Writing Science, Compiling Science: The *Coruña Corpus of English Scientific Writing*[1]

Isabel Moskowich and Javier Parapar
Universidade da Coruña

Abstract

The Coruña Corpus: 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. Its purpose is to facilitate investigation at all linguistic levels, though, in principle, phonology is not included among our intended research topics. A rough definition of our corpus would say it contains English scientific texts other than medical produced between 1600 and 1900. In order to retrieve information from the compiled data, we decided to create a corpus management tool. Loosely speaking the Coruña Corpus Tool (CCT) is an Information Retrieval (IR) system where the indexed textual repository is the set of compiled documents that constitutes the CC.

The Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing (henceforth, *CC*) 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. Its purpose is to facilitate investigation at

[1] 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*) (PGIDIT07PXIB104160PR) and Rede de investigación "Lingua e literatura inglesa e identidade" (Consellería de Educación e Ordenación Universitaria, 2007/000145-0). These grants are hereby gratefully acknowledged.

Proceedings 31st AEDEAN Conference. 2008. María Jesús Lorenzo Modia. Ed. 531-543. A Coruña: Universidade. ISBN: 978-84-9749-278-2.

all linguistic levels, though, in principle, phonology is not included among our intended research topics.

The *CC* is still a work in progress and we would like to present here, as we have been doing elsewhere, our main concerns about it, both theoretical and practical. To this end, we will first provide some issues we considered before taking the first steps. Therefore, section one will deal with principles of corpus compilation such as parameters of classification, time-span covered and representativeness. In section two we will then present some other technical and practical aspects related to the development of a search engine and other tools for the compilation and use of the *CC*.

A rough definition of our corpus project would say it contains English scientific texts other than medical produced between 1650 and 1900. Medical texts have been disregarded since they are being compiled by Taavitsainen and Pahta and their team in Helsinki and Jyvvaskyla. The Middle English part of it, *MEMT*, has been already released, and they are at present compiling the Modern English part.

Two of the ideas that triggered the whole project are the growing interest in the vernacularisation of Science in late-medieval and modern England as an understudied area, on the one hand[2], and the gradual increase in studies on genre conventions and special languages, on the other. Few dispute that scientific writing exhibits great variation and deserves study (Biber, 1988; Stubbs, 1996; Taavitsainen and Pahta, 1997a, b). As explained by Siemund and Claridge (1997: 67) when presenting their own work, our project proposes to complement other corpora pertaining to the history of what we nowadays call *ESP*, such as the well-known *Corpus of Early English Correspondence*, the *Corpus of Early English Medical Writing*, and the *Lampeter Corpus of Early Modern English Tracts*.

In line with Johansson (1991) and Atkins *et al.*'s (1992) claim that corpora must be principled and designed within certain constraints, several decisions were necessary prior to the compilation of texts itself. Such decisions are presented in the following paragraphs.

[2] As Pahta and Taavitsainen (2004: XV) have already pointed out: "Vernacular scientific writing in the late medieval and early modern periods is still an understudied area".

1. Principles of corpus compilation

1.1. Theoretical decisions

Among other considerations, three were the main principles according to which the samples of our corpus have been compiled, namely, parameters of classification, time-span and degree of representativeness of the text samples collected.

1.1.1. Classification

The selection of texts for our corpus was not random. On the contrary, it was made according to certain external parameters to ensure the possibility of fruitful linguistic analyses. As Atkins, Clear and Ostler (1992: 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. (1992: 5)

Therefore, one of the first theoretical decisions that we made concerned the concept of Science itself since it had a direct influence on textual selection.

Texts produced before and after the emergence of Empiricism and the generalisation of the scientific method need to be treated differently since this new method also entailed a change in the classification of knowledge and philosophy of science. Current *UNESCO* parameters have been used as a starting point for the selection of scientific texts produced after 1700, the part we have addressed first. Different criteria must be applied to texts prior to this date. For these, an inclusive perspective will be probably adopted to avoid the omission of texts dealing with those areas of knowledge which would not be considered science today (Alchemy). Following Biber, (1993: 244), we opted for a stratified sampling method where certain subgroups (in our case, scientific disciplines) are identified within the target population (scientific English in English).

Of the six areas into which *UNESCO* divides Science and Technology (see table 1), the first, "Exact and Natural Sciences", is also the first area we have selected, beginning with the compilation of the text-categories Mathematics, Astronomy, Physics (where we include Physics and Geophysics) and Life Sciences (where we include Biology mainly, but also Botanics, Zoology and others). Since some of the branches of human development have been considered science only very recently (Bugliarello, 2001), as is the case of Field II (Engineering), we have excluded them from our consideration to avoid skewing the corpus. The agricultural branches have been also included in Life Sciences.

