguistics as an individual discipline occurs comparatively late, and this affects the process of selection of samples particularly during the first decades of the eighteenth century. Looking at the opinions of authors here regarding the nature of their own works, as expressed in their abstracts and other introductory material, we can find labels such as ‘language’, ‘grammar’, or ‘etymology’, yet these are not always used in the same ways in which we might understand them today. To resolve this problem, we established the criterion that CETeL would be a corpus of texts on Language, rather than on Linguistics. Thus, CETeL goes beyond Linguistics as it would be considered nowadays, introducing several texts on the nature and philosophy of language, thus representing how scientific discourse on Language was considered at the time. Moreover, this also represents the

reality of scientific work during the period, as disciplines have become ever more specialised since the first decades of the eighteenth century (cf. Burke 2000: 132–137).

A second difficulty, although perhaps less important here than in other disciplines, was that Latin was still widely used in texts on Language and Linguistics well into the eighteenth century. This meant that a significant number of possible samples were ineligible, including not only samples written in Latin, but also samples in English which were translated from Latin. Identifying these translations was particularly difficult, because they were sometimes not advertised as such, particularly when the author was both the original writer and the translator. This made it necessary to conduct a comprehensive review of all the work of a given author in order to ascertain that a sample was indeed not a translation from a previous original in Latin or any other language.

Thirdly, some of the formats used during the period led to specific problems during the process of computerisation. For instance, dictionaries, with their organisation in entries which repeat the same grammatical structures, normally present little linguistic interest, whereas grammars of foreign languages, with a high number of examples in these languages, perhaps even in different alphabets, pose problems for transcription, as the rules of the *Coruña Corpus* qualify that the latter cannot be transcribed, and that the former, even if transcribed, have to be encoded in such a way that they do not count as words in the corpus (cf. Camiña and Lareo 2019: 22). This is also problematic in that whereas entire passages in a foreign language sometimes can easily be excluded, it is much more common to find foreign terms, endings, etc. inserted in the main text. In such cases, these are usually identified individually with editorial marks, but if they are so numerous as to impede transcription, the whole passage must be deleted, as it would not represent the real linguistic habits of the author. A further problem is the lack of a standard phonetic transcription for works on phonetics, and also in grammars for foreign languages where we find examples of how such words should be pronounced.

The final, and perhaps most notable, problem is the general lack of information about many authors at the beginning of the period. The principles of the *Coruña Corpus* state that it is preferable to select authors "about whom we could find basic biographical information and hence whose linguistic habits we could infer," (cf. Moskowich 2012: 48) and thus to be able to confirm that they are indeed native speakers. When this has

not been possible, samples were discarded, and this led to the rejection of a considerable number of otherwise valid samples from our initial inventory.

Further problems are related to the balance of the set of samples in the corpus. An important aspect of this balance is that it does not imply all the different categories being evenly represented throughout the period, which is, in any case, an impossible task with only two samples per decade. Rather, balance implies that the selection of samples should be representative of the reality of the discipline during the period, providing a good representation of both inter-disciplinary and intra-disciplinary differences across the different parameters.

This is best seen in the unequal distribution of genres over the period and across disciplines. For instance, as a result of the consideration of English as a means of social advancement, a large number of textbooks are included from both centuries.[5] This contrasts with other disciplines, such as Chemistry, in which textbooks were essential for the initial segment of the period, contributing to the dissemination of knowledge, but fell away in terms of importance during later stages.

However, sometimes not all categories are equally available, and this must also be taken into account when sorting the samples. For instance, particularly during the eighteenth century, grammars represent a very important proportion of all scientific production on Language. They reflect a widespread preoccupation with the correct use of language although, as noted above, some of these grammars also include a diachronic or stylistic perspective. Such works are featured in the corpus, but sometimes other content such as discussions on the correct use or the nature of language itself, both of which are also representative of the eighteenth century, are not as readily accessible as grammars. This leads to compilers having to choose samples from many valid grammars, whereas for other genres choosing among samples becomes impossible and it is sometimes necessary to include almost any valid sample. This in turn is complicated by the fact that sometimes genres are not easy to identify. A text might exhibit conflicting characteristics, being very broad and exhaustive in nature, and thus being potentially an example of either a treatise or a didactic work. In this sense, it could be classified as a textbook or as a manual but, in addition, it might have a question-and-answer format in a constructed dialogue form. In such cases, and if the

[5] It must be noted that the 'textbook' label includes a fuzzy textbook/handbook/manual category in this particular subcorpus.

author makes no reference to the genre itself, it is left to compilers to decide which genre is best represented in the sample. This is achieved by means of a close reading and a comparison of the texts with other, undoubted, texts to check the similarities between them. Such a comparison allows to assign the texts to a particular genre.

