

## Corpora and GenAI tools for learning language and content

Alex Boulton

ATILF – CNRS & Université de Lorraine, France, , [alex.boulton@univ-lorraine.fr](mailto:alex.boulton@univ-lorraine.fr)

How to cite: Boulton, A. (2025). Corpora and GenAI tools for learning language and content. In Y. Choubsaz, P. Díez-Arcón, A. Gimeno-Sanz, V. Morgana, A. C. Murphy & F. L. Seracini (Eds.), *Advancing CALL: New research agendas - EUROCALL 2025 Short Papers*. <https://doi.org/10.4995/EuroCALL2025.2025.21189>

---

### Abstract

*This paper brings together two strands of technology for CALL – corpus tools and generative AI – in an integrative approach to learning both language and content for ESP. Geography students in France in the first year of a master's degree worked in small groups to compile an academic corpus on a disciplinary topic of their choice, then explored this using AntConc, with GenAI to help present their work in written reports and spoken presentations. The paper outlines the course and draws on the post-course questionnaires to explore how students appropriated the tools and methodologies of both corpus linguistics and GenAI.*

**Keywords:** data-driven learning; GenAI; DDL; integration.

---

## 1. Introduction

Data-driven learning (DDL) exploits corpus tools and techniques for learning and using foreign or second languages (L2), especially via concordances, collocates, frequency lists and distributions. Though commonly used for lexicogrammar, it can also suggest nuances of meaning: the phrase *loved ones*, for example, may seem to be an obvious term for close friends and families, but it does not exist in many dictionaries, and examination of corpus data shows that it is typically associated with bereavement or other traumatic events; the top 10 collocates from the Corpus of Contemporary American English include *lost*, *victims*, *deceased*, *deaths* and *graves* (cf. Figure 1). DDL has been found to produce medium to large effect sizes in most areas where it has been tried extensively (Boulton & Cobb, 2017), but barriers to its implementation include training to use procedures and interfaces which are not always intuitive, formulating queries, and interpreting the output (Sun & Mizumoto, 2025). An obvious question in this context is whether generative artificial intelligence (GenAI) might be able to counter some of these difficulties.

" Well, why don't you ask the families whose **loved ones** were slaughtered by the ISIS Army and buried in sl... are the pictures of people who lost so many **loved ones** there is no one to remember them. " She found her... I've seen if they've been able to reach their **loved ones**, their family members, if they've been able to fil... unfortunate things going on in your life? Your **loved ones** will help guide you and your true friends will st... ting for us. Hundreds of their neighbors and **loved ones** are dead, injured or missing. Much of their city... of proof or because they fear they or their **loved ones** will be subject to bodily harm if they return a n... d them to, not out of choice. They and their **loved ones** deserve no less than our full support in this mom... is. Encourage your son or daughter to talk to **loved ones**, friends, or a counselor about any problems he or... w, you can get another complete set for your **loved ones** (just pay separate shipping and handling). What b... and especially when they're with their, their **loved ones** or these, or having these, being in these very in

Figure 1. Random concordance data for loved ones (COCA).

Following the public release of ChatGPT in November 2022, much attention has been focused on it as a “disruptive technology” (Todd, 2025) – in other words, to consider how to limit its effects so that teaching and research can continue as before. For educational aspects, the institutional response has concentrated on eliminating cheating (Freeman, 2025), largely involving a return to pen-and-paper evaluation in exam rooms. While this may work in the short term, it is essentially a capitulation to technology; a longer perspective might focus on its potential for positive change – potential to reduce the need for some learning (especially the boring parts), but potential also to help with learning. In research, the extreme scenario would be that entire fields of study become redundant and disappear. Surprisingly, perhaps, many researchers with vested interests and entire careers built in such areas have been eager to test GenAI and are open to the changes it may bring. For example, current generations of GenAI may be inferior to corpus linguistics in some respects, at least at this point in time, but some studies suggest that it can actually generate “much better data than even the most advanced corpora” (Davies, 2025, n.p.). Similarly, we cannot simply assume that DDL will endure in its current form when GenAI can do many of the same things faster, more efficiently, more intuitively (Crosthwaite & Boulton, 2025). But nor should we assume the worst. For DDL as in other fields, the answer may be: “not so fast!” (Crosthwaite & Baisa, 2023). What is needed is empirical evidence from actual classrooms in addition to researcher experimentation with GenAI (e.g. Liu & Ma, 2025).

