OPEN ACCESS

Artificial Intelligence in Language Education
ISSN
https://www.castledown.com/journals/aile/

Artificial Intelligence in Language Education, 104826 (2026)
https://doi.org/10.29140/aile.2026.104826

CorpusMate 2.0: Integrating AI-powered language analysis into a data-driven learning corpus tool

PETER CROSTHWAITE

University of Queensland, Australia

Abstract

This article presents *CorpusMate 2.0* (https://corpusmate.app/), a substantially rebuilt version of the *CorpusMate* corpus tool for data-driven learning (DDL) first described in Crosthwaite and Baisa (2024). The new version retains the core concordancing, n-gram and disciplinary variation functions of its predecessor while introducing a range of significant enhancements including a rebuilt Python/FastAPI backend, five collocation association measures (including MI², MI³, and log-likelihood), and – most substantially – a fully integrated large language model (LLM) assistant that reads users' actual search results to provide evidence-grounded linguistic commentary. As a novel feature specific for DDL, the AI assistant also operates across four user roles (Researcher, Teacher, Student, Language learner) each with a distinct pedagogical persona, while the platform also supports responses in eleven languages. This article describes the motivation for these changes, the new platform architecture, and enhanced functionality, while inviting continued use of the tool for DDL research and pedagogy.

Keywords: Data-driven learning, CorpusMate, concordancing, large language models, AI-assisted DDL, collocation, disciplinary variation

Introduction

Since its public release in 2023, CorpusMate (Crosthwaite & Baisa, 2024) has been adopted in language teacher training programmes across Asia-Pacific and Europe (e.g. Chung et al., 2024) while accumulating feedback from over 180 users via its embedded survey form. That feedback, alongside other ongoing developments in the fields of corpus linguistics and generative artificial intelligence (AI), motivated a comprehensive rebuild of the tool – CorpusMate 2.0 – which is described here.

The original CorpusMate addressed a persistent problem in data-driven learning (DDL) research, now well-established across three decades of empirical study (Boulton & Vyatkina, 2021),

namely the gap between the demonstrable learning gains achievable through corpus consultation and the practical barriers that prevent teachers and learners from sustaining corpus use after initial training. These barriers, namely complex user interfaces, unsuitable corpus data for younger or lower-proficiency learners, and the cognitive overhead of interpreting raw concordance output (Poole, 2022) led Crosthwaite and Baisa (2024) to develop CorpusMate as a purpose-built DDL tool prioritising simplicity, accessibility and discipline specificity.

Since the release of CorpusMate, however, the broader educational technology landscape has shifted significantly. The widespread adoption of generative AI tools, most notably ChatGPT, has raised learner and teacher expectations of what language support software should do, in that it is no longer apparently enough for tools to simply retrieve and display language data, but should assist with explaining it, contextualising it, and responding to natural language questions about it. Accordingly, research into AI and corpus linguistics has begun to explore how LLMs might complement rather than replace traditional corpus methods (e.g. Ebeling & Heuberger, 2025). As of 2026, it is apparent that the integration of LLMs into corpus tools is now a rapidly emerging development across the field, with Anthony (2025) already introducing a *ChatAI* module within AntConc v4 that allows users to pipe concordance output directly to an LLM for interpretation. Mizumoto (2023) has articulated a theoretical framework (Metacognitive Resource Use) for combining generative AI with corpus and other digital tools in ways that preserve learner agency, and has since developed LexiTracker, an AI-integrated vocabulary and writing support platform currently in active development (Mizumoto, in development). The first author has also collaborated on CorpusChat (Cheung & Crosthwaite, 2025), an AI-empowered platform embedding corpus-informed chatbots within an academic writing development context. CorpusMate 2.0 extends this line of development by integrating LLM assistance directly into a public-facing, multi-corpus DDL tool, where the AI assistant reads, interprets and explains users' actual search results rather than operating on its own logic or training data. This functionality is designed to address the long-standing 'so what?' problem in DDL whereby learners are able to retrieve concordance data but lack the linguistic knowledge to interpret what they have found (Boulton & Cobb, 2017).

The following sections describe the changes in platform architecture (Section 2), corpus data (Section 3), corpus functionality (Section 4), AI integration and its pedagogical rationale (Section 5) before a brief conclusion (Section 6).

