

Is DDL dead yet? Not so fast! Revisiting earlier claims on the state of corpus linguistics, DDL, generative AI, and Language for Specific Purposes.

Peter Crosthwaite

University of Queensland (Australia)

p.cros@uq.edu.au

Abstract

This forum article revisits earlier claims about the relationship between corpus linguistics, data-driven learning (DDL), and generative AI three years after initial discussions of these issues. Drawing on recent developments in large language model (LLM) architectures, prompting strategies, and corpus-AI integration, the article reflects on how current research practices still need to evolve beyond traditional concordancing and basic AI-assisted analysis. It highlights emerging opportunities for combining generative AI with corpus resources in EAP and LSP contexts, while also emphasising the need for greater methodological clarity regarding prompting, data integration, and the interpretation of AI-generated language. The article concludes by arguing that the future of corpus-informed language research and pedagogy will likely lie in even closer integration between empirical corpus evidence and AI-mediated language technologies.

Keywords: corpus linguistics; data-driven learning; generative AI; large language models; English for Academic Purposes; English for Specific Purposes; corpus-AI integration; language pedagogy

Resumen

¿Ha muerto ya el DDL? ¡No tan rápido! Revisión de afirmaciones previas sobre el estado de la lingüística de corpus, el DDL, la IA generativa y las lenguas con fines específicos

Este artículo revisa afirmaciones previas acerca de la relación entre la lingüística de corpus, el aprendizaje guiado por datos (*Data-Driven Learning*, DDL) y la IA generativa, tres años después de las primeras discusiones sobre estas cuestiones.

Con base en los avances recientes en las arquitecturas de los grandes modelos de lenguaje (*Large Language Models*, LLM), las estrategias de *prompting* y la integración entre corpus e inteligencia artificial, el artículo reflexiona sobre cómo las prácticas actuales de investigación aún necesitan evolucionar más allá de las concordancias tradicionales y del análisis básico asistido por IA. Asimismo, destaca oportunidades emergentes para combinar la IA generativa con recursos de corpus en contextos de inglés para fines académicos y de lenguas para fines específicos y enfatiza la necesidad de una mayor claridad metodológica en relación con las estrategias de *prompting*, la integración de datos y la interpretación del lenguaje generado por IA. Finalmente, el artículo defiende que el futuro de la investigación y la pedagogía lingüísticas basadas en corpus probablemente residirá en una integración aún más estrecha entre la evidencia empírica proporcionada por los corpus y las tecnologías lingüísticas mediadas por IA.

Palabras clave: lingüística de corpus, aprendizaje guiado por datos, IA generativa, grandes modelos de lenguaje, inglés para fines académicos, lenguas para fines específicos, integración corpus-IA, pedagogía lingüística.

Introduction

Most introductions to applied linguistics research involving generative AI now begin with some variation of the claim that “Generative AI has significantly changed the landscape of [insert topic or discipline here]”. But has it really? Or, at least, has the research conducted so far done much more than apply a new tool to longstanding questions? Has this tool been properly theorised or methodologically integrated into existing research traditions? And perhaps more importantly, has the technology itself already evolved into something quite different from what it was in late 2022 and early 2023?

Of course, one cannot deny the complete explosion in studies on generative AI (henceforth GenAI) in applied linguistics (and, let’s be clear – large language models *are* applied linguistics), with LSP no exception. In a forthcoming handbook chapter on the use of AI in EAP, I point out that, at the time of writing (early 2024), a search for the term “artificial intelligence” on the homepage of *Journal of English for Academic Purposes* only resulted in eighteen articles since 2010 – albeit with only two of these dealing directly with generative AI. A similar query resulted in only two articles from *English for Specific Purposes*. Since then, the unprecedented surge in all things GenAI has led the Editor-in-Chief of *Journal of English for Academic Purposes* in 2026 to pen an editorial (cleverly titled “GenAI! GenAI? GenAI...”, Hu, 2026)

bemoaning the deluge of repetitive submissions analysing GenAI-produced texts or stakeholder perceptions of GenAI “surveying perceptions and challenges of GenAI among every possible demographic: teachers, students, males, females, novices, experts, and, I suspect, even the family pet” (p.1).

