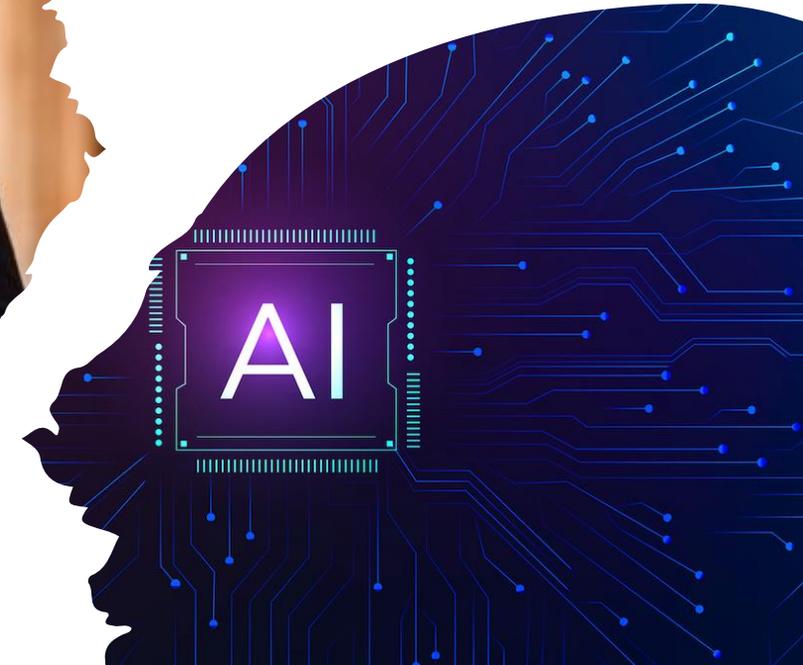
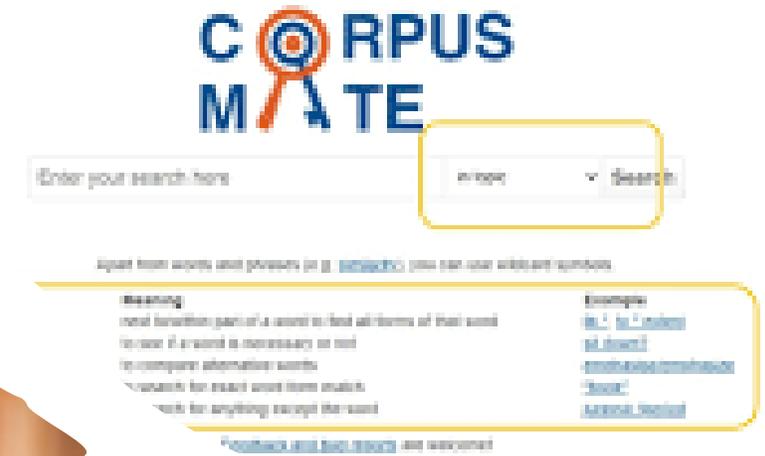


Corpora, Prompts, and Pedagogy: Human-AI Text Comparison in Applied Linguistics

- Assoc. Prof. Peter Crosthwaite
- The University of Queensland





WHO AM I?

- Associate Professor, University of Queensland, Australia
- Published 60+ journal articles, 3 textbook series, 4 books.
- Editor *Australian Review of Applied Linguistics*
- Editorial boards – *System, Journal of Second Language Writing, English for Academic Purposes, IRAL, Applied Corpus Linguistics, CALL, Studies in Corpus Linguistics*

What do I work on?

Research interests include:

- **Corpus linguistics,**
- Second Language Writing
- English for Academic Purposes
- Second Language Acquisition
- TESOL
- Generative AI

The intersection of these – Data-driven Learning (DDL)



CORPORA FOR LANGUAGE LEARNING

BRIDGING THE RESEARCH-PRACTICE DIVIDE

Edited by
PETER CROSTHWAITE



Today's talk

Corpora and English for Academic Purposes

GenAI vs 'human' academic writing – corpus studies

The problem with comparative AI-human corpus studies

Corpus-based Data-Driven Learning – is it dead? (Not quite...)

Where to next?



Learning the Language of Dentistry

Disciplinary corpora in the teaching of English for Specific Academic Purposes

Peter Crosthwaite and Lisa Cheung

Studies in Corpus Linguistics

Corpora and English for academic purposes (EAP) – my background

- Multidimensional analysis of a longitudinal learner corpus (Crosthwaite, 2016, JEAP)
- Error tagging said corpus to see impact of EAP instruction on L2 errors (finding – no impact!) (Crosthwaite & Jiang, 2017, System)
- Metadiscourse use in dentistry research articles vs. general academic reports by article subsection (Crosthwaite, Cheung & Jiang, 2017, ESPJ)
- Monograph on the language of dentistry case reports (Crosthwaite & Cheung, 2019, John Benjamins Studies in Corpus Linguistics).

