CETA in the Context of the *Coruña Corpus*

Begoña Crespo García and Isabel Moskowich-Spiegel Fandiño
Department de Filoloxía Inglesa, Universidade da Coruña, A Coruña, Spain

Abstract

The *Coruña Corpus* (CC): a Collection of Samples for the Historical Study of English Scientific Writing is a project on which the MUSTE group has been working since 2003 in the University of A Coruña (Spain). It has been designed as a tool for the study of language change in English scientific writing in general as well as within the different scientific disciplines (excluding medicine) between 1650 and 1900. Its purpose is to facilitate investigation at all linguistic levels, although, in principle, phonology is not included among our intended research topics. At the same time, we believe that the CC is an excellent tool for the study of scientific register/style at particular moments in history: it also offers the researcher the chance to analyse how this 'specific English' behaves from a synchronic point of view. To allow for socio-linguistic research using these scientific texts, we have included, when possible, some personal details about the author of each sample and, even, about the work from which the sample has been extracted in a separate file. From a technical point of view, all the texts have been keyed in following the Text Encoding Initiative conventions and saved in the XML format. The use of an extended mark-up language will make wide distribution and exploitation possible. Moreover, in order to retrieve information from the compiled data, we have decided to create a *corpus* management tool. Loosely speaking, the *Coruña Corpus* Tool is an Information Retrieval system, where the indexed textual repository is a set of compiled documents that constitutes the CC.

Correspondence:
Begoña Crespo García, Department de Filoloxía Inglesa, Universidade da Coruña, A Coruña, Spain.
E-mail:
bcrespo@udc.es

1 Introduction: The *Coruña Corpus* Project

MUSTE,[1] the group of historical linguists working in the English Department at the University of A Coruña (Spain) has traditionally studied Middle English and later texts from a morpho-syntactic, lexical, and semantic point of view in order to ascertain how change and variation had affected the evolution of the English language. Methodologically speaking, the group has resorted to *corpora* of different sorts (*Helsinki, Lampeter, BNC*) to obtain empirical data with which theoretical tenets can be exemplified.

At present, members of MUSTE are engaged in a major project—the compilation of the *Coruña Corpus (CC) of English Scientific Writing*.[2] We are also developing some tools for *corpus* linguistics investigation in collaboration with the Information Retrieval Lab team at the Department of Computer Science in the University of A Coruña.

In the pages that follow, it is our intention to explain the characteristics of our major project, the CC and, more specifically, to depict the salient

Literary and Linguistic Computing, Vol. 25, No. 2, 2010. © The Author 2009. Published by Oxford University Press on behalf of ALLC and ACH. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
doi:10.1093/llc/fqp038 Advance Access published on 19 November 2009

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

features of the first subcorpus to be released, *Corpus of English Texts on Astronomy* (CETA).

The samples in the CC, and, as is obvious, in CETA, have been assembled to meet the precise linguistic objectives set out in Leech (1992).

1.1 Aim and scope

A rough definition of our *corpus* project would say that it contains English scientific texts, other than medical, produced between 1650 and 1900.

Two main ideas have triggered the whole project. First, the one on scientific–thought styles developed in Helsinki by Irma Taavitsainen, Päivi Pahta, and Marti Mäkinen paved the way for the compilation of the CC, which, in a way, is an attempt to complement their compilation and works on medical writing.[3] Second, the *corpus* has been designed to facilitate investigation at all linguistic levels except phonology. Research on different linguistic levels could help outline the general characteristics of scientific English as well as its historical evolution from the very first 'scientific' writings in the vernacular until the end of the nineteenth century. Both synchronic and diachronic perspectives can be adopted for the study of English scientific writing. A diachronic point of view can be used since the CC covers a time-span of two and a half centuries. Another field of research that remains open is that of language variation within scientific texts production since several genres or text-types have been included in our compilation. As we will see later, the CC complies with the basic tenets of *corpus* compilation provided by Meyer (2002) and Crystal (2003).

Little attention was paid to scientific language until the 1990s, mainly because it was not regarded as an object of study in itself but as a vehicle to transmit knowledge to which a lexical, syntactic, and discursive uniformity was attributed. From the end of the twentieth century onwards, the increasing interest in English for specific purposes runs parallel to a similar interest in its historical description, evolution, and peculiarities.

1.2 Compilation principles

Since we understand that a *corpus* must be 'a large and principled collection of natural texts' (Biber *et al.*, 1998, pp. 4, 12), no random selection of texts was made, but it was based on certain external parameters to ensure fruitful linguistic analyses. Hence, our compilation principles can be described in the following terms:

1.2.1 *Discipline inventory*

If one of the aims of this principled collection is to offer the possibility of studying the evolution of scientific English as a mirror of the history of science, knowledge of post-medieval history was essential to decide which criteria should be used to determine the fields existing at the time. This means that the concept of science itself across history has had a direct influence on our textual selection. As Atkins *et al.* (1992, p. 5) claimed:

> the initial selection of texts for inclusion in a corpus will inevitably be based on external evidence primarily . . . A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation.