This paper reports on an attempt to integrate corpus techniques and GenAI into a course that goes beyond just language-oriented DDL to include discovery of discipline-specific content. In earlier work, Boulton (2024) asked five cohorts of his online master’s students of English to compile and explore a corpus relating to their disciplinary interests, reporting back as a final project. Charles produced a series of papers over several years of a course for master’s and doctoral students with an immediate need for disciplinary writing (course papers, dissertations, publications), also compiling a corpus of research articles (RAs) for exploration (see Charles & Hadley, 2022). Such studies have typically targeted relatively advanced and motivated students with specific needs, while there has been less work with other population types (Boulton & Vyatkina, 2021).

## **2. Method**

This paper explores the introduction of DDL and GenAI with master’s students studying Geography (n=33) in France who do not have immediate writing needs and are mostly at B1-B2 level of proficiency. It was guided by the following research questions: ±

- (1) How do these learners take to the approach and tools?
- (2) How do they combine corpus software and GenAI?
- (3) Do they learn about their topic and the language at the same time?

The first two-hour classes introduced the tools, assignments and objectives, with the following 16 hours allowing the students to explore all of this and produce their final projects, presented in the last two sessions. Mastering all tools was an integral part of the course, from their own explorations, in-class help and online video tutorials; the objective being to provide a original perspective on English in their field, which they discovered for themselves. Each sub-group of about three students chose 15 research articles (RAs) in pdf format (unedited except to exclude references and extraneous matter) to analyse using AntConc (Anthony, 2024); the total corpora consisted of 165 academic texts (1.2m tokens), mostly published in appropriate journals over the last 10 years, and relating to each group’s specific topic. The students completed a collaborative worksheet involving word lists (with and without a tailored stoplist) and *n*-gram lists of varying length, and exploring selected items in more detail (plot, cluster, collocates and KWIC tools). Each group’s written report included a corpus breakdown and full references (APA7) along with corpus screenshots to force use of AntConc, each accompanied by a short comment; they could use GenAI to help make sense of the results and to help write up the results. They were asked to provide 200-word outlines for what they felt they had learned (a) about the language, and (b) about the tools and approach (specifically, corpora and GenAI). A final group presentation was based more on the topic as featured in the RAs, with AntConc and GenAI helping with the analyses, scripts and visuals. The data collected

thus consisted of corpora, written assignments and visual supports for the oral from each of the 10 groups, along with individual course marks and questionnaires for all 33 participants, collected online after the course end and following best practice (e.g. Dörnyei & Dewaele, 2023; Mizumoto et al. (2016). Given space constraints, the results presented here are mainly derived from the questionnaires, somewhat condensed.

### 3. Results

The groups' final grades in this course ranged from 57% to 82% ( $M = 67\%$ ,  $SD = 8.3\%$ ). As pedagogical marks, this seems highly satisfactory, but further findings were more mixed. The final two sections of the worksheet involved free text so were submitted to an AI-detector, estimated to be between 6% and 100%. The mean is certainly high at 66% ( $SD = 0.36$ ), but classroom observation over the semester noted that they were for the most part producing original text and using GenAI to improve it, as requested. There was a slight inverse correlation ( $r = -.23$ ) between this and the written marks, suggesting that the 'weaker' students resorted more to AI, though the two best marks were for assignments estimated at 13% and 81% AI-generated text respectively. The students formed into groups spontaneously, mostly with friends, meaning that each was of mixed proficiency; overall, they self-estimated mainly at B1 ( $N=13$ ) or B2 ( $N=11$ ) level. More specifically, they rated various English-related abilities relatively low, speaking being the highest skill with a mean of just 3.1 (Figure 2). Though they had little previous experience of reading RAs even in French, by the end their average assessment was among the highest, as it was for corpus tools ( $M=3.21$  for both), which none had ever used before. In general, there was a medium correlation between their self-assessment and their enjoyment of English ( $M=3.5$ ;  $r=.55$ ) and with their daily use of the language ( $M=2.7$ ;  $r=.51$ ), but not their perception of the importance of English in their studies ( $M=3.4$ ;  $r=.22$ ), still less for their future careers ( $M=3.3$ ;  $r=.17$ ), as recorded on a 5-point Likert scale ranging from 0 = strongly disagree to 5 = strongly agree (intermediate scores were not labelled).