In order to faithfully represent the discipline of the period, several short texts have been included *in toto* even though they are shorter than the 10,000-word limit used across the *Coruña Corpus*, since they are characteristic of the production on linguistics in the period. However, special care has been taken in that the final number of words in any decade is roughly 20,000.

Finally, regarding the distribution of the samples according to the sex of their authors, it is worth mentioning that the majority of the samples of the *Coruña Corpus* are written by men, as was the case with science in general during the period. At that time, women faced considerable difficulties in accessing scientific knowledge and had to overcome a great many obstacles if they sought to become part of the social community of scientists. However, every subcorpus of the *Coruña Corpus* includes samples of texts written by women. Selecting female-authored texts has not always been easy, since publications by women lacked biographical information far more often than in the case of men, and women were also frequently obliged to write under pseudonyms or anonymously. Despite these difficulties, CETeL is among the subcorpora with the largest number of female authors.

6. DESCRIPTION OF CETeL

In this section, the beta version of CETeL will be described according to a number of parameters, namely: the distribution of the text samples over time, the topics and genres included in the overall set of samples, plus the sex and geographical origin and linguistic background of authors.[6]

CETeL contains a total of 44 samples, 24 from the eighteenth century and 20 from the nineteenth: the reason for this disparity lies in the inclusion of three, rather than two, samples in the following four decades: 1710s, 1720s, 1740s and 1780s. As shown in Table 1 below, most samples contain *c*.10,000 words, but in these four decades shorter

[6] It is important to note that the description provided here corresponds to the beta and not to the definite version of CETeL. Some classifications, particularly the word count of samples, may change during the process of revision, which is about to start.

texts were included. This was done partly due to the need to introduce some shorter texts *in toto* such as the 'Proposal for Correcting, Improving and Ascertaining the English Tongue' by Swift (1712) or Samuel Johnson's 'Plan of a Dictionary of the English Language' (1747), which were considered to be particularly representative of the period. In other cases, the quantity of text in a foreign language reduced the computable number of words in the selected text considerably, which was the case in 'The Rudiments of Grammar or the English-Saxon Tongue' by Elizabeth Elstob (1715). However, these issues do not affect the overall number of words, which is comparable in both centuries: 202,961 words in the eighteenth century and 203,062 in the nineteenth century.

Date	Author	Title	Words
1705	Lane, Archibald	A key to the art of letters, or, English a learned language, full of art, elegancy and variety. Being an essay to enable both foreiners, and the English youth of either sex, to speak and write the English tongue well and learnedly, according to the exactest rules of grammar, after which they may attain to Latin, French, or any other forein language in a short time, with very little trouble to themselves or their teachers: with a preface shewing the necessity of a vernacular grammar. Dedicated to His Highness the Duke of Glocester.	10,174
1706	Johnson, Richard	Grammatical commentaries: being an apparatus to a new national grammar: by way of animadversion upon the falsities, obscurities, redundancies, and defects of Lilly's system now in use.	9,908
1712	Swift, Jonathan	A proposal for correcting, improving and ascertaining the English tongue, In a letter to the most honourable Robert Earl of Oxford and Mortimer, Lord High Treasurer of Great Britain.	5,930
1714	Sheridan, Thomas	An easy introduction of grammar in English for the Understanding of the Latin Tongue. Compil'd not only for the ease and encouragement of youth, but also for their moral improvement; having the syntaxis examples gathered from the choicest pieces of the best authors. To which is added a compendious method of variation and elegant disposition of Latin.	7,777
1715	Elstob, Elizabeth	The rudiments of grammar or the English-Saxon tongue, first given in English: With an apology for the study of Northern antiquities. Being very useful towards the understanding our ancient English poets, and other writers.	6,839
1721	Gildon, Charles	The Laws of Poetry, as laid down by the Duke of Buckinghamshire in his Essay on Poetry, by the Earl of Roscommon in his Essay on Translated Verse, and by Lord Lansdowne on Unnatural Flights in Poetry, Explain'd and Illustrated.	6,161
1725	Stevens, John	A new Spanish Grammar, more perfect than any hitherto published. All the errors of the former being corrected, and the rules for learning that language much improv'd. To which is added, a vocabulary of the most necessary words: Also a collection of phrases and dialogues adapted to familiar discourse.	10,273
1728	MacCurtin, Hugh	The Elements of the Irish Language, Grammatically Explained in English. In 14 chapters.	5,140

Table 1: Samples included in CETeL and provisional word count in the beta version