Table 1. Fields of Science and Technology (*International Standardisation of Statistics on Science and Technology*, *UNESCO* 1978)

1. Natural Sciences.

Astronomy, bacteriology, biochemistry, biology, botanics, chemistry, entomology, geology, geophysics, mathematics, meteorology, mineralogy, computing, physical geography, physics, zoology and other allied subjects.

2. Engineering and Technology.

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

3. Medical Sciences.

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

4. Agricultural Sciences.

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

5. Social Sciences

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

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

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.

At the moment of writing this paper we have begun the selection of text-samples for the Humanities, namely, Philosophy and History and intend to compile the same number of samples for each scientific field in order to facilitate comparative studies on the language used in each discipline, and the evolution of particular features of each of them, confirming the wide range of variation within academic prose (Biber, 1988). With these premises we will obtain different sub-corpora that can be considered as independent entities though sharing a similar structure, organisation and mark-up.

1.1.2. Time-span

The second criterion applied concerns the selection of the time-span (1600-1900), which is based on some extra-linguistic considerations.

The seventeenth century marks the beginning of a new way of thinking in which old patterns are no longer repeated (Taavitsainen and Pahta, 1997b). Whereas medieval scholasticism conceived of science as deduction from assumed principles, later scholars began to devote themselves to induction, experimentation and mathematics. This way,

they began to develop the foundations of modern science in the seventeenth century.

Roughly speaking, there are three main differences between scholasticism and this modern stance: first, seventeenth-century science evolved independently, outside university circles, in many cases under the influence of the Royal Academy; second, it was not only concerned with types of knowledge and the relationship between science and theological matters, but with the practical application of scientific investigation; and, third, there was an attempt to reach precise conclusions by quantifying data.

The acceptance of this empirical view led to the modification of the corresponding discourse. This new school of scientific thought called for the creation of an *ad hoc* discourse which, as Stubbs (1996: 18) summarises from Swales (1990), "was consciously developed by scientists who required ways of expressing generally accepted knowledge about experimental matters of fact".

We have chosen 1900 as the other end of the time-span covered by our corpus due to no less important reasons. Facts such as 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 this same year, Planck's announcement of quantum mechanics, or Einstein's publication of a paper proposing what is today called the Special Theory of Relativity in 1905, must be viewed as milestones in the history of Science that probably established a turning point similar to the one which took place three centuries earlier. Besides, at the 1897 International Congress of Mathematics, Thomas Huxley outlined a new scientific style. From that moment onwards, scientific discourse changed dramatically again.

1.1.3. Representativeness

Another principle we have taken into account is that of the representativeness of texts and balance within the corpus. For each text category (discipline) we have selected two texts per decade, with each sample containing around 10,000 words, excluding tables, figures,

formulae and graphs[3]. Shorter texts have been included *in toto*. This decision is based on Kytö, Rudanko and Smitterberg's claim (2000: 92) that short-term change in diachrony can be safely studied over periods of thirty years. Each category is therefore represented by 600,000 words in each whole sub-corpus for eModE.

In the interests of thoroughness, first editions have been preferred; likewise, we have avoided using more than one text by the same author in order to avoid the proliferation of idiosyncrasies. For this particular issue, therefore, we have followed some of the compilation principles of the *Lampeter Corpus of Early Modern English Tracts*. However, we are conscious that the question of balance within the corpus, as a "small scale model of the linguistic material which the corpus builders wish to study" (Atkins *et al.*, 1992: 6), is at the discretion of the compilers.

At the moment of writing this paper, the categories of Astronomy, Philosophy and Mathematics have been completed for the eighteenth and nineteenth centuries and Natural History is being keyed in. Physics and History have been collected with availability, being an important determiner of choice and selection.

We have verified that, as the concept of Science alters over time, the associated textual typology must also change. We are still trying to find a more or less definitive classification for text types appearing in our categories, often based on their degree of technicality and target audience.

We are aware that register /style[4] are connected with certain social or extralinguistic variables that may permit sociolinguistic studies on the corpus. Though authors from the lower grades of society are not found for scientific English, more or less "colloquial" texts have been included. To the same end, the social background of authors together with some details about their lives will be provided where possible in separate metadata files. We also believe that the representativeness of

[3] We do not agree with what Claridge declares in her 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. Our samples have been selected so that all parts of texts (introductions, central chapters and conclusions) are more or less equally represented.

[4] As is well-known, Biber (1988: 70) uses "genre" to refer to textual categories defined from an extra-linguistic perspective. Also Taavitsainen (2001).

