Genre and change in the Corpus of History English Texts[1]

Isabel Moskowich, University of A Coruña

Abstract
This paper provides an overview of the Corpus of History English Texts, one of the component parts of the Coruña Corpus of English Scientific Writing (Moskowich and Crespo 2012), looking in particular at the communicative formats that it contains. Among the defining characteristics of the Coruña Corpus are that it is diachronic in nature, and that it can be considered either as a single- or multi-genre corpus, according to the theoretical tenets adopted (Kytö 2010; McEnery and Hardie 2013). The corpus has been designed as a tool for the study of language change in English scientific writing in general, and more specifically in the different scientific disciplines which have been sampled in each subcorpus. All the texts compiled were published between 1700 and 1900, thus offering a thorough view of late Modern English scientific discourse, a period often neglected in English historical studies (De Smet 2005). The analysis of this variety of English is also useful as a means of achieving a clear and detailed description of the origins of English as "the language of science".

Keywords: Coruña Corpus, genre, late Modern English, scientific discourse

1. Introduction

This paper offers a description of the *Corpus of History English Texts* (henceforth *CHET*), focusing mainly on the external factors of the compiled texts, such as sex, age and geographical provenance of authors, and genre/text-type. The paper is divided into four main sections, the first of which will present the history of the Coruña Corpus (henceforth *CC*), the core project within which *CHET* is found. This section will briefly describe some of the compilation principles adopted for the selection of samples for the *CC*, as well as a basic sketch of technical issues involved. Section two will focus on the description of *CHET* itself, paying special attention to those extra-linguistic factors which are peculiar to it, each one dealt with in its own subsection. Section three, in

[1] The research here reported on has been funded by the Spanish Ministerio de Economía, Industria y Competitividad (MINECO), grant number FFI2016-75599-P. This grant is hereby gratefully acknowledged.

Moskowich, Isabel. 2017. "Genre and change in the Corpus of History English Texts." *Nordic Journal of English Studies* 16(3):84-106.

turn, will explore one of these factors—that of genre or text-type—in greater detail, with the concepts of genre, text-type and textual category revisited and reconsidered in light of data gathered during the compiling of CHET and its sister subcorpora. Finally, section four will offer some closing remarks.

2. The Coruña Corpus and its family history

The *CC* project was initiated in 2003 with the intention of facilitating linguistic research into eighteenth- and nineteenth-century scientific texts at all levels. The novelty it offers is the possibility of using these texts for socio-historical as well as linguistic research, this achieved through the inclusion of metadata files containing personal details about the authors of each sample (age, sex, place of education) and about the works (date of publication, genre/text-type) from which the samples have been extracted (Crespo and Moskowich 2010; Moskowich 2012). This applies to all the subcorpora of the *CC* (Pahta and Taavitsainen 2010), both those already published, such as *CETA* (A *Corpus of English Texts on Astronomy*, Moskowich and Crespo 2012) and *CEPhiT* (A *Corpus of English Philosophy Texts*, Moskowich, Camiña, Lareo and Crespo 2016) and those currently under compilation, including *CECHeT* (*Corpus of English Chemistry Texts*) and *CHET* (*Corpus of History English Texts*). It is the latter subcorpus, *CHET*, which I will discuss here, in that its structure derives from the principles and parameters on which the whole compilation process of the *CC* has been based.

The historical period runs from 1700 to 1900, a timeframe motivated by the socio-historical context of scientific writing. It covers the rise of the scientific method (bringing about changes in discursive patterns) which coincided with the founding of the Royal Society and the beginning of the Restoration period. In a similar vein, many important events occurred in the final years of the nineteenth century, with the discovery of the electron, the publication of the Theory of Special Relativity by Einstein, and new calls for a renewal of scientific writing. Indeed, both at the beginning of the eighteenth century and the end of the nineteenth claims were made about the urgent need for a specialised language for the communication of science. These factors seem to be good indicators of a general change in society, science and the language

of science, and thus the period between 1700 and 1900 appears to be an appropriate timeframe for the project.

Among the characteristics of corpora, representativeness and balance are always mentioned. However, they are not always compatible. If we want to preserve balance, we must have the same number of words by men and women but this would not be representative of late Modern English scientific writing. This dilemma has come to us as compilers very often during the process. In terms of general compilation, two samples per decade of approximately 10,000 words each were extracted from original works, these extracts taken from different parts of the works, thus avoiding the repetition of the same rhetorical patterns typically found in introductions, commentaries on results, or conclusions. Likewise, in order to achieve an accurate representation of the author's own language, first editions were always used when available, and where this was not possible editions published within 30 years of the initial one were used (Kytö, Rudanko and Smitterberg 2000: 92). In order to ensure the representation of each author's particular linguistic habits, we included neither quotations by other authors nor translated texts, since in both cases these might lead to linguistic interference from the source language of the borrowed or translated text. To render the process of analysis for final users of the CC less cumbersome, tables, formulae, figures and graphs from the original texts have been eliminated, although their place in the original text is conveniently signalled in the electronic version.

All the subcorpora have been designed to share this general structure, organisation and mark-up, based both on intra-linguistic factors, as I have already noted, and extra-linguistic ones, such as the time delimitations used for compilation (Moskowich, 2016; Moskowich & Crespo, 2016).

From a technical point of view, all the texts have been keyed in following the Text Encoding Initiative (TEI 2) conventions and saved in XML format. Although some editorial decisions had to be made, due to the peculiarities found in some samples, the use of an extended mark-up language has made wide distribution and exploitation possible. We also decided to create a corpus management tool in order to retrieve both linguistic and non-linguistic information from the compiled data. Thus, the *Coruña Corpus* Tool (CCT) is an Information Retrieval system in which the indexed textual repository is a set of compiled documents that

constitutes the *CC* (Lareo 2009). Figure 1 below shows the interface of the CCT for metadata searches.

