

STATISTICIAN, PROGRAMMER, DATA SCIENTIST? WHO IS, OR SHOULD BE, A CORPUS LINGUIST IN THE 2020S?

ŁUKASZ GRABOWSKI

Institute of Linguistics, The Faculty of Philology, University of Opole,
Opole, Poland

GRABOWSKI, Łukasz: Statistician, Programmer, Data Scientist? Who Is, or Should Be, a Corpus Linguist in the 2020s?. *Journal of Linguistics*, 2023, Vol. 74, No 1, pp. 52 – 59.

Abstract: In this short essay, I aim to ruminate on the nature of a corpus linguist's work in the 2020s, a time marked by unprecedented advancements in the field of computer technologies and artificial intelligence. This seems to be particularly relevant considering the theme of the 12th International Conference Slovko 2023, which is “Natural Language Processing and Corpus Linguistics”. In the last two decades or so, corpus linguistics has drawn extensively from the fields such as statistics, computer science and data science. In many respects corpus linguistics has served as a significant source of inspiration for progress in the field of natural language processing (NLP), leading to the development of large language models (LLMs) as well as recent introduction of conversational artificial intelligence, among others. Thus, in this paper I will make an attempt at identifying the skills that may help rank-and-file or aspiring corpus linguists to survive and, hopefully, flourish in the research field in the 2020s.

Keywords: corpus linguistics, statistics, computer programming, natural language processing, artificial intelligence

1 INTRODUCTION

Linguistic analyses of extensive text volumes were performed well before the rise of computer technologies (McEnery – Wilson 1996, p. 9), yet the character of those studies was significantly different from modern practices. Back then, limited computer tools and data processing capabilities made such analyses, often conducted manually, laborious, time-consuming, and sometimes unfeasible. However, since the late 1980s and early 1990s, the emergence of “computer corpus linguistics” (Ooi 1998, p. 34) has revolutionized the field¹: it has become the research area that involves employing computer software and tools as well as various statistical techniques to study language through vast collections of naturally occurring text, enabling investigations into various linguistic phenomena, notably into the interplay

¹ Computer corpus linguistics “involves the study of language on the basis of textual or acoustic corpora, almost always involving the computer in some phase of storage, processing, and analysis of data” (Ooi 1998, p. 34).

of grammatical and lexical units. Today, however, simple descriptive statistics or concordance analysis is often not enough to explain more fine-grained patterns of language use in massive collections of linguistic data. In the last two decades, the range of statistical methods of corpus analysis (e.g. multivariate statistics) as well as data visualization techniques has expanded considerably. Today, in the 2020s, it is no surprise anymore that in many language studies researchers also make ample use of machine learning techniques popular in the field of natural language processing (NLP), also known as computational linguistics (Hirschberg – Manning 2015, p. 261).

In short, NLP researchers focus on the development of algorithms and models for processing, producing, learning and understanding human language (Hirschberg – Manning 2015; Dunne 2022; Jurafsky – Martin 2023). It is often the case that NLP researchers use text corpora compiled by corpus linguists to develop computational models aimed at performing various language-related tasks, such as text classification, text summarization, information retrieval, automatic speech recognition, machine translation, sentiment analysis, authorship analysis, semantic role labelling, coreference resolution, to name but a few (Dunne 2022; Jurafsky – Martin 2023). NLP methods facilitate automation of various routine research tasks performed by corpus linguists, namely corpus queries and corpus analyses, some of them earlier conducted manually, to identify and explore patterns of language use, that is, frequencies and distributions of linguistic forms and their classes in texts. Consequently, Dunne (2022) argues that “corpus analysis can be expanded and scaled up by incorporating computational methods from natural language processing”.

There is an important caveat, though. Dunne (2022, p. 1) observes that natural language processing and computational linguistics develop so fast that “these advances have become increasingly disconnected from corpus linguistics and linguistic theory”. This observation becomes even more prominent in the light of recent developments in the field of artificial intelligence. Crosthwaite and Baisa (2023, p. 1) argue that “our field [corpus linguistics – LG] risks being overshadowed by GenAI [generative artificial intelligence – LG] researchers who are essentially just doing what we as corpus linguists already do, but in a way that has finally captured the imagination of the public.” It seems that GenAI-assisted linguistic studies will appear sooner rather than later, and they will most probably contribute to the already exponentially growing number of researchers, journals and scholarly papers.² Hence, it can be expected that language researchers will face two scenarios.

² Hyland (2023, p. 2) cites UNESCO data according to which the number of researchers in the years 2014–2018 “grew three times faster than the world population”, and that in the year 2022 alone the publishing giant Elsevier received 2.7 million paper submissions, compared to 2.5 million in 2021; the number of published papers, which was over 600,000 in 2021, represented an increase of 89% compared to the previous decade.