Scattertext results summary

- Top ranked terms in the case report corpus come from a range of semantic domains.
- E.g. *states / conditions of patients or diseases*
 - *erythematous* (Redness of the skin caused by dilatation and congestion of the capillaries, often a sign of inflammation or infection)
 - *hypointense* (a 'less' intense abnormality on a magnetic resonance scan)
 - *pedunculated* (where a mass such as a cyst or polyp is said to be pedunculated if it is supported by a peduncle, or an elongated stalk of tissue).
- E.g. *diseases and conditions*
 - *osteosarcoma* (a common type of bone cancer)
 - *fibroma* (benign tumour)
 - *hemangioma* (noncancerous growths of blood vessels).
- E.g. *modifiers of position*
 - *mesiodistal* (a diameter joining the mesial and distal surfaces of a tooth)
 - *maxillary left* (to the left of the pyramid-shaped maxillary sinus)
 - *midline* (a median line or plane of bilateral symmetry).
- While numerous vocabulary elements are shared across both research report and case report genres of dentistry writing, Scattertext analysis has shown a number of semantic domains appear to be specific to the case report genre.



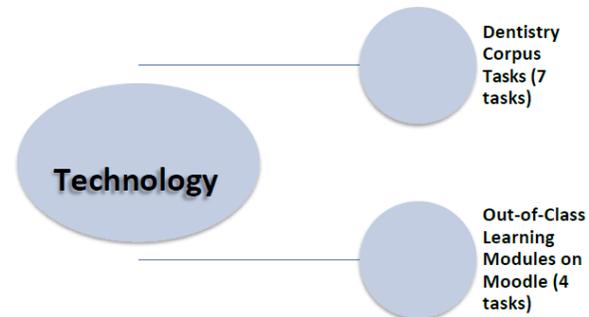
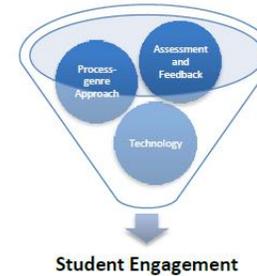
Hands-off DDL activities

- Insights gained from the dentistry corpus led to a range of hands-off DDL tasks embedded into coursebooks
- These were to be complimented with hands-on DDL activities sequenced later.
- In each case, exemplars were all sourced from disciplinary corpora.

Course Learning Objective



The learning objective of this course is to promote *student engagement through a multidimensional learning approach*. We believe the more you are engaged, the more you learn.



Integration of Technology in Self-directed Learning

GRSC6021 Introduction to Thesis Writing (Sciences and the Related Disciplines)



Centre for Applied English Studies
The University of Hong Kong

CAES9350 Student Course Booklet (2017-18)



Centre for Applied English Studies
The University of Hong Kong

So far so good - But
what now?
The elephant in the
room (Crosthwaite
& Baisa, 2023)

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

Suddenly, language data – corpora – are popular again! Yay!

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.

Corpus vs AI approaches to Academic language and register

- Despite claims that generative AI can write an entire academic research paper from scratch (Hsu, 2023), corpus linguistic research has already pointed out a range of differences between the academic register typically produced by software such as ChatGPT and that of human academic writers.
 - **Kong and Liu (2024)** – Significant differences across human/GAI in the types of rhetorical moves, move patterns, and frequency of move embedding
 - **Berber-Sardina (2024)** – Significant register variation between human/GAI under MDA approaches
 - **Mizumoto et al. (2024)** – Significant variation across lexical diversity, clausal syntactic complexity, embedded syntactic complexity, complex nominals, modals, epistemic markers, and discourse markers
 - **Yang et al. (2024)** – Significant variation across types of textual, interpersonal and marked topical themes for SFL
 - **Jiang and Hyland (2025)** – Significant variation across 3-word lexical bundles in a corpus of ChatGPT-produced academic essays with those of British students - ChatGPT used a narrower, repetitive range of such bundles, with less evidence of bundles used to convey epistemic stance and authorial presence compared with human student writers

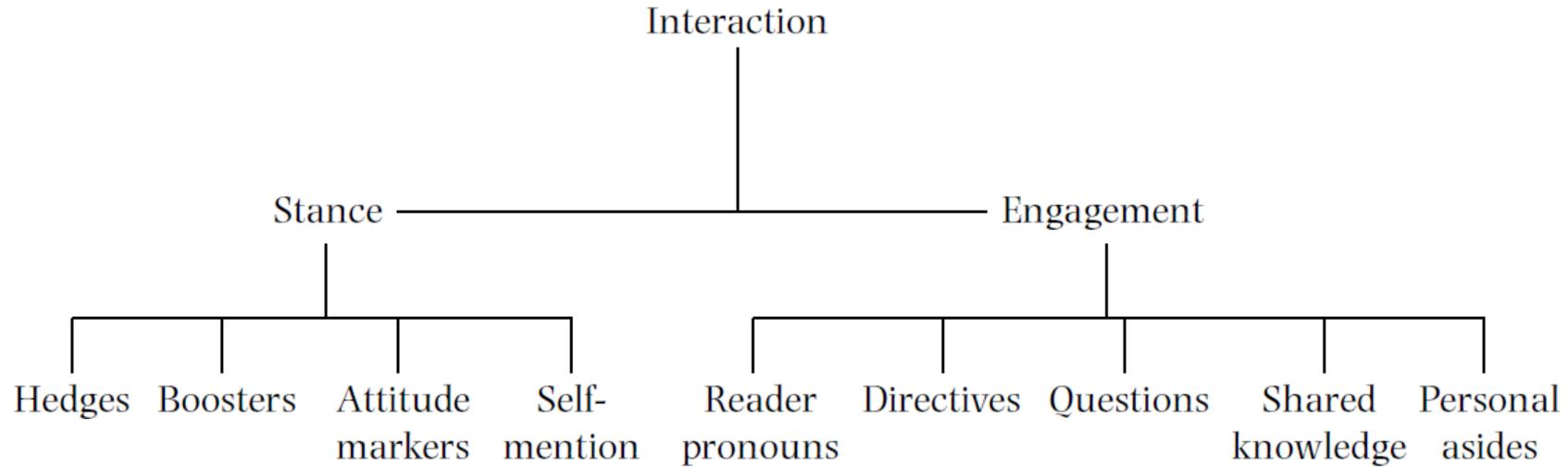


FIGURE 1. *Key resources of academic interaction*

Case study: Stance and Engagement between human and AI writers (Mo & Crosthwaite, 2025, JEAP)