Date	Author	Title	Words
1731	Stackhouse, Thomas	Reflections on the Nature and Property of Languages in General, and on the Advantages, Defects and Manner of Improving the English Tongue in Particular.	9,640
1737	Greenwood, James	The Royal English Grammar: containing what is necessary to the knowledge of the English tongue. Laid down in a plain and familiar way. For the use of young gentlemen and ladys.	10,014
1741	Squire, Samuel	Two essays, the former a defense of the Ancient Greek Chronology; to which is annexed, a new chronological synopsis; the latter, an enquiry into the origin of the Greek Language.	9,856
1747	Johnson, Samuel	The plan of a dictionary of the English language: addressed to the Right Honourable Philip Dormer, Earl of Chesterfield; One of His Majesty's Principal Secretaries of State.	6,909
1748	Martin, Benjamin	Institutions of Language; Containing, a physico-grammatical Essay on the propriety and rationale of the English tongue. Deduced from A general idea of the nature and necessity of speech for human society; A particular view of the genius and usage of the original mother tongues, the Hebrew, Greek, Latin, and Teutonic; with their respective idioms, the Italian, French, Spanish, Saxon, and German, so far as they have relation to the English tongue, and have contributed to its composition.	10,138
1751	Harris, James	Hermes: Or, a philosophical inquiry concerning language and universal grammar.	11,350
1753	Fisher, Anne	A new grammar, with exercises of bad English: or, An easy guide to speaking and writing the English language properly and correctly.	9,841
1762	Priestley, Joseph	A Course of Lectures on the Theory of Language and Universal Grammar.	8,855
1765	Elphinston, James	The Principles of the English Language Digested, or, English Grammar Reduced to Analogy.	11,604
1771	Fenning, Daniel	A New Grammar of the English Language; or, an easy introduction to the art of speaking and writing English with propriety and correctness: The whole laid down in the most plain and familiar manner, and calculated for the use, not only of schools, but of private gentlemen.	8,617
1776	Campbell, George	The Philosophy of Rhetoric.	9,082
1784	Nares, Robert	Elements of orthoepy: containing a distinct view of the whole analogy of the English Language; so far as it relates to pronunciation, accent, and quantity.	10,058
1784	Webster, Noah	A Grammatical Institute of the English Language, comprising, an easy, concise, and systematic method of education, designed for the use of English schools in America. In three parts.	10,040
1786	Jones, William	The Third Anniversary Discourse, on the Hindus. Delivered 2 February, 1786. By The President.	4,687
1797	Tytler, Alexander Fraser	Essay on the Principles of Translation.	10,068
1798	Fenn, Eleanor	The mother's grammar. Being a continuation of the child's grammar. With lessons for parsing. And a few already done as examples.	9,350
1810	Adams, John Quincy	Lectures on rhetoric and oratory: delivered to the classes of senior and junior sophisters in Harvard University.	11,913

Table 1 (continuation)

Date	Author	Title	Words
1810	Smart, B. H.	A practical grammar of English pronunciation: on plain and recognised principles, calculated to assist in removing every objectionable peculiarity of utterance, arising rather from foreign, provincial or vulgar habits; or from a defective use of the organs of speech; and furnishing, to pupils of all ages, the means of systematically acquiring that nervous and graceful articulation, which is the basis of a superior delivery: together with directions to persons who stammer in their speech, comprehending some new Ideas relative to English prosody.	9,611
1815	Richardson, Charles	Illustrations of English philology.	8,425
1819	Cobbett, William	A grammar of the English language: in a series of letters. Intended for the Use of Schools and of Young Persons in general; but, more especially for the Use of Soldiers, Sailors, Apprentices and Plough-boys.	12,713
1825	Cardell, William S.	Essay on language: as connected with the faculties of the mind, and as applied to things in nature and art.	15,040
1830	Booth, David	An analytical dictionary of the English language; in which the words are explained in the order of their natural affinity, independent of alphabetical arrangement; and the signification of each is traced from its etymology, the present meaning being accounted for when it differs from its former acceptation: the whole exhibiting, in one continued narrative, the origin, history, and modern usage of the existing vocabulary of the English tongue: to which are added, an introduction, containing a new grammar of the language, and an alphabetical index, for the ease of consultation.	11,026
1836	Allen, Alexander	An etymological analysis of Latin Verbs. For the use of schools and colleges.	10,128
1836	Bosworth, Joseph	The origin of the Germanic and Scandinavian languages, and nations: with a sketch of their literature, and short chronological specimens of the Anglo-Saxon, Friesic, Flemish, Dutch, the German from the Mœso-goths to the present time, the Icelandic, Danish, Norwegian and Swedish: tracing the progress of these languages, and their connexion with the Anglo-Saxon and the present English. With a map of European Languages.	10,601
1841	Latham, Robert Gordon	Elements of the English Language for the use of Ladies' Schools.	10,061
1845	Ellis, Alexander John	The Alphabet of Nature or contributions towards a more accurate analysis and symbolization of spoken sounds; with some account of the principal phonetical alphabets hitherto proposed.	10,237
1852	Rawlinson, Sir Henry Creswicke	Outline of the History of Assyria, as collected from the inscriptions discovered by Austin Henry Layard, Esq. In the Ruins of Nineveh. Printed from the Journal of the Royal Asiatic Society.	11,139
1854	Baker, Anne Elizabeth	Glossary of Northamptonshire Words and phrases, with examples of their colloquial use, and illustrations from various authors to which are added, the customs of the county.	10,069
1867	Whitney, William Wight	Language and the Study of Language: Twelve Lectures on the Principles of Linguistic Science.	10,196
1870	Steere, Edward	A Handbook of the Swahili Language as Spoken at Zanzibar.	10,066
1871	Earle, John	The Philology of the English Tongue.	10,639