Either the disconnect between corpus linguistic, linguistic theory, on the one hand, and NLP and GenAI, on the other, will grow even further, or the combination of theories, methodologies, tools and solutions from all these fields will produce a desired synergy for linguistic research.³

Hence, I believe that it is essential to ruminate over the skills that a (corpus) linguist A.D. 2023, recently plunged into the reality of AI chatbots, such as ChatGPT or Bard, to name but a few, should possess to facilitate realization of the second scenario (i.e., a useful synergy between corpus linguistics and NLP). To put this matter into perspective, I would like to start with a brief and very subjective overview of the changes in the field of corpus linguistics over the last two decades up to now.

2 CORPUS LINGUISTICS: GROWTH AND CHALLENGES

In general, linguists treat language corpora as a source of research material that they use to empirically verify previously formulated linguistic theories or develop new ones. In fact, every linguist is a bit of a corpus linguist, when using language corpora or elements of corpus methodology in his or her work. Above all, however, this group includes scientists who use previously developed language corpora in their research, compile and study new corpora of texts, and oftentimes – using programming skills – develop new tools for the analysis of the corpora they have collected by themselves. Given that the corpus linguistic methodology (corpus-informed, corpus-based or corpus-driven one) is also used in literary, pedagogical, forensic or sociological research, among others, new challenges arise that pertain to the interdisciplinary character of undertaken studies.

To put it very mildly, for corpus linguists the frequency of occurrence of various language units and their classes comes to the fore, which means that they treat the text as a certain probabilistic structure that should be examined using research methods typical of those fields of science dealing with the analysis of large amounts of empirical data, e.g. statistics, sociology, demographics. Here, it is also worth mentioning data science, which has become, in a sense, a separate discipline devoted to data analysis. It seems, therefore, that in order to understand how other researchers analyzed and interpreted linguistic, sociological and psycholinguistic data, it is necessary to have rudimentary statistical knowledge. Currently, this is no longer an isolated opinion, because for a long time we have been witnessing a shift in linguistics, and perhaps to some extent a change in the dominant research paradigm, towards empirical and quantitative methods. Gries (2013, p. 4) emphasized this very clearly, pointing to the three main goals of using quantitative methods: describing

³ For example, in a preliminary study Lew (2023, June 12, preprint) shows that the quality of CO-BUILD-style AI-generated (by ChatGPT Plus) lexicographic definitions is comparable to the ones written by human lexicographers.

data, explaining the relationships found in the data,⁴ and predicting the future state of affairs (also using other data for this purpose, but based on a model developed on the basis of previously obtained data).

The above observation related to the necessity of being familiar with statistics becomes even more important in the current situation, where more and more published studies provide us with the results of various statistical analyses, enriched with histograms, box plots or heat maps etc. At the same time, somewhat as an answer to this trend, many publications and textbooks devoted to statistics in linguistic research, also with elements of programming in R language (which until recently was the international standard in the field of statistical analysis and data visualization, now increasingly supplanted by Python), came out in print in recent years, e.g. Gries 2013; Cantos Gomez 2013; Levshina 2015; Navarro 2015; Desagulier 2017; Brezina 2018; Winter 2019. The increasingly advanced statistics used in corpus linguistic research has been accompanied by the growing use of machine learning methods popular in the field of NLP, which brings the corpus linguist even closer to the role of a data scientist, as also mentioned by Dunne (2022).

It can be argued that despite the complexity of some quantitative, statistical methods, including those used in the area of natural language processing (e.g. neural networks, word or sentence embeddings, transformers, cf. Jurafsky – Marin 2023), their use in linguistically-oriented should be based on strong humanistic premises. With the aid of these methods, coupled with theoretical linguistic foundations, we are now better equipped to detect and comprehend intricate textual, lexical or grammatical patterns present in huge amounts of texts. Such patterns might have been overlooked when relying solely on intuition or individual, subjective impressions, or even conducting rudimentary corpus analysis like reading concordances. It would have been also impossible to precisely classify the huge amounts of linguistic data (e.g. using support vector machines (SVMs), classification decision trees, random forests or other machine learning methods, both supervised and unsupervised). Without the use of different regression methods (depending on the type of dependent variable), it would have been impossible to verify the hypotheses explaining the occurrence of some linguistic features with a reliable degree of precision. Having gained insights into the data using methods popular in NLP, we can then go back to the text, take a closer look at the obtained linguistic patterns and conduct their in-depth interpretation in the full textual, cultural and social context, which is, after all, the essence of research in the humanities.

⁴ For example, multivariate methods allow us to examine the relationships between multiple independent variables (e.g. social ones, such as age, gender, time period, or textual ones, such as text type or genre) and their impact on the frequency and distribution of linguistic features (dependent variable), and to verify the obtained results based on a more advanced statistical apparatus (Cantos Gomez 2013, pp. 89–133).