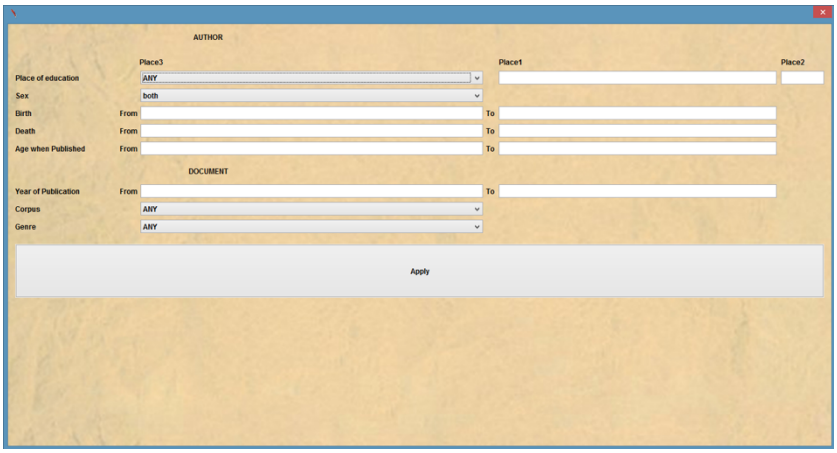


Figure 1. CCT interface for metadata

Searches by metadata can be made because information about external variables has been included in the corresponding files.

Other subcorpora in the *CC* have been described elsewhere, so in what follows I will consider the social variables that characterise historical texts in *CHET*.

3. CHET: discipline and external factors

CHET, as I have noted above, is the subcorpus of the *CC* containing texts pertaining to the realm of history, especially if we adopt an inclusive perspective (as the *CC* in general does), that is, taking into account the fact that fields of knowledge during the Modern Age were not as well-defined and discrete as they are today.

Over the years and centuries, different perspectives on History as a discipline have been seen. Thus, during the eighteenth century the author David Hume (himself included in the *CC*) defined History as "a collection of facts which are multiplying without end; and if they are to be made intelligible, they must, in some way, be abridged". Hume considered that History as a subject of study was justified due to its value as an instrument of education (1778: 116 in Black 1926). Likewise,

contemporary scholars such as Voltaire made clear that they saw history and historiography as a record of human activity in all its manifestations, and Gibbon (whose 1778 work is included in *CHET*) claimed that History was an organised sequence of cause and effect (Black 1926). However, other rationalist conceptions of human nature were proposed, ones which were intended to constitute the basis for an explanation of human action. Among these we can find Adam Ferguson's *An Essay on the History of Civil Society* (1767), John Millar's *The Origin of the Distinction of Ranks* (1771), and Adam Smith's *Wealth of Nations* (1776).

Following Stromberg (1951) and Okie (1991), Strangeman (2007) claims that the beginnings of historicism and history writing can be found in the Age of Reason, although other scholars in the twentieth century (Black 1926) pointed out the possibility that these historians perhaps dealt with documents in an amateurish and somewhat casual way, and as a consequence might have reached perverse conclusions. For example, Black stated that History did not have any standard nomenclature during the Enlightenment; rather, he argued, it was written using a jargon which varied from writer to writer, and was full of implicit assumptions. However, this is not the case; as early as the last quarter of the eighteenth century, Giambattista Vico published *New Science* (1782), a work that gave historians a fully-fledged theory of History, including proper methods of arriving at the truth (Breisach 1983). Current scholars consider Edward Gibbon equally influential, in that *The history of the decline and fall of the Roman Empire* (1788) was a methodological milestone for later historiographers. The importance of Gibbon's work (sampled in *CHET*) lies in the author's use of historical sources to organise and structure historical facts, thus arguing against previously accepted accounts of history.

The nineteenth-century rationalistic mode of thought accelerated the use of a scientific methodology based on working with existing documents. Throughout the nineteenth century, historiography completed its process of professionalisation in Western Europe and the United States, including the creation of academic chairs, degree-granting programmes, disciplinary associations and specialist journals (Ranke, 1982; Porter and Ross, 2003). Nineteenth-century scholars applied the scientific method previously described by John Locke (1690) as the "plain historical method" (Stromberg 1951), and contemporary authors

such as Humboldt (1822) corroborated such an approach when expressing his belief that History should in fact be exact, impartial and critical. This was precisely the origin of the present-day assumption that History is based on a collection of true and verified facts (Black 1926; Stromberg 1951). Indeed, more broadly, it was during the nineteenth century that historiography as a whole took its modern form (Olby, Cantor, Christie and Hodge 1996) and the difference between History (as the facts occurred in the past and somehow recorded) and Historiography (as the methods and techniques used to describe those recorded past events) appeared. Both terms are however often used interchangeably up to this day.

This century was also the period of biographies par excellence (Barnes 1962; Olby, Cantor, Christie and Hodge 1996). According to Barnes (1962), this was due to the individual now being seen as more glamorous, with biography readily adapted to such literary flights. As a matter of fact, towards the end of the previous century, Cornish (also included in *CHET*) defined biography in the preface to his 1780 work, contrasting it to other historical writings:

> Biography is a species of history which gives a writer some peculiar advantages, who would teach men to be good by examples. The historian must attend principally to great events, which affect Mankind only at large. But the biographer may enter into the walks of private life, and exhibit characters interesting to us as individuals (p. ii).

The evolution of both the discipline itself and its writing patterns can be seen in successive samples in *CHET*. In addition to being influenced by the idea of History itself, the extracts can also be seen in terms of external factors such as sex, age, geographical provenance of the author, plus the communicative format that he or she chooses to use.

The samples are of ca. 10,000 words each, as is the case in the *CC* as a whole, with a similar number of samples and words for both centuries, as set out in table 1 below:

Table 1. Words per century in the subcorpus under study

Century	Words
18th c.	201,938 words
19th c.	202,486 words
Total	404,424 words

When selecting the texts to be sampled a compound system was used as random sampling was preferred but certain canonical authors were also included. Although text selection is often determined by availability, extra-linguistic factors affecting this choice are also central to the metadata file accompanying each sample in all disciplines of the corpus. Figures 2 and 3 below illustrate the metadata file as seen in the CCT. All metadata files contain information about the author (sex, age, geographical provenance among others) and the text (date of publication and communicative format/genre), and here I set out the information relating to the author in Figure 2, and that pertaining to the text in Figure 3.

