Applied Corpus Linguistics 3 (2023) 100066

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp

Generative AI and the end of corpus-assisted data-driven learning? Not so fast!

Peter Crosthwaite [a,*], Vit Baisa [b]

[a] School of Languages and Cultures, University of Queensland, Australia
[b] Masaryk University, Faculty of Informatics, Natural Language Processing Centre, Czech Republic

ARTICLE INFO

Keywords:
Data-driven learning
generative AI
ChatGPT
DDL
Corpora

ABSTRACT

This article explores the potential advantages of corpora over generative artificial intelligence (GenAI) in understanding language patterns and usage, while also acknowledging the potential of GenAI to address some of the main shortcomings of corpus-based data-driven learning (DDL). One of the main advantages of corpora is that we know exactly the domain of texts from which the corpus data is derived, something that we cannot track from current large language models underlying applications like ChatGPT. We know the texts that make up large general corpora such as BNC2014 and BAWE, and can even extract full texts from these corpora if needed. Corpora also allow for more nuanced analysis of language patterns, including the statistics behind multi-word units and collocations, which can be difficult for GenAI to handle. However, it is important to note that GenAI has its own strengths in advancing our understanding of language-in-use that corpora, to date, have struggled with. We therefore argue that by combining corpus and GenAI approaches, language learners can gain a more comprehensive understanding of how language works in different contexts than is currently possible using only a single approach.

Introduction

The late 2022 release of OpenAI's *ChatGPT* and the subsequent explosion in generative artificial intelligence (GenAI) applications have already fundamentally changed the perception of the general public towards the possibilities of human interaction with large language data – something corpus linguists have been attempting to do for decades. Certainly, as corpus linguists dedicated to popularising what we do outside of academia, we've often had to describe what corpus linguists *do* with corpora and what corpus consultation can *bring*, for people who might not consider themselves corpus linguists. It generally goes something along the lines of the following:

"We use corpus tools to query large authentic, principled collections of electronically searchable text to discover the patterns of language-in-use across a wide variety of general and specific registers, so that we can better understand (and possibly teach) those patterns"

After the usual cursory nods of the head or the occasional "how interesting?", the conversation usually ends there, and the subject is changed.

But not so much these days. Recently, any talk of large language data is met with fervent interest – "oh, are you using that ChatGPT?", "I guess we are all out of a job soon", "my friend uses it to write all her work e-mails", etc. More astute observers by now even understand the underlying processes behind how such models and their user interfaces work, essentially by calculating – very quickly and at previously unprecedented scale – those underlying patterns of language-in-use that corpus linguists are so familiar with. Suddenly, language data – *corpora* – are back in vogue. Yet, the field of corpus linguistics is at a crossroads. Despite our best efforts, our field risks being overshadowed by GenAI 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.

This is certainly the case for those work on corpus-based data-driven learning (DDL). Corpora have now been used for over two decades to enhance the teaching and learning of languages, whether indirectly through the creation of dictionaries, wordlists or integration of corpus data into teaching materials, or directly through learners' hands-on corpus consultation (more commonly termed DDL). DDL is purported to promote language acquisition through exposure to the frequency and salience of patterns of language in use, whether individually through constructivist learning, or socioculturally when used in classroom contexts as learners, peers and teachers complete scaffolded activities

* Corresponding author.
E-mail address: p.cros@uq.edu.au (P. Crosthwaite).

https://doi.org/10.1016/j.acorp.2023.100066
Received 18 April 2023; Received in revised form 25 June 2023; Accepted 12 July 2023
Available online 13 July 2023

involving corpus data (O'Keeffe, 2021). A wealth of empirical studies has found advantages in using DDL across a variety of linguistic targets in experimental research, with generally positive accompanying perceptions (Boulton & Vyatkina, 2021), albeit tempered with criticisms about corpus tools and corpus data that we outline in more detail below.