Platform Design

The original CorpusMate was built on NoSke, the open-source version of Sketch Engine (Kilgarriff et al., 2014), with a minimalistic HTML/CSS frontend. While functional, this architecture was limited in the range of statistical measures that could be computed (by design, given the target audience of secondary age DDL users), and (of course, given the time it was produced) lacked the ability to integrate external services such as an LLM API.

CorpusMate 2.0 is a ground-up rebuild. The backend is now a Python FastAPI application, deployed as a Docker container on Hugging Face Spaces. On startup, the server parses all corpus files from their native CWB vertical format (*.vert*), builds an inverted index mapping every word form to the list of sentence indices containing it, and pre-computes global word frequency counts, per-corpus token totals, and per-discipline token totals. These data structures are compressed and cached to disk (gzip pickle) so that subsequent server restarts load the index in seconds rather than rebuilding it from source files. All four corpus search functions execute via index lookups rather than sequential file scans, maintaining the near-instant query response times of v1. The frontend is a single-page HTML/CSS/JavaScript application deployed on Cloudflare Pages, served via GitHub continuous deployment. Corpus files are fetched from the

backend on first visit and stored in the browser's Cache API as pre-parsed JSON chunks, so subsequent visits load entirely from local browser storage without any network requests to the backend.

The user interface has been substantially redesigned. In place of v1's combined search bar and separate results page, v2 presents a persistent top bar containing the search input, corpus selection buttons, discipline and mode filters, and query tabs, with results and the AI assistant panel displayed side by side in the main area, yet only when called upon. This layout allows users to move between search results and AI commentary without navigating away from their data.

Corpus Data

The corpus collection in CorpusMate 2.0 (Table 1) comprises the same eight sources that CorpusMate v1 had grown to incorporate by the time of v2's development. The six corpora described in Crosthwaite and Baisa (2024) – BAWE, Elsevier, TED Talks, BBC Teach, Simple English Wikipedia, and the BNC 2014 Spoken corpus, were subsequently augmented in v1 with a Movie Subtitles corpus and a Video Games corpus, added in response to user requests for broader register coverage, while the BNC 2014 Spoken corpus was replaced by BASE (British Academic Spoken English), a corpus of university lecture and seminar transcripts better suited to the tool's academic and EAP user base. CorpusMate 2.0 retains this eight-corpus collection, which together totals approximately 57 million words, while rebuilding the underlying indexing and retrieval infrastructure from scratch.

All corpora are provided in CWB vertical format with word, part-of-speech tag (Penn Treebank tagset), and lemma on each line, with document-level metadata encoded in <*doc*> tags including source identifier, filename, title, and discipline keywords. Discipline filtering in v2 retains the 20 subject area categories of v1, with some labels simplified for display (e.g. "Health and Medicine" → "Medicine", "Business and Economics" → "Economics", "Geography, Agriculture and Environment" → "Geography"). A further category (Film & Media) has been added to accommodate the Movie Subtitles corpus added post-publication of v1. All corpora were filtered for potentially inappropriate content during the preparation of v1, using the alt-profanity-check library with a threshold of 0.9 to remove sentences with high profanity

Table 1. *Corpus data in CorpusMate 2.0*

Source	Token count	Document count
BAWE (British Academic Written English)	7,402,159	2,761
Elsevier OA CC-BY Corpus	11,969,773	2,200
BASE (British Academic Spoken English)	~4,200,000	1,356
TED Talk Corpus	5,925,929	2,456
BBC Teach	436,610	520
Movie Subtitles	~9,800,000	~85,000
Video Games Corpus	~3,600,000	~2,100
Simple English Wikipedia	13,494,943	40,134
Total	**~56,800,000**	**~136,527**

scores prior to indexing. This filtering is retained in v2, as the pre-processed corpus files are carried forward unchanged. The tool therefore remains suitable for use with younger learners and in conservative teaching contexts, as was a design priority of the original tool.