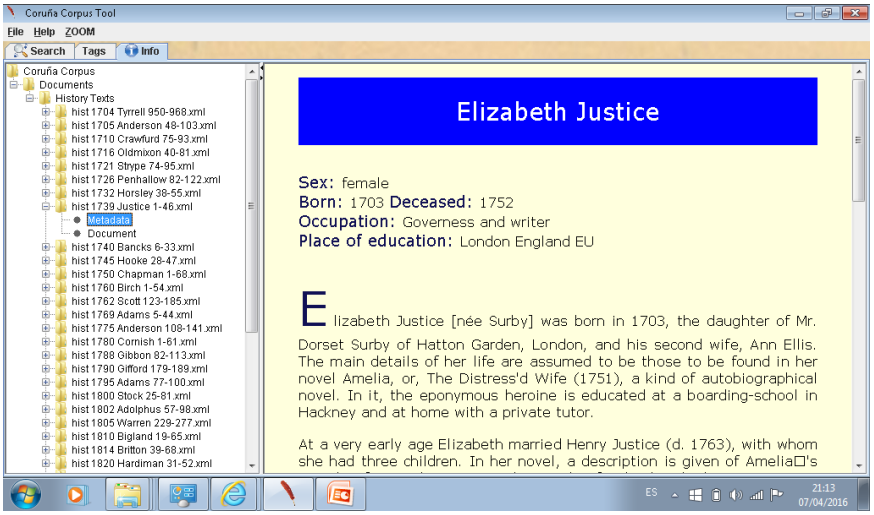


Figure 2. Metadata file: author.

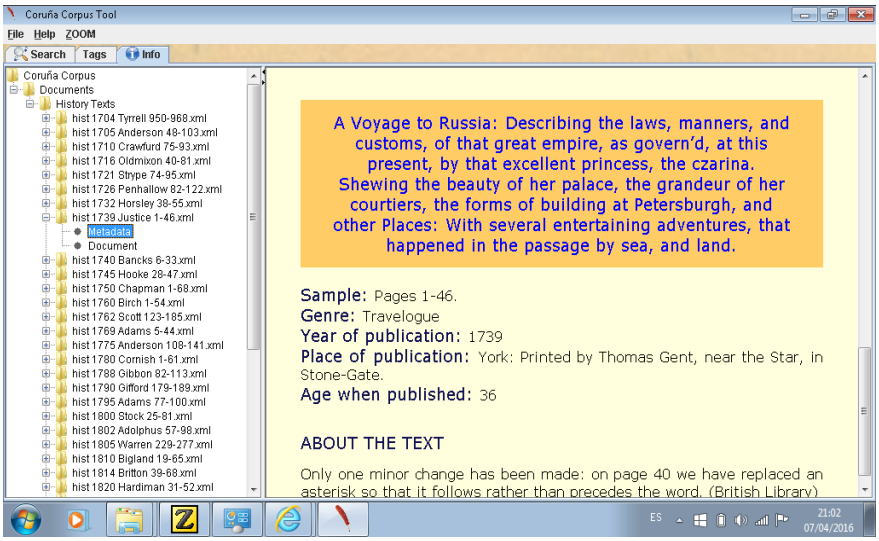


Figure 3. Metadata file: text.

Authors in *CHET* represent both sexes and include those educated on either side of the Atlantic. Indeed, in some cases their writing habits were acquired on both sides, as with Samuel Penhallow, who was born in Cornwall, studied in Middlesex and went to live in Massachussets at the age of twenty. Ages range from 26 years, in the case of Alice Cooke, to 78 years old for John Strype. All these author-specific factors will be dealt with in the following subsections, as well as that relating to text.

3.1. Sex

The *CC* attempts to reflect the real situation of scientific writing during the late Modern English period, and in this sense *CHET* conforms to this aim. Following the compilation of the text extracts, I noted that female authors are few in number, as was also the case in other disciplines. Besides the difficulty in accessing certain texts, this may be also due to the fact that women often worked in the shadows, as has often been observed (Crespo, 2016a; Moskowich, 2016). In fact, *CHET* contains eight samples written by women from a total of forty. However, women are even less well represented in the sister corpus *CETA* (*Corpus of English Texts on Astronomy*), with just two female authors, and also in the *CEPHIT* (*Corpus of English Philosophy Texts*) with three. The

different number of women found in the various subcorpora can perhaps be explained in terms of social factors, and also the kind of discipline in question. In the case of *CETA*, for instance, we should bear in mind that it was seen as inappropriate for women to observe the sky at night (Herrero 2007; Moskowich 2012). Similarly, women were not regarded as the ideal authors on topics concerning human understanding, politics or morals (Puente-Castelo and Monaco 2013, Crespo 2015, 2016a), the subjects typically dealt with in philosophical texts. On the contrary, writing about travel, or textbooks for schoolchildren that reproduced accepted historical accounts, were not seen as improper for ladies, and thus female authors are relatively well represented in this section of the *CC*.

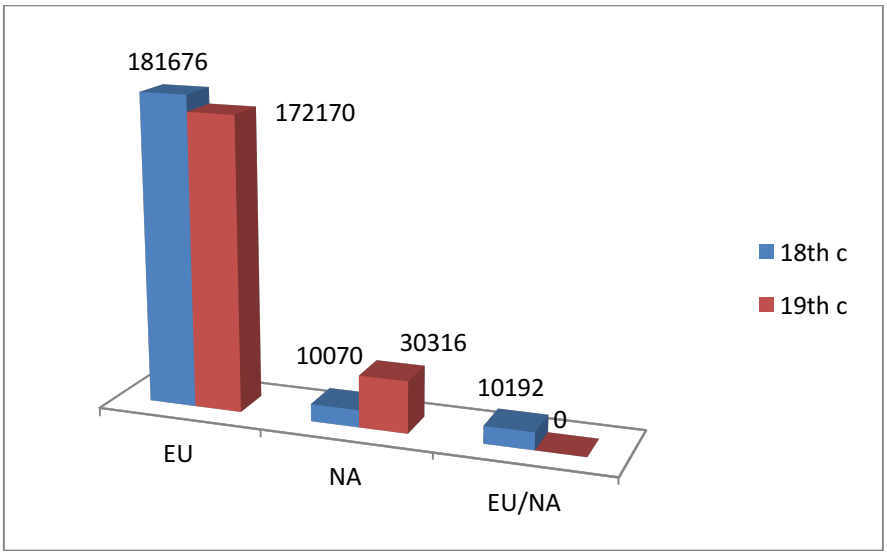
Similar social reasons may explain why of the eight female authors in *CHET*, only two (Sarah Scott and Elizabeth Justice) published their work in the eighteenth century, whereas history itself, specifically the history of the birth of the United States, may account for the presence of Mercy Otis Warren as the only American female author in this subcorpus. The issue of geography, however, will be dealt with in more detail in what follows.