- “put succinctly, every successful academic text displays the writer’s awareness of both its readers and its consequences” (Hyland, 2005 p.174).
- This is demonstrated through use of *Stance* and *Engagement* features

Stance

Hedges: Devices including 'possible', 'might' and 'perhaps' are commonly used in academic writing to weak the author's commitment towards a claim to be made, avoiding risk of critique while acknowledging the current state of knowledge of the discipline.

Boosters: Expressions such as 'clearly' and 'obviously' are used to strengthen the author's commitment to the claims they are making, while marking the degree of shared knowledge held by the discipline in question.

Attitude markers: The author's use of expressions including 'unfortunately', 'hopefully' and 'surprisingly' signal an affective dimension to the claims the author chooses to produce in their text, demonstrating a presupposed sharing of values between the writer, reader and the discipline.

Self-mention: The use of pronouns including 'I', and 'we' when used to refer specifically to the writer is a key component of disciplinary variation in the authors' own intrusion into their texts.

Engagement

- **Directives:** Terms including ‘must’, ‘consider’, ‘imagine’ and others directly instruct the reader to a specific action or conclusion by carrying out the writers’ directive as they encounter it in text. These can be textual (e.g. ‘see Lambert and Jones 2005...’), physical (e.g. ‘set the slider to 300mm’) or cognitive (e.g. ‘consider Walker’s claim that...’) in nature.
- **Reader pronouns:** The use of ‘we’, ‘our’ ‘you’, and ‘us’ mark a highly explicit method of drawing the reader directly into the author’s discourse, directly signalling co-membership of the discipline in question between the writer and the reader.
- **Questions:** Direct questions to the reader serve to increase writer/reader engagement as a form of dialogue, although in almost all cases are rhetorical (e.g. is it necessary to choose between....)
- **Appeals to shared knowledge:** Expressions such as ‘as we all know...’ serve as a form of solidarity between the writer, the reader, and the current state of knowledge of the discipline, asking the reader to accept or recognise any claim as familiar territory.
- **Personal asides:** Expressions such as ‘by the way’ mark the writers’ intention to break up the flow of the text to address the reader directly. While not typical in the sciences, their use in the soft disciplines as a form of writer/reader dialogue is well-established (Hyland, 2005).

Research questions

RQ1: How does human- and GenAI-produced academic writing convey academic stance and engagement?

RQ2: Can GenAI-produced academic writing approximate human-like production of academic stance and register features across disciplines?

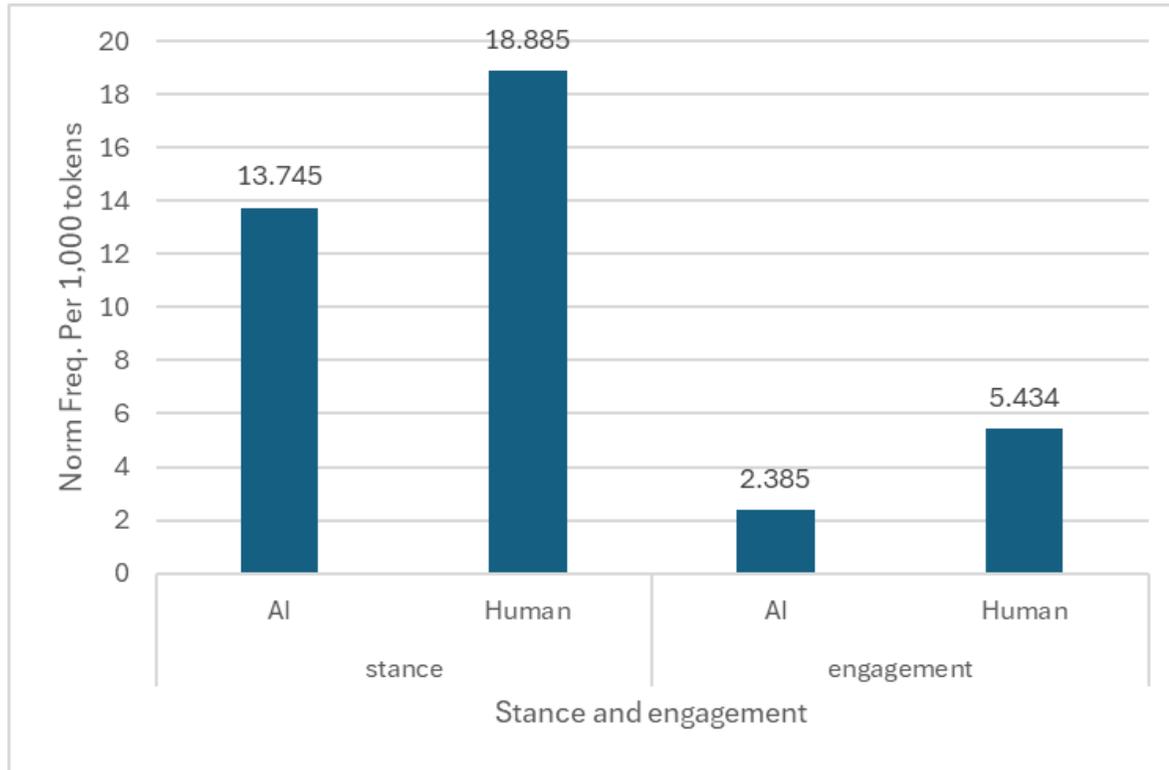
RQ3: To what extent do different GenAI models (i.e. ChatGPT, ERNIEBot, MetaAI) vary in their use of stance and engagement features in academic writing as compared with human academic writing and each other?