Corpus Search Functions

Concordance

The concordance function in CorpusMate 2.0 (Figure 1) is carried forward unchanged from v1, retaining what user feedback identified as the tool's core strength. Results are grouped by most frequent 4-gram starting with the query term, following the recommendations of Anthony (2018) as implemented in AntConc v4+ (Anthony, 2022). Each word in the concordance display is colour-coded by part-of-speech category derived from corpus annotation, with 15 words of context on each side. Results are sortable by frequency (most common 4-gram first, the default) or by positional sort (R1–R4, L1–L4). Clicking any concordance line expands an inline metadata panel displaying the full sentence, corpus source, discipline label(s), written/spoken mode, and document title; for lines drawn from the Elsevier corpus, a direct Google Scholar link pre-populated with the article title is provided. Drilling through from collocations or n-grams to filtered concordances is also retained from v1.

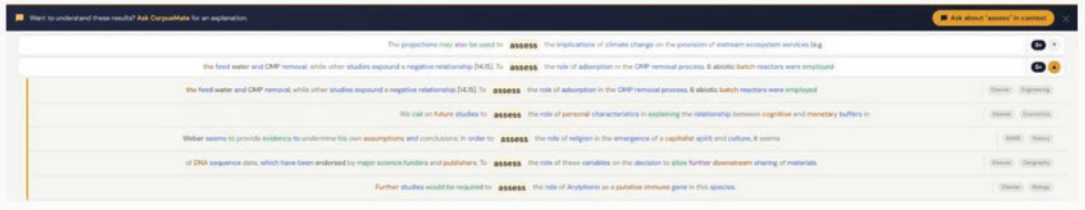

Figure 1. *CorpusMate 2.0 concordance display*

The principal novelty in v2 is not the concordance display itself but its integration with the LLM assistant. Following any corpus search, users are prompted to ask CorpusMate about what they have found; the AI receives the actual concordance lines with their sources, patterns, and metadata as context, and responds based on that specific corpus evidence rather than on generic linguistic knowledge (see section 5).

Collocations

CorpusMate 2.0 introduces a dedicated collocations module in a significant departure from v1, which offered co-occurrence information only in the form of document-proportion statistics showing the percentage of documents containing a query term in which a second term also appeared. While useful as a rough indicator of lexical association, this approach does not constitute collocation analysis in the corpus linguistics sense, as it is not sensitive to how frequently the two words appear together relative to their individual frequencies across the corpus (Hunston, 2022; Gries & Stefanowitsch, 2004).

The v2 collocations module (Figure 2) computes five standard association measures simultaneously from identical raw co-occurrence counts, with no additional query overhead:

- Frequency (raw co-occurrence count)
- MI (Mutual Information)

- MI² (frequency-weighted MI; the new default)
- MI³ (frequency-weighted MI with greater frequency emphasis)
- Log-likelihood / G² (significance-based measure)

All five scores are displayed simultaneously in a single table, with the active ranking measure highlighted. MI² is the default on the grounds that it provides a more balanced treatment of the distinctiveness–frequency trade-off than MI alone, reducing MI's known tendency to over-reward low-frequency collocates (Evert, 2008). Additional user controls include adjustable window size (±1 to ±5 words) and a minimum frequency threshold (options: 2, 5, 10, 25, 50, 100, 250) allowing researchers to apply the kind of frequency-based filters standard in corpus-based collocation studies. Clicking any collocate row opens a concordance filtered to sentences containing both the node word and the selected collocate, using an intersected inverted index lookup, and as with other search functions users are subsequently prompted to ask the LLM assistant about the collocational patterns retrieved.

Figure 2. *CorpusMate 2.0 collocation display with multiple association measures*

N-grams

The n-gram function is retained from v1, retrieving the most frequent 2-5-gram sequences containing the query term, ranked by frequency, with an example sentence displayed for each. As with the concordance, each n-gram result is clickable, taking users to a concordance filtered to that exact phrase. The function's integration with the LLM assistant follows the same pattern as described in Section 4.1 in that users are prompted following an n-gram search to ask the AI about the patterns retrieved.