Two different perspectives have been adopted for the selection of samples since the empiricist movement was a turning point in the treatment of scientific knowledge. For texts produced before 1700 an inclusive perspective will be adopted to embrace those fields that would not be considered science from our modern stance. This might be the case of Alchemy. However, a different point of view has guided our selection of texts after 1700, the task we have taken on in the first place. For this purpose, the current UNESCO Classification of Sciences was used as a starting point (see Appendix I). The first area we have selected is what UNESCO labels 'Exact and Natural Sciences'. We have begun with the collection of the disciplines of Mathematics, Astronomy, Physics (where we include Physics and Geophysics), and Life Sciences (where we include Biology mainly, but also Botanics, Zoology, and others belonging to Field IV, Agricultural sciences). We have excluded some of the branches of human development that have only very recently been considered science (Bugliarello, 2001) as is the case of Field II (Engineering).

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

While writing this paper, Field VI (Humanities) has been tackled: the selection of text samples for Philosophy is over, and we have initiated the selection of texts for History. The subcorpus of Life Sciences has also been collected and we are about to finish the collection of Physics. Our intention is to compile a more or less equivalent number of samples and words for each scientific field in order to facilitate comparative studies on the language used in each discipline, and the evolution of particular features, confirming the wide range of variation within academic prose (Biber, 1988). Using this methodology, we will obtain different subcorpora that can be regarded independent entities with a similar structure, organization and mark-up.

1.2.2 *Representativeness and balance*

Two other aspects have been taken into consideration: representativeness of texts and balance within the *corpus*. For each discipline we have selected two texts per decade, with each sample containing around 10,000 words, excluding tables, figures, formulae, graphs,[4] and all the quotations that are not representative of the author's speech. Shorter texts have been included *in toto*. This decision is based on claim by Kŷto *et al.* (2000, p. 92) that short-term change in diachrony can be safely studied over periods of thirty years. Each category is therefore represented by 500,000 words in each whole subcorpus for ModE (1650–1900).

In the interest of thoroughness, first editions have been preferred; likewise, we have tried not to use more than one text by the same author, even if in different disciplines, in order to avoid the proliferation of idiosyncrasies. We try to ensure balance by providing not only the same number of words per discipline but also the same number of disciplines per field (Exact Sciences and Humanities). We also believe that the representativeness of the CC is improved by not including any translations. Therefore, only English-speaking authors writing in English have been considered, though we are conscious that many of them also read and wrote in Latin, and this may have exerted an influence on their use of the vernacular. Moreover, in order to avoid recurrent rhetorical and linguistic patterns, samples have been extracted from different parts of the original texts.

We are aware that register/style is connected with certain social or extra-linguistic variables that may permit socio-linguistic studies on the *corpus*. To this end, the social background of authors together with some details about their lives and work will be provided in separate metadata files wherever possible (Moskowich-Spiegel Fandiño and Parapar López, forthcoming).

1.2.3 *Time-span*

The fourth criterion applied is the chronological limits of our collection (1650–1900), which are based on some extra-linguistic considerations.

The acceptance of an empirical view in the seventeenth century led to the modification of the corresponding scientific discourse. This marks the beginning of a new way of thinking. There are three main differences between scholasticism and this modern stance: first, seventeenth-century science evolved under the influence of the Royal Society; second, it was not only concerned with types of knowledge and theological matters, but also with the observation of Nature and the application of scientific research; and, third, there was an attempt to reach precise conclusions by using quantifiable data. As Boyle had already suggested in the seventeenth century, this new school of thought called for the creation of an *ad hoc* discourse (Stubbs, 1996, p. 18).

The date 1900 also establishes a turning point in the development of science as some extra-linguistic circumstances evince: the discovery of the electron by J. J. Thompson in 1896, the crisis of the grounds of mechanical physics announced by Mach, Kirchhoff, or Bolzmann in the same year, Planck's proposal of quantum mechanics or Einstein's publication of the Special Theory of Relativity in 1905. Moreover, Thomas Huxley, as Boyle had done a couple of centuries earlier, manifested the need of a new scientific style at the 1897 International Congress of Mathematics.

