Research in English Language Pedagogy (2024) 12(3): 469-477

 DOI: 10.30486/RELP.2024.1125835

Review Article

Digital Corpora in Language Study: Reviewing a Success Story in the Recent History of Linguistics

Christian Mair[*]

English Department, Albert Ludwig University of Freiburg, Freiburg im Breisgau, Germany

Submission date: 03-07-2024 Acceptance date: 14-07-2024

Abstract

The contribution provides an overview of the fifty-year success story of corpus linguistics. It acknowledges the enormous technical advancements and the excellent corpus resources available today – at least for some of the major languages in the world. A direct consequence of technological development is the rise of usage-based theoretical models in contemporary linguistic research. At the same time, the contribution points out the still deficient treatment of spoken spontaneous language in corpus linguistics and recommends giving more attention to the development of multilingual corpus resources in the future than has been done so far. Finally, it raises the question of the future function of corpus linguistics within the framework of Digital Humanities.

Keywords: Corpus linguistics, Digital humanities, Multilingualism, Progress in linguistics, Spoken language

[*] Corresponding Author's E-mail: Christian.Mair@anglistik.uni-freiburg.de

1. Introduction

In modern corpus-linguistic research (for an introduction, see, for example, McEnery & Hardie, 2012), a corpus has traditionally been defined as a digitally searchable collection of texts compiled by linguists for the purposes of linguistic research. Such digital language corpora have two advantages. Traditional philological pursuits, such as finding examples of lexical or grammatical usage, can be carried out more quickly and systematically. The typical output format of such searches is a Keyword in Context (KWIC) concordance. In addition, and more importantly, the new resources and methods have made sophisticated methods of quantitative analysis of linguistic data possible. Corpus linguistics and concordances have a long tradition that dates back well before the era of digitization. However, without technical support, this type of textual analysis was so labor-intensive that it was rarely undertaken. Concordances were typically only created when – as in the case of the Bible or the works of Shakespeare – the religious or cultural value of the corpus justified the effort. Researchers like the American structuralist Charles Carpenter Fries, who dedicated the same care to a corpus of transcribed conversations of average Midwestern Americans to empirically ground his grammar of English (Fries, 1952), remained the exception. Fries' corpus comprised approximately 250,000 words. Somewhat later, Randolph Quirk conceived the extremely daring plan to create a one-million-word corpus of British English, initially not considering any further technical aids apart from a tape recorder and typewriter. However, the Survey of English Usage corpus was partially digitized during its completion in the 1970s (Svartvik & Quirk, 1980), making it both the last representative of corpus linguistics' medieval era and a pioneering achievement of the digital corpus linguistic modern era.

As it is well known, since those days, corpus linguistics has claimed a central place not only in English language studies but also in research on many other languages. In Section 2, I will recapitulate the essential stages of development through which ideas and tools developed by a small group of partially distrusted, partially ridiculed eccentrics on the margins of their respective disciplines arrived at the center of the disciplines within just three decades. Section 3 will address some downsides of this success story, while the concluding Section 4 will weigh the role of technology and the human factor in shaping scientific progress in linguistics. The contribution concludes with the assessment that corpus linguistics, precisely because of its success, must reposition itself in the coming years within

the general development in the field of digitization in the humanities and social sciences (Digital Humanities).

2. Oh Pioneers! – From the Margins to the Mainstream

The year 1964 is commonly regarded as the birth year of modern corpus linguistics in English studies, when W. Nelson Francis, Henry Kučera, and their small team completed the Brown Corpus – officially titled "A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers." The corpus primarily served as a data source for usage-based approaches to describing the American standard variety of English, which is why, during a phase dominated by the early generative linguistics of Chomsky and the sociolinguistics of Labov, it did not receive enthusiastic reception. For generativists, corpus data were unnecessary because they suspected the real challenges for scientific language description lay beyond collecting relatively arbitrarily selected and structurally unremarkable performance data. For sociolinguists, on the other hand, the written standard was the least interesting variety, as there was nothing to gain here besides some stylistic variation.