Disciplinary Variation

The disciplinary variation function shows frequency per million tokens (rather than v1's document-proportion metric) for each corpus and each discipline, displayed as horizontal bar charts (Figure 3). In v2, clicking a corpus or discipline bar simultaneously: (1) filters the concordance to that corpus or discipline; (2) runs a concordance search for the query term within that filter; and (3) automatically poses a contextually appropriate question to the AI assistant thereby integrating quantitative frequency data, corpus evidence, and AI-generated linguistic commentary in a single user action, for example,

> *"Based on these concordance results, how is 'significant' used in the Medicine discipline? What patterns, collocations or register features stand out?"*

Figure 3. *CorpusMate 2.0 disciplinary variation display*

To test accuracy, an LLM query following a search for 'mate' correctly revealed the exact number of hits from each disciplinary subcorpora within its response, with other queries also matching exactly, further enhancing the accuracy of the LLM response while minimizing potential hallucination impact.

AI Integration

Design Rationale

The integration of a large language model assistant into CorpusMate 2.0 was motivated by a specific gap identified in DDL research and practice, namely an *interpretation problem*. Even when learners successfully retrieve corpus data, they often lack the linguistic knowledge or metalinguistic awareness to draw meaningful conclusions from what they have found (Boulton & Cobb, 2017; Flowerdew, 2015). Traditional DDL addressed this through teacher mediation or structured task design; CorpusMate 2.0 offers a complementary technological solution.

The tool uses the Cohere Command A model (command-a-03-2025), with usage rate-limited to 100 AI queries per user per day at no cost and no login required. Critically, the AI assistant does not operate independently of the corpus. When a user performs a corpus search and then asks the AI a question, the assistant receives the user's actual search results – namely concordance lines with their sources, collocation scores and frequencies, n-gram patterns, or dispersion data – as part of its context. The LLM thus reasons about specific corpus evidence rather than producing generic language advice and is explicitly instructed to distinguish between claims grounded in the retrieved data and those drawn from its general linguistic training. When corpus evidence is insufficient, the assistant flags this and clearly labels any response as drawing on its general knowledge base. The reliability of AI-generated commentary in CorpusMate 2.0 is supported by its retrieval-augmented design, in that the LLM receives the user's actual corpus results as context and is asked to interpret visible evidence rather than to retrieve facts from training data even matching the number of hits from each corpus mentioned in its response, substantially constraining the scope for confabulation relative to open-ended LLM use (Ji et al, 2023). Moreover, by presenting corpus and LLM results side-by-side, corpus-grounded claims from the LLM can be verified directly against the concordance lines, frequency counts and association scores displayed alongside each response. Indeed, this represents a distinctive strength of the tool, in that any claim made by the AI assistant can be immediately fact-checked by the user against the same corpus data that was sent to it. Nevertheless, as with all LLM-generated output, users are encouraged to treat AI commentary as a starting point for corpus-based inquiry rather than an authoritative linguistic account.

Search-to-AI Integration

The relationship between corpus search and AI commentary is designed to be mutually reinforcing. Following any corpus search (concordance, collocation, n-gram, disciplinary variation),