As an editor of a journal myself (*Australian Review of Applied Linguistics*), I can concur with Guangwei’s sentiment, and, upon personal reflection, as well as feedback on my own GenAI-related publications and conference talks, I also feel it is time to take a step back and reconsider what all this means and where we may be going. Accordingly, this forum article revisits some of my earlier claims about generative AI in applied linguistics and corpus-based pedagogy (with implications for LSP), examining whether the field has moved beyond technological novelty towards principled integration, or whether we are simply chasing (and even missing, in many cases) a rapidly shifting target. I also outline new opportunities for GenAI-related research in EAP/LSP, in the hope of revitalising work in both corpus linguistics and AI-assisted language research in this area.

Revisiting the DDL-GenAI Synthesis Debate

In 2023 (and so shortly after the advent of GenAI), I, together with Vit Baisa, had just released *CorpusMate*¹ (Crosthwaite & Baisa, 2024), a multi-disciplinary corpus query platform with a streamlined, Google like search interface and with accompanying data and query options spanning 20 disciplinary areas. *CorpusMate* was built on the previous work of the popular corpus query platforms *SketchEngine for Language Learning* (SKELL², Baisa & Suchomel, 2014) in terms of improving the user experience, and the *Michigan Corpus of Upper-Level Student Papers* (MICUSP³, Römer & O’Donnell, 2011) interface in terms of opening up a suite of disciplinary corpus query options.

However, mindful that my new creation - as well as the field of corpus-based data-driven learning (henceforth DDL, Johns, 1991), more generally - now faced a Gen-AI inducted existential crisis, I penned a similar forum article in *Applied Corpus Linguistics* (Crosthwaite & Baisa, 2023), since cited over 150 times since publication, that addressed certain recognised shortcomings of corpus approaches to DDL, certain evident shortcomings of GenAI (at the time), and a proposal to synthesise the two approaches to bring ‘the best of both worlds’ for language learners and teachers. My appeals to retain some degree of corpus-based DDL going forward centred on:

- a) Transparent, often open-access documentation and metadata underlying corpus files
- b) ‘Authenticity’ of corpus data (in terms of such data being produced by actual humans)
- c) Stable, replicable findings across corpus queries, as opposed to the variability observed under GenAI prompting and retrieval conditions
- d) Current state of the art corpus query tools (e.g. CorpusMate) providing multimodal concordances (over simple AI text output), and verifiable collocation statistics, frequency data and disciplinary distributional information based on actual corpus data.
- e) Safety, from ethical and teaching-focused perspectives
- f) Issues of hallucinations of GenAI output due to uncertainty around training data and minimal context windows.
- g) DDL’s emphasis on ‘active learning’ through learner discovery and hypothesis testing rather than passive adoption of GenAI output

Reflecting upon these premises for retaining corpus-based DDL in 2026, several are in need of amendment. Regarding access to LSP corpora, while many general corpora and some specialized corpora exist, domain-specific LSP corpora remain uneven and often scarce, especially for particular disciplines, languages, or pedagogical uses. Research on ESP/LSP corpus pedagogy still notes limited availability of specialised resources such as domain-specific or parallel corpora, which restricts empirical work in certain fields (El-Farahaty, 2025). Because of this gap, many studies recommend “DIY corpora” or small specialised corpora compiled by teachers and learners themselves for discipline-specific writing contexts (Charles, 2024). The ‘authenticity’ argument underpinning corpora is now equally problematic given the prevalence of AI-generated output permeating published research, social media and even oral forms of communication – it is highly unlikely now that any BootCat-like scraping of data sources to retrieve only human-produced text will result in only ‘authentic’ human produced data. Safety, in my opinion at least, is still an advantage from the perspective of corpus-based DDL from an ethical standpoint and for teachers and students in school contexts where AI is not permitted - although it also should be said that such contexts are disappearing rapidly. I