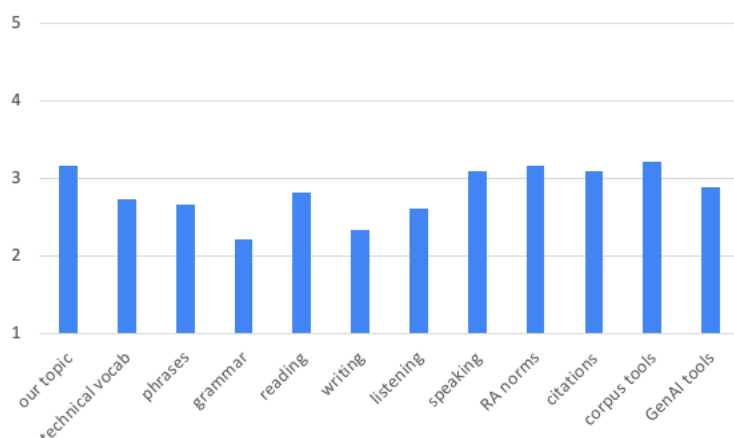


Figure 2. Mean ratings for English-related abilities (5pt Likert scale).

#### 3.1. Corpus tools

Appreciations of the corpus tools were also relatively low: the students rated the ease of use of AntConc at only 2.4 (Figure 3); of the various tools in the software, their favourite was the n-gram tool but still only at 3.0, while all the others scored between 2.7 and 2.8. This is partly because 7 students consistently ranked all items low, i.e. averaging less than 2.0. These preferences also do not seem to correlate with their overall English proficiency ( $r=.04$ ).

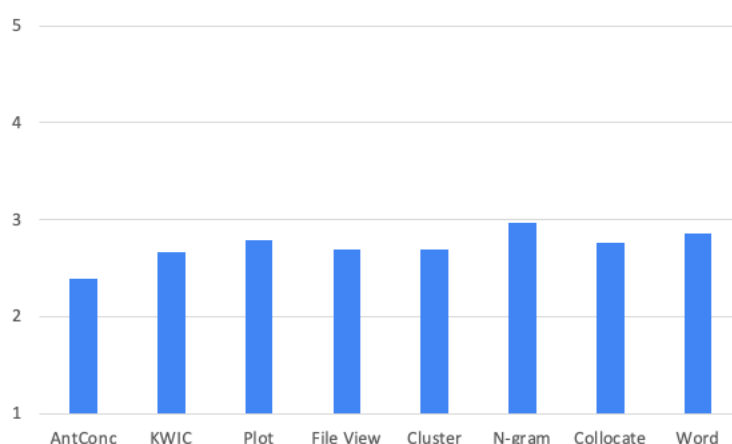


Figure 3. Mean ratings for corpus tools and use.

### 3.2. GenAI tools

Presented with a checklist, all students said they had used at GenAI tools at least once; the favourites were ChatGPT and DeepL–translate for 28 students, followed by Scribbr (14), DeepL–write (13), Gemini (7), DeepSeek (6), and a handful of others. For regular use, they cite ChatGPT (15), DeepL (10) and Scribbr (4); a few others were listed by one person only. They rate GenAI for personal use and for their studies (Figure 4), as well as for English, and claim to be aware of its advantages ( $M=4.0$ ) and especially the associated dangers  $M=(4.4)$ , though they do not always check results ( $M=3.0$ ). For the present course, they claimed mainly to use it for improving text ( $M=3.2$ ), more than summarising ( $M=3.2$ ), translating ( $M=2.5$ ) or generating text ( $M=2.4$ ). Though they generally found it easy to use ( $M=3.64$ ), it is interesting to note that they did not rely on it much to analyse the output from AntConc ( $M=2.4$ ); this suggests that they did not find the data output from the various tools particularly difficult to interpret.