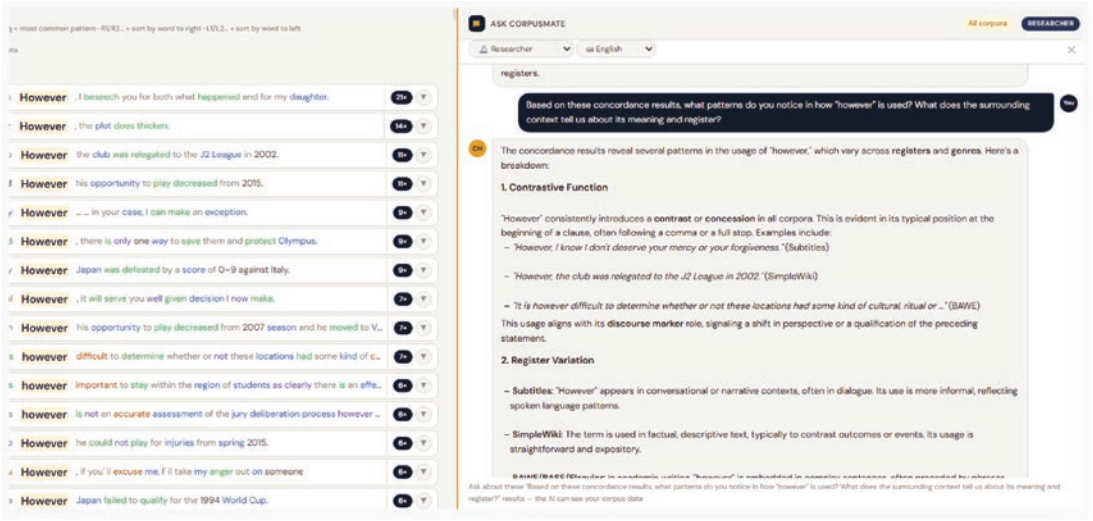

Figure 4. *CorpusMate 2.0 Search-to-AI prompt and response*

users are presented with a prompt inviting them to ask CorpusMate about what they have found. The prompt is tab-aware, pre-populating a contextually appropriate question, for example, asking about collocational patterns after a collocations search, or register features after a disciplinary variation search (Figure 4). As mentioned, the AI then receives the actual search results as its context, ensuring that responses are grounded in the specific data the user has retrieved rather than in generic linguistic knowledge.

This design reflects a pedagogical philosophy of corpus-grounded AI in that the assistant is intended to scaffold interpretation of real corpus data, not to substitute for engagement with that data. A planned future development is full bidirectionality, whereby a natural language question posed in Ask mode would automatically trigger a background corpus search, populating both panels simultaneously.

User Roles

A distinctive feature of the AI integration is a role system that fundamentally changes the AI's persona and pedagogical approach across the same corpus input context. Users may select one of four roles from a drop-down menu:

- Researcher – analytical and technical; uses corpus linguistics terminology freely; grounds every claim in the retrieved data
- Teacher – pedagogically focused; generates classroom activities, gap-fill exercises and discussion tasks from corpus patterns
- Student – Socratic and interactive; asks guiding questions and gives hints rather than direct answers; makes exercises genuinely interactive with one question at a time
- Language learner – plain language throughout; no unexplained jargon; practical focus on communicative relevance; warm and encouraging

The same corpus evidence is presented to the AI in each case, with the system prompt governing its response persona changes. This design means that a corpus search for 'however' can produce a technical discussion of the functional variation of 'however' for a researcher,