will discuss technical matters relating to c), d) and f) later in this article, although I can say here that such issues are rapidly becoming a thing of the past in the agentic AI era. Finally, I still strongly agree that corpus-based DDL represents a highly active form of engagement with language data with well-attested benefits for the acquisition of, among other things, disciplinary language. That said, studies seeking to combine DDL with GenAI in this regard are quickly gaining precedence (and preference) over standard concordance-based corpus-driven DDL, given reported usability issues with state of the art corpus tools are still finding their way into DDL literature. There is also an emerging body of research that while corpus-based DDL certainly ‘works’, it only really works with certain people (i.e. more on the intermediate to advanced end of the spectrum, and those already with a degree of technical experience, Lusta et al., 2023).

In terms of the case for bringing GenAI into DDL, I proposed the following:

- a) Lower required technical knowledge for corpus tools over interaction with GenAI chatbots
- b) The ability to easily generate synthetic disciplinary corpora
- c) Differentiation of output (i.e. the ability to tailor output for specific learners, e.g. beginner level vs advanced)
- d) Scale of data involved
- e) Iterative dialogue rather than one-shot corpus queries
- f) Ease of translation of output into teaching materials

While I have already discussed the ease of using GenAI tools over many open access corpus query tools used in LSP contexts, what I should say here is that as advances in GenAI have progressed, barriers to engaging with GenAI have reduced even further than what I proposed in 2023. As a case in point, I conducted a study with pre-tertiary teachers in Queensland, Australia in 2024 involving the construction of a ‘prompt taxonomy’ across several school subject areas including English, maths, science, music, physical education, and more (Crosthwaite et al. 2024). At that time, the art of producing prompts (or ‘prompt engineering’) was essential to meaningful interaction with GenAI as well as preventing hallucination. However, the advance in reasoning-based LLMs in state of the art GenAI applications (e.g.

Deepseek, ChatGPT version 5+) has largely negated the need for careful prompt construction as these systems can infer user intent, decompose tasks autonomously, and generate appropriate intermediate reasoning steps without relying on highly engineered prompts. We are now even asking LLMs to design and even redesign their own prompts, as I will discuss further in a later section.

Regarding the affordances of natural language dialogue-based AI systems, notably, the DDL community has also recognised that traditional corpus queries, and to some extent even traditional corpus consultation, might be better served through the integration of natural language prompts and explanations. Laurence Anthony (the creator of the *AntConc* corpus query platform commonly used in LSP studies, Anthony, 2024) blazed the trail here by integrating ChatAI into the latest versions of AntConc, giving users the power to query self-built corpora through natural language input. Similarly, Mark Davies' English-Corpora suite (English-corpora.org⁴, also commonly used in LSP) has also integrated LLMs, adding a catchphrase to his website to accompany this change – “Corpus first – AI second – Insights faster”. And, not wanting to miss out, I also co-developed an LLM driven corpus query platform (*CorpusChat*, Cheung & Crosthwaite, 2025) using retrieval augmented generation (RAG) techniques on disciplinary corpora to allow users to gain corpus-based insights through LLM-driven interaction. A number of special issues are now in chain for 2026 with major journals (e.g. *Computer-Assisted Language Learning*, *Language Learning and Technology*) to see where this brave new world takes us.

