

LINGUISTICS | MONOGRÁFICO
SOBRE LINGÜÍSTICA DE CORPUS

ISABEL MOSKOWICH, LUIS PUENTE-CASTELO,
BEGOÑA CRESPO

Universidade da Coruña

GONZALO CAMIÑA-RIOBOO

Université de Nantes

***The Coruña
Corpus of
English
Scientific
Writing:
Challenge
and Reward***

The Coruña Corpus Project is born

The Coruña Corpus of English Scientific Writing did not still have a name when it was first conceived of only two years after the new century started. At that moment, in 2003, research groups in Spanish universities were bunches of people with similar interests more than administrative units and thus, there were no formal lists of members of a research group. However, already at that point, we considered the need to have a name (MuStE)¹ and to have an identity. Administrative regulations only came later.

During the academic year 2002-03 the two staff members, senior Isabel Moskowich and junior Begoña Crespo, often met with other people at Isabel's tiny office in the building the Faculty of Philology occupied then in one of the campuses of the University of A Coruña (Spain). The researchers that grouped together then had been trained in English Historical Linguistics and that was their main common interest. In fact, by 2002 there had been some doctoral theses on different aspects of the evolution of English, none of which went beyond the seventeenth century (Begoña Crespo dealt with semantic change in 2001 using the Helsinki Corpus of English Texts, Elena Alfaya explored some new possibilities of the fit-technique in a medieval poem, *The Parlement of the Thre Ages*, Ana Montoya investigated the Paston Letters...). At this point, they all thought they were applying technology to philology to guarantee rigorous research.

Before that, in 2001 and in a sudden way, Isabel had become visually impaired. Doors seemed to be shut for a researcher who could not read but windows opened when the months of training and rehabilitation started. Rehabilitation for blind people includes computers and technology but most materials were not accessible yet as they were not in a machine-readable format that could turn that information into voice. After a couple of years and many meetings in that tiny office, the idea of doing something relating to machine-readable texts appeared in front of us. There is no way one can tell the precise day but at some point in 2003 the idea of compiling an electronic corpus became real.

The factors contributing to this, besides Isabel's need for technological support for simple daily activities, were various. One was an increasing interest of the junior researchers for periods in the History of English beyond the Middle Ages. Another relevant factor was the knowledge we had of the Helsinki Corpus of English Texts (1991) and of other recently published or ongoing compilation projects such as the Lampeter Corpus of

Early Modern English Tracts (Reiner et al. 1998), the Corpus of Early English Medical Writing (Taavitsainen and Pahta 1997) or the ARCHER, first constructed by Biber and Finegan in the 1990s (Biber et al. 1994).

We saw that these corpora were important collections of linguistic material for the study of the development of English, but we also detected that they left some empty spaces either in terms of their specificity or in terms of the periods covered. In fact, we observed that none of them addressed the systematisation of late Modern English scientific disciplines and that the Helsinki team (then starting the compilation of LMEMT) were focusing on medicine only and had (at that moment) no public intention to continue their previous endeavours up to the late Modern English period. Therefore, we thought it was worthwhile to investigate the development of scientific English in different disciplines, excluding medicine as the Helsinki team was working with it and considering that both projects could complement each other and offer a broad picture on the history of English scientific writing.

The idea was to investigate the evolution of English scientific discourse paying special attention to the late Modern English period as we confirmed it was an understudied period in the history of English. With that, we could both contribute to increase knowledge on that particular period and delve into the specific features of scientific writing. It was at this moment that we first believed compiling a corpus could be an interesting research line and Isabel Moskowich applied for external funding through the research project "Etiquetación electrónica de textos científico-técnicos en lengua inglesa entre los siglos XVI y XX: Coruña Corpus" (2003-2006). We also thought it could be a good idea to do some research on a specific variety of English as the language of science and the then junior researcher Begoña Crespo applied for some funding at the University of A Coruña to develop this general idea: "Lengua y ciencia: Corpus de textos científico-técnicos en lengua inglesa (ss. XVI-XVIII)." This project was funded from 2003 to 2005.

Where to start? Initial challenges

Once we decided it would be good to cover the development of scientific English for the lapse of time that other corpora had left blank, we had to face several challenges. Initial design tasks were carefully undertaken so as to avoid inconsistencies. We drew the picture of what we wanted to do by establishing our principles for corpus compilation, the criteria we would follow.

