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

Isabel Moskowich & Javier Parapar

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. Medical texts have been disregarded since they are being compiled by Taavitsainen Pahta and their team in Helsinki.

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, 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 intends 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.

1. ***Principles of corpus compilation***

1.1. Classification

The selection of texts for our corpus has been made according to external parameters to ensure the possibility of fruitful linguistic analyses. As Atkins, Clear and Ostler (1992: 5) claimed:

> *"A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation."*

We think that 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.

Following Biber, (1993: 244), we opted for a stratified sampling method where certain subgroups (in our cse, scientific disciplines) are identified within the target population (scientific English).

Different criteria must be applied to texts prior to this date. For these, an inclusive perspective will be probably adopted.

Of the six areas into which UNESCO divides Science and Technology, the first, "Exact and Natural Sciences", is also the first we have selected, beginning with the compilation of the text-categories Mathematics, Astronomy, Physics (where we include Physics and Geophysics) and Natural History (where we include Biology mainly, but also Botanics, Zoology and others). Since some of the branches of human deveolpment 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 included elsewhere.

We have already 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).

1.2. Time-span

The second criterion concerns the selection of the time-span (1600-1900), which is also based on 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 17th century.

There are three main differences between scholasticism and this modern stance:

- seventeenth-century science evolved independently, outside university circles, in many cases under the influence of the Royal Academy
- 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.
- there was an attempt to reach precise conclusions by quantifying data.

The acceptance of an 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 tthree centuries earlier. Besides, at the 1897 International Congress of Mathematics, Thomas Huxley outlined a new scientific style. From that moment, scientific discourse changed dramatically again.

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. 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.

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. Here we have followed some of the compilation principles of the *Lampeter Corpus*. 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 18th and 19th c. and Natural History is being keyed in. Physics and History have been collected with availability being an important aspect of 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.

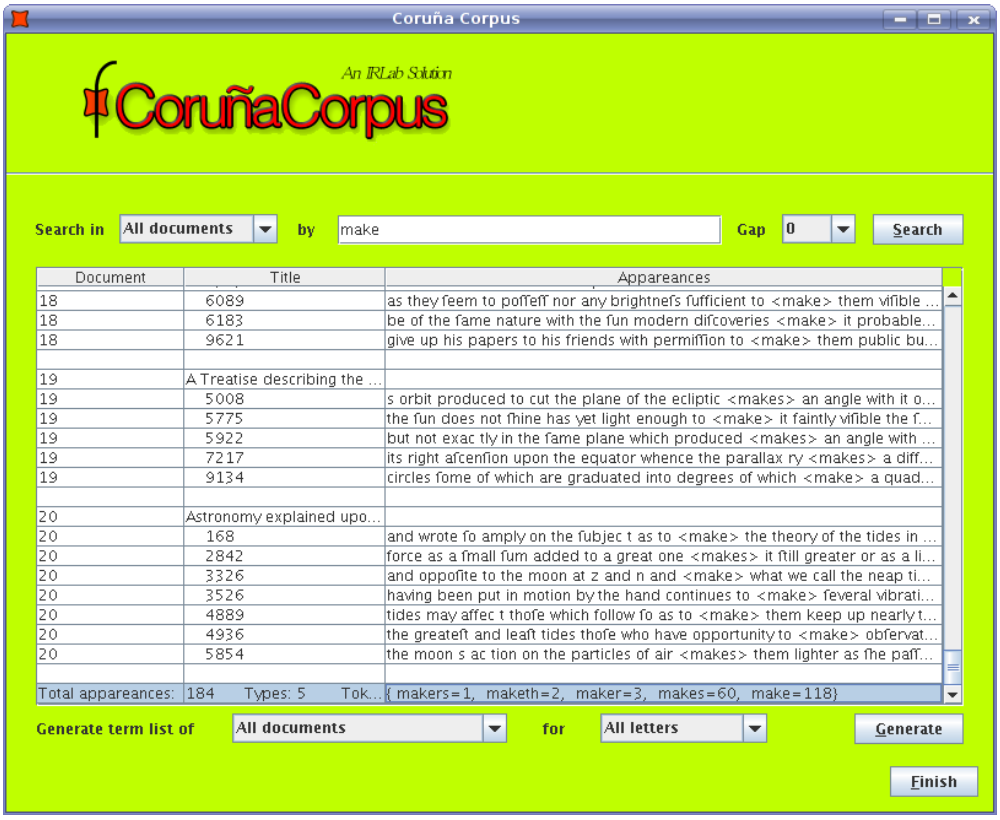
We are aware that register/style[1] are connected with certain social or extralinguistic variables that may permit sociolinguistic studies on the corpus. Though authors from the lower grades of society

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

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 files.

We also believe that the representativeness of 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.