	HUMAN	GENAI	CHATGPT	ERNIEBOT	METAAI
Texts	30	90	30	30	30
Tokens	108200	157122	60123	43997	53002
Words	89662	130930	49014	38295	43621

Corpus

- Human corpus sourced from British Academic Written English Corpus (BAWE, Alsop & Nesi, 2009)
- GenAI uses ChatGPT, ErnieBOT, MetaAI

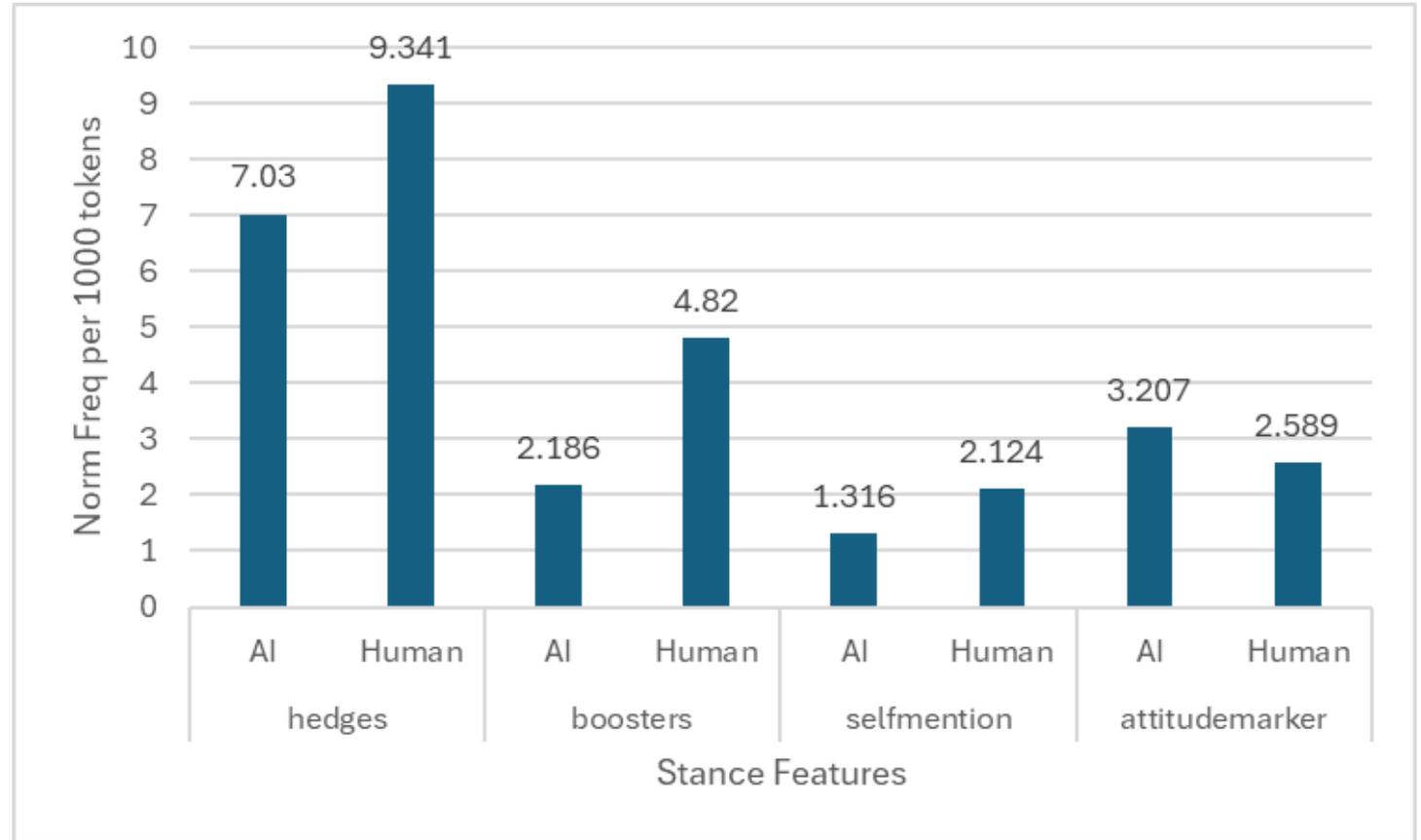
Results RQ1 (Stance and Engagement in human vs GAI production)



- Human writers use more explicit stance ($H=10.16$, $z=-3.18$, $p<.001$, Rank $\epsilon^2 = 0.085$) and engagement features ($H=5.19$, $z=-2.28$, $p=0.023$, Rank $\epsilon^2 = 0.040$) than GAI

