RESEARCH PAPER | LINGUISTICS

ISABEL MOSKOWICH Y BEGOÑA CRESPO
Universidade de A Coruña

MuStE: the Dimensions of Linguistic Research at UDC

Research Project: Grupo de investigación Research Group for Multidimensional Corpus-Based Studies in English

MuStE stands for Research Group for Multidimensional Corpus-based Studies in English. Of course, it is also an old-fashioned form of the verb must, which may reflect some of the characteristics of the people in the team. The group's logo also represents history and time —from old hand-written letters to newest computer fonts—, one of the main axes of the research carried out: diachrony.

The group was formally constituted in 2005 when the University of A Coruña, where it belongs, demanded that research should be clearly organised and officially recognised once the requisites for the creation of research groups were fulfilled. MuStE, as is the case with other groups at the University of A Coruña, is very dynamic in what refers to its composition. At the moment of writing this report it is formed by four Faculty members, three of them PhD, one postdoc researcher, three PhD students and eight occasional collaborators from other institutions. That is, MuStE is a small but very active group with different research lines, all of them stemming from the idea that language cannot be viewed in isolation from its speakers.

Among those lines, and within the frame of language change and variation, lexical as well as semantic aspects in the evolution of English have been paid attention to, especially their development during the Old and Middle English periods. This research line is well represented by Crespo (2002), Crespo and Moskowich (2004), Crespo and Moskowich (2005), Moskowich and Crespo (2007), Crespo (2008), Crespo (2013) and Crespo (2016).

Other aspects of the medieval stages of English were also studied in some depth, as is the case of language contact of Old and Middle English with other linguistic varieties (Moskowich and Seoane, 1995; Moskowich and Seoane, 1996; Crespo, 2000; Moskowich, 2012). The interest at this point was to show how the lexicon of English coming from Old Norse or Latin revealed a different view of the relation between their respective speech communities.

Derivational morphology is another of the lines developed by the team. However, and as already mentioned, our conception of language as a living being that depends on its environment, led us to consider morphological processes as socio-historical context dependent. Therefore, aspects such as the etymological origin of both bases and affixes have been taken into account in many of our publications (see, for instance, Crespo and Moskowich, 2005-2006; Moskowich and Crespo, 2006; Moskowich, 2010; Crespo, 2011a; Crespo, 2011b; Moskowich, 2012).

Some of the researchers in the group have devoted their efforts to the study of syntactic topics also in more recent times. Thus, word order within the phrase was studied in Moskowich and Crespo (2002) and Moskowich (2002). The analysis of nominalisations was addressed by Bello (2016), complex predicates —mainly the structures called collocations in Mel'čuk's (1994) terminology— were dealt with in Lareo and Esteve (2008) and Lareo (2009) or conditional structures were delved into by Puente-Castelo and Monaco (2013) or Puente-Castelo (2016).

Our interest for the socio-historical dimension of the English language has recently grown into several and gradual forays into the wide field of discourse analysis. In those we have studied written texts from various discursive perspectives such as stance, persuasion, abstraction, involvement, modality and women's scientific writing. The triggering effect of all this was the creation of what has been and still is MuStE's flagship, the *Coruña Corpus of English Scientific Writing* (CC for short). Designed to be a generic or specific corpus —as opposed to a general corpus—, it is now well known and respected within the academic community. An electronic corpus is not a mere juxtaposition of texts —as sometimes understood in the field of literary studies. It is not a simple bunch of scanned images either as these formats cannot possibly be read and processed by a computer. On the contrary, the same as Biber (1993), Meyer (2002) and Crystal (2003), we agree that a corpus should be briefly defined as a "principled" collection of machine-readable texts.

The truth is that the idea of creating a corpus, a specialised one focusing on scientific English, first arose in 2003 when some members of the MuStE group were awarded funding from the University of A Coruña to explore the historical background of English as the language of science. We soon realised that the compilation of a corpus of scientific texts from the eighteenth and nineteenth centuries would fill a gap in the field of English historical linguistics. At that moment, we had the examples of the *Helsinki Corpus of English Texts* (Rissanen et al. 1991) and the *Lampeter Corpus of English Tracts* (Schmied et al, 1999). In Helsinki, Prof. Taavitsainen and her colleagues were working on the compilation of MEMT (*Middle English Medical Texts*) and we thought our corpus would complement theirs in the history of scientific English as, initially, the Helsinki project was intended to cover the Middle Ages and the early Modern period, focusing on medical texts.