3.2. Geographical distribution

The metadata files in the *CC* and hence in *CHET* include details of a maximum of three geographical places where an author acquired his or her scientific writing habits, that is, the places of education rather than where they initially learned to speak. The three possibilities included in these metadata files range from the very general labels of "North America" (NA) or "Europe" (EU) to a particular territory (England, Scotland, Canada, among others) or a specific place (Cambridge, Edinburgh, Cork, etc.) where authors were educated. Place of birth has not been considered, since in the analysis of scientific writing the place of education is a great deal more relevant than where someone was born.

Graph 1 below illustrates the geographical distribution for the samples in *CHET* according to whether authors were educated in North America, Europe or both. As can be seen, samples were mostly produced by authors educated in Europe, in both the eighteenth and nineteenth centuries, although the end of the American War seemed to lead to more

authors educated in the Americas writing about history between 1800 and 1900.



Graph 1. American vs. European authors in CHET

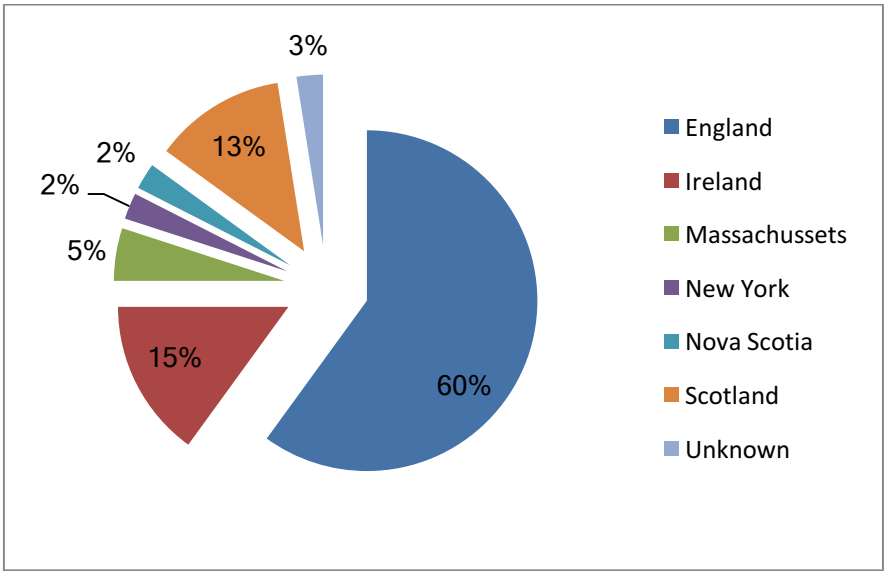
As regards the label "Place 2" in the metadata, that is, the territory where an author acquired his/her academic writing habits, we find that most of the European authors were educated in England, followed by Ireland and Scotland. The four North American authors in *CHET* were all educated in the Eastern states, as might be expected. To these, we could perhaps add Penhallow, who studied both in England (Cornwall and London) and Massachusetts (Middlesex).

Having graduated from Harvard as a priest, Amos Adams on one occasion moved his audience to some kind of revolt during the General Fast. It is precisely this lecture in 1770 we have sampled in *CHET*. The Canadian author John Hamilton Gray (1814–1889) was educated in King's College (Nova Scotia) and became a jurist and a politician. His professional background is reflected in the work *Confederation; or, The Political and Parliamentary History of Canada, from the Conference at Quebec, in October, 1864, to the Admission of British Columbia, in July, 1871*, an excerpt from Volume One of which figures in *CHET*. Sidney Breese (1800–1878) was also a jurist, as well as Chief Justice of the

Illinois Supreme Court, and a U.S. Senator for Illinois. He came to occupy these positions thanks to a formal education received at Hamilton and Union Colleges. On the contrary, Mercy Otis Warren (1728–1814) received no formal schooling but was allowed to attend the lessons received by her brothers at home.

As regards their training, authors educated in North America seem to follow the same pattern as those from Europe. This implies that men received formal education and were often either priests or lawyers, whereas most female authors did not receive a systematic training but learnt somewhat casually.

Graph 2 below sets out information about the provenance of authors in more detail. As can be observed, American authors seem to concentrate on the Eastern Coast whereas those from Europe are slightly more scattered. This may be due not only to a longer cultural tradition of writing in Europe but also to the socio-historical events in America during the period, where the population tended to concentrate in the Eastern states, with the West still being explored and colonised.

