B. Crespo, I. Moskowich

A CORPUS OF HISTORY TEXTS (*CHET*) AS PART OF THE *CORUÑA CORPUS* PROJECT[1]

Abstract. The aim of this paper is to offer a description of the *Corpus of Historical English Texts* (*CHET*), one of the several sub-corpora within the *Coruña Corpus of English Scientific Writing* (*CC*). The compilation principles behind it as well as the sociolinguistic variables considered in the process of text selection will be explained.

Keywords. Scientific English, corpus linguistics, Late Modern period.

1. Introduction

Since every scientific field has its own writing traditions and restrictions, we have decided to compile different sub-corpora forming the *Coruña Corpus of English Scientific Writing* (*CC*). Each of them contains samples of texts published between 1700 and 1900 which correspond to a different scientific discipline. Overlapping of disciplines constitutes a basic difficulty in the selection of representative samples of scientific language, mainly when it is not present-day science we are dealing with. Instead of designing our own taxonomy of disciplines when compiling the *CC*, we resorted to the one published by UNESCO [1988] as a starting point. The first sub-corpus compiled was *CETA, Corpus of English Texts on Astronomy*. The second was *CEPhiT, Corpus of English Philosophy Texts* and the third is the one we are presenting here, *CHET, Corpus of Historical English Texts*.

[1] The research which is here reported on has been funded by the Spanish Ministerio de Economía y Competitividad (grant number FFI2013-42215-P) within its programme "Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia". This grant is hereby gratefully acknowledged.

2. Compiling *CHET*[1]: principles and parameters

The compilation principles applied to *CHET* are those applied to the whole *CC*. We have tried to compile two 10,000 words text files per decade, so that each of the centuries represented contains approximately 200,000 running words. Some pilot studies with our corpus have shown that 1,000-word samples are not really enough for the study of variation within the scientific register [Biber 1993] mainly because the scientific register was not as standardised at that time as it is nowadays. This corpus shares the structure and mark-up conventions used for the whole project which have proved to be extremely useful and valid for research since the sampling methods avoid authorial idiosyncrasies and any sort of interference caused by translation.

We have also born in mind the principles of representativeness and balance [McEnery and Wilson 1996; Biber et al. 1998: 251–253] most specialists in corpus linguistics care about. In addition, it was our conscious decision to include only edited and printed texts in prose. As with the other sub-corpora, first editions have been used whenever possible and this addresses mainly the issue of availability. Otherwise, and taking for granted that language change can be observed within 30-year periods (Kytö et al. 2000: 92], texts published within a thirty-year span from the first publication date were selected.

In order to have a complete representation of stylistic and pragmatic devices, we have collected extracts from different parts of the works sampled so that introductions, central chapters and conclusions are more or less equally represented. Similarly, prefaces or dedications which are not scientific in themselves have been excluded. The final word counts are those shown in Table 1.

[1] Since *CHET* has not been released yet, the data we present here correspond to those of a beta version.

Table 1. Words in *CHET*

Eighteenth century	201,794
Nineteenth century	202,823
Total	404,617

Some non-linguistic factors such as age, sex, place of education and genre/text-type of each author and text which may prove useful for sociolinguistic purposes are part of the information in the metadata files accompanying text files [Crespo and Moskowich 2010].

A more detailed explanation of the general principles of compilation applied to the *CC* can be consulted in [Moskowich and Crespo 2007; Moskowich and Parapar 2008; and Crespo and Moskowich 2010].

3. Time-span represented

CHET, like the rest of the subcorpora, has been compiled to cover the Modern English period. Considerations outside the purely linguistic discipline, but of a more historical nature, determine the period covered by the *Coruña Corpus* and, therefore, by *CHET*. History has demonstrated that changes in scientific thought bring about changes in scientific discourse [Moskowich 2011]. Therefore, we have used landmarks in scientific thought rather than those in language change to set the time limits of our selection.

The time-span chosen begins with the outburst of the scientific revolution, the foundation of the Royal Society and with the publication of the basic guidelines on how to present scientific works to its members with the ideas of clarity and simplicity behind it all. *CHET* earliest texts date back to 1704 (James Tyrrell) and 1705 [James Anderson), a moment at which medieval scholastic patterns undergo a radical transformation [Taavitsainen and Pahta 1997] and, therefore, the best moment to start our compilation.

At the other end of the time-line, several events which were really important for the History of Science occurred around 1900, the last year covered by *CHET* [Alice Cooke, 1893 and Montagu Burrows, 1895]. Some of these events were the discovery of the electron by J.J. Thompson in 1896, the crisis of the grounds of mechanical physics announced in this same year, Planck's proposal of quantum mechanics, or Einstein's publication of the Special Theory of Relativity in 1905 [Moskowich and Crespo 2010; Moskowich 2011]. All these discoveries, as in the seventeenth century, were also accompanied by the need to change the discursive patterns of science announced by Thomas Huxley at the 1897 International Congress of Mathematics.