1.3 Computing devices

First and foremost, to ensure wide distribution and exploitation (cross-platform) an extended mark-up

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

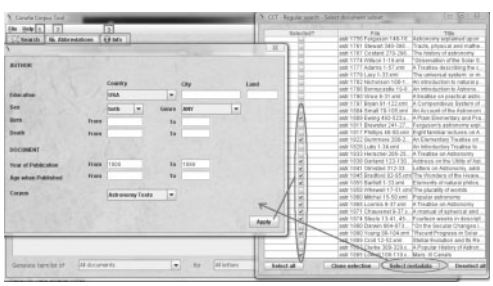
language has been used. Accordingly, samples of texts have been keyed in using e-macs as an XML editor following the Text Encoding Initiative (TEI) conventions. UNICODE standard has also been used for symbols and old characters. Moreover, poems, quotations, editors' additions and numbers or symbols that do not develop a syntactic function on the phrase have been deleted and some other changes have been made: unnecessary blank spaces and correction of obvious spelling mistakes. Likewise, some editorial marks written between square brackets have been added to include information about omitted parts or to disambiguate formulae or numbers that could be indexed as an English word.

The CC incorporates a tool (*Coruña Corpus* Tool or CCT) that has been specially designed to carry out linguistic analyses with the samples contained in the *corpus*. As a matter of fact, the CCT is an Information Retrieval (IR) system that creates an index from the set of compiled documents (the CC or any of its parts), and it is this index that functions as the textual repository on which searches are based.

It works like most concordance programmes but offers some special features adapted to the characteristics of CC (e.g. possibility to search old-fashioned characters, tags in texts or in metadata files).

The CCT includes three principal windows: Search window, Abbreviations window and Info window.

In the Search window, a subset of texts can be selected using metadata parameters. This window also offers the possibility of generating word lists of all the texts or the words selected or to look for a particular linguistic aspect.

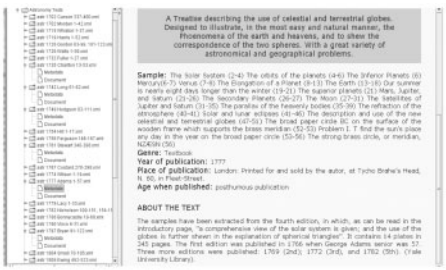


A Results Summary window appears on clicking the bottom row; it shows percentages, tokens, examples, the whole title of the sample, etc.

This Info window offers two possibilities:

(i) Clicking on 'document', a closer version to the original text appears. It means that the deleted parts are included in red.

(ii) Clicking on 'metadata', all the data found about authors and texts are displayed. Also, here are the parameters that can be automatically used for the selection of a subset of texts that fulfil the same extra-linguistic criteria.



2 CETA

CETA paves the way for studies on language variation and language change within a particular genre (Kohnen, 2007) or special language, scientific English, or, even more particularly, astronomical writing, during the so-called Late Modern period. As language and society are inextricably connected, the use of *corpus* linguistics as the methodological pillar for the creation of a source of philological study entails certain knowledge of other disciplines

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

such as History of the English-speaking world, the History of Science (basically Astronomy), and other related areas (Moskowich-Spiegel Fandiño, I. and Crespo García, B., 2007).

2.1 Astronomy as a growing science

Astronomy, or the observation of stars, the heavens, and planetary motions, as part of Nature itself, has always been one of man's prime concerns. Ancient people such as the Chaldeans or the Babylonians devoted themselves to find explanations for the effects of the Sun on the Earth, to observe the stars and to venture upon possible relationships between human beings and celestial bodies, which, on many occasions, were used for predictions and calculations (what is called astrology today).

Later in time, the Greeks tackled the field of stars and their influence upon the Earth with a more systematic approach. In fact, Aristotle's assumption of the geocentric model was challenged by Aristarchus of Samos resorting to logical reasoning. He was the first to point to a heliocentric model of the Cosmos, which, in the fifteenth century, would be assumed as revolutionary by Copernicus' contemporaries.

In the Middle Ages learning was enclosed in the so-called *Trivium* and *Quadrivium* and Astronomy formed part of the latter. Being regarded as a 'specific subject of study', Astronomy was conceived of deduction from well-established principles under the influence of medieval scholasticism; written texts were of a dialogic nature and oral debates developed basically as academic discussions in a university environment.

Writers on Astronomy employed Latin as the conduit of learning and culture and, although it would not cease to be so until the eighteenth century, English began to be used in parallel. From the thirteenth century onwards a timid movement in favour of rational demonstrations of facts can be traced. This movement began to settle during the last quarter of the fourteenth century when scholars started working with the key aspects of modern science: induction, experimentation, and mathematical language.

The first written texts in the vernacular were modelled on the rhetorical strategies on classical texts. Even the linguistic devices and the formats employed by authors were a calque of those found in Latin originals. Latin syntactic constructions as well as vocabulary items were adopted by translators during this period. Two early texts on Astronomy, Chaucer's *A Treatise on the Astrolabe* and *The Equatorie of the Planetis*, testify to this (Banks, 1997;Taavitsainen, 2004). The retention of the classical format was merely a linguistic necessity to cover the vacuum in English technical writing; at the same time, it served socio-political interests in endowing the vernacular with prestige.

