Corpus studies of language through time

Introduction to the special issue

Tony McEnery,[i, ii] Gavin Brookes,[i] and Isobelle Clarke[i]
[i] Lancaster University | [ii] Xi'an Jiaotong University

The study of language through time has long been an area where the corpus approach to the analysis of language has been an important method. The possibility of using other methods, such as elicitation, introspection or psycholinguistic experiments to investigate language in the past is, effectively, eliminated by a simple fact – there is no direct access possible to speakers of language beyond those generations that are living. Our only access to them is through the traces of language they left – recordings, for language spoken in the past 160 years or so or, more commonly, written records. That access is deeply skewed. For some languages, such as English, data is available which would, in principle, allow researchers to study the language in some depth through changing varieties over a long stretch of time. Language spoken in the past which had no written form or where written records have been lost in whole or part are, in effect, lost to us. Research on those languages and varieties is, essentially, impossible.

Yet the skew extends beyond simple existence or non-existence of records for some (nominal) standard form of a language. What was chosen to be written in the past, which texts it was deemed important to preserve, the varieties of the speakers with power who had access to the ability to have their language recorded, inter alia, are all factors which skew what we may be able to study when looking at language in the past. So even where we find that much exists, as is the case with English, more is lost. In this thinning-out of the totality of language in the past to the reality of what sources are available now, we also know that some things that we may consider to be constants impact on records, irrespective of language – the survival of texts is linked to cultural value, representations of speech are less commonly found than writing, and literacy levels are likely to greatly influence the types of texts produced. To pick up on that last point, consider a society in which literacy levels are low – the production of many literacy events will be limited accordingly. So, the production of letters to friends, the production of shopping lists and other aide-memoires, the keeping of diaries and even the production of written graffiti are all examples of forms of literacy which

https://doi.org/10.1075/ijcl.00050.edi
International Journal of Corpus Linguistics 27:4 (2022), pp. 393–398. ISSN 1384-6655 | E-ISSN 1569-9811

are, of necessity, suppressed in a society with low literacy levels. If the majority of people cannot write, then they need to hire scribes to produce such writing – adding to the expense of the task. Yet such expense would be pointless, or is enhanced, in a context in which the person for whom the message is intended is illiterate also and would thus have to find or hire someone to read the message. In short, in societies where literacy is widespread, the value of certain types of writing is boosted as the ability to produce and comprehend them becomes widespread.

Nonetheless, in spite of the skew pressing on the written record, a great deal of data survives for many languages, though the linguist needs, because of the skew in that record, to be cautious in drawing conclusions from it. However, the linguist also needs to either work piecemeal with the data, in a qualitative fashion, or to find or create digital versions of the data. Words on a page are not the stuff of which corpus linguistics is made. Unless those words can be rendered as machine readable text, then the archive remains a source of data only for those linguists who are willing to work directly with the written records using what we might term 'hand and eye' techniques.

This bottleneck is well known and pioneering work in the digital humanities and corpus linguistics has set about expanding that bottleneck for almost as long as data processing equipment that allowed for it has existed. From Roberto Busa's pioneering work to produce concordances of medieval Latin, which started in 1949 (see McEnery & Hardie, 2012: 37), continuous efforts have been made to digitise the texts of the past. A milestone in corpus linguistics was the production of the Helsinki corpus (see Rissanen et al., 1993). A team led by Matti Rissanen and Ossi Ihalainen at the University of Helsinki began this mammoth task in 1984. The corpus stretches from 750 to 1700 AD, covering Old English, Middle English and Early Modern English. It attempted to produce, through the texts chosen for the different time periods it covered, a representative sample of English through time. Given its size (1.5 million words), the corpus was always going to be of most importance in studying the most frequent aspects of the language. But for the researcher wishing to study such features, the fact that it was a machine-readable corpus, and that it had been collated with a claim to representativeness in mind, allowed users of the corpus to gain insights into English. Those insights, using hand and eye methods, would have been difficult to produce and, perhaps, harder to justify.

Historical corpora have pushed on a long way since the Helsinki corpus and now truly vast stores of information are available. For some languages, such as Chinese, the scale of digitized resources is great, though access to them is often highly limited (Zinin & Xu, 2020). For languages like English, French and Spanish, very large collections of data are easy to access in different time periods, including the recent past (for example, the Reference Corpus of Current Spanish, 1975–2004; https://www.rae.es/banco-de-datos/crea), and the more distant past (for example, the publicly available elements of the Early Modern French FreEMmax corpus; Gabay et al., 2022). In unlocking these rich seams of data, skew arises again of course – we get to study what those digitizing allow us to study and we need to be mindful of how their choices limit and define the research questions we may ask of the data. We also need to be aware of another issue – the fidelity of the data. When corpora such as the Helsinki corpus were produced, an emphasis was put on data quality. Painstaking efforts were made to make the fidelity of the transcriptions, in terms of them being a fair representation of the original texts for example, as high as possible. With some historical corpus sources, scale is achieved at the cost of fidelity. Some large sources of machine readable text, while not constructed with the intention of producing a corpus as such, nonetheless provide vast volumes of historical language data that can, with caution, be used. Probably the best example of this arises from the British Library's release of machine-readable texts derived from their nineteenth century newspaper collections. These have been produced using optical character recognition (OCR) and were originally intended not for corpus research as such, but to act as a way of allowing users to type in search words and to retrieve a scanned image of pages in newspapers in which that word appeared. The OCR text was used as a way to mediate between the user's search and the pages to be displayed. With such a task a degree of inaccuracy might be acceptable – if a word is mentioned 20 times on a page and only 5 examples have actually been rendered faithfully by the OCR, then it is still possible to use the OCR to determine that the page in question should be displayed as it contains at least one example of the word that the user is looking for. This is a way of assisting hand and eye analyses. To take the OCR data and use it as corpus data is fraught with risk, as though the dataset is huge – composed of complete runs of a large number of papers spanning the century, amounting to billions of words – errors are frequent. The challenge for researchers, given the low likelihood of the scanning of the texts being done again, is to make, as Nevalainen (1999) suggests pragmatically, "the best use of bad data". Hence the very nature of the data itself has become the spur for experimentation and methods development as researchers have worked, successfully, to overcome the limitations of the data and to get the benefit of the vast body of historical language data that OCR has unleashed (see Joulain-Jay, 2017, for an in-depth study of the limitations and potential of this data source).