The power of the prompt: Issues with approaches to generating synthetic corpora with GenAI

Turning to the affordances of GenAI regarding synthetic corpora, this is one area where I advise caution for corpus linguists and LSP practitioners. A great deal of research has rapidly turned to the question of whether AI-generated academic writing resembles that produced by human writers. Much of this work has adopted a corpus-based comparative design where researchers generate a corpus of AI-produced texts using prompts modelled on authentic human-produced writing tasks matched for genre, topic, and length. Using established frameworks covering lexical choice, register variation, discourse organisation, and rhetorical structure, such studies then

attempt to identify similarities and divergences between human and AI writing in order to “shed light on the rhetorical preferences of LLM-generated writing and enhance our understanding of automated text production” (Jiang & Hyland, 2025, p.18). My own research (e.g., Mo & Crosthwaite, 2025) found that GenAI handles synthetic production of academic metadiscourse use broadly in line with disciplinary expectations, at least in Arts and Humanities related disciplines, without specific instruction to produce such metadiscourse. However, across different analytical traditions, findings tend to converge on similar conclusions, namely that while generative AI output can approximate the surface features of academic register, it may not fully capture what has been described as the “human touch” in academic writing.

That said, there is a need for greater caution in how differences between human- and AI-generated texts are interpreted. Much of the evidence comparing human and AI writing is highly context-dependent, not only in relation to task and genre, but also in relation to the elicitation conditions under which AI outputs are generated. In particular, prompt design constitutes a highly explicit and manipulable methodological variable that can substantially influence the linguistic properties of generated texts - far more so than might be usually present in human-produced corpora. Unlike human writing, where linguistic patterns emerge implicitly from genre and task conditions, large language models are *highly* sensitive to prompt wording, framing, and instructional detail, meaning any property of AI text can be shaped or co-constructed through prompting. Therefore, treating generative AI output as directly comparable across studies or conditions without reference to prompt specifications risks conflating properties of the underlying model with the specific instructions used to elicit the data.

To test this hypothesis, building on the work of Jiang and Hyland (2025) on corpus-based analyses of metadiscourse, and Kuteeva and Andersson (2025)’s investigation into the impact of prompting strategy (i.e. zero-shot, few-shot, and iterative role-based prompting) in relation to stance metadiscourse production, we created an AI-generated synthetic corpus of academic essays using a ‘baseline’ prompt that specified certain task conditions, role and genre, creating a rhetorically underspecified yet broadly ecologically valid corpus following Jiang and Hyland’s original work:

You are a competent senior high school student writer of English texts for academic purposes. Write 16 argumentative essays with a persuasive writing style on the topic of [transport]. Each essay is about 500 words long.

This corpus was then annotated for interactive and interactional metadiscourse features according to Hyland's framework, including stance markers (e.g. hedging, boosting, self-mention, attitude markers), engagement markers (e.g. questions, appeals to shared knowledge, endophoric markers, etc.), and text organising markers (e.g. transitions, frame markers, etc.). The frequencies of these features were compared with data from a human-produced reference corpus (LOCNESS), annotated using the same taxonomy. The findings were largely in line with those reported for previous research, namely that clear differences emerged between human writing and GPT-generated texts under the baseline prompt in terms of stronger rhetorical control over organisation in human writing than in GPT texts, while human writers used more hedging, stance markers, and reader engagement strategies with GPT writing more impersonal and monologic.

So far, so good? However, in the next stage of the study, we created a third corpus to be generated using AI, this time specifically adding mention of the metadiscoursal features that were derived to be significantly different between human vs AI writing from the first round of analysis. This metadiscourse-specific prompt is seen below:

"You are a competent university-student writer of English texts for academic purposes. Write [5] argumentative essays with a persuasive writing style on the topic of [transport]. Each essay is about 500 words long. You need to perform a human-like employment of metadiscourse markers from the interactive perspective by using more transitions (including contrastive, consequence, causal and conditional transitions), endophoric markers, code glosses and evidentials (including reporting evidence and reference to scholarship), while using fewer frame markers. From the interactional perspective, you need to use more hedges, self-mentions, and attitude markers. From an engagement perspective, you need to use more reader pronouns, directives, questions to the reader, appeals to shared knowledge, and personal asides."