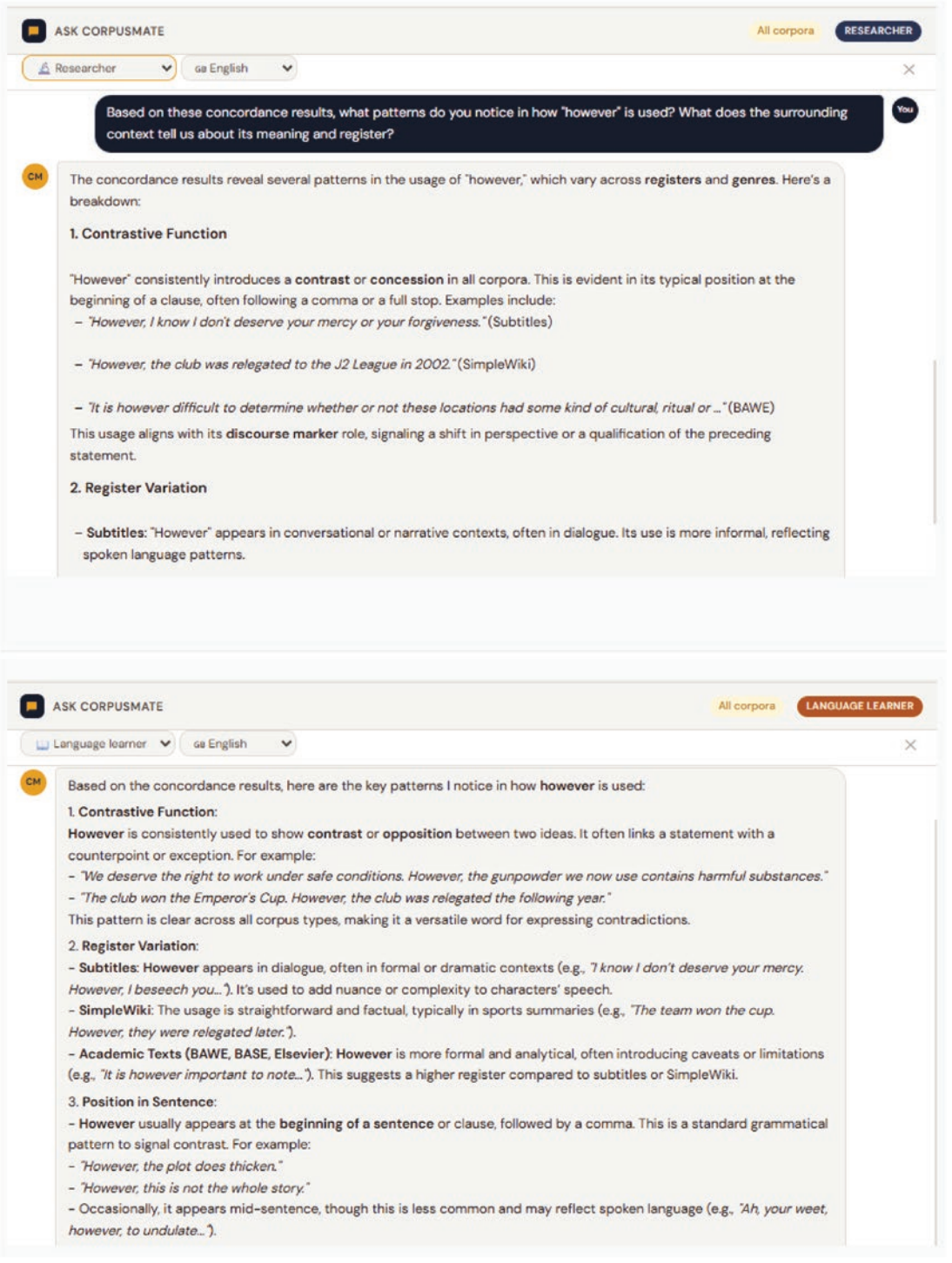

Figure 5. *CorpusMate 2.0 AI assistant panel with role selector and corpus-grounded response*