In pre-empiricist times a different structural organization of science was conveyed by a non-specific language. In contrast, in the eighteenth century the empiricist movement helped establish different taxonomies of reality. This, in turn, generated several semantic or conceptual categorizations with a linguistic counterpart in various jargons. An instance of this can be found in the distinction between Astronomy and Astrology, though the separation will not be effective until the so-called 'Age of Reason'. Thus, medieval scholastic works, full of references to authoritative statements, were replaced; first, by erudition in the Renaissance and, later, by a new empirical method grounded in direct observation. Great astronomers such as Galileo and Newton worked under these new trends. As could be expected, the method changed and, as a consequence, so did the discursive strategies found in Astronomy texts.

In the late seventeenth and early eighteenth centuries there was also an increase in the readers to include not only rich or professional groups but also the middle class who demanded instruction. Hence, the growing relevance of the use of the vernacular to transmit 'science', even though the style employed by writers was not always the same. In some instances, it was simple and clear, as recommended by the Royal Society; in others, it was more literary, complex, and dense[5]. These two styles coexisted in the construction of scientific discourse during the eighteenth century.

Optiks by Newton and other scientific treatises by the members of the Royal Society were the precursors of the technological innovations of the Industrial Revolution later in the eighteenth century. The use of instruments for astronomical

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

observation (such as the reflecting telescope and the spectrometer) triggered a major advancement of learning in this field. Some of the texts also incorporate for the first time descriptions of the instruments used for observation as well as an account of methodology.

Mathematical logic and the experimental procedures of science definitively merged with society in the nineteenth century. Industrial technology was redesigned by engineers and technicians working in tandem, and reformulated, once again, into new disciplines with specialized vocabularies. As society progressed and became more compartmentalized, the language used by the corresponding speech community also became more specialized.

2.2 Place of CETA in the CC

As mentioned earlier, CETA is the *Corpus of English Texts on Astronomy* containing samples from 42 different authors publishing texts on Astronomy from 1700 to 1900.

One of the main characteristics of the discipline is that it can be classified as both modern and pre-modern: the former refers to the fact that it is part of the 'Natural and Exact Sciences' subgroup in UNESCO; the latter implies it was included in the medieval *Quadrivium*. Being already studied in the Middle Ages, it permits diachronic studies crossing the chronological limits of CETA itself.

The features of CETA as well as its compilation principles are the same to be applied to the whole CC. To those we could add the fact that no tagging or parsing is provided. However, it includes encoded information about spelling, paragraphs, page numbers, abbreviations and notes as well as information sources. Apart from the tags based on TEI, some editing of samples was necessary for which special marks have been used.

As for general figures, there is a very similar amount of words per century (208,083 words in the eighteenth century; 202,533 words in the nineteenth century), which guarantees rigorousness to carry out comparative analyses. This can be seen in Fig. 1.

Some extra-linguistic parameters have been considered for text selection, namely, sex, geographical provenance and text-types/genres.

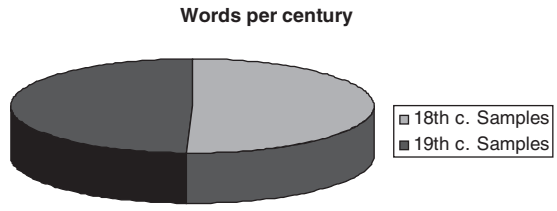


Fig. 1 Words per century

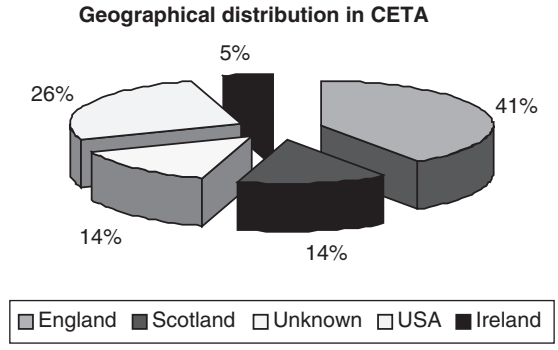


Fig. 2 Geographical distribution in CETA

As for sex, only two out of forty-two writers are women. This low rate of female authors can be accounted for by the difficult accessibility to some texts as well as for the fact that, though women were involved in scientific tasks, they worked 'in the shadow'.[6] Men were the only ones who were prominent as scientists. The predominant uniformity in the sex variable contrasts with the wide variety present in geographical provenance as shown in Fig. 2

Both American and European English-speaking authors have been included, and in this aspect CETA follows the CC principles and criteria for selection. The proportion of writers from England (41%) surpasses by far those from Scotland (14%) or Ireland (5%). North American authors occupy the second position in this ranking of nationalities (26%), mainly in their nineteenth-century production. The 14% figure that corresponds to 'Unknown' must be interpreted as the percentage of authors for whom no biographical details have been found.