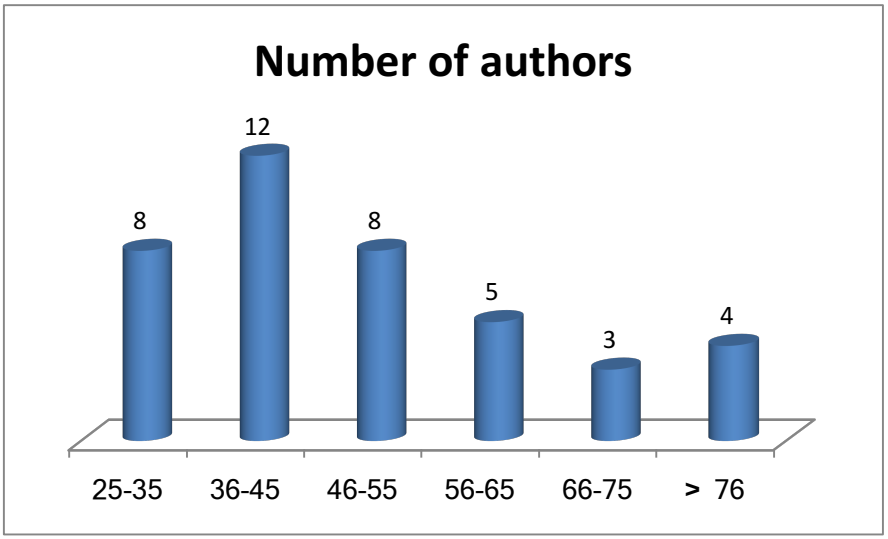


Graph 2. Detailed distribution of authors' place of education

The third external variable, age of authors, is discussed in the following subsection.

3.3. Age

Age is generally regarded as a significant independent variable, indeed a very notable one, in the study of language change (Kerswill 1996) and language variation (Wagner (2012), and for this reason it would be desirable to have corpus samples by authors from a wide range of ages. However, the Coruña Corpus contains extracts pertaining to the academic register, that is, texts that require a prior education and training to be written. This, in turn, imposes age limitations as the authors need to take a time to get that training. This may account for the distribution of authors according to their ages. For this description I have grouped authors in ten-year gaps. As a result, the age-group predominating in the samples under survey is that between 36 and 45 (with samples by 12 authors).



Graph 3. Age of authors in CHET

Authors in the age groups prior to and immediately following the predominant range (8 authors in both cases) are also well represented, as can be seen in Graph 3 above; the remaining three age groups, that is, authors older than 56, are relatively consistent. Age on its own, however, is not enough to conduct any complete form of sociolinguistic or discursive analyses. Variables such as sex or geographical provenance of authors are often taken as obvious complements in sociolinguistic

studies. Moreover, the age variable can be combined with others, such as genre/text-type, leading to useful insights. Information describing the genre/text-type variable in *CHET* will be discussed in section 3.4, below.

3.4. Genre, text-type or others

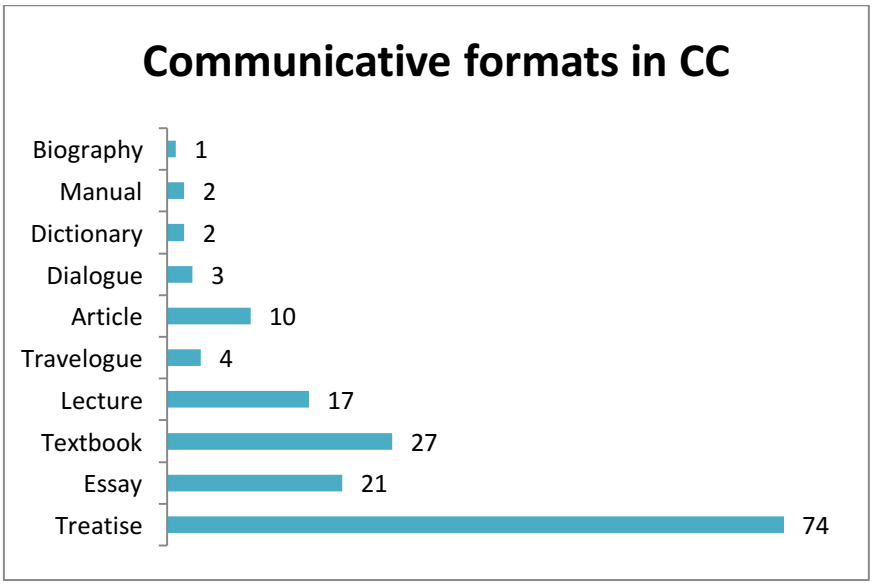
Previous studies have noted a kind of terminological chaos when dealing with notions of genre, register, text-type and textual category (Lee 2001). Genre has been seen to refer to function and external criteria (Biber 1988; Lee 2001; Crespo 2016b) or to communicative purposes (Swales 1990; Martin 2000), whereas text-type has been more closely related to form (Biber and Finegan 1989; Lee 2001; Alonso Almeida 2008). Textual category, in turn, is a more neutral term often used to refer to a more general or even perhaps unclear characterisation of texts.

Given that a clear dependency between form and function seems to exist in texts, the term "communicative format" has been preferred here to encapsulate the symbiosis between form and function which is intrinsic to any text. It is undoubtedly the case 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. That in turn makes the receiver an important element within the communication process. However, depending on the kind of response the sender/addresser 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 divided.

The *CC* contains many different communicative formats[2] adapted to the social and functional needs of a particular period and discipline. During the compilation process we have seen that certain disciplines appear to be more clearly associated with specific formats, almost as if they were inherent to the discipline itself. This description will, hopefully, shed some light on the tendency to use suitable communicative formats in late modern scientific writing according to disciplinary idiosyncrasies. Graph 5 below shows the distribution of the

[2] This can be considered a provisional list of categories, since some of the subcorpora in CC are still beta versions under revision.

different subcorpora of the *CC* compiled thus far (*CETA, CEPhiT, CHET* and *CECHeT*[3]) in terms of communicative formats:



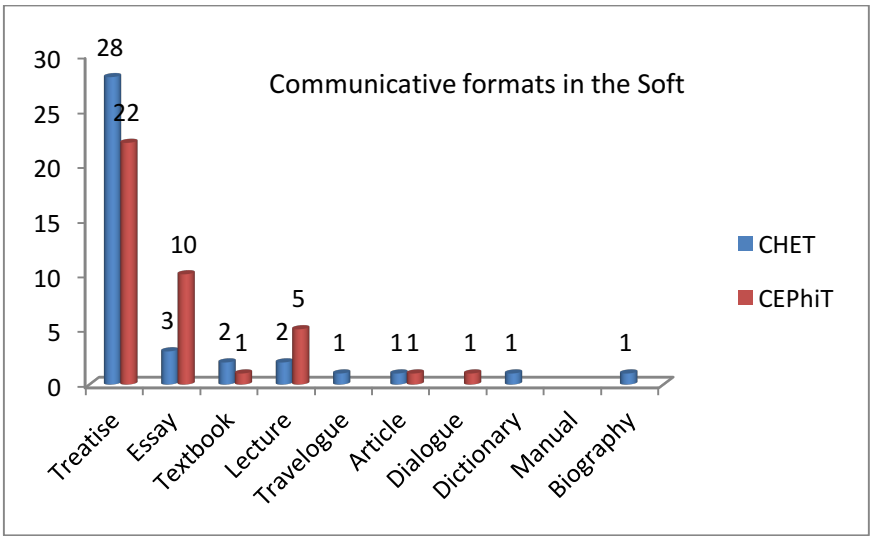
Graph 5. Communicative formats in the subcorpora of the CC