It is in this context that this special issue, looking at time in corpus linguistics, has been produced. There has never been a better time to look at the past using corpus methods, but when we do so we must proceed with caution and accept that our methods are likely to be challenged by the data and will therefore need to adapt. Sometimes that adaptation will be required because of the nature of the data we are dealing with, sometimes because of its scale, and sometimes because of the unfolding possibilities of looking at language change through corpus data. But throughout, time itself is a controlling factor – a crucial variable to consider as we explore our data. In a research context that is dynamic and challenging, time is an ever-present issue.

In the first paper in this special issue, Clarke et al. look at changes in discourse over time. The data used deals with the recent past, but has the virtue of being well structured, composed of a range of newspapers, and plentiful. The density of mentions of the words examined in the study lends itself to a fine-grained examination of change over time which is an advance on previous, similar, studies. By harnessing techniques used to cluster short texts with similar patterns of co-occurring keywords, this paper shows how well, and in what detail, the impact of time on discourse can be observed.

Fitzmaurice and Mehl take a look at the issue of shifts in English discourse in the early modern era. They use a technique looking at quads of lemmas co-occurring in a wide window to gain insight into processes of change in discourse. By so doing, they find effects such as secular usage emerging from religious usage over time and find that vague meaning may act as an engine for such change.

Taylor looks at one issue, the representation of migration, over a very long period of time (over 200 years), in one continuous source, *The Times* newspaper. Taylor addresses some important issues that such studies face, for example the identification of stability as well as change, how to interpret results in a changing historical context, and the variable nature of the scale of data available across time.

Alexander and Struan examine a single source, the Hansard record of the UK parliament, and use that to explore the changing meaning and associations of the concept of the *uncivil* in two centuries of Parliamentary debate. The study aligns this investigation with a reference resource, the *Historical Thesaurus of English*, which is used by the analysts to answer a question at the root of corpus studies which take on a Protean aspect over time – what words to look for. By relying on the description of the concept of the uncivil in the *Thesaurus*, the study is able to deal with the task of analysing this shifting concept and its associated lexis.

Rodriguez-Puente et al. take a morphological perspective on change over time, exploring the unfolding competition of two suffixes, *-ity* and *-ness*, across a range of Early Modern English corpora. In doing so, they critically examine previous claims made about these suffixes. Yet to do so, they are required to innovate, both in terms of statistical analyses and data visualization. As a consequence, the insight they achieve into the use of these suffixes is far more robust than that provided by previous studies in this area.

Stifter et al. take us to the heart of some of the most difficult issues in the study of language through time: sparse evidence, difficulties in identifying the extent of the surviving archive, and challenges in understanding the language used. In this paper, an innovation designed to deal with the degree of variation in the data, Bayesian Language Variation Analysis, is introduced. This shows how, even in such a challenging context, the corpus method can permit insights into change over time.

There is so much more that could be done to look at the intersection of time and corpus linguistics than one issue of this journal can achieve. However, in this single issue we hope we have shown how challenging this area is, and yet how rich in promise and insights it is also.

References

Gabay, S., Suarez, P., Bartz, A., Chagué, A., Bawden, R., Gambette, P., & Sagot, B. (2022). From FreEM to D'AlemBERT: A large corpus and a language model for early modern French. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 13th Conference on Language Resources and Evaluation* (pp. 3367–3374). ELRA. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.359.pdf

Joulain-Jay, A. (2017). *Corpus Linguistics for History: The Methodology of Investigating Place-Name Discourses in Digitised Nineteenth-Century Newspapers*. [Doctoral dissertation, Lancaster University]. Research directory, Lancaster University. https://doi.org/10.17635/lancaster/thesis/143

McEnery, T., & Hardie, A. (2021). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Nevalainen, T. (1999). Making the best use of 'bad' data: Evidence for sociolinguistic variation in Early Modern English. *Neuphilologische Mitteilungen*, *100*(4), 499–533.

Rissanen, M., Kytö, M., & Palander, M. (Eds.) (1993). *Early English in the Computer Age: Explorations through the Helsinki Corpus*. Mouton.

Zinin, S., & Xu, Y. (2020). Corpus of Chinese dynastic histories: Gender analysis over two millenia. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference 2020* (pp. 785–793). ELRA. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.98.pdf

Address for correspondence

Tony McEnery
Department of Linguistics and English Language
Lancaster University
Bailrigg, Lancaster LA1 4YW
United Kingdom
a.mcenery@lancaster.ac.uk

Co-author information

Gavin Brookes
Department of Linguistics and English Language
Lancaster University
g.brookes@lancaster.ac.uk

Isobelle Clarke
Department of Linguistics and English Language
Lancaster University
i.clarke@lancaster.ac.uk