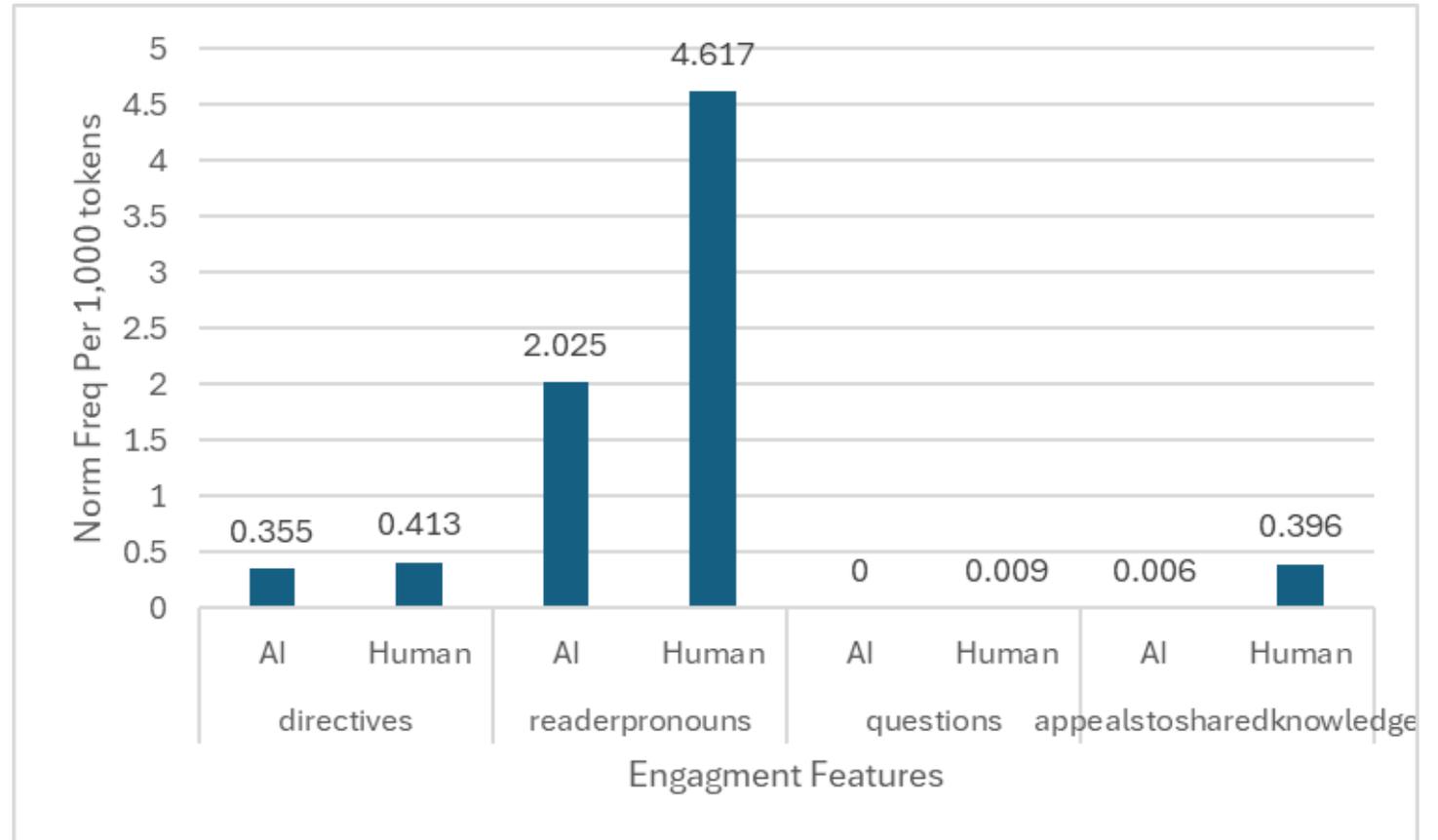
Stance features

- Humans are significantly more likely to use hedges (U=55.83, z=-2.58, p=0.012, Rank ϵ^2 = 0.053), and boosters (U=51.91, z=-4.68, p=<.001, Rank ϵ^2 = 0.185) than GAI



Engagement Features

- Significant differences were found however in the use of appeals to shared knowledge (U=52.15, z=-7.33, p<.001, Rank $\epsilon^2 = 0.452$) with a strong effect size



Wordings

Human Hedges (p<.0001)	GenAI Hedges (p<.0001)	Human Boosters (p<.0001)	GenAI Boosters (p<.0001)
Would, perhaps, seems, might, quite, seem, probably, assumed, appear, suggested, almost, mainly, apparently	Often, typically, suggesting	In fact, clearly, indeed, never, believe, actually, thought, shows	
Human Attitude Markers (p<.001)	GenAI Attitude Markers (p<.001)	Human self-mention (p<.001)	GenAI self-mention (p.<001)
Clearly, interesting, obvious, interestingly, unfortunately, correctly	Significant	I, my, this paper,	Our, this essay

RQ2 – Disciplinary variation (Stance)

- No significant differences between human/GAI
- GAI can handle disciplinary variation in stance features well enough