As can be seen, all 161 samples compiled in the four subcorpora currently forming the *CC* can be classified into ten different communicative formats[4]: Letter, Manual, Dictionary, Dialogue, Article, Travelogue, Lecture, Textbook, Essay, and Treatise. As for frequency of use, the format Treatise is recorded in 74 samples, that is, in 45.96% of the samples. Textbook is the second most common format, used in 27 of the samples compiled (16.77%), followed by Essay (21 samples; 13.04%), Lecture (17; 10.55%) and Article (10; 6.21%). This illustrates broad tendencies in the use of communicative formats within late Modern English scientific discourse (Moskowich and Crespo 2016).

[3] *CECHET*, *Corpus of English Chemistry Texts*.

[4] As has been done for the other subcorpora samples have been assigned to particular genres or communicative formats by taking into consideration not only the author's self-labelling but also the adequacy of the actual characterisitics of the text to the ones expected (see Moskowich, 2012)

On the lines of previous research, and in order to go a step further in this description, I will classify the four disciplines in the *CC* into two different subgroups: the so-called soft sciences (philosophy and history) and the hard sciences (astronomy and chemistry), as seen in Graphs 6 and 7.



Graph 6. Communicative formats in the Soft Sciences in the CC

In the case of the soft sciences, and following the general tendency, Treatise is the most common format across the two disciplines, with 50 samples. The term "treatise" refers to

> A book or writing which treats of some particular subject; commonly (in mod. use always), one containing a formal or methodical discussion or exposition of the principles of the subject; formerly more widely used for a literary work in general". However, there is a more general meaning, now obsolete: "A descriptive treatment, description, account (of something).

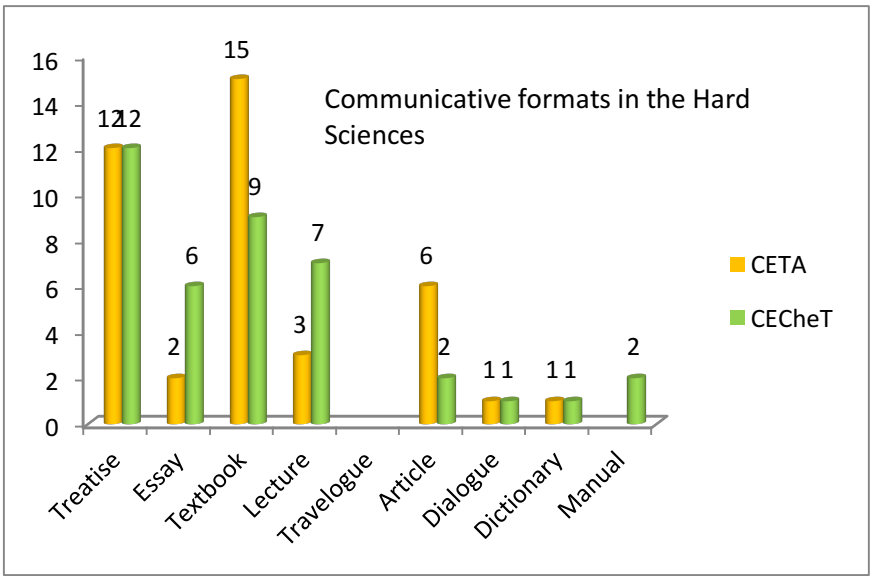
This is one of the senses provided by the *Oxford English Dictionary*, and has also been used by later authors to classify English text-types (Görlach 1994).

Both philosophy and history are theoretical or descriptive fields that constitute a good fit for this format. Neither is a procedural discipline in which an applied goal is sought. Besides, given the period under survey,

some of the authors in the *CC* may have had this very sense of the term in mind when naming and describing their works. Such is the case with Olmsted, one of the authors included in *CETA* (1841: vii), who considers that in a treatise "the deepest research is united with that clearness of exposition which constitutes the chief ornament of a work intended for elementary instruction".

Essay is defined in the *OED* as "A composition of moderate length on any particular subject, or branch of a subject; originally implying want of finish, 'an irregular undigested piece' (Johnson), but now said of a composition more or less elaborate in style, though limited in range. The use in this sense is app. taken from Montaigne, whose *Essais* were first published in 1580", and is the second most common format in the soft sciences. Nevertheless, there are only 13 samples using it, 10 in philosophy texts and 3 in history. Whereas Essay can perhaps be considered a philosophy-specific format in the period under survey here, in history writing there are other typical formats, such as Travelogue and (biographical) Dictionary. Coincidentally, although discipline-specific, both Travelogue and Dictionary are examples of underrepresented formats. Equally significant is the underrepresentation of Article, Dialogue and Textbook, as well as the total absence of Manuals, in that this may also indicate some kind of disfavouring of less obviously appropriate formats for the expression of particular content. Consequently, either the presence or absence of particular communicative formats might be useful in determining the kind of constraints underlying format selection.

As for the hard sciences, different selection preferences have been found. Graph 7 below illustrates the distribution of formats in astronomy and chemistry texts.



Graph 7. Communicative formats in the Hard Sciences in the CC

Textbook and Treatise are the two most frequently used formats within the group of the hard sciences, with 24 samples each. Whereas the graph shows the same number of treatises in both astronomy and chemistry, there seems to be some kind of preference for Textbook in the case of samples from *CETA* (15 instances). Only 9 have been recorded for *CECHeT*.