Therefore, having just released our own DDL corpus tool early 2023 (https://corpusmate.com), we were concerned that our work was all in vain – who would want to use corpus tools to consult corpora for DDL, when the world's largest corpus (in a sense) was now publicly available for query, with the latest interactive chatbot available to quickly and recursively query that corpus, and all (currently) for free? Does this mean the end of DDL as we know it? Upon reflection, it certainly *does* mean the end of DDL as we know it if corpus linguists and DDL practitioners do not take steps to rectify ongoing issues that have continually plagued the field and begin to consider how GenAI may help us overcome them. Crosthwaite and Boulton (2023) described these issues for DDL at some length, including (amongst other factors) expanding the boundaries of what constitutes DDL outside of concordancing, the complexity of most publicly available corpus datasets used for DDL, and the high level of Technological, Pedagogical and Content Knowledge (TPACK – Koehler & Mishra, 2009; see Meunier, 2019 for a description of its potential for DDL) required for practicing teachers to make corpus-based DDL work for them. Now the genie is out of the proverbial bottle when it comes to ChatGPT/GenAI, the field of DDL needs to address these issues if more 'traditional' corpus consultation is to remain viable (funnily enough, in most DDL studies, corpora are seen as the trendy alternative to 'traditional' teaching). Alternatively, can DDL researchers now harness the power of GenAI to bring the kind of learning espoused under DDL to a wider audience? In what follows, we briefly make the case for both possibilities.

The continued case for corpora in DDL

The potential affordances of GenAI for language learning include real-time conversation, immediate formative and corrective feedback, natural language explanations of vocabulary in contexts, instant generation of texts of specific registers and genres, dictionary definitions and examples, and machine translation (Kohnke et al., 2023). However, given the delay between the release of GenAI applications e.g., ChatGPT, such affordances remain as yet largely empirically untested at the actual classroom level. We do, however, have a lengthy body of research pointing out the learning gains for genre/register awareness, vocabulary, grammar and error correction that corpus consultation can bring to end-users, as suggested in numerous meta-analyses and bibliometric reviews (e.g., Boulton and Cobb, 2017; Lee et al., 2019; Dong et al., 2022). At a fundamental level, we also know that principled, curated consultation of corpus data can lead to meaningful learning at both individual and sociocultural levels of engagement (O'Keeffe, 2021).

There is absolutely no reason why this should not continue in the GenAI age. Let us consider here some of the advantages that corpora and corpus tools still hold over chatbot-like GenAI applications:

1) **Knowing the data** – One of the main 'selling points' of corpora as a research and later teaching tool was that we know, exactly, the domain of texts from which the corpus data is derived, something that we cannot track from current large language models underlying applications e.g., ChatGPT. We know the texts that make up large general corpora e.g., the BNC2014, the BAWE, etc., and can even go to the trouble to extract the full texts from these corpora should we need to (see Crosthwaite et al., 2021 for a pedagogical example). CorpusMate, for example, has a 'citation' function whereby one can click any concordance result to see which corpus the text was derived from, the title of that text, and a link to said corpus. The ability to use DIY corpora (e.g., Charles, 2012) is a considerable advantage in this regard, with learners able to claim complete ownership over the corpus used for subsequent queries.
2) **Authenticity** – together with the first point above, there is a significant difference between corpus data as produced by humans, and GenAI output as generated through a statistical procedure. GenAI may generate sentences which are grammatically correct, but which might rarely be used in actual writing or conversation, or which may not be contextually or register-appropriate. Subsequent prompt tweaking may be required to ensure issues of this nature are minimised when using GenAI, but the authenticity of corpus data – that is, language data *actually produced by humans* – should be seen as a more reliable indicator of real language-in-use, particularly for second language learners who do not have the benefit of being able to easily authenticate whether a given output would match what a native speaker of the target language might produce in the same context.
3) **Replicability** – While GenAI applications 'generate' text through complex statistical procedures, an end-user currently cannot see, nor be able to replicate, the statistical procedures that lead to that generated text. Even if you could, the answers are randomly sampled leading to a unique answer for each subsequent identical query. The benefit of corpora in this regard is that one can easily replicate a given finding with the same query on the same data, in this respect producing 'hard evidence' that word X belongs with word Y, for example. Such evidence is incredibly powerful for language learners (and their teachers) – even better if you can replicate a finding in Corpus X in Corpus Y. You can also do this time and time again, without limitation.
4) **Multimodality** – Several recent corpus tools present multiple pathways into corpus data, be this in the form of (coloured) concordances, statistical tables (e.g., collocation scores), and, increasingly, visual charts and maps of relationships between words and lexico-grammatical units (e.g., *Voyant Tools*, Sinclair and Rockwell, 2016). These improvements to corpus tool functionality work to specifically target and highlight patterns in corpus data for improved visibility and learnability for groups of users for whom more traditional concordancing may be difficult. Additionally, some pedagogical corpora e.g., SACODEYL (Pérez-Paredes & Alcaraz-Calero, 2009) utilise video and audio files in tandem with concordancers. At the time of writing, most GenAI tools do have the ability to generate tables or detailed images from text prompts, but it can be difficult to operationalise this within a chatbot context and usually requires integration of one tool (e.g., ChatGPT) with another (e.g., Midjourney), making this difficult for non-technical users.
5) **Safety** – In working closely with educators at the primary and secondary levels, an oft-reported concern is the lack of clarity about how GenAI tools and companies use user data. Any data provided by pre-tertiary age users is commonly subject to numerous ethical and legal safeguards, as is sensitive personal data, data on curricula, assessments, and many more. Educational bodies are as yet hesitant to allow even staff access to ChatGPT for these reasons, never mind younger learners. As most corpus tools only require very little in terms of user data (perhaps only initial registration details), corpus linguists looking to use this opportunity to push corpus use in schools can stress that corpus consultation is currently a 'safe' option.
6) **Hallucinations** – Due to the way GenAI works as a predictive language model, the accuracy of its output can often leave a lot to be desired. For example, recent tests on ChatGPT's ability to generate non-latin script languages is poorer than its ability to understand them (Bang et al., 2023), and has also been shown to invent terms that lie outside of its training data (Shen et al., 2023). While corpus data can also be rife with typos, punctuation and spacing issues, and lexico-grammatical errors and innovations (albeit this is what you want in learner corpora, for example), many of these issues can be overcome through appropriate data cleaning and preparation, unlike GenAI where no such access to the data is available and where users are a slave to the algorithm.
7) **Active vs. passive learning** – The literature on the 'L' in DDL suggests that the type of constructivist, usage-based learning that takes