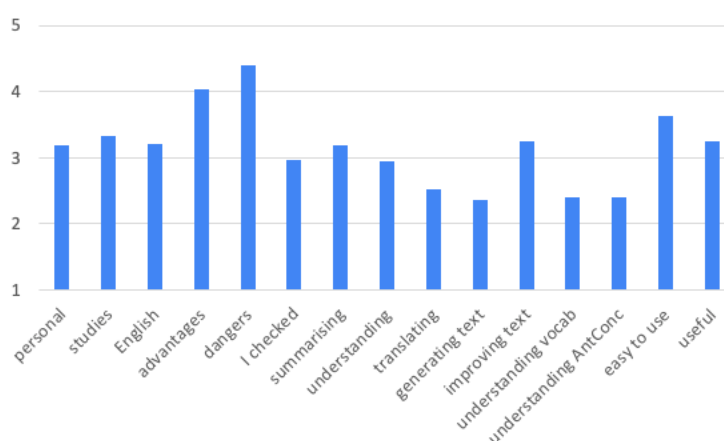


Figure 4. Mean ratings for GenAI tools and use.

### 3.3. Overall course appreciation

Feedback on the course as a whole was quite mixed (Figure 5): the students particularly appreciated freedom to choose their own topic ( $M=4.6$ ) and the trust they felt from the teacher in this regard ( $M=4.18$ ); they also liked working in groups for the most part ( $M=3.8$ ), but not so much that this required them to work actively rather than ‘being taught’ ( $M=2.8$ ). They preferred working on their oral presentations ( $M=3.6$ ) to exploring RAs ( $M=3.0$ ), and preferred using GenAI ( $M=3.0$ ) to AntConc ( $M=2.15$ ), though a combination of the two was perhaps an appropriate mix ( $M=2.8$ ). Open-ended questions indicate a certain distrust of GenAI, which is the position taken

by many of their teachers, which may sit uncomfortably with being encouraged to use it in this class; they would also have appreciated more extensive explanation on why a corpus approach is useful, and how to use AntConc.

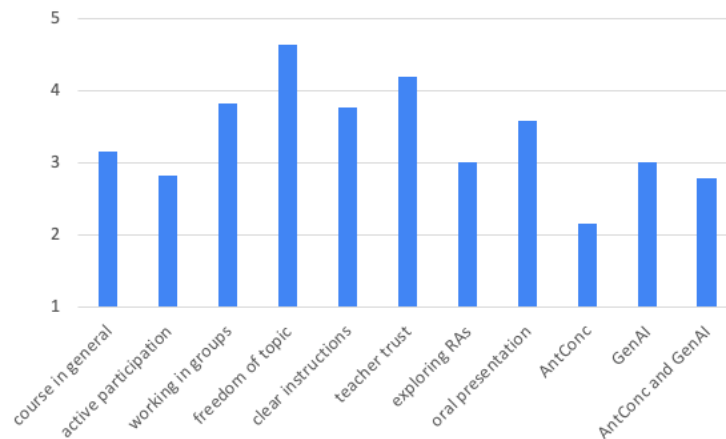


Figure 5. Mean ratings for course appreciation.

#### 4. Discussion and Conclusion

This paper outlined a course that required ESP students in France to work collaborative on a DDL approach to analysing a self-compiled corpus relating to a topic they chose themselves, alongside GenAI. Other data sources have yet to be fully analysed (coursework and corpora in particular), but the post-course questionnaire results are mixed: the students appreciated the freedom to work on a topic of their own choice, but did not fully understand the rationale underlying the approach; they felt they were left too much to their own devices and would have liked more specific guidance in using the corpus tool, and more ‘being taught’ as they were used to. Of course, this reflects only their stated perceptions and preferences – which can tend towards the negative for almost any approach to ESP in HEI in France – and in any case, the teacher is not there just to do what the students want; however, it would be arrogant simply to ignore them. It may be an indication that as a teacher also researching DDL, it is easy to assume that students will immediately see why the approach should be helpful, and to underestimate the difficulties in introducing a new approach to students at these levels of proficiency and with limited ESP needs, both current and future.