As already mentioned, each file is accompanied by another metadata file containing information about the text and the author: full name, profession,

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

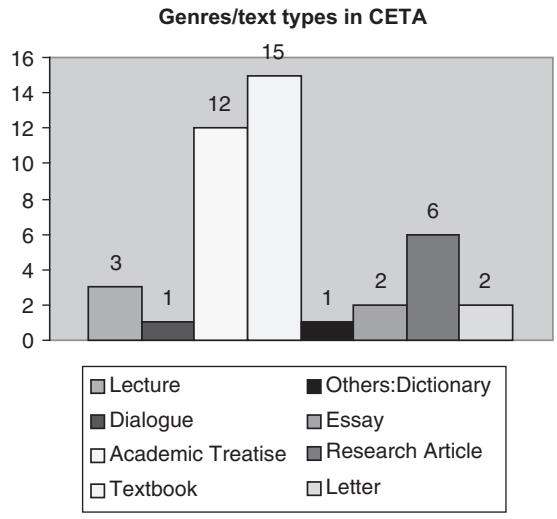


Fig. 3 Genres/text-types in CETA

dates of birth and death, place of education, sex, and age of the author when the text was published are some of the fields included.

As for text-types/genres, samples have been classified as lectures, dialogues, academic treatises, textbooks, letters, essays, articles, and others.

Figure 3 registers the eight different genres contained in CETA. They correspond to the traditional formats of scientific and, by extension, astronomy writing. These eight categories can be further subcategorized according to the type of audience the text is addressed to. The general aim of spreading knowledge lies behind the writing of dialogues, lectures, letters, and textbooks that are frequently written in a simple and plain manner, and whose contents are more or less elementary and informative. Essays and articles are more precise and field-specific.

Treatises in late Modern English do not exactly coincide with our modern interpretation. As a matter of fact, some eighteenth- and nineteenth-century authors make explicit in the prefaces to their books that they are written with educational purposes.

Olmsted (1841), for example, offers the following definition of treatise:

> [. . .] treatise, in which the deepest research is united with that clearness of exposition, which constitutes the chief ornament of a work intended for elementary instruction.

In the category 'Others' we have included those samples that do not exactly fit into any of the seven main categories. In the particular case of CETA, we have only included one sample that is part of a dictionary: John Hill's *Urania*. In accordance with our idea of using a stratified sampling method (Biber, 1993), we decided to include this particular text because it reflects both the growing lexicographic production of the seventeenth and eighteenth centuries and the Royal Society's urgent need to have not only models but also a specific lexical stock that writers could use in an accurate way.

3 Final Remarks

As compilers we believe that CETA as part of a major project, the CC, will be useful to verify the increasing relevance and the transformations undergone by the so-called 'vernacular scientific register' after its initial stages, that is, during the late Modern English period.

Our intention was to offer a *corpus* that complied with all the basic requirements *corpora* must meet. In McEnery *et al.*'s (2006, p. 5) words, a *corpus* that was

> *machine-readable*
> [contained] *authentic texts*
> *sampled* to be
> *representative* of a particular language or language variety

In our case, scientific English from 1650/1700 to 1900.

However, we also wanted it to follow Kohnen's (2007) recommendations when advocating for the construction of genre-based micro-corpora, since they seem to facilitate the tracking of change of those manifestations he labels as 'hidden' and which very often constitute the really distinctive feature of genres.[7]

Finally, as for publication, we intend the *corpus* to come out early in 2010 in a CD-ROM format

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

together with a book in which some pilot studies by different authors will be included.

References

Atkins, B., Clear, J., and Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, **7**(1): 1–16.

Banks, D. (1997). Your very first ESP text: wherein Chaucer explaineth the astrolabe. *La Revue du Geras*, **15/18**: 451–460.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

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

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

Bugliarello, G. (2001). Science, the Arts and the Humanities: Connections and Collisions (accessed 6 October 2004) Available at http://www.poly.edu/news/speech/newTQ.cfm.

Crystal, D. (2003). *A Dictionary of Linguistics and Phonetics*. 5th edn. London: Blackwell.

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

Kohnen, T. (2007). Text types and the methodology of diachronic speech-act analysis. In Fitzmaurice, S. and Taavitsainen, I. (eds), *Methods in Historical Pragmatics*. Berlin/New York: Mouton de Gruyter.

Leech, G. (1992). Corpora and theories of linguistics performance. In Svartvik, J. (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*. Berlín: Mouton de Gruyter, pp. 105–47.

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London/New York: Routledge.