2. *The Coruña Corpus Tool*



Search in: All documents by: make Gap: 0

Document	Title	Appareances
18	6089	as they ſeem to poſſeſſ nor any brightneſs ſufficient to <make> them viſible …
18	6183	be of the ſame nature with the ſun modern diſcoveries <make> it probable…
18	9621	give up his papers to his friends with permiſſion to <make> them public bu…
19	A Treatise describing the …	
19	5008	s orbit produced to cut the plane of the ecliptic <makes> an angle with it o…
19	5775	the ſun does not ſhine has yet light enough to <make> it faintly viſible the ſ…
19	5922	but not exac tly in the ſame plane which produced <makes> an angle with …
19	7217	its right aſcenſion upon the equator whence the parallax ry <makes> a diff…
19	9134	circles ſome of which are graduated into degrees of which <make> a quad…
20	Astronomy explained upo…	
20	168	and wrote ſo amply on the ſubjec t as to <make> the theory of the tides in …
20	2842	force as a ſmall ſum added to a great one <makes> it ſtill greater or as a li…
20	3326	and oppoſite to the moon at z and n and <make> what we call the neap ti…
20	3526	having been put in motion by the hand continues to <make> ſeveral vibrati…
20	4889	tides may affec t thoſe which follow ſo as to <make> them keep up nearly t…
20	4936	the greateſt and leaſt tides thoſe who have opportunity to <make> obſervat…
20	5854	the moon s ac tion on the particles of air <makes> them lighter as ſhe paſſ…
Total appareances:	184 Types: 5 Tok…	{ makers=1, maketh=2, maker=3, makes=60, make=118}

Generate term list of: All documents for: All letters

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 selected samples 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.

2.1. Technical considerations

The application was 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 of 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.

Some of the used technologies were:

- Lucene: a Java indexing library developed by the Apache Foundation. This tool makes the index construction transparent to the developers. The indices are the structures that allow the efficient processing of users queries.
- JDom: to deal with the reading, transformation and writing of XML documents.

Previous to the construction of the index on the corpus texts, we have a preprocessing step over the collection of documents. In this phase several tagged fields that we desire to index are extracted from the documents. We store, for example, information about authors, date, scientific field, corpus document identifier, etc.

2.2. System features

As a result of the linguists requirements, the initial version of the system allows:

- Document validation: The XML-TEI 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. To avoid these failures the platform offers a syntax validator for the XMLs that shows coders the errors.
- 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 an query all the occurrences of a word are showed. For each one the following data is available:

    - Document identifier.
    - Word position.
    - Word concordance.
- 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 *deſcribed.*
    - 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.
- Term list generation: the system offers the lexicon list of the whole corpus or inside each document (as chosen). An alphabetical sorted list of words with the number of appearances is generated.
- Report generation: the system allows to export the search results to a plain text format editable by users.

We would like to point out that the user queries are stemmed following the well-known Porter's algorithm. Therefore in the search process every word whose stem matches the stemmed query will be included in the final results.

The tool is being developed according to the corpus aims. Simultaneous development ensures text changes may be added to improve both. The fact that it is scalable provides the opportunity to enlarge the corpus. So this is a case of perfect symbiosis.

References

Atkins, Sue, Clear, Jeremy and Ostler, Nicholas. 1992. “Corpus Design Criteria”, *Literary and Linguistics Computing*, 7/1: 1-16.

Biber, Donald. 1988. *Variation Across Speech and Writing*. Cambridge: CUP.

Biber, D. 1993.Representeativeness in Corpus Design. Literary and Linguistic Computing, 8/4: 243-257.

Bugliarello, George. 2001. “Science, the Arts and the Humanities: Connections and Collisions”. http://www.poly.edu/news/speech/newTQ.cfm. Accessed 6.10.2004.

Crespo, Begoña. 2004. “General Survey of the Growth of Scientific Culture”, in Woodward, E. ed. *About Culture*. Universidade da Coruña: Servicio de Publicacións. (157-165).

Johansson, Stig. 1991. “Computer Corpora in English Language Research”. In Johansson, Stig, Stenstróm, Anna-Brita. eds. *English Computer Corpora. Selected Papers and Research Guide*. Berlin/New York: Mouton de Gruyter. 3-6.

Kÿto, Merja, Rudanko, Juhani and Smitterberg, Erik. 2000. “Building a Bridge between the present and the past: A Corpus of 19-century English”, *ICAME Joumal* 24: 85-97.

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.

Nevalainnen, Terttu & Raumolin-Brunberg, Helena. 1989. “A Corpus of early Standard Modern English in a Socio-Historical Perspective”, *Neuphilologische Mitteilungen*, 90/1 (67- 110).

Siemund, Rainer & Claridge, Claudia. 1997. “The Lampeter Corpus of Early Modern English Tracts”. *ICAME*, 21: 61-70.

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

Taavitsainen, Irma & Pahta, Päivi. 1997a. “Corpus of Early English Medical Writing 1375-1750”. *ICAME*, 21: 71-78.

____. 1997b. “The Corpus of Early English Medical Writing: Linguistic variation and prescriptive collocations in scholastic style”. In Nevalainen, Terttu & Leena Kahlas-Tarkka (eds.) *To Explain the Present: Studies in Changing English Language in Honour of Matti Rissanen.*

(Mémoires de la Société Néophilologique de Helsinki, 52. Helsinki: Société Néophilologique (209-225).

____. eds. 2004. *Medical and Scientific Writing in Late Medieval English.* Cambridge: CUP.

Taavitsainen, Irma. 2004. “Transferring classical discourse conventions into the vernacular”. In Taavitsainen, Irma and Pahta, Päivi (eds.). (37-72).

____. 2005. “On Corpus Linguistics: Computers and the History of English”. In Moskowich & Crespo (eds.), *Re-Interpretations of English: Essays on Languages, Linguistics and Philology, (II)*. A Coruña: University of A Coruña.

Taavitsainen, Irma, Pahta, P., Leskinen, N., Ratia, M. & Suhr, C. 2002. “Analysing Scientific Thouht-styles: What can Linguistic Research Reveal about the History of Science?”. In Raumolin- Brunberg, H., Nevala, M., Nurmi, A. & Rissanen, M. eds. *Variation Past and Present. VARIENG Studies on English for Terttu Nevalainen*. Helsinki: Société Néophilologique. (251-270)