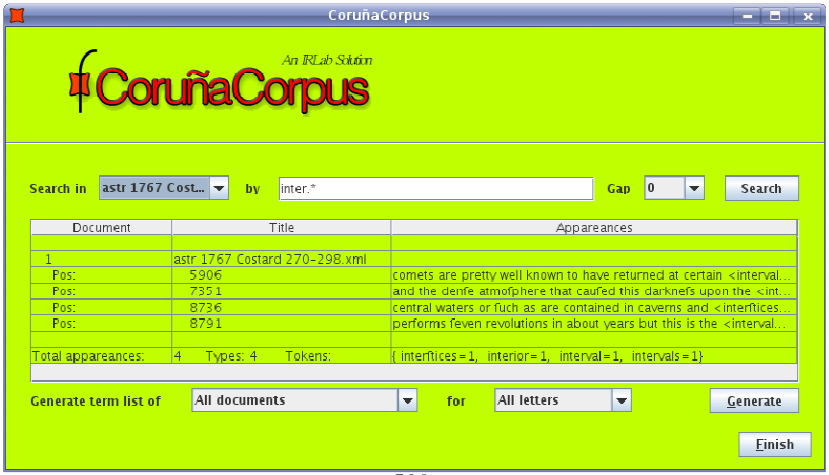
the *CC* is improved by not including any translations. Only English-speaking authors writing in English have been considered, though we are conscious that many of them also used Latin, and this may have had an influence on their use of their native language.

1.2. Practical considerations

In the last few years the corpus has been already tested and several pilot-studies have been published. On a practical level, that is to say, as far as the process of compilation itself is concerned, the coding of texts is still being carried out. At this moment, texts are being encoded in *XML*. We are not including tagging or parsing, though we have provided encoded information about spelling, paragraphs, page numbers and notes (marginal, foot or end notes), as well as the necessary information about the sources.

In order to retrieve information from the compiled data, we decided to create a corpus management tool to facilitate the use of the *CETA* (*Corpus of English Texts on Astronomy*) as well as the rest of the sub-corpora to be contained in the *Coruña Corpus* when compilation is completed. This software application, designed to help linguists to extract and condense valuable information for their research, is currently in the testing phase.

Figure 1. The Coruña Corpus Tool



Loosely speaking the *Coruña Corpus Tool* (*CCT*) is an Information Retrieval (*IR*) system where the indexed textual repository is the set of compiled documents that constitutes the *Coruña Corpus* or any of its parts. As already explained the selected texts were coded and stored as *XML* documents. We chose to tag the information following the recommendations and rules of the *TEI* (Text Encoding Initiative) standard, and the defined *DTD* (Document Type Definition) that fixes the strict structure and key-words used in the *XML-TEI* file.

The application is divided in two parts. On the one hand, we built an administration module where the authorised users or compilers can create new repositories from the *XML* documents, add new texts to an existing repository or edit and delete documents already present in the index. On the other hand, the user module offers other users the main tool functionalities such as basic searching, creation of concordances or term list generation.

1.2.1. Technical considerations

The system was conceived of following two main design concepts. Firstly, we chose a Model-View-Controller (*MVC*) architectural pattern in order to isolate the logical layer of the application from the view layer. Secondly, we planned a component-based software to allow the easy implementation of future improvements, the adding of new features, and enabling the reuse of the software.

The application was also designed considering the computational efficiency of the system execution and to be scalable, i.e., enabling the possibility of increasing the number of texts that conforms the Corpus without producing a degradation in the performance. We designed a desktop (standalone) application due to the needs of the target users and to allow the easy tool packing and redistribution. For its development we used Java as a programming language since, this way, we obtained a platform independent software. It is fair to mention here that we resorted to some existing open-source libraries and *COTS* (Components Of The Self) for the system implementation. Among them we would like to point out the following two:

- Lucene: It is a Java indexing *API* (Application Programming Interface) developed by the Apache Foundation. It is a library widely used in the development of *IR* applications. Indices are the structures that allow the efficient processing of users queries and this tool makes index construction transparent to the developers.
- JDom: This is another Java *API* to deal with the reading, transformation and writing of *XML* documents. This will facilitate us the load in memory of the tree structures associated with the *TEI* documents and the extraction of the desired content from them.

Previous to the index construction on the corpus texts, we have a pre-processing step over the collection of documents. In this phase several tagged fields that we desire to index are extracted from the documents. In this sense we have to mention that we build a multi-field index to allow searches using different criteria; we can store, for example, information about authors, date, scientific field, corpus document identifier, etc.