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

Moskowich-Spiegel Fandiño, I. and Crespo García, B. (2007). Presenting the *Coruña Corpus*: a collection of samples for the historical study of English scientific writing. In Pérez Guerra, J. *et al.* (eds), *Of Varying Language and Opposing Creed: New Insights into Late Modern English*. Bern: Peter Lang, pp. 341–57.

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

Olmsted, D. (1841). *Letters on Astronomy, Addressed to a Lady in Which the Elements of the Science are Familiarly Explained in Connexion with its Literary History*. Boston: Marsh, Capen, Lyon and Webb.

Siemund, R. and Claridge, C. (1997). The Lampeter Corpus of Early Modern English Tracts. *ICAME Journal*, **21**: 61–70.

Stubbs, M. (1996). *Text and Corpus Analysis. Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.

Taavitsainen, I. (2004). Transferring classical discourse conventions into the vernacular. In Tavitsainen, I. and Pahta, P. (eds), *Medical and Scientific Writing in Late Medieval English*. Cambridge: Cambridge University Press, pp. 37–72.

Notes

1 MUSTE (Research Group for Multidimensional Corpus-based Studies in English) is a research group at the University of A Coruña, though some researchers from other institutions (Universitat Jaume I de Castellón, University College Cork, and University of Galway) also collaborate with us. Its members have been involved in different projects funded by the University of A Coruña, the Government of the Autonomous Community of Galicia and Spanish Ministry of Education and other private entities.

2 The research here reported on has been funded by Programa de Promoción Xeral de Investigación do Plan Galego de Investigación, Desenvolvemento e Innovación Tecnolóxica (Incite) (PGIDIT07PXIB 104160PR) and Red de Grupos de Investigación 'Lingua e literatura inglesa e identidade' (Consellería de Educación e Coordenación Universitaria, 2007/000145-0). These grants are hereby gratefully acknowledged.

3 We are indebted to both Päivi Pahta and Irma Taavitsainen for their counsel and encouragement.

4 We do not agree with what Siemund and Claridge (1997) declare in their introduction to the *Lampeter Corpus* when stating that they have taken complete texts because any other option would have been 'arbitrarily cut-out smaller text chunks' put together.

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

Our samples have been selected so that all parts of texts (introductions, central chapters, and conclusions) are more or less equally represented.

5 Boyle himself used elaborated sentences in his writings, though, at the same time, he advocated a plain style in a puritanical vein.

6 That was the case of Carolina Herschel, William Herschel's sister, and many others who may have written but not signed their works.

7 To this end, the members of MUSTE have carried out several pilot studies that contribute to verify the validity of the *corpus* as a research tool. Ongoing research should also be pointed out here. Pilot studies are listed by author in alphabetical order:

Bello Piñón, N. (2007). *Code-switching and Borrowing in the Renaissance: A Study of Scientific English from the 17th Century*, *17th SEDERI (Sociedad Hispano-Portuguesa de Estudios Renacentistas Ingleses) Conference*, Cádiz, Spain.

Bello Piñón, N. (2006). *Code-switching and Borrowing: A Study of English Scientific Texts from the 18th Century, 14 ICEHL*, Universita degli Studi di Bergamo, Italy.

Bello Piñón, N. and Méndez Souto, D. E. (2005). *Complex Predicates in Early Scientific Writing*, *17th SELIM International Conference*, A Coruña: Universidade da Coruña.

Camiña Rioboó, G. (2005). Tmesis and In(ter)fixation: the unity of words and/or morphemes in question. In *Reinterpretations of English. Essays on Language, Linguistics and Philology (II)*, A Coruña: Universidade da Coruña, pp. 71–81.

Crespo García, B. (2004). The scientific register in the history of English: a corpus-based study. *Studia Neophilologica* **76**: 125–39.

Crespo García, B. (forthcoming). Specific and non-specific nouns in lME: when Robert grows from man to herb. *English Studies*.

Crespo García, B. and Moskowich-Spiegel Fandiño, I. (2004). Enlarging the lexicon: the field of technology and administration from 1150 to 1500. *Studia Anglica Posnaniensia*, **40**: 163–80.

Crespo García, B. and Moskowich-Spiegel Fandiño, I. 2005. *Latin Forms in Vernacular Scientific Writing: Code-Switching or Borrowing? En Selected Proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (Hel-Lex)*, Sommerville, MA: Cascadilla Press, pp. 51–9.

González Orta, M. (2006). *The Device of Nominalizing in English Scientific Register: Diachronic Analysis of Late Modern English Philosophical Writing, 14 ICEHL*, Università degli Studi di Bergamo, Italy.

Lareo Martín, I. (forthcoming). Analysing a type a collocations. Make-complex predicates in 19th century science and fiction. Atlantis.

Lareo Martín, I. (forthcoming). Make-collocations in nineteenth-century scientific English. *Studia Neophilologica*.