place during corpus consultation requires much intention on the part of the learner to succeed, with concordancing a cognitively demanding task that requires significant inductive learning processes (Sun and Wang, 2003). So far, ChatGPT claims its use promotes inductive learning - "*When a user interacts with GPT, they are essentially making use of the patterns and associations learned by the model to generate new text or provide relevant information. The process is inductive because it involves drawing conclusions based on the patterns and associations observed in the input data, rather than applying pre-existing knowledge or rules (as in deductive reasoning)*(OpenAI, 2023). However, to what extent is ChatGPT or the learner doing the induction? Without the learners' knowledge of these 'patterns' or 'associations' in the data, it appears that GPT is the one doing the induction, and there is a significant risk that users simply copy-paste ChatGPT output into their own work with little actual 'learning' taking place.

To summarise, corpora and DDL still offer advantages over large language models such as ChatGPT when it comes to language learning and teaching, which should come as a relief for the field. That said, there are still many reasons why the field needs to embrace this technology if we are to stay relevant, which we outline in the following section.

The case for bringing GenAI into DDL

DDL researchers are *acutely* aware of the shortcomings of the current state-of-the-art related to available corpora and corpus tools. Following DDL training sessions, we hear the same qualitative comments from teacher trainees and language learners again and again, summarised in the form of "*Corpora and DDL sound really cool, but…*". Crosthwaite and Boulton (2023) note some of these "buts" include:

- The level of technical knowledge needed to use the tool (for DDL).
- Complicated or unintuitive user interfaces at odds with how modern learners typically access digital information through resources such as Google (and now GenAI).
- Unsuitable corpus data for the target learners (e.g., COCA for younger or less proficient L2 learners)
- A general inability to track users' corpus use (and any associated learning gains) over time.

GenAI stands as a potential solution to almost all these concerns if we can take the kind of inductive, discovery learning approach that we currently employ for corpus based DDL and apply that to our use of GenAI. In other words, if we work on "bringing a DDL mindset to our learners, rather than expecting them to come to corpus linguistics" (Crosthwaite & Boulton, 2023), we can open the DDL field to new possibilities rather than gatekeeping DDL behind concordancing and concordancers. While this was necessary before the advent of GenAI, it is now crucial for the DDL field to take this step if we are to remain relevant going forward.

Let us discuss the possibilities in turn.