¹ This stands for Research Group on Corpus-based Multidimensional Studies in English. The adjective multidimensional refers here to the multiple perspectives and dimensions of language analysis that can be adopted and not to Biber's approach alone. Besides, the resulting acronym, MuStE, with its medieval shade, reflected the determination of the people in the group to carry out whatever task.

We would compile a series of subcorpora, each dealing with one discipline, covering the late Modern English period. The delimitation of this period was a matter of careful deliberation and, although in its first conceiving the Coruña Corpus would cover the period 1650-1900, this would later evolve and published versions of finished subcorpora actually cover the period 1700-1900 to ensure we were providing texts published after the effects of the foundation of the Royal Society and the scientific method were visible.

As we consider corpora as small-scale mirror images of the register, variety or state of the language it aims at reflecting, it was also necessary to decide at this moment about the size of both subcorpora in general and individual samples in particular. There was not a clear criterion on these issues at that moment. Some compilers were gathering *samples in toto* (especially short samples) or samples with an arbitrary number of words and Biber (1993) had indicated that 1000-word samples were enough to detect variation in a representative corpus. However, we concluded that so short samples were not appropriate for such a specific corpus as ours. At that moment, we decided to set the limit in around 10,000 words per sample (as we claimed as early as 2004 during the LModE Conference in Vigo and later published in a chapter in Pérez-Guerra *et al.* (2007)) and two samples per decade and subcorpus, so that there would be 200,000 words per century and discipline.

Discipline selection and the concept of science. On what grounds?

One of the first problems that we found was the selection of disciplines itself. We decided that we would not build a simple inventory of disciplines but rather resort to one that was internationally accepted. We had recently applied for funding and one of the pieces of information we had to provide was the project's UNESCO code. This led us to search for the whole catalogue of the fields of science and technology UNESCO had published in 1988 and we included its use for discipline selection.

Knowing the importance mathematics had acquired after the scientific revolution, we wanted to start our compilation with CEMaT (Corpus of English Mathematical Texts) but samples did not present enough written text so as to allow for linguistic studies. In fact, samples were full of figures, tables and formulae, especially in nineteenth-century works. That forced us to search for more and more texts so as to cover two 10,000-word samples per decade. Getting material was

one of the main difficulties at the very beginning (more on this below) and, as finding texts was expensive and time-consuming and we lacked both sufficient funding and the necessary human resources, we decided to try with a different discipline. This is why, knowing that mathematics was behind all sciences from Newton onwards, Astronomy seemed to be a good choice. Nowadays, CEMaT has not been completed yet.

Discussions on the concept of science appeared again when reading early seventeenth-century texts on Astronomy, for instance, and we discovered that the contents there were not what we had expected. The theoretical debates were crucial for text selection depending on their contents. Finally, we decided to adopt an inclusive perspective, that is, to regard science as it was seen in every period. With this we avoided leaving aside certain texts that could be considered "scientific" in the eighteenth and nineteenth centuries, but not nowadays. This allowed the inclusion in CETA of samples where Astronomy and Astrology intermingle.²

Text-mining

Obtaining the samples was a challenging task, not only because finding the books and obtaining the permissions from the corresponding libraries and institutions was hard work but also because we needed to learn about the history of the discipline to be compiled in order to know what kind of problems scholars were facing at the time, which were the relevant names, if there existed any controversies, etc. Learning about the history of a particular science, however, requires learning about the history of science as such and this, in turn, required a deeper knowledge of the (external) history of English. Therefore, all this process consumed many hours of study accompanied by some research extending to late Middle English, to the period when some had claimed science started to appear in English.

In the early 2000s, when you wanted to read a book that was not at your own library, you could ask your librarian to use the interlibrary loan service. However, this did not apply, of course, to works in Special Collections and Archives such as those published between 1700 and 1900, the ones we needed. The answer often given by lending libraries was that the pages you wanted to read could be either microfilmed or photocopied and sent to you. But this was of little or no use as it was necessary to read the complete text in order to decide which extracts would be representative. This involved going to a particular library to access their collections and archives to know whether a particular work was valid.

² While looking for texts for CETA (Moskovich *et al.* 2012) we realised that we could not apply the UNESCO classification directly, as the eighteenth and nineteenth centuries had a different vision of science, a different taxonomy, to the one used nowadays, so we agreed to use the UNESCO classification of the fields of Science and Technology (1988) as a starting point rather than to follow it thoroughly.