Another contextual characteristic in the development of the CC worth mentioning is that in the early years of the twenty-first century it was infrequent to find linguists that were at the same time computer specialists, as computers had come into our lives only one decade earlier and we were hardly coming out of a MS-DOS-based universe. We nevertheless decided that we would like to compile a well-structured corpus and we spent a couple of years thinking about its design.

This design began to be tested while the group was searching for the necessary samples in the different libraries worldwide —the INTERNET was not what it is today— and it was precisely during the compilation process that we detected some technical and non-technical barriers to overcome. This obviously forced the gradual introduction of changes in the original design. Ours was a flexi-

ble design. That allowed us to compile ca. two hundred thousand-word subcorpora with the same structure, each devoted to a particular discipline in the realm of scientific knowledge by adopting an inclusive perspective. From then on, any scientific discipline, except for medicine, is welcome to be included in the CC. Yet, we are working on a symmetrical compilation of disciplines —from both the Soft and the Hard sciences— to generate a final balanced product that allows for comparison among sister corpora.

As the body of material gathered grew, we also had to take some decisions about how and where to store our samples. For that, we needed a protocol to record what was being done, who was doing it and in which stage of the whole process the corresponding file was. Since different scientific disciplines were stored in different folders, one spreadsheet per discipline was used to keep records. The fields in each spreadsheet grew in number and became more sophisticated in order to include all the variables —about the text and its author, the state of the file, etc.— that would help us keep track of all details in a quick and efficient way. Each sample had to be readily identified as unique.

Of course, initially, we had to contact many libraries asking for permission to reproduce extracts of works. On many occasions, once the fragments —in paper— were in our hands we detected they were not suitable to be included in the corpus as they did not comply with the CC compilation principles. One of them required compiling ten-thousand word samples as we detected that shorter extracts would not be very useful for language analysis in the period (Crespo and Moskowich, 2010).

Those extracts that could be included had to be processed so as to flee from the photocopy and reach a computer-readable file. Fortunately, nowadays the Internet provides us with .pdf files and paper copies are no longer needed. The corpora that we knew were formed by .txt files but we wanted to go one step further and decided not only to encode our samples as .xml files, but also to follow the Text Encoding Initiative (TEI). That was as early as 2007. Since then, the use of TEI and the ten thousand-word samples have been adopted by other gr oups.

Once all these parts of the process were completed, we realised we had been following the five steps mentioned by Kennedy (1998, 70-85) for corpus compilation:

1. Corpus design

2. Planning a storage system and keeping records

3. Obtaining permissions

4. Text capture

5. Markup

As a result, we have found that scientific writing in general is not as objective as initially thought of. Such is the case of stance markers of different sorts, mainly adverbs —*perhaps, indeed*— , modal verbs or even personal pronouns that demonstrate that our idea of an object-centred, aseptic scientific discourse may not be completely true. Some pilot studies showed that women often have to resort to certain linguistic features —in the case of persuasive strategies— that seem to function as an over-reaction as they must convey scientific knowledge —and convince their readership— in an androcentric world. Moreover, the kind of strategies women use tend to be more subtle than those used by men, more direct in their use of language. In the same vein, we have also found, for instance, that women, considered to be more sensitive than sensible, indeed use lots of structures typical of a highly-abstract frames of mind. May this serve as an example of ongoing research.

The Coruña Corpus project has not always received funding but we have advanced in its compilation, although more slowly at times. Typing and xml-encoding late Modern English texts that are revised three times each, gathering information about each author and the work to prepare xml metadata files that allow searches by variables, working with Information Retrieval researchers to design a specific search engine able to discriminate eighteenth-century spellings is not easy but time-consuming. On top of that, as researchers, we are evaluated by what we publish, what makes academic life even more complex.

The experience in corpus compilation gained by the team has also brought about the creation of other corpora covering nearby areas of study in modern science or that could provide a deeper analysis on aspects already signposted in the CC.