Table 1 (continuation)

Date	Author	Title	Words
1879	Findlater, Andrew	Language (Chambers's Elementary Science Manuals)	10,812
1880	Bain, Alexander	Higher English Grammar.	10,109
1886	Bell, Alexander Melville	Essays and postscripts on elocution.	10,108
1891	Dickson White, Andrew	New Chapters in the Warfare of Science, XI. From Babel to Comparative Philology	10,230
1892	Sweet, Henry	A Short Historical English Grammar.	10,135
		TOTAL	**406,023**

Table 1 (continuation)

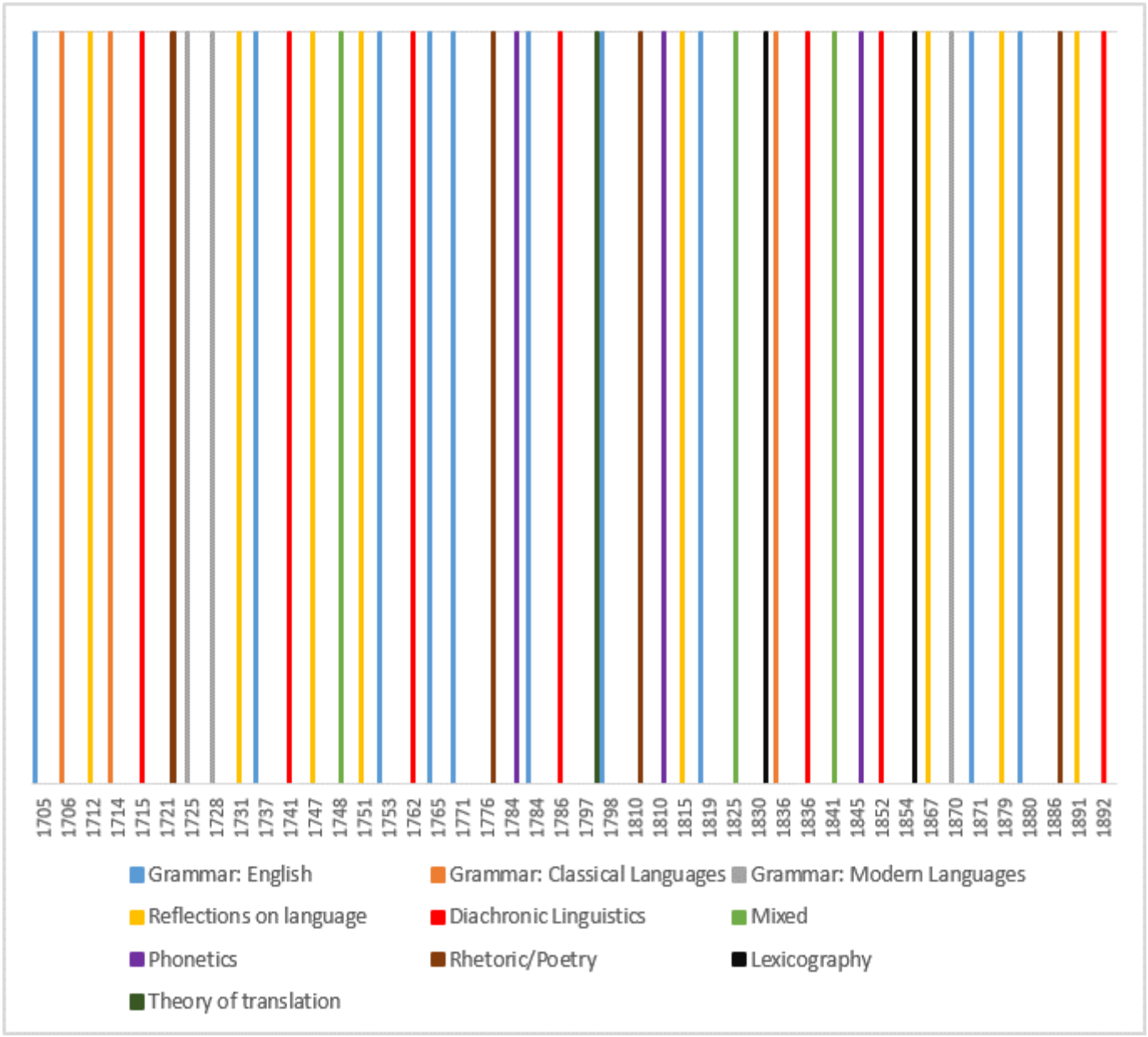

Figure 2: Distribution of topics in the samples of CETeL over time

The distribution of topics in the samples, as shown in Figure 2 above, allows us to see the evolution of the discipline over time. For instance, English grammars, shown in light blue, are found throughout the whole period, whereas grammars of classical languages (orange) are concentrated in the early 1700s, with only another example in the 1830s. As mentioned in Section 4 above, some of those grammars also contain diachronic explanations, and they are labelled as 'mixed topic' (light green). There is also a notable number of grammars of modern languages (grey) which also appear in both centuries.

Other subjects, such as works on Phonetics (purple), Rhetoric (brown) and Lexicography (black), emerge somewhat later in the period. The first of these, of which we have three samples, only begins in the 1780s, just a decade later than works on Rhetoric (1770s, as the first sample in the 'rhetoric and poetry' group, in the 1720s, deals with Poetry). The two lexicographical works – a dictionary and a phrasebook of localisms – are both from the nineteenth century. A work dealing with Theory of translation (dark green) appears at the very end of the eighteenth century.