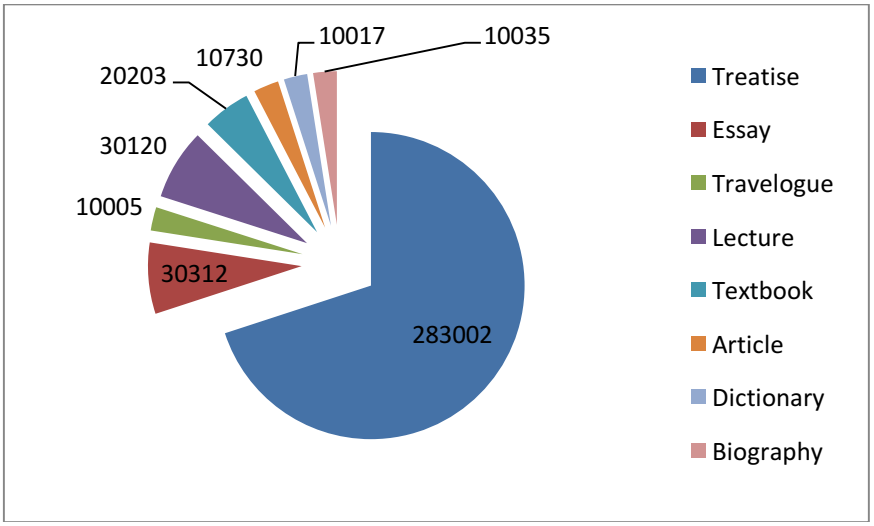
The *OED*, from which the following definition was taken, dates the first use of the term "textbook" to 1779: "A book used as a standard work for the study of a particular subject; now usually one written specially for this purpose; a manual of instruction in any science or branch of study, esp. a work recognized as an authority".

The frequent use of the Textbook format (plus a couple of Manuals) within the hard sciences may reflect a response to the growing social demand for knowledge which characterised post-empiricist times and the practical/applied nature of those fields. Likewise, a manual is defined as "A handbook or textbook, esp. a small or compendious one; a concise treatise, an abridgement. Also in extended use" (*OED*).

The absent formats in the hard sciences (Travelogue and Letter) differ from those for the soft sciences, as might be expected. This

reinforces the idea that there is a clear dependency between discipline (that is, content), function (which is audience-related) and format.

As regards the particular case of *CHET*, I noted above that the information which history texts typically provide seems to be conveyed mainly through a format which narrates previous facts or past events as a timeline or sequence; it evinces the voice of a distant third person narrator who seeks only to present straightforward facts through expository writing. In fact, in *CHET* we find a predominance of treatises (with 28 samples, 283,002 words) as well as some formats which are completely absent in other disciplines (such as Biography and Travelogue). The existence of formats peculiar to certain disciplines may indicate that the symbiosis between form and function I argued for may indeed be observed here. Graph 8 below illustrates how samples are distributed across different communicative formats in *CHET* according to number of words.



Graph 8. Communicative formats in CHET.

A clear example of the symbiosis between the form and the function of texts can be seen in the Travelogue format. The knowledge and communicative practices shared by travellers are precisely the elements which turn travelogues into an efficient communicative format in historical writing (Moskowich and Crespo 2016), whose expository

nature (describing various kinds of travel events) is different from that of treatises. In the same way that Travelogue and Biography seem to be typical and exclusive of *CHET*, no samples of Manual or Dialogue are found in the history corpus. Therefore, it seems reasonable to conclude that the presence or absence of certain formats in particular disciplines can be considered a determinant factor in the characterization of those scientific disciplines.

4. Final remarks

This description of *CHET* from the perspective of the different variables characterising the samples, together with the results obtained from previous studies of other CC subcorpora, seem to reveal that some of these variables are constrained by subject matter. Such is the case with the sex of the author and with format selection. In this paper I have proceeded from the general to the particular, looking first at the *CC* as a whole, then narrowing down to the two main sets of fields represented (hard sciences and soft sciences) and finally focusing on *CHET*. Through this we have seen that communicative formats are potentially discipline-dependent in late modern scientific writing, perhaps more so than nowadays. The information communicated in a text necessarily demands a particular format and this seems to explain their presence or absence in specific subcorpora. Similarly, particular disciplines or subject matter may also imply constraints on the sex of the author due to external factors, these being mainly social and cultural. Therefore, both variables pertaining to the text (communicative format) and variables pertaining to the author of the text (sex) seem to be related to subject matter during the late Modern English period, although only a comparison with similar corpora for present-day English would reveal whether this tendency has persisted or changed.

References

Alonso Almeida, Francisco. 2008. "The Middle English medical charm: Register, genre and text type variables." *Neuphilologische Mitteilungen* 109/1: 9-38.

Moskowich, Isabel. 2012. "CETA as a tool for the study of modern astronomy in English". *Astronomy 'playne and simple': The Writing*

of Science between 1700 and 1900. Eds. Isabel Moskowich *et al*. Amsterdam: John Benjamins. 35-56.

Moskowich, Isabel. 2016. "Lexical richness in modern women writers: Evidence from the Coruña Corpus of History English Texts." *Revista Canaria de Estudios Ingleses* 72: 111-128. http://ruc.udc.es/dspace/bitstream/handle/2183/16922/Moskowich__2016_Lexical_richness_Modern_Women_Writers_CHET.pdf?sequence=2&isAllowed=y.

Moskowich, Isabel, Gonzalo Camiña, Begoña Crespo and Inés Lareo. 2016. *The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy*. Amsterdam/Philadelphia: John Benjamins.

Moskowich, Isabel and Begoña Crespo (Eds.). 2012. *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900*. Amsterdam/Philadelphia: John Benjamins.

Moskowich, Isabel and Begoña Crespo. 2016. "Classifying communicative formats in *CHET*, *CECHeT* and others." *EPiC Series in Language and Linguistics* 1: 308-320. http://easychair.org/publications/download/Categories_and_Genres_in_CHET_and_CECHeT.