Inés Lareo travelled to Berkeley and Berlin; Begoña Crespo and Isabel Moskowich travelled to London and Cambridge; Gonzalo Camiña stayed for some years in Cork and had the opportunity to revise some books there as well. Nuria Bello travelled to Edinburgh for the same purpose and other more recent members of the group (M^a José Esteve from Universitat Jaume I) also visited Glasgow. Although many texts are now available on the Internet in PDF format, as if the whole book had been photocopied or microfilmed, personal visits to libraries are still required: as late as 2019 Isabel visited the Special Collections and Archive section of the Library in the University of Liverpool and Luis Puente-Castelo took advantage of the collections at the University of Birmingham during his post-doc stay there.

Drawing from these experiences, we gradually came to the conclusion that our way of proceeding had to do with our concept of corpus compilation as it was much more than putting texts together. It required linguistic and computing skills but also historical and philological knowledge.

Creating a representative and balanced corpus

It is true that the two main virtues of a corpus are balance and representativeness. However, in the Coruña Corpus it was difficult to achieve both virtues at once: as explained above, it was our intention to compile corpora that could be mirrors of each disciplinary register at the period, but this made it difficult to achieve a balanced corpus on account of the different sociolinguistic variables used (mainly genre of the text and sex and geographical origin of the authors), as scientific writing was not that widespread at the moment.

The best examples of this were the obvious difficulties in the gathering of scientific writings by female authors. Generally speaking, during the late Modern period women were not allowed to learn about scientific issues and their “scientific activities” were mostly limited to those relating to everyday life in the household and anything that had to do with the female body: pregnancy, birth... Thus, it was not that surprising that it was so difficult to find published scientific texts authored by women dating back to the eighteenth century. Only a few privileged women could have access to learning scientific topics (Abir-Am & Outram 1987, Schiebinger 1987) and even fewer could write on that and publish their works under their own names (Herrero 2007). Add to this reality the complexity of mining (particularly eighteenth-century texts) we have already mentioned and you have a perfect storm of female invisibility. In any

case, the few women whose works have been included in the subcorpora published so far do represent the reality of the period and, hence, make sure that the corpus reflects the social peculiarities of those times and that they are, indeed, representative.

However, we have made up for this problem by providing extensive information about each of the authors. It is a characteristic element of the Coruña Corpus since the beginning of the project (perhaps due to previous research and training) that both staff and students thought that language could not be studied in isolation from its users. Just as the socio-historical events in the English-speaking world may have affected the development of scientific English, we were sure that events in the individual lives of authors, especially all that was related to their training and education, may have an influence on their linguistic habits too.³ We also thought that studying in Cambridge or in Dublin or reading such and such works had to have an impact on the linguistic habits of particular authors. On top of that, having information on their lives might be an asset in the depiction of the socio-external circumstances behind late Modern scientific English, interesting also for historians of science. This information is the one contained in the accompanying metadata files which have proved one of the wisest choices in our project.

XML format

The technical aspects of the corpus posed a dilemma, as we had the choice to take two diverging paths in our quest: the first and easier would mean to stick to the prevailing trend back in 2004, which implied transcribing originals into raw .txt samples and then rely on a search tool that would do its best to make searches based on the resulting .txt files; the second was more challenging and involved encoding the texts using the incipient eXtended Markup Language (XML), which meant a lot more work, but promised greater flexibility to further extend our range of possibilities when searching for diverse linguistic items in the samples.

In an act of sheer determination, we decided to follow the second path. But since XML was *terra incognita* for us, we wisely adhered to the existing text coding standard, the so-called Text Coding Initiative (TEI), so we would check the validity of our files against a pre-defined set of rules laid out in a document called *teixlite.dtd* (where *dtd* stands for Document Type Definition), which restricted the virtually unlimited, and often misleading, freedom offered by XML. Needless to say, no one in the research group knew how to code in XML, so Gonzalo Camiña

³ At this point, we saw it was useful to have some knowledge of the Middle English period and the scriptoria: we knew one could identify the particular place where a particular scribe had acquired his writing abilities.

volunteered to self-teach the new language first and then train the rest of the team.

Once the mark-up language was set, typing the first corpus posed yet another obstacle, as many characters in the original books were rather uncommon and, for this reason, difficult to represent because everyday fonts did not include them. This is the case, for example, of the long <s>, <f> and ligatured <ct>, <ct> in eighteenth-century texts, along with various other graphic symbols present in many samples of the Coruña Corpus. This could only be solved by resorting to specific fonts provided by Unicode.org. In short, before we ever managed to transcribe a single line of our corpus, we had already committed to using three conventions: XML, TEI and Unicode at a moment when more experienced corpus compilers were adamant to .txt. But there's more.