Red bars represent works on what is nowadays considered Diachronic Linguistics. As can be seen in Figure 2, this was a topic which received attention throughout the period, although the treatment of the topic changed considerably from the early eighteenth century, with an interest in pureness against corruption, to the end of the nineteenth, with efforts towards reconstruction using the comparative method. On the other hand, yellow bars represent what has provisionally been labelled 'reflections on Language', and once more these are present throughout the whole period. These are texts which deal with concepts of Language and Linguistics, and thus samples are drawn from texts that discuss the correction of English (or, rather, denounce its corruption), as well as from theoretical works on the nature of language, its origins, or the philosophical matter behind them, which appear to be the first works on the discipline of Linguistics as we understand it today.

Regarding the genres of the samples, a preliminary classification (cf. Figure 3, below) shows that the most frequently represented genre – twelve examples in total – is that of textbooks, followed by treatises and essays, this reflecting the didactic and, up to a point, philosophical nature of the works written on Language during the period. Likewise, there is a relatively high number of lectures (five), several of which correspond to speeches written to be delivered at meetings of the various societies

established for the study of Language in both the eighteenth and the nineteenth centuries. The number of letters (four) reflects the well-known use of this genre in scientific discourse during the period, particularly in the eighteenth century, and indeed the subcorpus keeps with historical use here in that the last included letter dates from 1819.