Lareo Martín, I. (2006). *Collocations in Nineteenth-century Science and Fiction, 14 ICEHL*, Università degli Studi di Bergamo, Italy.

Lareo Martín, I. and Moskowich-Spiegel, I. (2007). *Comparison ''Made Possible'': Collocation of Make Plus Adjective in 18th Century Science and Fiction, Symposium on the Language of Science in the Time of Linnaeus*. Sweden: Uppsala University.

Lareo Martín, I. and Estéve Ramos, M. J. (2007). *Scientific Registers in the 18th Century: The Case of Make-collocation in the Coruña Corpus, Symposium on the Language of Science in the Time of Linnaeus.* Sweden: Uppsala University.

Moskowich-Spiegel Fandiño, I. (2008). 'To lerne sciences touching nombres and proporciouns': the proportion of affixation in early scientific writing. *English Studies*, **89**: 39–52.

Moskowich-Spiegel Fandiño, I. and Crespo García, B. (2006). Lop-webbe and henne cresse: morphological aspects of the scientific register in Late Middle English. *Studia Anglica Posnaniensia*, **42**: 133–45.

Parapar López, J. and Moskowich-Spiegel Fandiño, I. (2007). The *Coruña Corpus* tool. *Revista del Procesamiento de Lenguaje Natural*, **38**: 289–90.

Sánchez Riveiro, V. (2006). *Compounding and Derivation in ModE: Some Morphological Issues of Scientific Writing during the 19th Century, 14 ICEHL*, Università degli Studi di Bergamo, Italy.

Sánchez Riveiro, V. (2007). *Scientific English in ModE: Compounding and Derivation during the 17th Century, 18th SEDERI Annual Conference*, Universidad de Cádiz.

Appendix I

1 UNESCO Classification

1.2 Fields of Science and Technology (*International Standardization of Statistics on Science and Technology*, UNESCO 1978):

1.2.1 *Natural Sciences*

Astronomy, bacteriology, biochemistry, biology, botanics, chemistry, entomology, geology,

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

Table A1 Authors included in CETA

Author	Year	Title	No. of words
Henry Curson	1702	*The theory of sciences illustrated, or the grounds and principles of the seven arts; grammar, logick, rhetorick, musick, arithmetick, geometry, astronomy. Accurately demonstrated and reduced to practice*	9,846
Robert Morden	1702	*An Introduction to Astronomy, geography, navigation, and other mathematical sciences made easie by the description and uses of the coelestial and terrestrial Globes. In seven parts*	10,006
William Whiston	1715	*Astronomical lectures, read in the publick schools at Cambridge*	9,757
John Harris	1719	*Astronomical dialogues between a gentleman and a lady: wherein the doctrine of the sphere, uses of the globes, and the elements of Astronomy*	9,884
George Gordon	1726	*An introduction to geography, astronomy, and dialling. Containing the most useful elements of the said sciences, adapted to the meanest capacity, by the description and uses of the terrestrial and celestial globes with an introduction to chronology*	10,003
Isaac Watts	1726	*The knowledge of the heavens and the earth made easy: or, the first principles of astronomy and geography explain'd by the use of globes and maps*	10,049
Samuel Fuller	1732	*Practical astronomy, in the description and use of both globes, orrery and telescopes wherein the most useful elements, and most valuable modern discoveries of the true astronomy are exhibited, after a very easy and expeditious manner, in an exact account of our solar system*	10,035
Jasper Charlton	1735	*The ladies astronomy and chronology in four parts*	10,082
Roger Long	1742	*Astronomy, in five books*	10,045
James Hodgson	1749	*The theory of Jupiter's satellites: with the construction and use of the tables for computing their eclipses*	9,929
John Hill	1754	*Urania: or, a compleat view of the heavens; containing the antient and modern astronomy, in form of a dictionary: illustrated with a great number of figures (Vol.I. Being the first of a compleat system of natural and philosophical knowledge)*	10,044
James Ferguson	1756	*Astronomy explained upon Sir Isaac Newton's principles and made easy to those who have not studied mathematics*	10,040
Matthew Stewart	1761	*Tracts, physical, and mathematical. Containing an explication of several important points in physical astronomy; and a new method for ascertaining the sun's distance from the earth, by the theory of gravity*	9,881
George Costard	1767	*The history of astronomy, with its application to geography, history, and chronology; occasionally exemplified by the globes.*	9,959
Alexander Wilson	1774	*Observations on the solar spots*	4,137
George Adams	1777	*A treatise describing the construction and explaining the use of celestial and terrestrial globes*	9,901
John Lacy	1779	*The universal system: or mechanical cause of all the appearances and movements of the visible heavens: shewing the true powers which move the earth and planets in their central and annual rotations with a dissertation on comets, the Nature, cause, matter, and use of their tails, and the reasons of their long trajectories; likewise and attempt to prove what it is that moves the sun around its axis.*	5,845
William Nicholson	1782	*An introduction to natural philosophy*	9,932
John Bonnycastle	1789	*An introduction to astronomy in a series of letters from a preceptor to his pupil*	9,947
Samuel Vince	1790	*A treatise on practical astronomy*	9,993
Margaret Bryan	1797	*A compendious system of astronomy*	10,064
Robert Small	1804	*An account of the astronomical discoveries of Kepler: including an historical review of the systems which had successively prevailed before his time*	10,033