Further initiatives, such as the seemingly sensible expansion of the data basis with a British parallel corpus (the LOB [= Lancaster/Oslo-Bergen] corpus), progressed slowly so that the founding of ICAME (International Computer Archive of Modern English) in 1977 in Oslo rather was a promise on a great future rather than the presentation of present achievement. The marginal status of corpus linguists at the time can be gleaned from a telling episode in a best-selling contemporary campus novel, David Lodge's (1984) *Small World*. Apparently, John Sinclair, another pioneer of early corpus linguistics, provided the real-world inspiration for the character of the language technologist Robin Dempsey in the novel. Here, Dempsey presents his research program to an astonished visitor with a mixture of fanaticism and missionary zeal:

> I'd like to take you over to our Computer Centre this afternoon," he said. "We've got something set up for you that I think you'll find interesting." He was sort of twitching in his seat with excitement as he said it, like a kid who can't wait to unwrap his Christmas presents. […] "Anyway," he went on, "when we heard that the University was going to give you an honorary degree, we decided to make yours the first complete corpus in our tape archive." "What does that mean?" I said. "It means," he said, holding

> up a flat metal canister rather like the sort you keep film spools in, "It means that every word you've ever published is in here." His eyes gleamed with a kind of manic glee, like he was Frankenstein, or some kind of wizard […] "What's the use of this?" I asked. "What's the use of it?" he said, laughing hysterically, "What's the use? Let's show him, Josh." And he passed the canister to the other guy, who takes out a spool of tape and fits it to one of the machines. […] "What's your favourite word?" "My favourite word? I don't have one." "Oh yes, you do!" he said. "The word you use most frequently." "That's probably the or a or and," I said. "He shook his head impatiently. "We instruct the computer to ignore what we call grammatical words – articles, prepositions, pronouns, modal verbs, which have a high-frequency rating in all discourse. Then, we get to the real nitty-gritty, what we call the lexical words, the words that carry a distinctive semantic content. Words like love or dark or heart or God. Let's see. (Lodge, 2011, pp. 183-184)

Today's student readers recognize the basic constellation: searching for words in digitized texts. However, the details are surprising. Having to go to a computer center to conduct a trivial word search in a relatively small corpus, the published works of a single author? Data storage on a film-like medium? What is the technician for?

This brief passage from a novel illustrates the impressive technological advancements of corpus linguistics. Many of the problems Dempsey and his technician had to struggle with are solved today. At least for written language data, computer storage capacity has not been an issue for decades – although, unlike the pure text of early corpora, today's data typically come in XML-compatible format and with part-of-speech annotation. Access to most corpora is readily available – from the user's desk or mobile device. In the novel, the act of transferring text from the print to the digital format is revolutionary. Today, we have become used to born-digital raw material for compiling corpora, and there are few living authors around whose complete published works are not available in digital format anyway.

Progress in search queries and statistical analysis of data is also enormous. Searching digital text data is no longer specialized knowledge of a small avant-garde of experts but an essential cultural technique without which nobody can exist who considers themselves part of the modern world. In this sense, Dempsey's successors are found not only in the community of corpus linguists but far beyond. This includes anyone, for example, innocently using a search engine for information queries in databases, as well as the growing circle of

those working for various intelligence agencies in the widespread surveillance of global communication.

Corpus linguistic methods have revolutionized lexicology as well as practical lexicography within a very short time. In grammatical research, they were an essential prerequisite for the current boom in usage-based theoretical models. There have also been significant impulses for language history as well as for text linguistics and discourse analysis. These are questions that great scholarly minds considered impossible to answer 50 years ago. For an illustration, consider the great Dutch language historian, Fredericus Theodorus Visser, who about half a century ago complained, "Today begin + form in -ing is used with striking frequency alongside of begin + infinitive. Which of the two alternatives predominates cannot be ascertained because of the lack of statistical data" (Visser, 2002, p. 1888). Now, any intelligent undergraduate can go to the english-corpora.org website and produce the relevant statistics for a wide range of contemporary and historical varieties of English.

3. The Dark Side of Success

The success story described above certainly has its dark side as well. Corpora have become easily accessible, and unlike the early ones, their modern counterparts come with sophisticated annotation (part-of-speech-tagging, syntactic parsing, even semantic annotation) and integrated powerful search facilities. This is a tremendous advantage for expert users who are aware that no system of automatic annotation is error-free and know where the pitfalls are for their particular search queries. For inexperienced users, all this may be a black box, and they lack the capacity to assess the quality of their search output.