To put it simply, the difference in results was staggering (Figure 1, Crosthwaite et al., 2026). Under a revised prompt that explicitly encouraged metadiscourse, GPT increased its use of transitions, engagement markers, and stance features, in some cases matching but in many cases completely exceeding the frequencies observed from the baseline prompt and the human reference corpus. Qualitative analysis also revealed that these resources were frequently overproduced or used in formulaic ways, suggesting that the model responded to prompt instructions in an amplificatory manner rather than deploying them strategically within the argument.

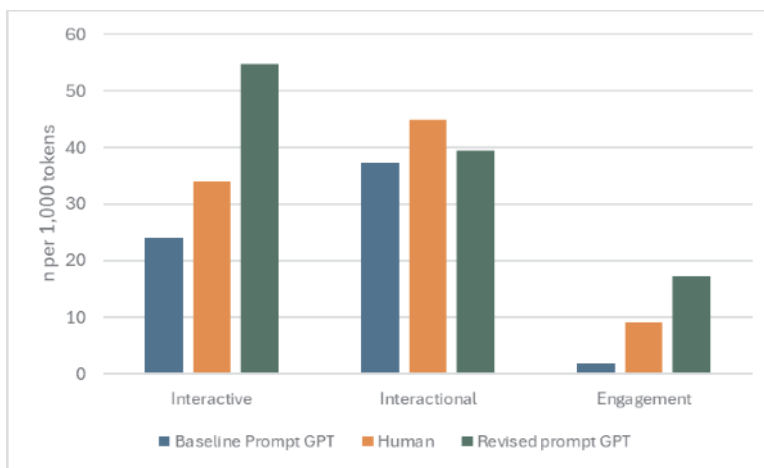


Figure 1 – Metadiscourse in baseline vs human vs specific metadiscourse prompt conditions (taken from Crosthwaite et al., (2026))

Methodologically, the results highlight the importance of treating prompt design as an explicit variable in corpus-based analyses of AI-generated writing. As AI outputs are heavily interaction-dependent artefacts that may emerge entirely explicitly through the relationship between user instructions and the model, future EAP/ESP research should therefore document prompting procedures much more carefully, and examine how discourse features vary across different prompts and iterations, combining quantitative corpus analysis with qualitative examination of rhetorical function.

From a pedagogical perspective, these findings suggest that because the rhetorical profile of AI output varies substantially depending on prompt design, learners who treat such texts as exemplars of academic style may adopt distorted patterns of stance, engagement, or textual organisation depending on the prompt used to generate them. Instead, generative AI may be more productively used as an object of critical analysis within EAP instruction, with teaching therefore focusing on developing students' AI literacy by helping them evaluate AI-generated drafts, recognise the rhetorical effects of prompt design, and iteratively revise AI output so that it better aligns with disciplinary expectations of argumentation and reader engagement.

The future (for now?)

Returning to matters around safety, variability of LLM findings over static corpus query results, hallucination, and advantages of purpose-built corpus tools over GenAI chatbots, it is now becoming apparent that even these issues are no longer as serious as they once were.

One of the biggest changes in the GenAI landscape from 2025 onwards is the availability of open-source LLMs that can run locally even on modest hardware, with more powerful versions available for fine-tuning relatively cheaply through cloud services e.g. Amazon AWS. This allows for users to be able to upload data, for example learner corpora of student texts, for LLM consultation and/or training without requiring users feed their data into third party training mechanisms, a frequent complaint against GenAI adoption from those in tertiary and particularly pre-tertiary education contexts (at least in Australia where I am based).

Moreover, the increasing availability of cloud-based GenAI platforms has significantly reduced the technical barriers and costs associated with adapting large language models. Many providers now allow researchers to customise models through fine-tuning or retrieval-based approaches (RAG) using their own datasets, including open-access corpora of human-produced texts. This was the approach taken in *CorpusChat* as discussed earlier. Integrating curated corpora in this way also helps mitigate some limitations associated with smaller context windows in earlier or free versions of GenAI systems, which could otherwise lead to hallucinated responses when models attempt to reason over external data. At the same time, anchoring model outputs to a fixed corpus can reduce one of the most common criticisms of generative AI in language research, namely that responses vary across prompts and repeated queries, thereby bringing AI-assisted analysis closer to the stability and reproducibility traditionally associated with corpus-based approaches.