1) **Required technical knowledge** – Already GenAI chatbots have significantly reduced the levels of technical knowledge required to successfully consult large language data. The simple ability to use natural language inputs to receive (generated) natural language output is really the main gamechanger, even before the release of OpenAI's ChatGPT chatbot. For example, we no longer need to use complex corpus query language syntax to isolate parts of speech within our corpus retrievals, as we can now simply 'ask' for what we want, e.g., "can you list a few example sentences including 'phrase X'"?, "what is a more formal equivalent of word Y", or "what are some words similar to the word "Z". This significantly reduces the degree of metalinguistic knowledge required to successfully query a corpus platform for DDL, and for those with such knowledge, more advanced queries can unlock even more patterns (e.g., "list 10 sentences containing the past participle of the word 'do').
2) **User experience** – The single biggest reported complaint we hear following DDL training sessions is related to the complexity of the corpus tools, with most comments reserved for the user interface (UI). So much DDL training has been done with highly complex UIs such as BYU/English-corpora.org, which accounted for 31% of all DDL studies up to 2019 (Boulton & Vytakina, 2021). Admittedly, such corpus tools commonly used for DDL are research rather than teaching tools, but even attempts to streamline complex corpus tool UIs have done little to help (e.g., SketchEngine's recent 'overhaul' of its UI). Interacting with GenAI applications such as ChatGPT could not be easier in form or function, hence its current popularity amongst the general population.
3) **Differentiation** – while we already discussed the rigidity of corpus data as an advantage earlier in this paper, such rigidity is also a major problem for users without the required proficiency to understand the concordance results they are getting. What GenAI applications do very well at present is the ability to 'differentiate' language intended for, say, advanced users of the target language and reproduce that using simpler structures and vocabulary (e.g., "re-write this text for a 9th grader"). English as an additional language/dialect teachers we work with in Australia see this function as game-changing for their everyday work, particularly for mixed-proficiency classes. The potential impact of GenAI's ability to generate results from almost any register, domain or even *language* cannot be overstated, and can greatly widen the scope of DDL from its current almost invariable focus on tertiary academic English language.
4) **Data size** - Another factor to consider is simply the size of current large language models such as ChatGPT (using GPT-4), with training data token counts in the billions or trillions. Being able to quickly query models of this size and at speed – online - is previously unprecedented with even the best available corpus tools.
5) **Prior user input as input** – What separates OpenAI's success from rivals with older large language models is the ability to take previous inputs to the model and use these to refine future outputs, with these 'chats' saved for future, currently unlimited use. While numerous DDL studies have sought to track how learners engage with corpora in the form of screen recordings (e.g., Kotamjani et al., 2017) and query logs (e.g., Pérez-Paredes et al., 2011), easily capturing this longitudinal data using existing corpus tools is currently not possible. Revisiting user interactions with these language models and using chatlogs as evidence of 'learning' will likely be a major research methodology in the coming years as researchers investigate to what extent GenAI use promotes language acquisition and even 'better learning' in general, something DDL research has largely yet to achieve.
6) **Translation to teaching materials** – It generally takes a teacher with a high degree of corpus literacy, content and pedagogical knowledge to convert corpus findings into actual teaching materials, as seen in a number of recent studies exploring DDL lesson planning (e.g., Ma et al., 2022). However, GenAI/ChatGPT's ability to take a previous language-focused query (e.g., "what are some nouns that fill slot X in this sentence") and seamlessly convert that finding into a teaching task or assessment item (e.g., "take the list of nouns you generated and create a sequence of multiple-choice questions based on them") or even a full lesson plan (e.g., "build a lesson plan for 2nd graders based on acquiring these nouns") is simply amazing, and a complete game-changer for mainstream pre-tertiary teachers with little time or resources.
7) **Funding** – typically, most corpus builders and corpus tools creators are one-person shows or small university-based teams, working on limited grants, donations, or in many cases for free as a hobby. This can leave tools unfinished or with no foreseeable upgrades in data or functionality, leading to a limited shelf-life beyond with end-users move on. Current major GenAI companies such as OpenAI are

already multi-billion-dollar enterprises, with new updates and upgrades coming out on a weekly or even daily basis. We, as a field, cannot hope to compete with the speed and scale of GenAI, but if you can't beat them, join them.

Conclusion

To conclude, it may be time to consider whether we continue to gatekeep DDL behind concordancers, or open up the 'D' in DDL to new, GenAI-assisted possibilities. We are keen to stress that this does not have to be a zero-sum game – there are situations where corpora will continue to be advantageous over GenAI for some time to come, while GenAI can already solve certain issues that have continued to plague corpus-based DDL for years. Combining corpus-based DDL with GenAI appears to be 'a useful methodological synergy', similarly to when corpora began to be used for critical discourse analysis, for example (Baker et al., 2008). We are also keen to stress, however, that *not* leveraging GenAI risks leaving DDL practitioners and DDL as an enterprise behind, given the current depth of interest in GenAI for education. DDL researchers are well-placed to take advantage of this renewed mainstream interest in language data as we understand the power of such data for language teaching as well as the conditions required for meaningful learning using such data to proceed. Let's therefore put our money where our mouth is and get started.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., Wodak, R., 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse Society 19 (3), 273–306. https://doi.org/10.1177/0957926508088962.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q., Xu, Y. & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint* arXiv:2302.04023.