The most significant deficit in the history of corpus linguistics so far, however, is the handling of spoken data, where there is still much to be done for the next 50 years. The corpora collected in the International Corpus of English (ICE) family do contain spontaneous spoken language, but almost all of them offer the user only very simple orthographic transcription rather than access to the original audio recordings (see Greenbaum and Nelson 1996). It is thus not possible to investigate the grammar-prosody interface, and many other things besides. Newer standard corpora such as the Corpus of Contemporary American English (COCA) do offer significantly more transcribed spoken material, but mostly from secondary sources such as transcripts created by radio and television stations with the help

of linguistically mostly untrained personnel. Perhaps in ten years, we will reach a situation where every spoken corpus is available multimodally: as a video or audio file, aligned with machine support using an orthographic transcription enriched in additional versions with further grammatical and prosodic information (or can be enriched by users according to their own preferences).

The glaring underrepresentation of spontaneous speech data in corpora points to a general danger in the development of scientific infrastructures. What can be done without great technical and with reasonable financial effort is done. In recent decades, that has been the creation of ever larger and ever better annotated standard written language corpora.

Ultimately, the treatment of spoken language in corpus linguistics is a case in point for a more general problem. The development of research infrastructures always takes place in the tension between technical innovation and the further development of the theoretical frameworks of linguistic research. In this situation, we need a level playing field and dialogue between the people who develop the corpora and other digital resources for language study and the scholars and students who use them. If we do not promote this dialogue, we may end up with digital language corpora that distort linguistic reality rather than represent it.

In a world language such as English, which is widely used in multilingual contexts as a second language (ESL), foreign language (EFL) or lingua franca (ELF), this is a very real problem. Developers prefer monolingual corpora, because they are much easier to digitise, annotate and search than multilingual corpora. For the technicians compiling a corpus of Indian or Philippine English, it may, therefore, be an attractive strategy to target texts that do not have a lot of Hindi or Tagalog in them. But, is this an accurate picture of the use of English in India or the Philippines? In yet-to-be-created multilingual corpora of educated spoken usage in India or the Philippines, entirely new insights into a complex linguistic reality could be gained – by not marking the remaining borrowings from the local contact languages as "extra-corpus" material but by including them comprehensively and without pre-censoring as legitimate part of multilingual repertoires including English and local languages.

The French Revolution of 1789 is often blamed for spreading an ideology of the unity of language and nation in its aftermath in Europe as well as in other parts of the world colonized by it. Given the predominance of monolingual corpora and the still-strong national

traditions in corpus research, one might suspect that the long shadow of the French Revolution extends to today's digital research infrastructures: a nation with a language needs a monolingual national corpus. Perhaps the time has come to challenge the dominance of monolingual written standard corpora and to establish working with non-standard varieties, spontaneous spoken language, and multilingual data as new research priorities in corpus linguistics.

4. Progress in Linguistics: Technology versus the Human Mind

Looking back on the history of linguistics, we see that some scientific revolutions in the field occurred independently of technological development. For example, Ferdinand de Saussure needed nothing more than paper and pencil for the conception of the structuralist approach. Sociolinguistics and discourse analysis, on the other hand, would be inconceivable without the invention of recording equipment and the storage of acoustic data. And every improvement in technical equipment – such as increased recording and playback quality or the reduction in the size of devices for mobile use – immediately expands the horizon for scientific work. Michael Halliday did not exaggerate much when he argued that "[p]erhaps the greatest single event in the history of linguistics was the invention of the tape recorder, which for the first time has captured natural conversation and made it accessible to systematic study" (Halliday 1994, p. xxiii).

Linguistics also owes a comparable debt to the computer. Everything that can be done with tape recorders and video recorders works even better and faster with digital methods. In addition, we now have luxurious corpus-linguistic working environments – at least for the small number of languages that traditionally attracted the most scientific attention. The result was an unprecedented upswing in usage-based descriptive approaches and statistical analysis methods. Where technical and conceptual innovation is in balance, almost perfect conditions prevail. For many questions, the efforts associated with data acquisition and processing before digitization have been drastically reduced. "First throw away your evidence!" was the slogan with which John Sinclair (1986) described the new linguistic work practice vividly. Corpora enable new data-driven (bottom-up) approaches and increase the potential and efficiency of established conceptually-theoretically motivated (top-down) procedures. A good example of this is provided by the multitude of recent corpus linguistic studies on

grammaticalization, in which a theoretical concept over a hundred years old is revitalized in a new technical environment (See the research overview in Mair, 2011).