Descriptives - stance

HumanvAI	Discipline	N	Mean	SD	SE	Coefficient of variation
AI	Archaeology	15	15.29	6.853	1.769	0.448
	Classics	15	10.34	4.384	1.132	0.424
	English	15	10.33	4.671	1.206	0.452
	History	15	12.31	7.428	1.918	0.603
	Linguistics	15	11.65	3.987	1.030	0.342
	Philosophy	15	22.51	5.629	1.453	0.250
Human	Archaeology	5	22.60	5.442	2.434	0.241
	Classics	5	9.240	4.159	1.860	0.450
	English	5	16.62	5.317	2.378	0.320
	History	5	13.31	1.115	0.499	0.084
	Linguistics	5	21.56	7.688	3.438	0.357
	Philosophy	5	29.95	7.592	3.395	0.253

RQ2 – Disciplinary variation (Engagement)

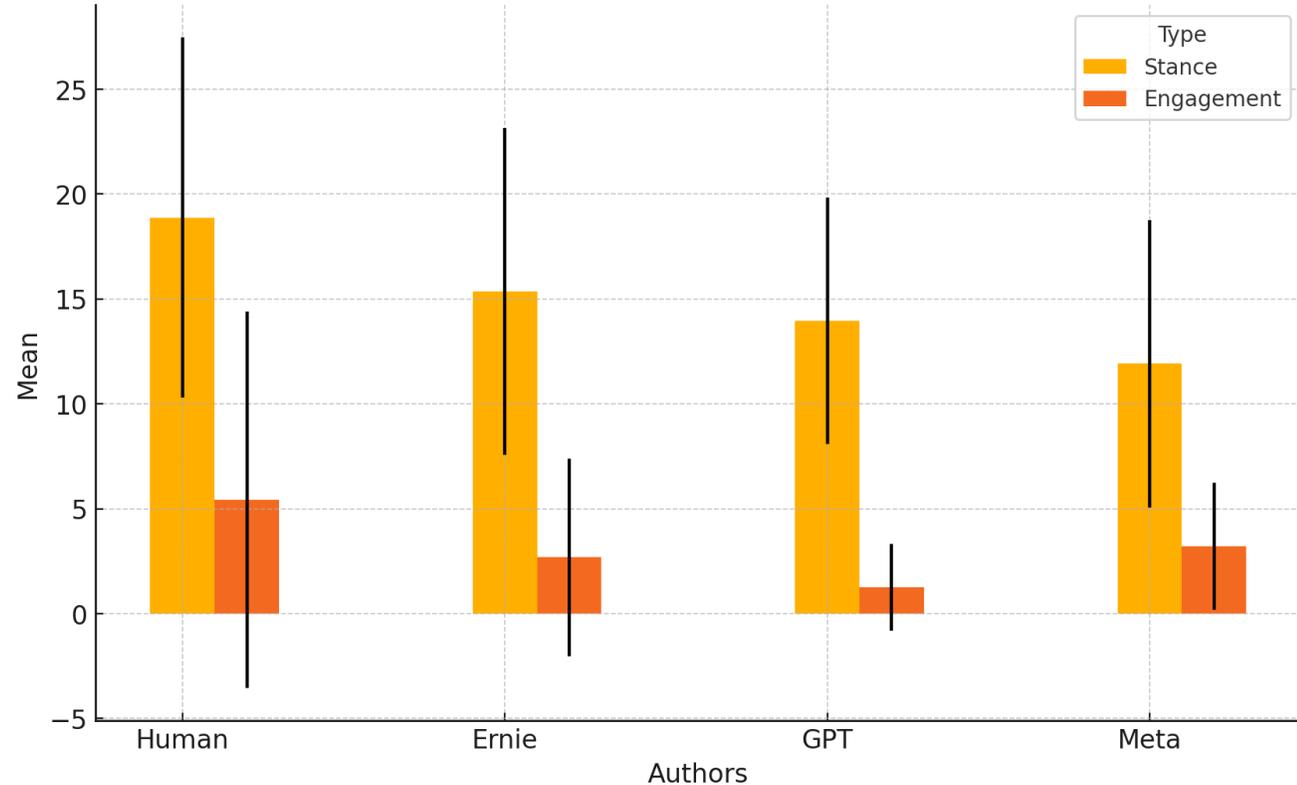
- Significant differences found for Philosophy discipline in reader pronouns and appeals to shared knowledge
- These are prominent in that discipline (Hyland, 2005)
- GAI was unable to capture this – not enough training data?

	discipline	N	Mean	SD	SE	Coefficient of variation
AI	Archaeology	15	2.45	2.987	0.771	1.219
	Classics	15	0.741	0.882	0.228	1.191
	English	15	1.843	2.595	0.67	1.408
	History	15	0.99	1.465	0.378	1.48
	Linguistics	15	2.099	1.63	0.421	0.777
	Philosophy	15	6.19	6.009	1.552	0.971
Human	Archaeology	5	1.344	0.667	0.298	0.496
	Classics	5	0.324	0.48	0.215	1.482
	English	5	3.494	1.402	0.627	0.401
	History	5	2.792	1.563	0.699	0.56
	Linguistics	5	2.61	1.59	0.711	0.609
	Philosophy	5	22.042	12.49	5.586	0.567

RQ3 – Comparisons across three LLMs

- Use of stance features significantly lower in MetaAI texts, particularly for hedges
- Use of engagement features varies across humans and the three LLMs individually

Comparison of Mean Across Stance and Engagement by Authors (with Error Bars)



Summary

Overall, the study has found significant variation in the production of stance and engagement features between human- and machine-produced academic discourse

This includes disciplinary variation in the use of said features, as well as variation across the three large language models (LLMs) used in the present study.