In the sections that follow, we will enumerate and explain all the extra-linguistic variables that play a part in the corpus.

4. Genres/Text types

Either text type (the internal characteristics of texts) or genre (as a way of socialising and, therefore, with certain external functions) [García-Izquierdo and Montalt 2002] can cause variation within academic writing.

The classification we have used in the CC is based on Görlach [2004]. All the categories proposed by this author were already used during the Modern Period.

Görlach [2004: 88] claimed that proper definitions of each genre are necessary prior to text collection so as to ensure that corpora contain representative samples of the material under analysis. He mentions eight categories, proposing the following definitions:

Table 2. Görlach's classification of text-types

Article	Non-fictional composition or dissertation in a newspaper, journal or read at a conference
Essay	Short prose composition, first draft
Lecture	Formal discourse delivered to students. Piece of writing intended to be read aloud

Treatise	Discussion of a topic including some methodological issues
Dialogue	Literary work in conversational form
Textbook	Book used as a standard reference work
Letter	Written communication (not necessarily sent by post)
Encyclopaedia	Book containing information in all branches of knowledge, arranged alphabetically

This classification has been used as a starting point in our compilation of scientific texts, but as mentioned above, other parameters have been also taken into account such as the explicit mention of the author indicating the genre the work belongs to.

Table 3 below represents the number of samples compiled belonging to each genre:

Table 3. Genres in History Texts

Genres in *CHET*	**Samples**
Treatise	28
Essay	3
Textbook	2
Lecture	2
Others	4
Article	1

The ascription of texts to genres may be arguable [Fowler 1982], but we have examined very carefully both the whole texts from which samples had been extracted and their prefaces thus bearing in mind how the author himself would classify her/his text according to contemporary standards. This allowed us to conclude that *CHET* contains samples of the six genres/text-types in the table above. In turn, this may be due to restrictions imposed by subject-matter: certain disciplines or domains seem to prefer just a few types of texts whereas others manifest themselves in a more varied way [Moskowich 2011].

Modern authors writing about History seem to prefer Treatise by large followed by the general category of Other in which there are three biographies and one travelogue. Essays come next with three samples, which points to a real liking for formal genres. Other categories are also represented through the instructive in the shape of Lecture and Textbook.

Fig. 1 displays the different genres gathered in the whole history corpus samples where 70% corresponds to Treatise.

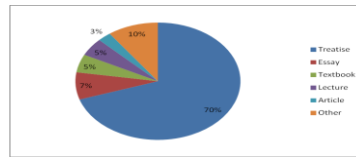

Fig. 1. Distribution of words per genre

However, such distribution is not identical in the two centuries compiled. The graphs below show these differences reflecting the external reality which influenced text production in the field.

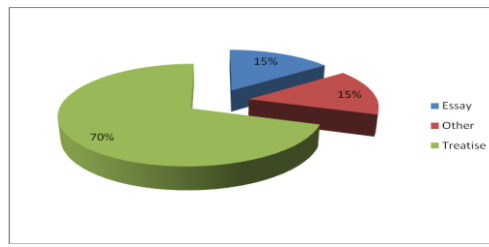

Fig. 2. Words per genre in 18th c. History texts

Both figs. 2 and 3 illustrate a wider variety of genres used in the nineteenth century as compared with those used by authors in the preceding century. The fact that History was considered to deserve dissemination at different social and cultural levels may have caused this.

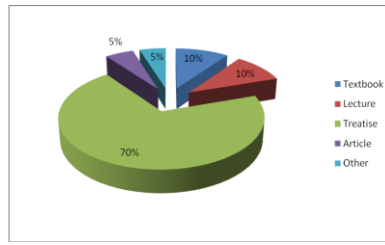

Fig. 3. Proportion of words per genre in 19th c. History texts

The information contained in *CHET* metadata files suggests that History opens itself to a larger readership from 1800 and does so by resorting to a wider range of genres as was the case with Philosophy.

5. Sex

CHET contains only two samples of eighteenth-century female writing. These women are Catherine Justice (1700) and Sarah Scott (1783). A higher number of women writing history have been collected for the nineteenth-century section of the corpus: Mercy Otis Warren (1805), Mary Calcott (1828), Lucy Aikin (1833), Elizabeth Sewell (1857), Martha Freer (1860) and Alice Cooke (1893).

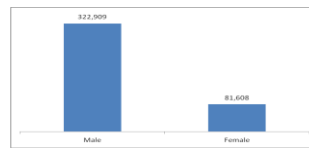

Fig. 4. Words written by male and female authors

CHET reflects this scarcity of overt female activity which is, nonetheless, higher than in other sub-corpora [Moskowich, 2011; 2012].