Since we wished to provide extralinguistic information about our authors and samples, XML alone would not suffice to compile the corpus. Linking the samples with the data referring to authors such as sex, place of education and date of birth required yet another type of coding called eXtensible Stylesheet Language Transformations (XSLT), which basically modifies existing XML files. This is when we considered liaising with IrLab, a research group on Information Retrieval at UDC, for some advice and even some programming, please.

In the meantime, seeing that the resulting tagged samples were a pain to the eye, as XML tags scattered all over the texts made them difficult to understand and spotting errors in the code became even more difficult, we were compelled to beautify the way the samples appeared on screen by implementing a Cascade Style Sheet (CSS). Thus, users of our corpus could see “clean” and easy-to-read versions of the texts that respect the old-fashioned spelling paragraph structure of the original documents.

Our experience with early domain-specific texts in English showed us that variation would be better approached in microscopic studies. That is why the corpus was intentionally designed to be used for analyses that do not completely rely on automatic searches, counting and statistical tests, but for those that require close reading of texts, manual disambiguation and interpretation. Our aim and another of our challenges was to create a corpus that helped solving the philologist's dilemma (Rissanen 1989) as our experience with domain-specific early texts in English had taught us that the lack of standardisation was also a factor to be taken into account. However, as the application of a search engine could very much ease the task of searching, the creation of a corpus query tool was seen as necessary. Besides, since it was also our intention to render versions of late Modern English texts where,

for instance, orthographic variation could be studied as well, we wanted to produce those special characters with Unicode and we also wanted a software able to deal with them. This “philological” desire gave as a result the need not to use plain text formats but something else as well as to design our own software able to manage, display and search such rich texts. At this point, we also wanted the tool to be able to select samples in our corpora by resorting to certain external and sociolinguistic variables, the ones in the metadata files accompanying each sample.

We were all happy and excited about the outcome of our technical decisions until the moment Javier Parapar, at that moment a pre-doc in Computing (from IrLab) and attending one of Isabel's PhD courses at the UDC and our first real programming expert, produced a first working version of the Coruña Corpus Tool (CCT). To our surprise, our first searches showed strange results. For example, the CCT interpreted math variables (A) as articles (a), while roman number one (I) magically became the first-person pronoun (I). Also, the tag we used to delete text quoted by authors, and therefore not representative of their own language, ended up removing chunks of content that would make understanding texts easier. Fixing this took us a while, because apparently minimal changes required tweaking the whole corpus, text by text. Eventually, we devised a reasonable solution and laid out the editorial tags described in the editorial policy included in the manuals to all the subcorpora and also published later by Camiña and Lareo (2016). However, it has become a bone of contention among members of the research group, as some of them see editorial tags simply as a temporary solution till further XML tagging is implemented in the samples, which would in turn demand further work in the CCT. And then more work on the samples...and then...

The maturity of the Project

After this initial process of planning and design, the project advanced slowly but surely. Isabel Moskowich successfully applied for external funding in several public calls both by the Spanish and the Galician autonomous governments up to the present day (2020-2022),⁴ as did Inés Lareo with her project “El Coruña Corpus: Compilación y análisis de textos de Astronomía. El uso de los predicados complejos en textos científico-técnicos en el período del inglés moderno,” funded by the Provincial Government of A Coruña from 2006 to 2008.

In 2009, we finished our first subcorpus, CETA, the Corpus of English Texts on Astronomy and started working on possibilities for its publication, finally deciding on starting

⁴ For the completion of these projects we had the opportunity to count on researchers from other universities as well: María-José Esteve-Ramos (UJI), Francisco Alonso Almeida (ULPG), Marina Dossena (U. Bergamo), Margarita Mele Marrero (ULL) and Andrew Hardie (U Lancaster).

our fruitful collaboration with John Benjamins, which would culminate in its publication in 2012 (Moskowich et al. 2012) together with a book (Moskowich and Crespo 2012).