In addition, new methods of incorporating external data into LLM pipelines such as emerging recursive or iterative language modelling approaches may offer potential advantages over more widely used retrieval-augmented generation (RAG) systems. In a typical RAG architecture, documents from an external corpus are segmented, represented through semantic vector embeddings, and retrieved for insertion into the model's context window during generation. Although the underlying corpus itself remains unchanged, this process alters how the corpus is accessed and represented, as only small retrieved segments are provided to the model at each step – a

complaint made against *CorpusChat* in that user queries were not exactly based on the original corpus data, but an abstraction of that data. By contrast, recursive approaches allow the model to iteratively query, evaluate, and refine its use of external data across multiple reasoning cycles, with the potential to support more structured reasoning and deeper integration of corpus evidence. For corpus-based language research, such architectures may therefore offer a promising pathway for integrating specialised datasets into AI systems while maintaining closer alignment between model outputs and the linguistic evidence contained in the corpus itself, although this remains largely untested at present in our field.

Another emerging development concerns the growing use of machine-generated prompting rather than prompts written manually by users. Early work with generative AI often assumed that effective use depended on carefully crafted human prompts, leading to the rise of “prompt engineering” as a key skill. However, recent systems increasingly generate, refine, or optimise prompts automatically through iterative interaction with the model itself. In these approaches, the system can generate multiple candidate prompts, evaluate their outputs against predefined criteria, and recursively improve them. This shifts prompting from a purely human activity to a computational process that can systematically explore large prompt spaces and identify instructions that produce more reliable or task-appropriate outputs. Such machine-generated prompting may allow researchers to systematically test large numbers of prompt variants, helping to identify which instructions produce outputs most closely aligned with the linguistic patterns observed in reference corpora. Secondly, automated prompt optimisation could improve the reproducibility of AI-assisted analyses by reducing the variability associated with individual users’ prompting styles. Furthermore, such approaches may enable the development of corpus-informed prompting systems, where prompts are dynamically generated on the basis of patterns observed in curated corpora, potentially strengthening the link between generative AI outputs and empirical linguistic evidence (“corpus-driven prompting”).

Finally, since late 2025, there has been a marked shift in the affordances of ‘vibe-coded’ or ‘no-code’ AI solutions to creating almost any type of web-based application, made possible through software such as Anthropic’s *Claude*. While previous interaction with GenAI for DDL was most entirely through text-only chatbots (and later chatbots that allowed image and video input/output), we now have the ability to actually create our own corpus

query tools similar in function to *CorpusMate* or *AntConc*. As a case in point, using Claude in early 2026, I was able to ‘vibe-code’ a fully online corpus query platform (*CorpusMate* BAWE⁵) based on the functionality of *CorpusMate* and the BAWE disciplinary corpus, only this time incorporating natural language query input options, as well as natural language explanations of corpus query output, made possible through Anthropic API key calls, broadly emulating the *ChatAI* function in the latest versions of *AntConc*. The website (Figure 2) contains options to query BAWE files using standard corpus queries (e.g. “however”) or natural language inputs (e.g. “how is ‘however’ used in academic writing”), view concordances, filter queries by BAWE metadata (e.g. disciplinary group, text type, etc.), and generate results for n-grams and disciplinary distribution as seen in *CorpusMate*.