(continued)

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

Table A1 Continued

Author	Year	Title	No. of words
John Ewing	1809	*A plain, elementary and practical system of natural experimental philosophy: including astronomy and chronology*	9,731
David Brewster	1811	*Ferguson's astronomy explained upon Sir Isaac Newton's principles: with notes and supplementary chapters*	9,492
William Phillips	1818	*Eight familiar lectures on astronomy: intended as an introduction to the science: for the use of young persons and others not conversant with the mathematics*	10,056
John Gummere	1822	*An elementary treatise on astronomy. In two parts. The first, containing a clear and compendious view of the theory. The second, a number of practical problems. To which are added, solar, lunar and some other astronomical tables.*	10,059
Thomas Luby	1828	*An introductory treatise on physical astronomy*	10,012
John Frederick William Herschel	1833	*The cabinet encyclopedia. Conducted by the Rev. Dionysius Lardner . . . Assisted by eminent literary and scientific men. Natural philosophy. Astronomy. A treatise on astronomy*	10,120
Landon Cabell Garland	1838	*An Address on the utility of astronomy*	9,510
Denison Olmsted	1841	*Letters on astronomy, addressed to a lady in which the elements of the science are familiarly explained in connexion with its literary history.*	8,644
Duncan Bradford	1845	*The wonders of the heavens, being a popular view of astronomy, including a full illustration of the mechanism of the heavens; embracing the sun, moon, and stars*	10,078
William Holms Chambers Bartlett	1855	*Elements of natural philosophy (spherical astronomy)*	10,054
William Whewell	1858	*The plurality of worlds. With an introduction by Edward Hitchcock.*	10,066
Ormsby McKnight Mitchel	1860	*Popular astronomy. A concise elementary treatise on the sun, planets, satellites, and comets*	10,048
Elias Loomis	1868	*A treatise on astronomy*	9,987
William Chauvenet	1871	*A manual of spherical and practical astronomy, embracing the general problems of spherical astronomy, the special applications to nautical astronomy, and the theory and use of fixed and portable astronomical instruments, with an appendix on the method of least squares.*	10,005
Dorman Steele	1874	*Fourteen weeks in descriptive astronomy*	9,849
George Howard Darwin	1880	*On the secular changes in the elements of the orbit of a satellite, revolving about a tidally distorted planet*	5,076
Charles Augustus Young	1880	*Recent progress in solar astronomy*	6,376
James Croll	1889	*Stellar evolution and its relation to geological time*	9,243
Agnes Mary Cerke	1893	*A popular history of astronomy in the nineteenth century*	10,004
Percival Lowell	1895	*Mars: Canals*	8,486

geophysics, mathematics, meteorology, mineralogy, computing, physical geography, physics, zoology, and other allied subjects.

1.2.2 *Engineering and Technology*

Engineering sciences such as chemical, civil, electrical, and mechanical engineering and their specialized subdivisions; forest products; applied sciences such as geodesy, industrial chemistry, etc.; architecture; the science and technology of food production; specialized technologies of interdisciplinary fields, e.g. systems analysis, metallurgy, mining, textile technology, and other allied subjects.

1.2.3 *Medical Sciences*

Anatomy, stomatology, basic medicine, paedriatics, obstetrics, optometry, osteopathy, pharmacy, physiotherapy, public health services, technical health assistance, and other allied subjects.

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012

1.2.4 *Agricultural Sciences*

Agronomy, zootechnics, fisheries, forestry, horticulture, veterinary medicine, and other allied subjects.

1.2.5 *Social Sciences*

Anthropology (social and cultural) and ethnology, demography, geography (human, economic, and social), law, linguistics, management, political sciences, psychology, sociology, organization and methods, miscellaneous social sciences and interdisciplinary, methodological and historical S&T activities relating to subjects in this group.

Physical anthropology, physical geography, and psychophysiology are normally classified under the natural sciences.

1.2.6 *Humanities*

Arts (history of art and art criticism, excluding artistic 'research'), ancient and modern languages and literatures, philosophy (including the history of science and technology), prehistory and history, together with auxiliary historical disciplines (such as archaeology, numismatics, paleography, genealogy, etc.), religion, other subjects and humanistic branches as well as other methodological and historical S&T activities relating to the subjects in this group.

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on January 17, 2012