Boulton, A., Cobb, T., 2017. Corpus use in language learning: a meta-analysis. Lang. Learn. 67 (2), 348–393. https://doi.org/10.1111/lang.12224.

Boulton, A., Vyatkina, N., 2021. Thirty years of data-driven learning: taking stock and charting new directions over time. Lang. Learn. Technol. 25 (3), 66–89. http://hdl.handle.net/10125/73450.

Charles, M., 2012. Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. English Specific Purposes 31 (2), 93–102. https://doi.org/10.1016/j.esp.2011.12.003.

Crosthwaite, P., Boulton, A., 2023. DDL is dead? Long live DDL! Expanding the boundaries of data-driven learning. In: Pérez-Paredes, P., Tyne, H. (Eds.), Discovering Language: Learning and Affordance in press.

Crosthwaite, P., Sanhueza, A.G., Schweinberger, M., 2021. Training disciplinary genre awareness through blended learning: An exploration into EAP students' perceptions of online annotation of genres across disciplines. J. Engl. Acad. Purp. 53, 101021. https://doi.org/10.1016/j.jeap.2021.101021.

Dong, J., Zhao, Y., Buckingham, L., 2022. Charting the landscape of data-driven learning using a bibliometric analysis. ReCALL 1–17. https://doi.org/10.1017/S0958344022000222.

Koehler, M., Mishra, P., 2009. What is technological pedagogical content knowledge? Contemp. Issues Technol. Teacher Educ. 9 (1), 60–70.

Kohnke, L., Moorhouse, B.L., Zou, D., 2023. ChatGPT for language teaching and learning. RELC J. https://doi.org/10.1177/00336882231162868, 00336882231162868.

Kotamjani, S.S., Razavi, O.F., Hussin, H., 2017. Online Corpus Tools in Scholarly Writing: a Case of EFL Postgraduate Student. English Lang. Teach. 10 (9), 61–68. https://doi.org/10.5539/elt.v10n9p61.

Lee, H., Warschauer, M., Lee, J.H., 2019. The effects of corpus use on second language vocabulary learning: a multilevel meta-analysis. Appl. Linguist. 40 (5), 721–753. https://doi.org/10.1093/applin/amy012.

Ma, Q., Yuan, R., Cheung, L.M.E., Yang, J., 2022. Teacher paths for developing corpus-based language pedagogy: a case study. Comput. Assist. Lang. Learn. 1–32. https://doi.org/10.1080/09588221.2022.2040537.

Meunier, F., 2019. A case for constructive alignment in DDL: rethinking outcomes, practices and assessment in (data-driven) language learning. In: Crosthwaite, P. (Ed.), Data-driven Learning For the Next generation: Corpora and DDL For Pre-Tertiary Learners. Taylor & Francis, Routledge, pp. 13–31. https://doi.org/10.4324/9780429425899-2.

OpenAI. (2023). ChatGPT (19/4/23) [Large language model].

O'Keeffe, A., 2021. Data-driven learning: a call for a broader research gaze. Lang. Teach. 54 (2), 259–272. https://doi.org/10.1017/S0261444820000245.

Pérez-Paredes, P., Alcaraz-Calero, J.M., 2009. Developing annotation solutions for online data driven learning. ReCALL 21 (1), 55–75. https://doi.org/10.1017/S0958344009000093.

Pérez-Paredes, P., Sánchez-Tornel, M., Alcaraz Calero, J.M., Jiménez, P.A., 2011. Tracking learners' actual uses of corpora: guided vs non-guided corpus consultation. Comput. Assist. Lang. Learn. 24 (3), 233–253. https://doi.org/10.1080/09588221.2010.539978.

Shen, Y., Heacock, L., Elias, J., Hentel, K.D., Reig, B., Shih, G., Moy, L., 2023. ChatGPT and other large language models are double-edged swords. Radiology 307 (2), e230163.

Sinclair, S. & Rockwell, G. (2016). *Voyant Tools*. http://voyant-tools.org/.

Sun, Y.C., Wang, L.Y., 2003. Concordancers in the EFL classroom: cognitive approaches and collocation difficulty. Comput. Assist. Lang. Learn. 16 (1), 83–94. https://doi.org/10.1076/call.16.1.83.15528.