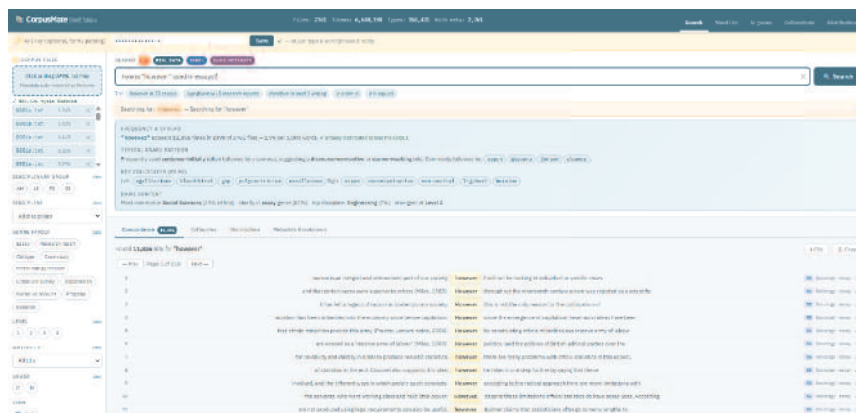


Figure 2: CorpusMate BAWE interface

Users only need a copy of the raw BAWE text files (available open access), and an Anthropic API key for LLM queries (normal corpus queries do not even require the API key). The most impressive thing about constructing this platform was that it was done without any knowledge of coding, based entirely a set of on natural language prompts that sought simply to add LLM query and analysis functions to an existing corpus tool (*CorpusMate*). While writing this article, using the same approach, I have now completely redeveloped the original Corpusmate (Corpusmate v2, <https://corpusmate.app>) with full LLM support, multiple user roles (teacher, researcher, language learner) and support for 12 languages. The implications of this are potentially massive for LSP and applied linguistics

more generally, in that researchers and teachers can now prototype sophisticated corpus-AI tools without specialised programming expertise, lowering the technical barriers to developing customised resources for discipline-specific language analysis, pedagogy, and research. The sky really is the limit here.

Conclusion

I hope this forum article has given readers plenty of food for thought regarding the current state of corpus-AI research and teaching for LSP and more broadly. Reflecting on previous claims regarding the affordances of integrating GenAI into corpus linguistics and DDL, I am now even more positive about the future potential in this space for researchers, teachers and students. Developments in LLM pipelines, automated prompting, and new ways of integrating corpus data into AI systems suggest that we are entering a period where corpus-informed language analysis and support can become more interactive, scalable, and accessible than was previously possible.

At the same time, these developments invite the field to reflect on how corpus-based approaches themselves are evolving. Much current practice in corpus linguistics and DDL remains centred on relatively traditional modes of interaction with corpus data, particularly concordancing and frequency-based exploration. These methods remain extremely valuable and have formed the backbone of corpus-informed pedagogy for decades. However, the rapid development of AI-assisted analytical tools suggests that there may now be opportunities to extend these approaches further, for example by allowing users to query corpora in more natural language terms, explore rhetorical or discourse-level patterns more easily, or integrate corpus evidence directly into writing support systems.

If corpus linguistics and DDL are to remain central to the future of technology-enhanced language learning, the field may therefore need to continue experimenting with new forms of interaction between users, corpora, and intelligent systems. Rather than replacing established corpus practices, generative AI offers the possibility of augmenting them, helping learners and researchers move more fluidly between concordance evidence, linguistic analysis, and applied language tasks. The challenge ahead will be to ensure that these emerging tools remain grounded in the empirical strengths

that have always defined corpus linguistics, while also embracing the new methodological and pedagogical opportunities that AI-mediated language technologies now make possible.