At the same time, some of our students (who had become full part of the team) during the process of setting up the Coruña Corpus Project finished their dissertations in which they used the very corpora they were helping compile. Such is the case of Inés Lareo, Gonzalo Camiña, Iria Bello Viruega, Estefanía Sánchez Barreiro or Paula Lojo. MuStE is a dynamic group and new members joined it, such as Leida María Monaco, Luis Puente-Castelo or Anabella Barsaglini, some of which have since finished their dissertations as well.

However, even as things changed and results started to appear, the tiresome reality of compiling the corpus did not really change: The process still consisted in finding a huge amount of texts, selecting among those we deemed valid to achieve a representative set of texts, typing and XML marking each of them and then subjecting the texts to several rounds of careful revision so that any mistakes are corrected and the XML version is a faithful representation of the original. This is a task that the members of the team still take personal responsibility for, as the editorial decisions involved require careful consideration and a deep knowledge of the period and of the scientific register.

Most of this process remained (and still remains) the same throughout the project, but text mining became somewhat easier with the expansion of copyright-free online repositories such as Internet Archive, which allow us to check the suitability of potential texts before actually travelling to libraries to obtain them.

Expansion and popularisation

At the same time, the members of our team were travelling all over Europe presenting both the Corpus and their work on it in Conferences and Workshops such as Corpus Linguistics (Lancaster, Birmingham, Cardiff), CILC (from its first Conference in Murcia in 2009), ICAME, ISLE, etc., and also establishing working relationships with some of the best research groups on Corpus Linguistics during their doctorate stays in Universities such as Lancaster, Birmingham or Liverpool, among others.

Fruit of this effort, the Coruña Corpus became well-known among researchers as a very useful resource for the study of late Modern English scientific register and we had the pleasure to count with the contributions of scholars such as Douglas Biber, Bethany Gray, Joan C. Beal, Marina Dossena, Andrew Hardie, Pascual Cantos, David Banks, Stephania Degaetano-Ortlieb, María-Luisa Carrió Pastor, Katrin Menzel and Elke Teich in the different books we were publishing, for, after CETA

(Moskowich et al. 2012), we finished and published the Corpus of English Texts on Philosophy (Moskowich et al. 2016a) together with its book (Moskowich et al. 2016b).

All throughout this project, a series of issues such as the concept of science and the differences among text-type, genre and register have remained as topics of discussion both within the group and in different fora and particular work addressing them has been produced by members of the team.

Currently the CC ...

Nowadays in 2020, the Coruña Corpus is a fully-fledged project, with a very important body of publications (<https://www.udc.es/grupos/muste>) and now both starting the compilation of its seventh subcorpus and working in the publication of its fourth, CELiST (Lareo et al. forthcoming). Even though we still follow the same principles that were set up in the old days, we, nevertheless, have introduced some new developments over the last five years.

Conscious of the necessity of providing value-for-money in the investments the public governments were making on us, from 2019 we decided to publish our corpora in open-access. After quite a lot of searching, we decided that the easiest path to follow was to publish them in our University's institutional repository (RUC) which we first did with the Corpus of History English Texts (CHET), (Moskowich et al. 2018) accessible at <https://doi.org/10.17979/spudc.9788497497091>.

We also held discussions with John Benjamins to reissue in open-access the corpora on astronomy and philosophy already published in CD-ROM, which they kindly agreed on, leading to the re-issue of CETA (Moskowich et al. 2019a, <https://doi.org/10.17979/spudc.9788497497084>) and CEPHiT (Moskowich et al. 2019b, <https://doi.org/10.17979/spudc.9788497497077>) as well as to the publication of new accompanying books for CHET (Moskowich et al. 2019c) and CELiST (Moskowich et al. forthcoming).

Some forked projects also appeared, leading to the first steps towards the development of two different POS-tagged versions of the Coruña Corpus, by Andrew Hardie at Lancaster University and Stephania Degaetano and colleagues at Universität Saarland.

The Coruña Corpus Tool received some upgrading as well, making it both more reliable and faster and allowing for the use of several subcorpora at once (Barsaglini and Valcarce 2020).

However, the old process of “discovering some difficulties-finding a solution-correcting all the corpora” still plagues our development. A good example of this is one of our main headaches at the moment: During the

compilation of the corpora on Chemistry and Linguistics we discovered that subindices were not recognised by the CCT and that some texts talked substantively about punctuation symbols, which the CCT (as many concordance programs) was coded to disregard. This led to a still ongoing search for a solution which would change the treatment of subindices so that H₂O is read as a single token by our tool, and which would offer an

option to mark up those cases in which punctuation is used substantively and thus not to be ignored.