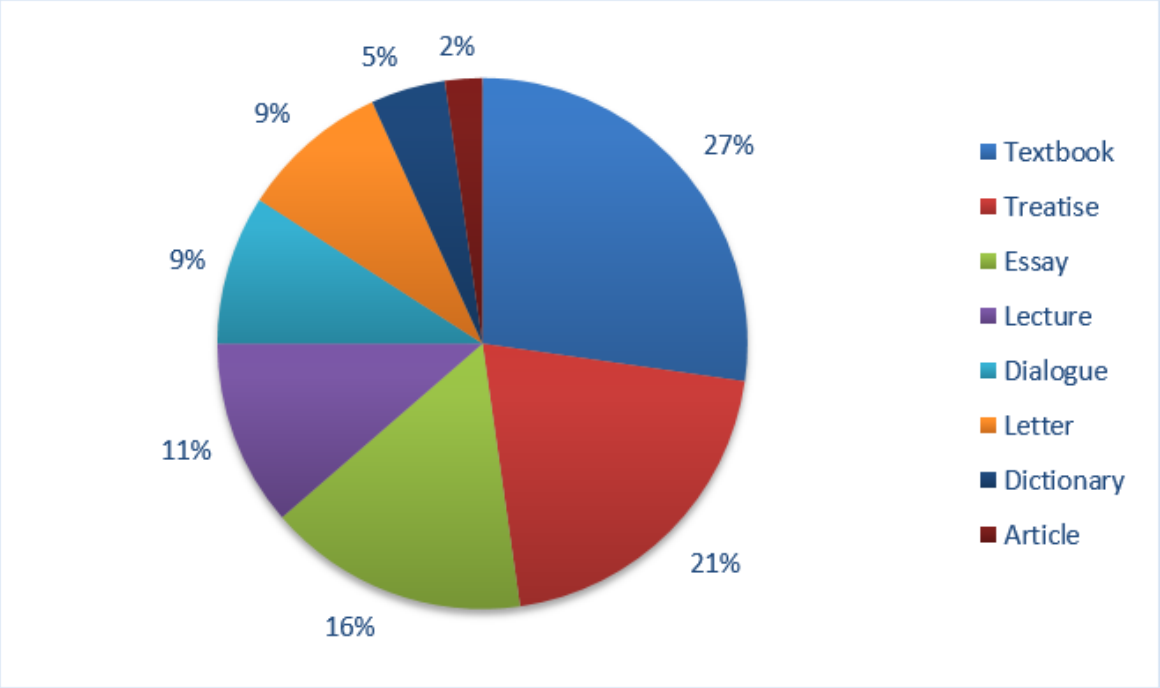

Figure 3: Genres in the samples in CETeL

Among the least frequent genre are two dictionaries, a very specific genre which appears to be relatively frequent in the discipline compared to others, as there are only three other examples of dictionaries (for Astronomy, History, and Chemistry) in the rest of the *Coruña Corpus* to date. Finally, there is only one article, dating from 1891, this reflecting the comparatively late emergence of this genre in the discipline.

Figure 3 also shows four dialogues, which merit special attention. These are different from other dialogues included in other subcorpora of the *Coruña Corpus* (cf. CETA and CEPhiT), in that rather than presenting a conversation between two or more characters, they comprise series of questions and answers, similar to catechisms, albeit of a non-religious kind. All four dialogues follow this format, which raises the question as to whether they should be considered dialogues or, rather, it might be necessary to create a new category for this putative genre. However, since they contain similar

structures to other dialogues in the *Coruña Corpus*, and there are no contrasting samples (either dialogues with this format in other disciplines, or dialogues with any other format), it was decided to classify them as part of the category 'dialogue'.

Regarding the sex of the authors, Figure 4 shows that only four of the 44 samples included in CETeL were written by women. This represents 9.09% of the total samples, which seems representative of the discipline in the period under study. This represents a higher proportion than in other subcorpora (cf. CETA 4.76%, CEPhiT 7.5%, CECheT 7.31%), but a far lower proportion than in CELiST and CHET, both with 20% of female-authored samples. These percentages are in keeping with the aim of representativeness, both in the whole corpus and in each discipline, for Life Sciences and History were among the disciplines which were most open to female practitioners.

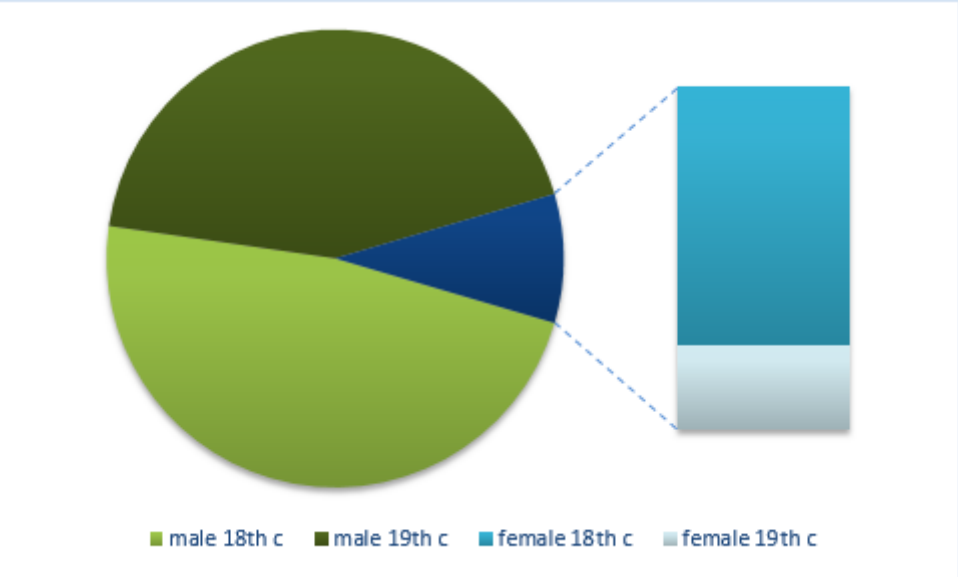

Figure 4: Samples per sex in CETeL

Finally, in terms of the geographical origin and linguistic background of authors, Figure 5 below shows that most samples were written by English authors, followed by Scottish, North American and Irish ones. The four samples marked 'other' include authors for whom little or no information has been found, or who were educated in more than one place, making it very challenging for compilers to ascertain where they might have acquired their linguistic habits.