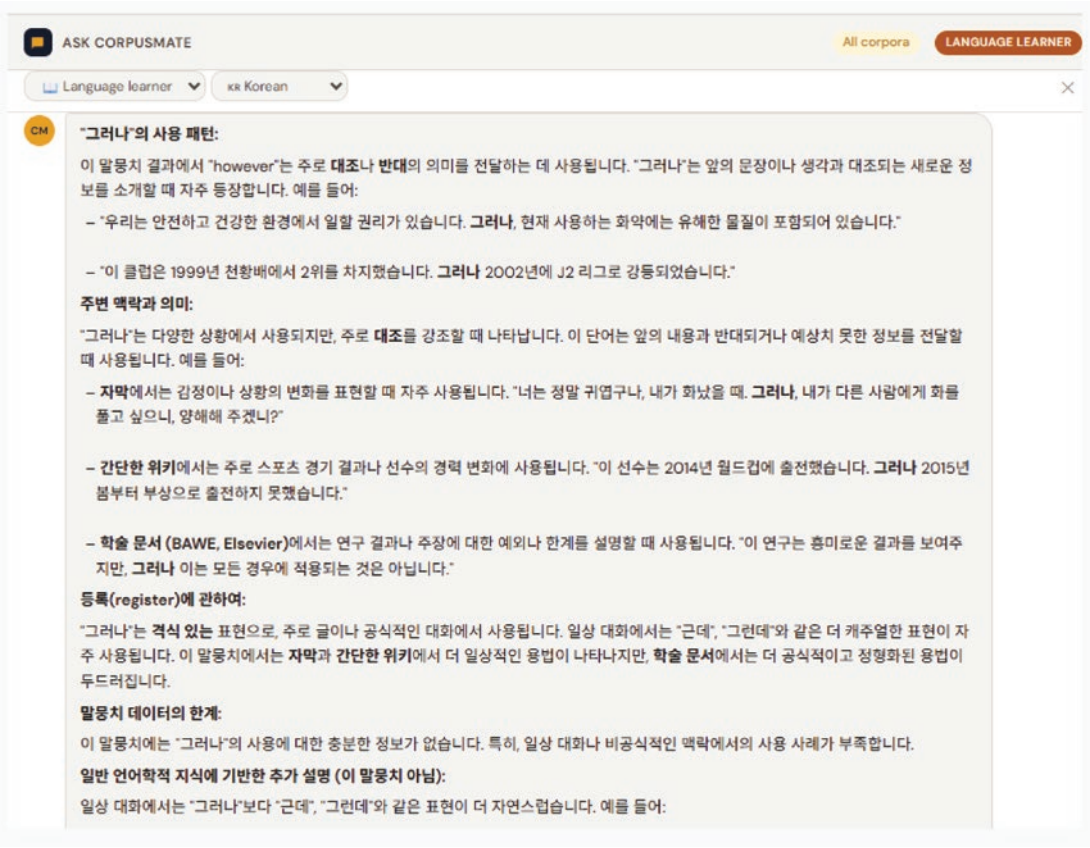

Figure 6. *CorpusMate 2.0 multilingual AI response (Korean)*

a gap-fill worksheet for a teacher, a Socratic dialogue for a student, and a plain-language explanation of how to use the word correctly for a language learner – each from identical underlying data (Figure 5). This role system directly addresses the concern raised in Crosthwaite and Baisa (2024) that learners and teachers require different kinds of corpus tool support.

Multilingual Support

Following on from v1's English-only interface, CorpusMate 2.0 supports AI responses in eleven languages: English, French, Spanish, German, Italian, Portuguese, Dutch, Chinese, Japanese, Korean, and Arabic (Figure 6). The corpora themselves remain in English; the language selector controls the language of the AI's explanations and commentary. This allows, for example, a Korean university student to perform corpus searches in English and receive linguistic explanations in Korean, supporting the growing use of English-language corpora in EFL and EMI contexts outside English-speaking countries (Boulton & Vyatkina, 2021).

Conclusion

CorpusMate 2.0 represents a substantial evolution motivated by user feedback from v1, advances in LLM technology, and an expanded understanding of what DDL-focused corpus tools need to do for diverse user communities. The two principal advances, namely a dedicated

collocations module offering five association measures simultaneously, and a corpus-grounded AI assistant with a role-based pedagogical persona system operating across eleven languages, both address specific limitations identified in the v1 paper and in subsequent DDL research.

Of these advances, the AI integration is the most novel and the most requiring of empirical investigation. At present, CorpusMate 2.0 offers a technologically grounded implementation of corpus-AI integration, but research into its efficacy for language learning outcomes, learner engagement with AI-mediated corpus interpretation, and the comparative value of the four role personas remains to be conducted. We invite the corpus linguistics and SLA community to use the tool in classroom and self-study settings and to share empirical data on its effectiveness. Feedback can be submitted via the tool's embedded feedback form or by contacting the author directly. CorpusMate 2.0 is freely accessible at https://corpusmate.app/, requires no login or download, and offers 100 free AI queries per day. The tool's source code is available on request for research and institutional deployment purposes.

References

Anthony, L. (2018). Visualisation in corpus-based discourse studies. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 197–224). Routledge.

Anthony, L. (2022). What can corpus software do? In A. O'Keeffe & M.J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (2nd ed.) (pp. 103–125). Routledge.

Anthony, L. (2025). Integrating AI technology into corpus-based language learning through ChatAI. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2025.2589747

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*(2), 348–393.

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology, 25*(3), 66–89.

Cheung, L., & Crosthwaite, P. (2025). CorpusChat: Integrating corpus linguistics and generative AI for academic writing development. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2025.2506480

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

Ebeling, S.O., & Heuberger, R. (2025). Large language models and corpus linguistics: Issues and opportunities. *International Journal of Corpus Linguistics* [advance online].

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 2, pp. 1212–1248). Mouton de Gruyter.

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15–36). John Benjamins.

Gries, S.T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics, 9*(1), 97–129.

Hunston, S. (2022). *Corpus approaches to evaluation: Phraseology and evaluative language* (2nd ed.). Routledge.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

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

Mizumoto, A. (2023). Data-driven learning meets generative AI: Introducing the framework of Metacognitive Resource Use. *Applied Corpus Linguistics, 3*(3), 100074. https://doi.org/10.1016/j.acorp.2023.100074

Poole, R. (2022). "Corpus can be tricky": Revisiting teacher attitudes towards corpus-aided language learning and teaching. *Computer Assisted Language Learning, 35*(7), 1620–1641.