One of the variables that we have been interested in exploiting is that of the sex of the author in order to find out whether men and women showed different communicative strategies when writing science. It was precisely this interest that gave rise to the Prefaces of Women Writers of Science (Crespo, forthcoming) or PreWoS project, still under way, in which the aim is to compile prefaces of scientific works by some of the women writers selected for the CC, together with many others.

Another idea, still in its design phase, is that of compiling a Corpus of (Pseudo)scientific Language (Puente-Castelo, forthcoming), currently called like that. Somehow following the CC, it will contain six twin subcorpora, all with the same design and principles of compilation, and one per each of the pseudoscientific doctrines: Homeopathy, Antivaccination, Climate change denialism, Flat-earth movement, Creationism and Holocaust denialism. The idea is that this corpus should contain texts from the last thirty years and belonging to a wide variety of genres.

There is another work in progress, carried out by Barsaglini-Castro (forthcoming), which consists of a corpus that comprises a carefully planned selection of 50 texts from 1950 to 2017. Those are both fiction and non-fiction texts about transhumanism, posthumanism, transcendence, technology and artificial intelligence. Fiction texts include 18 sci-fi novels whereas the non-fiction section has 16 articles and 16 book chapters.

In terms of criteria, non-fiction texts have been chosen randomly but considering balance and covering fields such as education, philosophy, medicine, technology, and life sciences. It contains around one million, eight hundred thousand words from texts written by male and female authors.

As happens with all MuStE's work, all these corpora pay special attention to the principles of representativeness and balance and honestly try to be thorough. For the future, we intend to continue applying the principles of rigorousness, honesty and hard work to the compilation of new CC subcorpora as well as to the writing of academic papers that contribute to the study of the English scientific discourse from a historical perspective.

REFERENCES

Barsaglini-Castro, Anabella. Forthcoming. *Posthumanism and the Art of Persuasion: How Stance, Hedging and Stylistics Influence the Reader*. PhD Dissertation. A Coruña: Universidade da Coruña.

Bello, Iria. 2016. "Nominalizations and female scientific writing in the late Modern period". *Revista Canaria de Estudios Ingleses*, 72 (1): 35-52.

Biber, Douglas. 1993. "Representativeness in Corpus Design, Literary and Linguistic Computing, 8 (4): 243-257.

Britton, Derek. 1994. *English Historical Linguistics.* Amsterdam: John Benjamins.

Clas, André and Pierrette Bouillon. 1994. *TA-TAO recherches de point et applications immédiate*s. Montréal: Les Presses de l'Université de Montréal.

Crespo, Begoña and Isabel Moskowich. 2004. "Enlarging the Lexicon: The field of technology and administration from 1150 to 1500". *Studia Anglica Posnaniensia*, 40: 163-180.

—. 2005-06. "Medicine, Astronomy, Affixes And Others: An Account Of Verb Formation In Some Early Scientific Works". SELIM, 13: 179-198.

—. 2005. "Latin Forms in Vernacular Scientific Writing: Code-Switching or Borrowing?" In McConchie et al. 2005, 51-59.

—. 2009. "CETA in the context of the Coruña Corpus". *Literary and Linguistic Computing*, 25/2: 153-164. doi:10.1093/llc/fqp038.

—. 2009. "The limits of my language are the limits of my world: the scientific lexicon from 1350 to 1640". *SKASE Journal of Theoretical Linguistics*, 6/1: 45-58.

Crespo, Begoña. 2000. "Historical Background of Multilingualism and its Impact on English". In Trotter 2000, 23-35.

—. 2002. "A preliminary approach to the semantic study of socio-economic terms in the History of English". *Quaderni di Semantica*. 257-272.

—. 2008. "Specific and non-specific nouns in late Middle English: when Robert grows from man to herb". *English Studies*.

—. 2011a. "Rosewater, Wheel of Fortune: Compounding and Lexicalisation in Seventeenth-century Scientific Terminology". *Nordic Journal of English Studies*, 10 (1): 135-153.

—, ed. 2011b. "A Study on Noun Suffixes: Accounting for the Vernacularisation of English in Late Medieval Medical Texts". *Linguistik Online* 57 (7): 27-42.

—. 2013. *Change in Life, Change in Language: A Semantic Approach to the History of English*. Bern: Peter Lang.