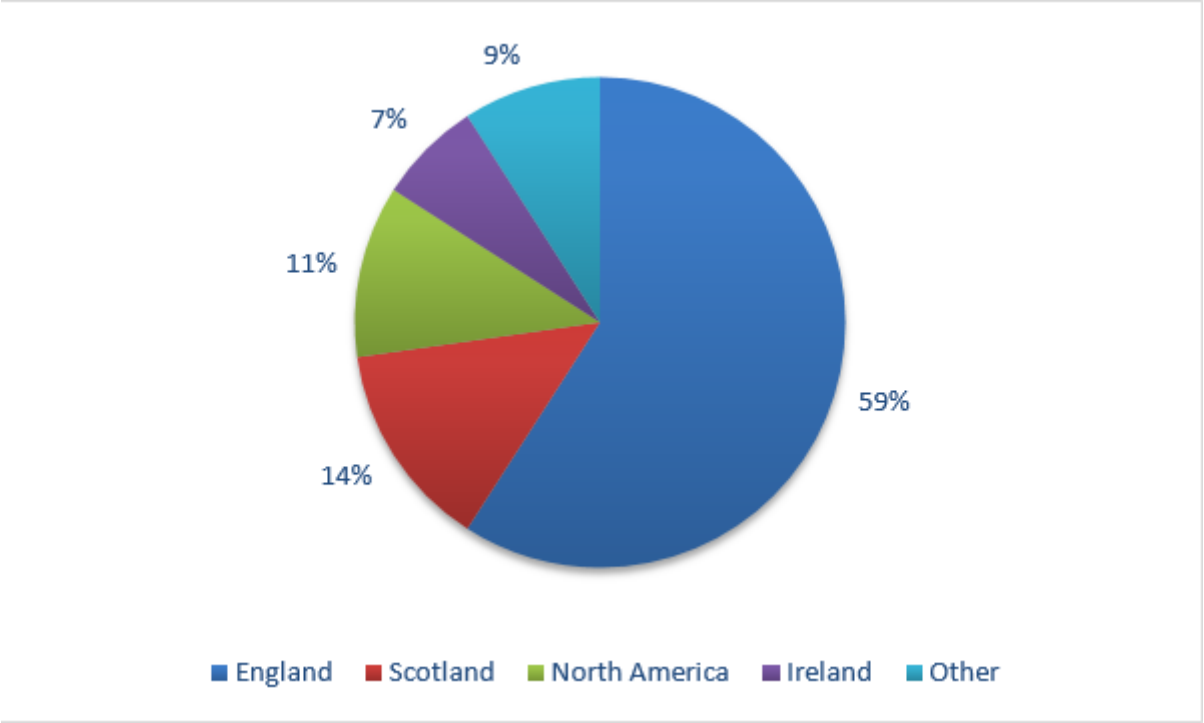

Figure 5: Geographical origin per sample in CETeL

7. Concluding remarks

This paper has presented a new subcorpus of the *Coruña Corpus*, namely CETeL, focusing on its main characteristics regarding the timespan, topics, and genres of its text samples, and the sex and linguistic background of the authors. It has also pointed out the main drawbacks and challenges faced in the process of compilation. Once published, CETeL, the first corpus of its kind on Language and Linguistics during the eighteenth and nineteenth centuries, is expected to be a reliable source of linguistic data for research on the evolution of the English linguistic subregister throughout the Late Modern English period, as well as a valuable illustration of historical scientific writing. With the process of computerisation being now complete, a process of revision is underway, in which each of the samples will be manually revised three times by different reviewers in order to guarantee the most faithful representation of the original. CETeL is scheduled to be completed over the 2020–2022 period, although final beta versions for testing will be made available sooner.

REFERENCES

Allen, Bryce, Jian Qin and Frederik Wilfrid Lancaster. 1994. Persuasive communities: A longitudinal analysis of references in the philosophical transactions of the Royal Society, 1665–1990. *Social Studies of Science* 24/2: 279–310.

Atkinson, Dwight. 1996. The philosophical transactions of the Royal Society of London, 1675–1975: A sociohistorical discourse analysis. *Language in Society* 25/3: 333–371.

Bailey, Richard W. 1985. The conquests of English. In Sidney Greenbaum ed. *The English Language Today*. Oxford: Pergamon Institute of English, 9–19.

Beal, Joan. 2004. *English in Modern Times*. London: Arnold.

Beal, Joan. 2008. Shamed by your English? The market value of a 'good' pronunciation. In Joan Beal, Carmela Nocera and Massimo Sturiale eds. *Perspectives on Prescriptivism*. Bern: Peter Lang, 21–40.

Beal, Joan. 2012. Late Modern English in its historical context. In Isabel Moskowich and Begoña Crespo eds. *Astronomy 'Playne and Simple.' The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins, 1–14.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8: 243–257.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style.* Cambridge: Cambridge University Press.

Boyle, Robert. 1661 (1965). Proemial essay. In Thomas Birch ed. *The Works of Robert Boyle*. Vol. I. Hildesheim: Georg Olms, 192–204.