Moskowich, Isabel *et al*. 2012. *Corpus of English Texts on Astronomy* (*CETA*). Amsterdam: John Benjamins.

Moskowich, Isabel *et al*. (in preparation). *Corpus of English Chemistry Texts* (*CECheT*).

Moskowich, Isabel *et al*. (in preparation). *Corpus of History English Texts* (*CHET*).

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas and Edward Finegan. 1989. "Drift and evolution of English style: A history of three genres." *Languages* 65: 487-517.

Black, John Bennett. 1926. *The Art of History: A Study of Four Great Historians of the Eighteenth Century*. London: Methuen and Co. Ltd.

Brodie, Benjamin. 1880. *Ideal Chemistry. A Lecture*. London: Longmans, Green and Co.

Crafts, James. 1870. *A Short Course in Qualitative Analysis, with the New Notation*. New York: John Wiley and son. https://archive.org/details/idealchemistryle00broduoft.

Crespo, Begoña. 2015. "De como la mujer participó en el desarrollo científico del mundo anglosajón." *Cuadernos del CEMYR* 23: 105-119.

Crespo, Begoña. 2016a. "On writing science in the Age of Reason." *Revista Canaria de Estudios Ingleses* 72: 53-78. https://riull.ull.es/xmlui/bitstream/handle/915/4504/RCEI_72_%282016%29_03.pdf?sequence=1&isAllowed=y.

Crespo, Begoña. 2016b. "Genre characterization in *CEPhiT*." *'The Conditioned and the Unconditioned'. Late Modern English texts on Philosophy*. Eds. Isabel Moskowich *et al*. Amsterdam: John Benjamins. 25-44.

Crespo, Begoña and Isabel Moskowich. 2010. "*CETA* in the context of the Coruña Corpus." *Literary and Linguistic Computing* 25/2: 153-164. https://academic.oup.com/dsh/article-abstract/25/2/153/942823/CETA-in-the-Context-of-the-Coruna-Corpus?redirectedFrom=fulltext.

De Smet, Hendrik. 2005. "A Corpus of late Modern English Texts". *ICAME* 29: 69-82.

Görlach, Manfred. 2004. *Text Types and the History of English*. Berlin: Walter de Gruyter.

Herrero, Concepción. 2007. "Las mujeres en la investigación científica". *Criterios* 8: 73–96

Kallel, Amel. 2002. "The age variable in the rise of periphrastic 'do' in English". *Reading Working Papers in Linguistics* 6: 161-185. http://www.reading.ac.uk/internal/appling/wp6/kallel.pdf.

Kerswill, Paul. 1996. "Children, adolescents and language change." *Language Variation and Change* 8: 177-202

Kytö, Merja. 2010. "Data in historical pragmatics." *Historical Pragmatics*. Eds. Andreas Jucker and Irma Taavitsainen. Berlin: Mouton de Gruyter. 33-67.

Lareo, Inés. 2009. "El Coruña Corpus. Proceso de compilación y utilidades del *Corpus of English Texts on Astronomy* (*CETA*). Resultados preliminares sobre el uso de predicados complejos en CETA". *A Survey on Corpus-based Research Panorama de investigaciones basadas en corpus*. Eds. Pascual Cantos and Aquilino Sánchez. Murcia: Asociación Española de Lingüística de Corpus. 267-280. http://www.um.es/lacell/aelinco/contenido/pdf/19.pdf.

Lee, David. 2001. "Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle." *Language Learning & Technology* 5/3: 37-72. http://www.llt.msu.edu/vol5num3/pdf/lee.pdf.

Martin, J.R. 2000. "Analysing genre: functional parameters." *Genre and Institutions. Social Processes in the Workplace and School*. Eds. Frances Christie and J. Martin. London/New York: Continuum. 3-39.

McEnery, Tony and Andrew Hardie. 2013. "The history of corpus linguistics." *The Oxford Handbook of the History of Linguistics.* Ed. Keith Allan. Oxford: Oxford University Press. 707-726.

Oxford English Dictionary. http://www.oed.com. (2016, February 11).

Packe, Christopher. 1708. *Medela Chymica: or, an Account of the Vertues and Uses of a select Number of Chymical Medicines Adapted to the Cure of the most Chronick and Rebelious Diseases To which is Subjoyned a Brief History of Cures Effected by Them...* London: Printed for John Lawrence at the Angel in the Poultrey.

Pahta, Päivi and Irma Taavitsainen. 2010. "Scientific discourse." *Historical Pragmatics*. Eds. Andreas Jucker and Irma Taavitsainen. Berlin: Mouton de Gruyter. 549-586.

Rissanen, Matti. 1989. "Three problems connected with the use of diachronic corpora." *ICAME* 13: 16-19. http://clu.uni.no/icame/archives/No_13_ICAME_Journal_index.pdf.

Puente-Castelo, Luis and Maria Monaco. 2013. "Conditionals and their functions in women's scientific writing." *Procedia — Social and Behavioral Sciences* 95: 160-169. http://www.sciencedirect.com/science/article/pii/S1877042813041554

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

Sinclair, John. 2004. "Corpus and text — Basic principles." *Developing Linguistic Corpora: A Guide to Good Practice.* AHDS Literature, Languages and Linguistics. Ed. Martin Wynne. University of Oxford. https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm.

Strangeman, Charles. 2007. *Strange Allies? English Catholicism and the Enlightenment*. Phd Dissertation. Southern Illinois University Carbondale. Ann Arbor: University of Michigan

Taavitsainen, Irma. 2012. "Discourse forms and vernacularisation processes in genres of medical writing 1375–1550." *Studies across Disciplines in the Humanities and Social Sciences* 7. Eds. Annelli

Aejmelaeus and Päivi Pahta, Helsinki: Helsinki Collegium for Advanced Studies. 91-112. https://helda.helsinki.fi/bitstream/handle/10138/34748/7_07_Taavitsainen.pdf?sequence=1.

Wagner, Suzanne Evans. 2012. "Age grading in sociolinguistic theory." *Language and Linguistics Compass* 6/6: 371–382. http://onlinelibrary.wiley.com/doi/10.1002/lnc3.343/abstract.