The latest publications also provide us with a growing body of evidence that currently those linguists who employ quantitative methods when conducting empirical research (both descriptive and experimental one) have become increasingly aware of the limitations of custom-designed tools for text analysis (e.g. range of options and functionalities, user-friendliness of the user interface, required degree of technical expertise).⁵ Some linguists, literary scholars or specialists in culture studies – identifying themselves with the trend currently known as digital humanities – increasingly learn to write text-analysis software by themselves (e.g. in the form of developing short scripts in R or Python, enabling the extraction of the desired linguistic material from their collections of text or to conducting appropriate analyses). Alternatively, they closely cooperate with specialists in the field of computational linguistics or NLP. In the near future, conversational artificial intelligence (or generative artificial intelligence, GenAI), which – if properly asked (*prompt engineering*) – may lend them a helping hand when pursuing their research, by facilitating comprehension of the intricacies of programming code or a statistical method. For example, in the paper devoted to data-driven language learning (DDL), Crosthwaite and Baisa (2023, pp. 2–3) list a number of advantages that corpora and custom-designed corpus tools hold over GenAI applications. At the same time, however, they also reflect upon the ways GenAI chatbots may overcome some of the limitations of language corpora and corpus software, notably in aspects such as technical complexity, lack of flexibility, data size, etc.

So, should a corpus linguist of the 2020s, specializing, for example, in phraseology, language pedagogy or sociolinguistics, also be an NLP and machine learning expert? It seems that the answer to this question depends largely on the specific needs (research questions) of the researcher as well as the culture of collaboration between researchers from different disciplines. It seems that when working with authentic empirical data (linguistic, sociological, etc.) it is worth understanding (at least conceptually rather than in technical detail) the methods typical of both computational linguistics and NLP simply in order to assess their usefulness. In other words: with a leading research question in mind, it is worth knowing in the first place what we get thanks to using one method rather than another. For example, it is good to know what the use of a selected statistical measure in relation to word co-occurrence brings us (e.g. a particular measure of lexical attraction or repulsion, statistical significance test, measure of effect size), rather than just know the mathematical formula each and every metric. In practice, the answer to the question: “Would it better if I learned how to program?” may also depend on the possibility of finding computer programmers who will develop (and

⁵ Popular corpus software among corpus linguists include, among others, e.g. WordSmith Tools (Scott 2022), AntConc (Anthony 2022) or SketchEngine (Kilgariff et al. 2014).

maybe even run) an appropriate programming script for us, taking into account the specificity of our linguistic dataset, research questions and hypotheses.

At this point, it is extremely important to encourage a culture of collaboration between researchers in the humanities (including linguists) and specialists from other fields of science, especially those outside the humanities (computer scientists, chemists,⁶ biologists). Thanks to the research and development projects carried out in recent years, the situation has changed for much better in this respect. For example, thanks to the European project CLARIN,⁷ researchers in the humanities and social sciences from all over Europe and beyond gained access to tailor-made computational tools and solutions designed to handle large textual datasets (including their processing, analysis and visualization). It is also possible to contact the helpdesk and developers of such tools, affiliated in national CLARIN consortia, for specialist advice on the use of statistical and NLP methods, and even to ask for assistance during the data analysis itself. After all, the essence of creating a pan-European research infrastructure is the very usefulness of such infrastructure for researchers.

3 CONCLUSIONS: A CORPUS LINGUIST IN THE 2020S

It is important to highlight that my observations pertain to corpus linguistics, a set of methods that involve analyzing extensive empirical textual data. However, this is not the sole approach to linguistic research. Hence, the arguments presented here may not be equally applicable to rationalistic “armchair” linguistics, where researchers themselves construct language examples to test linguistic theories: this practice continues to endure and is also likely to evolve in the future. Also, this brief reflection-overview paper, focusing on the desired skills of contemporary corpus linguists, reveals that certain questions remain without a definitive answer, and the arguments presented here are intentionally selective.

It is evident that the development of both NLP and language engineering is extremely fast, and today corpus linguistics has become increasingly embedded in statistics, data science and machine learning. Further advances in research on artificial intelligence, especially in the area of application of large language models (LLMs), will soon bring further challenges for researchers. In fact, the volume of data that surround us will grow fast: according to the forecast provided by Statista⁸ (2023), it will grow from around 125 ZB (zetabytes) in 2023 to more than 180 ZB in 2025, where 1 ZB equals a billion TB (terabytes). Most probably, a significant proportion of this will be electronic textual data produced naturally or artificially generated. In addition, the exponentially growing number of digital publications

⁶ An interesting example of such collaboration can be found in Woźniak et al. (2017).