The findings are likely to have implications for stakeholders involved on EAP courses, EAP students themselves, LLM developers, and those looking to detect LLM-produced academic text.

HOWEVER...

Hubbard (2025) notes the “temporal limits” of much ‘corpus-based’ research involving GenAI

GenAI *is not* a corpus – improvements to the model (e.g. gpt4 to gpt5) or changes to the prompt can yield vastly different results...

Many studies (including mine!) have made claims e.g. “ChatGPT cannot/does not mimic human-like production of X, Y, Z”

However, will it do so, *if you tell it to do so?*

The answer is forthcoming....

If you do not like the results, change the prompt! Exploring issues with corpus-based analyses of human vs generative AI produced academic writing (Submitted 2026, MLJ).



- Typical method of AI/Human corpus comparison:
 - Generate an AI corpus matched to a human corpus (topic, discipline, size)
 - Conduct comparative corpus analyses
 - Analyses focus on features such as:
 - Evaluation and stance (appraisal)
 - Discipline-specific vocabulary
 - Register and multidimensional patterns
 - Rhetorical structure (move-step analysis)
 - In our case - metadiscourse
 - Goal: understand how LLM writing differs from human academic discourse
 - Claim: Human vs AI writing is significantly different!
- “In summary, current research appears to show that while AI has come a long way in appropriating an academic-looking register in its output, it still has a long way to go in really capturing the essence of ‘the human touch’ in academic writing.” (Crosthwaite, 2026, p. 5, emphasis in original)

The Problem: Prompt Sensitivity

- Existing claims may be **context-bound**, not generalizable
- AI output depends heavily on:
 - Task and genre
 - Prompt wording and framing
 - Instructional detail
- Even small prompt changes can substantially alter:
 - Lexical choices
 - Discourse features
 - Rhetorical organisation
-  What appears to be “AI style” may simply be **prompt-specific output** (Kuteeva & Andersson, 2025)

What's the problem? (2)



Challenge: what constitutes a representative or 'default' prompt in line with the perceived prompting abilities of students and/or teachers.



Lin and Crosthwaite (2024), for instance, attempted “to replicate the prompting strategies typical of a practicing L2 language teacher with only minimal training in prompt engineering [...] which at the time of writing would *presumably* represent the abilities of the vast majority of language teachers” (p. 5, emphasis ours).



However, emerging evidence suggests such assumptions may underestimate teachers' (or students') actual, practical prompting expertise given the rapid diffusion of informal professional learning and communities of practice around GenAI use (Moorhouse et al., 2024).

(underspecified) Single shot prompts vs iterative dialogue



Kuteeva and Andersson (2025) compared single-shot, few-shot, and iterative role-based prompting for stance metadiscourse.

Zero-shot prompts produced risk-averse epistemic stance and over-assertive attitudinal stance;

few-shot prompts led to context-neutral stance and unsolicited content;

iterative role-based prompts, while more aligned with expectations, still “displayed a rather rigid understanding of stance” (p. 17).



In Jiang and Hyland (2025d), the following prompt was selected after multiple trials because it “mimicked the format and tone of the [human] writing tasks while providing ChatGPT with clear expectations” (p. 20):

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.



Yet, while this prompt specifies writer role, genre, length, and rhetorical style, it provides no guidance on metadiscourse, including stance-taking, reader engagement, self-mention, or evidentiality.



As Kuteeva and Andersson caution, “stance expressions in LLM-generated texts often require adjustment: in this case, the dialogue takes place between the author and the model *via prompts*” (p. 5, emphasis ours)

Methodological Implications

Treating AI output as a fixed corpus assumes stability that may not exist

LLM texts are better understood as **Dynamic, interaction-dependent artefacts**— co-constructed through prompting

Corpus comparisons risk reifying prompt-specific results as intrinsic model abilities

- e.g., “ChatGPT can/cannot do X”

This may overstate both limitations and capabilities of AI writing

Implication: We must reconsider how AI texts are generated, sampled, and interpreted in applied linguistics research

Present study

Stage 1

- partial replication of Jiang and Hyland (2025d), using the same academic essay prompt used in that study to generate a corpus of AI-produced texts comparable in size and topic to a human-produced corpus of academic essays. Both corpora are then annotated, analysed and contrasted for metadiscourse use

Stage 2

- On the basis of the differences observed under this baseline prompting condition, a revised prompt explicitly targeting the metadiscourse features underrepresented in the initial AI corpus, generating a new AI-produced corpus

RQ1. To what extent does a baseline, rhetorically underspecified academic essay prompt constrain GPT's use of interactive, interactional, and engagement metadiscourse when compared with human-produced academic writing?

RQ2. How does targeted manipulation of the prompt affect GPT's production of interactive, interactional, and engagement metadiscourse, and to what extent does this reshape observed differences between GPT-generated and human-produced academic texts?

Method

Stage 1 - Two corpora annotated for interactive, interactional, and engagement metadiscourse using UAM CorpusTool v3.3 (O'Donnell, 2008).

- Corpus 1: 145 essays from the Louvain Corpus of Native English Essays (LOCNESS; Granger, 2014), written by British university students.
- Corpus 2: Newly constructed parallel corpus of the same number of essays on identical topics generated by ChatGPT version 4.0, which at the time of the Jiang and Hyland (2025d) study represented the state-of-the-art model.
- The prompt used to generate the AI corpus was identical to that reported in the literature review and is referred to hereafter as the *baseline prompt*.