Article history:

Received 12 March 2026

Received in revised form 28 March 2026

Accepted 18 May 2026

References

- Anthony, L. (2024). AntConc (Version 4.3.1) [Computer software]. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software/AntConc>
- Baisa, V., & Suchomel, V. (2014). SkELL: Web interface for English language learning. In A. Horák & P. Rychlý (Eds.), *Proceedings of the Eighth Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2014)* (pp. 63–70). NLP Consulting. https://www.sketchengine.eu/wp-content/uploads/SkELL_2014.pdf
- Charles, M. (2024). Benefits and Challenges of Using Do-it-yourself Corpora for Academic Writing Development. *TESOL Quarterly*, 58(3), 1205–1214. <https://doi.org/10.1002/tesq.3276>
- Cheung, L., & Crosthwaite, P. (2025). CorpusChat: Integrating corpus linguistics and generative AI for academic writing development. *Computer Assisted Language Learning*. Advance online publication. <https://doi.org/10.1080/09588221.2025.2506480>
- Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3), Article 100066. <https://doi.org/10.1016/j.acorp.2023.100066>
- Crosthwaite, P. and Baisa, V. (2024): Crosthwaite, P., & Baisa, V. (2024). A user-friendly corpus tool for disciplinary data-driven learning: Introducing CorpusMate. *International Journal of Corpus Linguistics*, 29(4), 595–610. <https://doi.org/10.1075/ijcl.23056.cro>
- Crosthwaite, P., Sun, S. & Jiang, FK. (2026). Prompt Sensitivity in AI Writing: A Methodological Caution for Corpus-Based Analysis of AI-Generated Language. *The Modern Language Journal*, to appear.
- Crosthwaite, P., Smala, S., & Spinelli, F. (2025). Prompting for pedagogy? Australian F-10 teachers' generative AI prompting use cases. *The Australian Educational Researcher*, 52, 1795–1825. <https://doi.org/10.1007/s13384-024-00787-0>
- El-Farahaty, H. Integrating Corpora and AI Tools in the Teaching, Learning, and Assessing of Arabic/English Legal Translation. *International Journal of Semiotics of Law* 38, 2031–2060 (2025). <https://doi.org/10.1007/s11196-025-10281-0>
- Hu, G. (2026). GenAI! GenAI? GenAI... *Journal of English for Academic Purposes*, 80, Article 101650. <https://doi.org/10.1016/j.jeap.2026.101650>
- Jiang, F. (Kevin), & Hyland, K. (2025d). Rhetorical distinctions: Comparing metadiscourse in essays by ChatGPT and students. *English for Specific Purposes*, 79, 17–29. <https://doi.org/10.1016/j.esp.2025.03.001>
- Kuteeva, M., & Andersson, M. (2025). Negotiating stance expressions in research writing with GenAI: A three-step prompt approach for ChatGPT. *Research Methods in Applied Linguistics*, 4(3), 100267.
- Lusta, A., Demirel, Ö., & Mohammadzadeh, B. (2023). Language corpus and data driven learning (DDL) in language classrooms: A systematic review. *Heliyon*, 9(12), Article e22731. <https://doi.org/10.1016/j.heliyon.2023.e22731>
- Mo, Z., & Crosthwaite, P. (2025). Exploring the affordances of generative AI large language models for stance and engagement in academic writing. *Journal of English for Academic Purposes*, 75, Article 101499. <https://doi.org/10.1016/j.jeap.2025.101499>
- Römer, U., & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 6(2), 159–177. <https://doi.org/10.3366/cor.2011.0011>

Peter Crosthwaite is an Associate Professor in Applied Linguistics at the university of Queensland. His research focuses on second language writing, corpus linguistics, English for Academic Purposes, computer-assisted language learning, and the role of generative AI in language education and academic writing. He has published widely on corpus-based approaches to language learning, learner corpora, written corrective feedback, academic discourse, and AI-mediated writing support. He is Editor-in-Chief of the *Australian Review of Applied Linguistics* and serves on the editorial boards of several leading journals in applied linguistics. His current work examines AI, feedback, and evidence-based language pedagogy.

NOTES

¹ <https://www.corpusmate.com>

² <https://skell.sketchengine.eu/>

³ <https://elicorpora.info/>

⁴ <http://english-corpora.org/>

⁵ <https://pcros1980.github.io/CorpusMate-BAWE-with-LLM-support/>