⁷ <https://www.clarin.eu/>

⁸ <https://www.statista.com/statistics/871513/worldwide-data-created/> (NCES. (n.d.))

means that better and better (more accurate, mathematically-sophisticated etc.) research methods may tip the scales in the race for getting published and becoming visible, even for a brief moment, in the reality of attention economy that has permeated academic publishing, to use the phrase taken from Hyland (2023).

Thus, it can be concluded that a corpus linguist working with linguistic data in the 2020s, be it a stylometrist, sociolinguist or otherwise, should have sufficient skills to understand, at least conceptually, descriptive and inferential statistics as well as the methods used in computational linguistics and NLP. This will allow him or her to better understand the specificity of linguistic data (e.g. recognize the type of variables used in the study, understand what vector text representation is etc.), compare performance of different methods (e.g. in relation to topic modelling or text classification), interpret different visualizations of linguistic data, and maybe even estimate the quality and complexity of a statistical model (e.g. using metrics such as accuracy, precision, recall, F-score, mean error rate, AIC or BIC etc.). At the same time, the knowledge of language theories will still play a particularly important role, as it enables one to interpret the complex patterns and regularities that exist in naturally occurring language. Ultimately, the goal of any corpus analysis is to obtain new knowledge of language, be it a language-system or language in use. For these reasons, a corpus linguist should remain a linguist in the first place. On the technical front, it can be expected that the work of corpus linguists will to a greater degree overlap with the tasks performed by a computer programmer, statistician or data scientist.⁹ Hence, at least basic programming skills may come in useful, especially in terms of running programming scripts (e.g. written in Python) either locally or using collaborative programming platforms, such as Jupyter Notebook or Google Collab. It might be also useful to know how to extract linguistic data from various resources or infrastructures (lexical databases, language corpora, etc.) that offer access to researchers via API (Application Programming Interface). The informed use of low-code programming platforms or conversational artificial intelligence (and other AI-assisted solutions) is likely to offer assistance as well. As mentioned earlier, collaboration skills will play a vital role in advancing interdisciplinary research at the intersection of (corpus) linguistics, statistics, and NLP. Now that the period of AI-assisted research has already started, it is essential for corpus linguists to navigate through it with an open-mind and caution.

References

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan. Waseda University. Accessible at: <https://www.laurenceanthony.net/software>.

⁹ At this point, I do not consider whether artificial intelligence will replace corpus linguists altogether, as this is a completely separate issue to me. For the same reason, I do not consider limitations in development of NLP AI-research (consumption of energy, funding etc.).

Brezina, V. (2018). *Statistics for Corpus Linguistics*. Cambridge: Cambridge University Press, 314 p.

Cantos Gomez, P. (2013). *Statistical Methods in Language and Linguistic Research*. London: Equinox, 256 p.

Crosthwaite, P., and Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3). Accessible at: <https://doi.org/10.1016/j.acorp.2023.100066>.

Desagulier, G. (2017). *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. Berlin: Springer, 366 p.

Dunne, J. (2022). *Natural Language Processing for Corpus Linguistics (Elements in Corpus Linguistics)*. Cambridge: Cambridge University Press, 96 p.

Gries, S. (2013). *Statistics for Linguistics with R*. Berlin: De Gruyter, 374 p.

Hirschberg, J., and Manning, Ch. (2015). *Advances in natural language processing*. Science, 349(6245), pages 261–266.

Hyland, K. (2023). Academic publishing and the attention economy. *Journal of English for Academic Purposes*, 64. Accessible at: <https://doi.org/10.1016/j.jeap.2023.101253>.

Jurafsky, D., and Martin, J. (2023). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. (Third edition e-book: draft of January 7, 2023). Accessible at: <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>. (accessed on 19 July 2023).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pages 7–36.

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins, 454 p.

Lew, R. (2023, June 12). ChatGPT as a COBUILD lexicographer. Accessible at: <https://doi.org/10.31219/osf.io/t9mbu>.

McEnery, T., and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: University Press, 256 p.

Navarro, D. (2015). *Learning Statistics with R: A tutorial for psychology students and other beginners*. (Version 0.6), 599 p. Sydney. University of New South Wales. Accessible at: <http://compcogscisysdney.org/learning-statistics-with-r/>.

NCES. (n.d.). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes). In *Statista - The Statistics Portal*. Accessible at: <https://www.statista.com/statistics/871513/worldwide-data-created/>.

Ooi, V. (1998). *Computer Corpus Lexicography*. Edinburgh: University Press, 224 p.

Scott, M. (2022). *WordSmith Tools version 8 (64 bit version)* Stroud: Lexical Analysis Software.

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. London: Routledge, 310 p.

Woźniak, M., Wołos, A., Modrzyk, U., Górski, R. L., Winkowski, J., Bajczyk, M., Szymkuć, S., Grzybowski, B., and Eder, M. (2018). Linguistic measures of chemical diversity and the ‘keywords’ of molecular collections. *Scientific Reports*, 8(1), page 7598.