6. Authors' provenance in *CHET*

It has been already mentioned that the corpus is valid not only for the diachronic study of English scientific writing but also for that of variation depending on other variables such as geographical origin. This is why we have included texts by authors whose linguistic habits could be traced.

In compliance with the *CC* principles, we have selected English-speaking authors writing in English, avoiding any sort of translation. When referring to «geographical distribution of authors» we are not considering the places where they were born but, instead, those where they received formal education, and where they acquired the linguistic habits to be found in their writings. An overview of the different places (either Europe or North America) where the authors contained in *CHET* learned to write is offered in fig. 5.

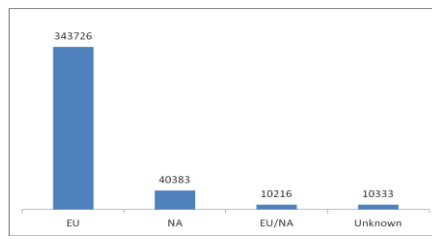

Fig. 5. The provenance of authors in *CHET*

A few American authors have been included in this subcorpus though they abound in other parts of *CC*. It was Europe that was producing most works on history, whereas North America had lived a convulsive eighteenth century and was, in the nineteenth, more worried about the practical application of scientific advances and about the forge of its own history than about the narration of past facts. In this sense, *CHET* is a small-scale mirror of reality.

7. Final Remarks

Like in the case of the previous releases [*CETA*, 2012; *CE-PhiT*, in press] the intention behind the creation of the *Corpus of Historical English Texts* (*CHET*) has been to allow the scholarly community to conduct research into the historical underpinnings of English for Specific Purposes. This interest was reinforced by a gradual increase in the number of studies on genre conventions and special languages from the final decade of the 20th century. In line with the principles established by corpora experts we have attempted to adhere to those of balance, representativeness, stratified sampling methods and delimitation of period covered under the guidance of extra-linguistic facts. Pilot studies with astronomy and philosophy texts have proved to be useful to describe the characteristics of academic writing and disciplinary conventions. We hope this new sub-corpus will be a step forward in pursuing this goal.

References

Biber D. (1988), Variation across speech and writing. Cambridge, UK: Cambridge University Press.

Biber D. (1993), Representativeness in corpus design. Literary and Linguistic Computing 8(4), pp. 243–57.

Biber D. (1995), Dimensions of register variation: A cross-linguistic comparison. Cambridge, UK: Cambridge University Press.

Biber D., *Conrad S.*, *Reppen R.* (1998), Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Crespo B., *Moskowich I.* (2010), *CETA* in the context of the *Coruña Corpus*. Literary and Linguistic Computing 25(2), pp. 153–64.

Fowler A. (1982), Kinds of literature. An introduction to the theory of genres and modes. Oxford: Clarendon Press.

Görlach M. (2004), Text types and the history of English. Berlin/New York: Mouton de Gruyter.

Kytö M., Rudanko J., Smitterberg E. (2000), Building a bridge between the present and the past: A corpus of 19-century English. ICAME 24, pp. 85–97.

McEnery T., Wilson A. (1996), Corpus linguistics. Edinburgh: Edinburgh University Press.

Monzó Nebot E. (2002), La profesió del traductor juridic i jurat: Descripció sociòlogica del professional i anàlisi discursiva del trasgenere. Unpublished PhD Dissertation. Universitat Jaume I.

Moskowich I., Crespo B. (2007), Presenting the *Coruña Corpus*: A collection of samples for the historical study of English scientific writing. In *Pérez Guerra et al. (eds.),* 'Of varying language and opposing creed': New insights into Late Modern English, pp. 341–357. Bern: Peter Lang.

Moskowich I., Parapar López J., (2008), Writing science, compiling science. *The Coruña Corpus of English Scientific Writing.* In *Lorenzo Modia, MJ (ed.),* Proceedings from the 31st AEDEAN Conference, pp. 531–544. A Coruña: Universidade da Coruña.

Moskowich I. (2012), *CETA* as a tool for the study of modern astronomy in English. In *Moskowich I., Crespo B (eds.)* Astronomy 'playne and simple'. The writing of science between 1700 and 1900. Amsterdam/Philadelphia: John Benjamins, pp.35–56.

Taavitsainen I., Pahta P. (1997), The Corpus of Early English Medical Writing. ICAME Journal 21, pp. 71–78.

UNESCO. (1988), Proposed international standard nomenclature for fields of science and technology. UNESCO/ROU257 rev. 1. Paris.

Crespo Begoña
Isabel Moskowich
University of A Coruña
E-mail: bcrespo@udc.es