Such a cyclical rhythm of development is sometimes frustrating, but at the same time it is an integral part of our project. For, if the Coruña Corpus is a living project, it is precisely because it keeps offering such new challenges we have to face and overcome.



REFERENCES

- ABIR-AM, Pnina and Dorinda Outram. 1987. "Introduction". In Abir-Am & Outram 1987, 1-16.
- ABIR-AM, Pnina and Dorinda Outram, eds. 1987. *Uneasy Careers and Intimate Lives: Women in Science (1789-1979)*. New Brunswick, NJ: Rutgers UP.
- BARSAGLINI, Anabella and Daniel Valcarce. 2020. "The Coruña Corpus Tool: Ten years on". *Revista de Procesamiento del Lenguaje Natural* 64: 13-19.
- BIBER, Douglas. 1993. "Representativeness in Corpus Design". *Literary and Linguistic Computing* 8/4: 243-257.
- BIBER, Douglas, Edward Finegan and Dwight Atkinson. 1994. "ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers". In Fries et al. 1994, 1-13.
- CAMIÑA-RIOBOÓ, Gonzalo and Inés Lareo. 2016. "Editorial policy in the Corpus of English Philosophy Texts: Criteria, conventions, encoding and other marks." In Moskowich et al. 2016a, 45-60.
- FRIES, Udo, Peter Schneider and Gunnell Tottie, eds. 1994. *Creating and using English language corpora: Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993*. Amsterdam: Rodopi.
- HERRERO, Concepción. 2007. "Las mujeres en la investigación científica". *Criterios* 8: 73-96.
- LAREO, Inés et al. Forthcoming. *Corpus of English Life Sciences Texts (CELiST)*. A Coruña: Universidade da Coruña. <https://doi.org/10.17979/spudc.9788497497848>.
- MOSKOWICH, Isabel and Begoña Crespo. 2007. "Presenting the Coruña Corpus: a collection of samples for the historical study of English scientific writing". In Pérez-Guerra et al. 2007, 341-357.
- , eds. 2012. *Astronomy 'playne and simple'. The writing of science between 1700 and 1900*. Amsterdam: John Benjamins.
- MOSKOWICH, Isabel, Inés Lareo and Gonzalo Camiña-Rioboo, eds. Forthcoming. *"All families and genera": Exploring the Corpus of English Life Sciences Texts*. Amsterdam: John Benjamins.
- MOSKOWICH, Isabel et al. 2012. *Corpus of English Texts on Astronomy (CETA)*. Amsterdam: John Benjamins. <https://doi.org/10.17979/spudc.9788497497084>.
- , eds. 2016a. *'The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy*. Amsterdam: John Benjamins.
- 2016b. *Corpus of English Philosophy Texts*. Amsterdam: John Benjamins. <https://doi.org/10.17979/spudc.9788497497077>.
- 2018. *Corpus of History English texts (CHET)*. A Coruña: Universidade da Coruña. <https://doi.org/10.17979/spudc.9788497497091>.
- 2019a. *Corpus of English Texts on Astronomy. A Coruña*: Universidade da Coruña. <https://doi.org/10.17979/spudc.9788497497084>

- 2019b. *Corpus of English Philosophy Texts*. A Coruña: Universidade da Coruña. <https://doi.org/10.17979/spu-dc.9788497497077>
- , eds. 2019c. *Writing History in Late Modern English, Explorations of the Coruña Corpus*. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.225>.
- PÉREZ-GUERRA, Javier et al., eds. 2007. *“Of Varying Language and Opposing Creed”: New Insights into Late Modern English*. Bern: Peter Lang
- RISSANEN, Matti. 1989. “Three Problems Connected with the Use of Diachronic Corpora”. *ICAME Journal* 13: 16-19.
- SCHIEBINGER, Londa. 1987. “The history and philosophy of women in science: A review essay”. *Signs* 12 (2): 305-332.
- SIEMUND, Reiner. 1998. *The Lampeter Corpus of Early Modern English Tracts*. Oxford Text Archive [Accessed online on July 27, 2020].
- SIEMUND, Rainer and Claudia Claridge. 1997. “The Lampeter Corpus of Early Modern English Tracts”. *ICAME Journal* 21: 61-70.
- TAAVITASAINEN, Irma and Päivi Pahta. 1997. “Corpus of Early English medical writing 1375–1750”. *ICAME Journal* 21: 71-78.