—. 2016. "Specialised language varieties: when a cognitive framework can explain semantic changes". *Anuari de Filologia. Estudis de Lingüística*, 6: 63-83.

—. Forthcoming. *PreWoS: A Corpus of Prefaces to Women Scientists Works*.

Crystal, David. 2003. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge U P.

REFERENCES

Kennedy, Graeme D. 1998. *An introduction to corpus linguistics*. London: Longman.

Lareo, Inés and María José Esteve-Ramos. 2008. "18th Century Scientific Writing: a Study of make Complex Predicates in the Coruña Corpus". *ICAME*, 32: 69-96.

Lareo, Inés. 2009. "Make-collocations in nineteenth-century scientific English". *Studia Neophilologica*. 81 (1): 1-16.

McConchie, Rod et al. 2005. *Selected Proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (Hel-Lex)*. Somerville: MA Cascadilla P.

Mel'čuk, Igor. 1994. "Les fonctions lexicales dans le traitement du language naturel." In Clas and Bouillon 1994, 193-219.

Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge U P.

Monaco, Leidamaria. 2016. "Cognitive implications of nominalizations in the advancement of scientific discourse". *International Journal of English Studies*, 16 (2): 1-23.

Moskowich, Isabel and Begoña Crespo. 2002. "Adjectival forms in Late Middle English. Syntactic and Semantic Implications". *Studia Neophilologica*, 74: 161-170.

—. 2006. "Lop-webbe and henne cresse: Morphological Aspects of the Scientific Register in Late Middle English". *Studia Anglica Posnaniensia*, 42: 133-145.

—. 2007. "Different paths for Words and Money: The Scientific field of *Commerce and Finance* in Middle English". In Moskowich and Crespo 2007, 101-115.

—. 2007. *Bells Chiming from the past. Cultural and Linguistic Studies on Early English*. Amsterdam/Philadelphia: Rodopi.

—. 2012. *Astronomy "playne and simple": The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins.

Moskowich, Isabel and Elena Seoane. 1995. "The Lexical Scandinavian Element in Early Modern English. Some Preliminary Considerations". *Neuphilologische Mitteilungen*, 4 (XCVI): 399-415.

—. 1996. "Scandinavian Loans and Processes of Word-Formation in ME: Some Preliminary Considerations". In Britton 1994, 185-198.

Moskowich, Isabel et al. 2016. *"The Conditioned and the Unconditioned": Late Modern English Texts on Philosophy*. Amsterdam/Philadelphia: John Benjamins.

Moskowich, Isabel. 2002. "The adjective in English. *The French type* and its place in the history of the language". *Folia Linguistica Historica*, 23 (1-2): 59-71.

—. 2010. "Morphologically complex nouns in English Scientific Texts after Empiricism". *Linguistik Online*, 43 (3).

—. 2012. "Patterns of English Scientific Writing: adjectives and other building-blocks". In Moskowich and Crespo 2012, 79-92.

—. 2012. *Language contact and vocabulary enrichment. Scandinavian elements in Middle English*. Bern: Peter Lang.

Puente-Castelo, Luis and Leidamaria Monaco. 2013. "Conditionals and their functions in Women's Scientific Writing". In Vargas-Sierra 2013, 160-169.

Puente-Castelo, Luis. 2016. "Explaining the use of If...then... structures in CEPhiT". In Moskowich et al. 2016, 167-181.

—. Forthcoming. *The Corpus of (Pseudo)scientific Languag*e.

Rissanen, Marri et al. 1991. *The Helsinki Corpus of English Texts*. Department of Modern Languages, University of Helsinki.

Schmied, Josef, Claudia Claridge and Rainer Siemund. 1999. The Lampeter Corpus of Early Modern English Tracts. ICAME Collection of English Language Corpora (CD-ROM), Second Edition, eds. Knut Hofland, Anne Lindebjerg, Jørn Thunestvedt, The HIT Centre, University of Bergen, Norway.

Trotter, David. A. 2000. *Multilingualism in Later Medieval Britain*. Cambridge: D.S Brewer.

Vargas-Sierra, Chelo. 2013. *Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013)*. Vol 95, *Procedia. Social and Behavioral Sciences*.