Despite the mixed results, the intention is to continue but with a number of changes. Future iterations will offer a more gentle lead-in to DDL and its rationale, presenting more varied and accessible examples at the start of the course which the students can mirror on their own laptops, before requiring them to form their groups, choose their topic and start work. AntConc and GenAI will be presented in tandem, to be integrated rather than treated separately; the written assignment will provide more specific guidance in the form of a template; and the groups will be monitored more closely rather than waiting for students to request help. This reflects a conviction that students can be trusted to make important decisions about their language learning, provided that scaffolding is made available; that DDL can bring real benefits to such students, for the specific language and topic covered, as well as a novel perception of what language is and how it functions in specific cases, which should help in the future; and that GenAI can and should be integrated in language education rather than ignored (Pérez-Paredes & Boulton, 2025).

As outlined in Crosthwaite and Boulton (in press), given that GenAI exists, it would be blinkered to assume that DDL can or should continue as before, entirely unaffected. The other extreme – that DDL should be abandoned altogether – seems equally undesirable. More promising is to play to the strengths of each, though how exactly remains to be seen. None of the groups in the present study showed interest in using the ChatAI tab in AntConc, and only some of them used their chosen GenAI tools to help interpret the corpus output, preferring to use it to refine their writing and produce visuals for the oral presentations. Whether the DDL and GenAI tools can merge more fruitfully remains to be seen.

## Acknowledgements

I would like to thank the students involved, as well as Carmenne Kalyaniwala, Alessandra Cacciato, Jarvis Looi and Marc Schirès for discussion and support.

## References

- Anthony, L. (2024). *AntConc* (Version 4.3.1) [Computer Software]. Waseda University. <https://www.laurenceanthony.net/software/AntConc>
- Boulton, A. (2024). Corpus à volonté : À l'écoute des étudiants. *Humanités, Didactiques, Recherches*, 4, 117-136.
- Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 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. *Language Learning & Technology*, 25(3), 66-89. <https://doi.org/10.125/73450>
- Charles, M., & Hadley, G. (2022). Autonomous corpus use by graduate students: A long-term trend study (2009–2017). *Journal of English for Academic Purposes*, 56, 101095. <https://doi.org/10.1016/j.jeap.2022.101095>
- Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3), 100066. <https://doi.org/10.1016/j.acorp.2023.100066>
- Crosthwaite, P., & Boulton, A. (2025, in press). DDL is dead? Long live DDL! Expanding the boundaries of data-driven learning. In H. Tyne, M. Bilger, L. Buscail, M. Leray, N. Curry & C. Pérez-Sabater (Eds.), *Discovering language: Learning and affordance*. Peter Lang.
- Davies, M. (2025, May). *Integrating AI / LLMs into English-Corpora.org*. <https://www.english-corpora.org/ai-llms>
- Dörnyei, Z., & Dewaele, J-M. (2023). *Questionnaires in second language research: Construction, administration and processing*. Routledge. <https://doi.org/10.4324/9781003331926>
- Freeman, J. (2025). *Student generative AI survey 2025*. Higher Education Policy Institute / Kortext. <https://www.hepi.ac.uk/2025/02/26/student-generative-ai-survey-2025>
- Liu, J., & Ma, Q. (2025). Supporting low-proficiency L2 learners' vocabulary learning with custom GPT-scaffolded corpus-based language pedagogy: a case study. *Computer Assisted Language Learning*, Advance Online Publication. <https://doi.org/10.1080/09588221.2025.2539152>
- Mizumoto, A., Chujo, K., & Yokota, K. (2016). Development of a scale to measure learners' perceived preferences and benefits of data-driven learning. *ReCALL*, 28(2), 227-246. <https://doi.org/10.1017/S0958344015000208>
- Pérez-Paredes, P., & Boulton, A. (2025). *Teaching and data-driven learning in and out of the language classroom*. Cambridge University Press. <https://doi.org/10.1017/9781009511384>

- Sun, A. X., & Mizumoto, A. (2025). Exploring the barriers to data-driven learning in the classroom: A systematic qualitative synthesis. *Applied Corpus Linguistics*, 5(2), 100126. <https://doi.org/10.1016/j.acorp.2025.100126>
- Todd, R.W. (2025). Generative AI as a disrupter of language education. *International Journal of TESOL Studies*, 250127. <https://doi.org/10.58304/ijts.250127>