1.2.2. System features

The *CCT* offers several services to provide information. In this sense all the linguist staff was present during the requirements engineering phase. As a result of this, the initial version of the system allows:

- Document validation: An important issue derived from the text coding in *XML-TEI* is that tagging rules are very strict so it is very easy to breach the correctness of the document, i.e., if there is some tag missing, the document will be said not to follow the *DTD* rules. Therefore, to avoid these failures the platform offers a syntax validator for the *XML*s that shows the compilers/coders the errors present in the document so that they can be fixed.
- Basic term search: i.e., looking for a word across the collection. This can be applied to the whole set of indexed documents or at individual document level. As the result of

a user query all the occurrences of a word are shown. For each one the following information is available:

  - Document identifier: the id of the text where the word was found.
  - Word position: the place where the term is located into the document.
  - Word concordance: The terms preceding and following the query term. They are exposed to allow disambiguation; the length of the concordance is a system parameter.

- Advanced search: over the basic word search a certain number of custom search characteristics are implemented to facilitate the extraction of research results:

  - Wild card use: the inclusion of wild card characters are allowed to specify the searching of spelling variations of the same form e.g. *de.cribed* will match with *described* and *defcribed.*
  - Regular expression searching: to allow searching using patterns, it is useful to search for example by suffixes or prefixes *inter.** will match for example with: *intervenes, intercalary, interrupted, intercept, intervallorum, interrupt, internal, interception, interruption, etc.*
  - Phrase search: combinations of words can be specified as a query indicating the gap between the words. This can be used for example to look for expressions or verbal forms.

- Term list generation: the system offers the lexicon list of the whole corpus or inside each document (as chosen). An alphabetically sorted list of words with the number of appearances is generated. The user can also choose a letter to filter only the words that start with the selected character.

- Report generation: the system allows exporting the search results to a plain text format editable by users.

We must also mention that some extra processing is done over the user queries to improve the final results. For instance, user queries are stemmed following the well-known Porter's algorithm. Thus in the search process every word whose stem matches the stemmed query will be included in the final results. Besides, the interface is designed to facilitate the input of special characters not present in the traditional keypads as *æ, ſ, λ, a*, etc.

2. Final Remarks

Though still a work in progress at the moment of writing this lines, not only *CETA*, but the whole *Coruña Corpus* reflect a well-planned process. Both text selection and the implementation of an Information Retrieval tool have been carefully planned. It is the desire of the compilers that it will be useful as a means to offer a new perspective on the evolution of the language of Science.

References

Atkins, S., J. Clear and N. Ostler. 1992. "Corpus Design Criteria". *Literary and Linguistics Computing* 7: 1. 1–16.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: CUP.

---. 1993. "Representativeness in Corpus Design". *Literary and Linguistic Computing*, 8: 4. 243–257.

Bugliarello, G. 2001. "Science, the Arts and the Humanities: Connections and Collisions". October 2004. http://www.poly.edu/news/speech/newTQ.cfm.

Crespo, B. 2004. "General Survey of the Growth of Scientific Culture". *About Culture*. E. Woodward, *Ed.* 157–165. Universidade da Coruña: Servicio de Publicacións.

Johansson, S. 1991. "Computer Corpora in English Language Research". *English Computer Corpora. Selected Papers and Research Guide*. S. Johansson, Stig, A-B. Stenström. *Eds.* 3–6. Berlin/New York: Mouton de Gruyter.

Kÿto, M., J. Rudanko, and E. Smitterberg. 2000. "Building a Bridge between the Present and the Past: A Corpus of 19-century English", *ICAME Journal* 24. 85–97.

Lee, D. 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.

Nevalainnen, T. and H. Raumolin-Brunberg. 1989. "A Corpus of Early Standard Modern English in a Socio-Historical Perspective" *Neuphilologische Mitteilungen* 90: 1. 67–110.

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

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Taavitsainen, I. 2004. "Transferring Classical Discourse Conventions into the Vernacular". Taavitsainen and Pahta. 2004. 37–72.

---. 2005. "On Corpus Linguistics: Computers and the History of English". *Re-Interpretations of English: Essays on Languages, Linguistics and Philology, (II)*. I. Moskowich & B. Crespo. *Eds.* 325-345. A Coruña: University of A Coruña.

Taavitsainen, I. and P. Pahta. 1997a. "Corpus of Early English Medical Writing 1375–1750". *ICAME*, 21. 71–78.

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

---. Eds. 2004. *Medical and Scientific Writing in Late Medieval English*. Cambridge: CUP.

Taavitsainen, I., P. Pahta, N. Leskinen, M. Ratia, and C. Suhr. 2002. "Analysing Scientific Thouht-styles: What Can Linguistic Research Reveal about the History of Science?" *Variation Past and Present. VARIENG Studies on English for Terttu Nevalainen*. H. Raumolin-Brunberg, M. Nevala, A, Nurmi, A., and M. Rissanen. *Eds*. 252–270. Helsinki: Société Néophilologique.