Burke, Peter. 2000. *Historia Social del Conocimiento: De Gutemberg a Diderot.* Vol. I Barcelona: Paidos Ibérica.

Camiña, Gonzalo and Inés Lareo. 2019. Editorial policy in CHET. In Isabel Moskowich, Estafanía Sánchez-Barreiro, Inés Lareo and Paula Lojo-Sandino comps eds. *Corpus of History English Texts (CHET)*. A Coruña: Repositorio Universidade da Coruña. https://ruc.udc.es/dspace/handle/2183/21849 (29 September, 2019)

Campbell, Lyle. 2001. The history of linguistics. In Mark Aronoff and Janie Rees-Miller eds. *The Handbook of Linguistics*. Oxford: Blackwell, 81–104.

Claridge, Claudia, Josef Schmied and Rainer Siemund. 1999. The Lampeter Corpus of Early Modern English tracts. In Knut Hofland, Anne Lindebjerg and Jørn Thunestvedt eds. *ICAME Collection of English Language Corpora* (CD-ROM). Norway: The HIT Centre, University of Bergen.

Crespo, Begoña. 2004. The scientific register in the history of English: A corpus-based study. *Studia Neophilologica* 76/2: 125–139.

De la Cruz Cabanillas, Isabel. 2001. Lexicografía y semántica del inglés moderno. In Isabel de la Cruz Cabanillas and Francisco Javier Martín Arista eds. *Lingüística Histórica Inglesa*. Barcelona: Ariel, 699–727.

Di Cesare, Donatella. 1990. The philosophical and anthropological place of Wilhelm von Humboldt's linguistic typology: Linguistic comparison as a means to compare the different processes of human thought. In Tullio De Mauro and Lia Formigari eds. *Leibniz, Humboldt, and the Origins of Comparativism*. Amsterdam: John Benjamins, 157–179.

Gotti, Maurizio. 1996. *Robert Boyle and the Language of Science*. Milano: Guerini Scientifica.

Gotti, Maurizio. 2001. The experimental essay in Early Modern English. *European Journal of English Studies* 5/2: 221–239.

Gotti, Maurizio. 2003. *Specialized Discourse: Linguistic Features and Changing Conventions*. Bern: Peter Lang.
Gotti, Maurizio. 2005. *Investigating Specialized Discourse*. Bern: Peter Lang.
Gray, Bethany. 2011. *Exploring Academic Writing through Corpus Linguistics: When Discipline Tells only Part of the Story*. Flagstaff, AZ: Northern Arizona University (Unpublished PhD dissertation).
Hickey, Raymond. 2010. Attitudes and concerns in eighteenth-century English. In Raymond Hickey ed. *Eighteenth-Century English*. Cambridge: Cambridge University Press, 1–19.
Kytö, Merja, Juhani Rudanko and Erik Smitterberg. 2000. Building a bridge between the present and the past: A corpus of 19th-century English. *ICAME Journal* 24: 85–97.
Millward, Celia M. and Mary Hayes. 2012. *A Biography of the English Language.* Boston: Wadsworth, Cengage Learning.
Moskowich, Isabel. 2012. CETA as a tool for the study of modern astronomy in English. In Isabel Moskowich and Begoña Crespo eds. *Astronomy 'Playne and Simple.' The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins, 35–56.
Moskowich, Isabel and Begoña Crespo eds. 2012. *Astronomy 'Playne and Simple.' The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins.
Moskowich, Isabel, Gonzalo Camiña-Rioboo, Inés Lareo and Begoña Crespo eds. 2016. *The Conditioned and the Unconditioned: Late Modern English Texts on Philosophy*. Amsterdam: John Benjamins.
Moskowich, Isabel, Begoña Crespo, Luis Puente-Castelo and Leida Maria Monaco eds. 2019. *Writing History in Late Modern English: Explorations of the Coruña Corpus.* Amsterdam: John Benjamins.
Robins, Robert H. 1978. The Neogrammarians and their nineteenth-century predecessors. *Transactions of the Philological Society* 76/1: 1–16.
Robins, Robert H. 1997. *A Short History of Linguistics*. London: Routledge.
Schmidt, Siegfried. 1975. German philosophy of language in the late 19th century. In Herman Parret ed. *History of Linguistic Thought and Contemporary Linguistics*. Berlin: de Gruyter, 658–684.
Taavitsainen, Irma and Päivi Pahta. 1998. Vernacularisation of medical writing in English: A corpus-based study of scholasticism. *Early Science and Medicine* 3/2: 157–185.

Corresponding author
Leida Maria Monaco
University of Oviedo
Department of English, French and German
Calle Amparo Pedregal, 5
33011 Oviedo
Spain
e-mail: lmonaco@uniovi.es

received: August 2018
accepted: October 2019