One last aspect should not be forgotten in the balance sheet. Across all individual topics, corpora also lead to a new work culture in linguistics. They promote cooperative research and cumulative knowledge gain because many scientists start from a common database. Regarding the future of the field, it must be clear that multilingual corpora, whose creation was formulated as a desideratum above, can only be compiled and analyzed in teamwork. Corpora also have the potential to undermine certain (linguistic) scientific hierarchies, such as that between linguists as native speakers and non-native speakers of the language under investigation. Finally, the next generation in corpus linguistics often gains authority very early on through technical affinity and familiarity with complex statistical analysis and visualization procedures.

5. Self-dissolution of Corpus Linguistics and its Integration into Digital Humanities

After fifty years of tumultuous development with an overall very positive outcome, corpus linguistics is facing its next upheaval. Until the early 1990s, only very few linguists used digital corpora, and the majority of the community had little understanding of what such tools were good for and how they worked in detail. This situation has fundamentally changed. Corpora are everywhere. Entry into corpus linguistics is low-threshold. The fundamental benefits of the new methods are not questioned. However, due to rapid technological development, it has become more difficult for newcomers today to develop an understanding of how the tools work in detail and what their potentials and limitations are.

Corpus linguistics has even lost a central identifying feature, namely the consensus on what a corpus is. The narrow definition of the linguistic corpus, as a digital text collection created by linguists for the purposes of linguistic analysis, is becoming irrelevant. Many corpora are no longer compiled using traditional methods and then digitized but are generated through the recycling of existing digital materials. Corpus-linguistic methods are no longer restricted to linguistics but are used across the humanities and social sciences. To use the metaphor coined by Gatto (2011), the solid foundation of a clearly defined corpus (“body”) has been lost, and in its place has emerged a diverse, rich, but also confusingly complex digital language network (“web”). At the end of an impressive success story, corpus linguistics is thus in the process of dissolving and integrating into the Digital Humanities

movement. It is such a revamped and repositioned corpus linguistics that will be fit to face the big challenge of the early 21st century, domesticating the power of Artificial Intelligence (AI) and Large Language Models (LLMs) for linguistic research and employing linguistic research methods to understand better a world that will be profoundly transformed by these new affordances. As publications such as Rudnicka (2023) and Torrent et al. (2023) show, the work has just begun.

References

Bartlett, J. (1913). *A complete concordance or verbal index to words, phrases, and passages in the dramatic works of Shakespeare*. Macmillan.

Fries, C. C. (1952). *The structure of English: The construction of English sentences*. Harcourt Brace.

Gatto, M. (2011). The 'body' and the 'web': The web as corpus ten years on. *ICAME Journal 35*, 35–58.

Greenbaum, S. & Nelson, G. (1996). The International Corpus of English (ICE) project. *World Englishes, 15*(1), 3–15.

Halliday, M. A. K. (1994*). An introduction to functional grammar* (2nd ed). Edward Arnold.

Leitner, G. (Ed.), *The English reference grammar: Language and linguistics, writers and readers*. Niemeyer.

Lodge, D. (2011). *Small world*. Random House.

Mair, C. (2011). Grammaticalization and corpus linguistics. In H. Narrog & B. Heine (Eds.), The Oxford handbook of grammaticalization (pp. 239–250). Oxford University Press.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Narrog, H. & Heine B. (Eds.). (2011). The Oxford handbook of grammaticalization. Oxford University Press.

Sinclair, J. (1986). First, throw away your evidence! In G. Leitner (Ed.), *The English reference grammar: Language and linguistics, writers and readers* (pp. 56–64). Niemeyer.

Rudnicka, K. (2023). Can Grammarly and ChatGPT accelerate language change? AI-powered technologies and their impact on the English language: Wordiness vs. conciseness. *Procesamiento del Lenguaje Natural 71*, 205-214.

Torrent, T. T., Hoffmann, T., Almeida, A. L., & Turner, M. (2023). *Copilots for linguists: AI, constructions, and frames*. Cambridge University Press.

Visser, F. T. (2002). *An historical syntax of the English language*. Brill Archive.