Topics are UK-focused and included British transport, boxing, computers, the UK monarchy, Britain and Europe, in vitro fertilisation, the lottery, fox hunting, and beef consumption in the context of 'mad cow disease'.

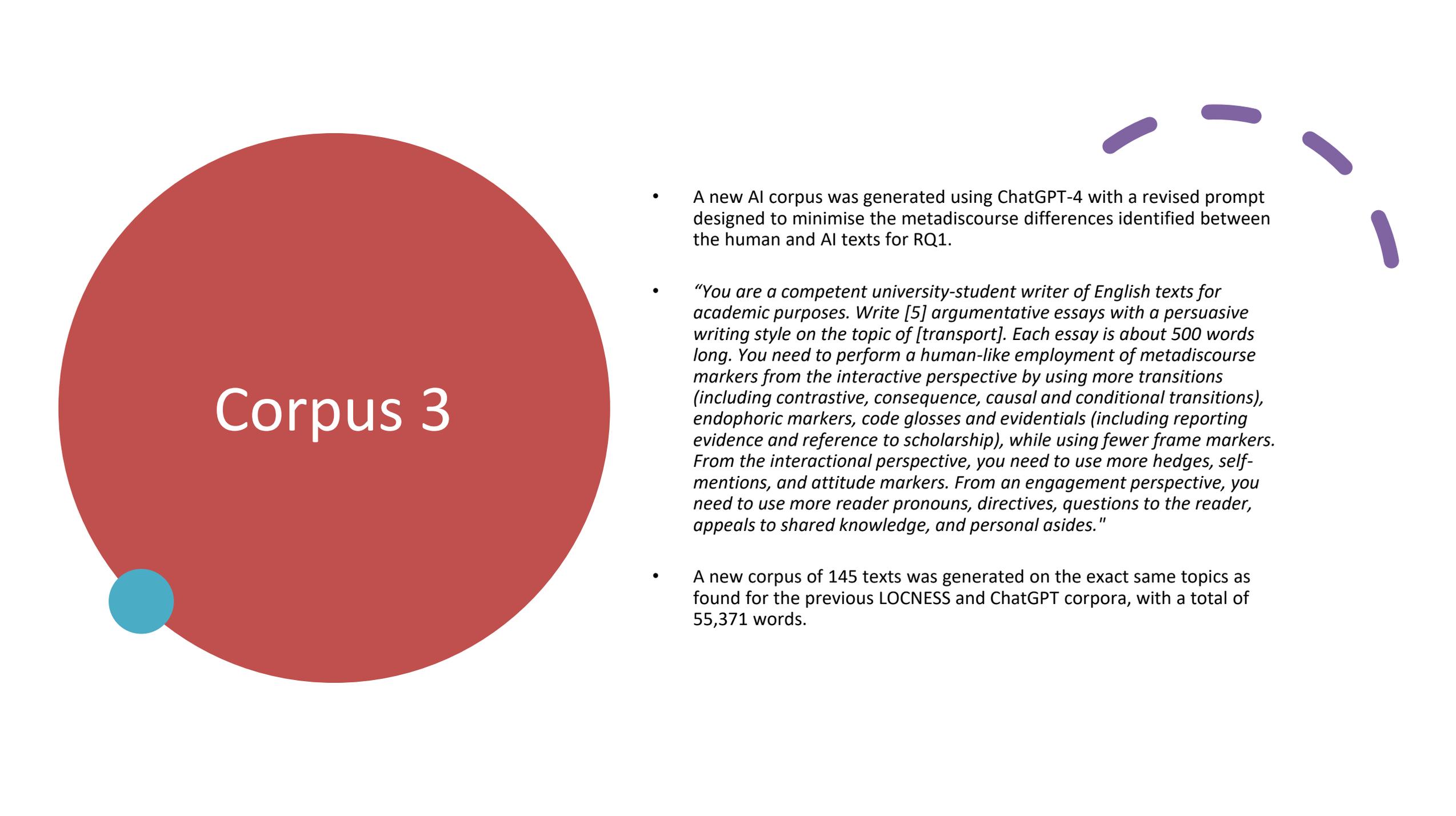
These topics and the baseline prompt were used to generate the ChatGPT corpus in line with best practice for matching writing conditions across human and AI corpora (Zhang, 2025).

The LOCNESS corpus comprised 78,060 words, while the ChatGPT corpus contained 62,819 words.

Coding (interactional metadiscourse)

Table 1: Interactional metadiscourse

Category	Sub-Category	Function	Typical Examples
INTERACTIONAL	Hedges	Mark uncertainty/open space for alternatives	<i>might, perhaps, it seems that</i>
	Boosters	Mark certainty/close alternatives	<i>clearly, obviously, demonstrates</i>
	Attitude Markers	Express affect or evaluation	<i>importantly, unfortunately</i>
	Self-mentions	Signal author presence	<i>I argue, we suggest</i>
Category	Sub-Category	Function	Typical Examples
ENGAGEMENT	Reader Pronouns	Include reader explicitly	<i>you can see, as we will observe</i>
	Directives	Tell reader what to do/attend to	<i>note that..., consider..., remember that...</i>
	Appeals to Shared Knowledge	Build common ground	<i>of course, we all know that</i>
	Questions	Invite reflection or anticipate objections	<i>What does this imply?</i>
	Personal Asides	Address the reader in a conversational aside	<i>as you might expect..., incidentally..., by the way...</i>

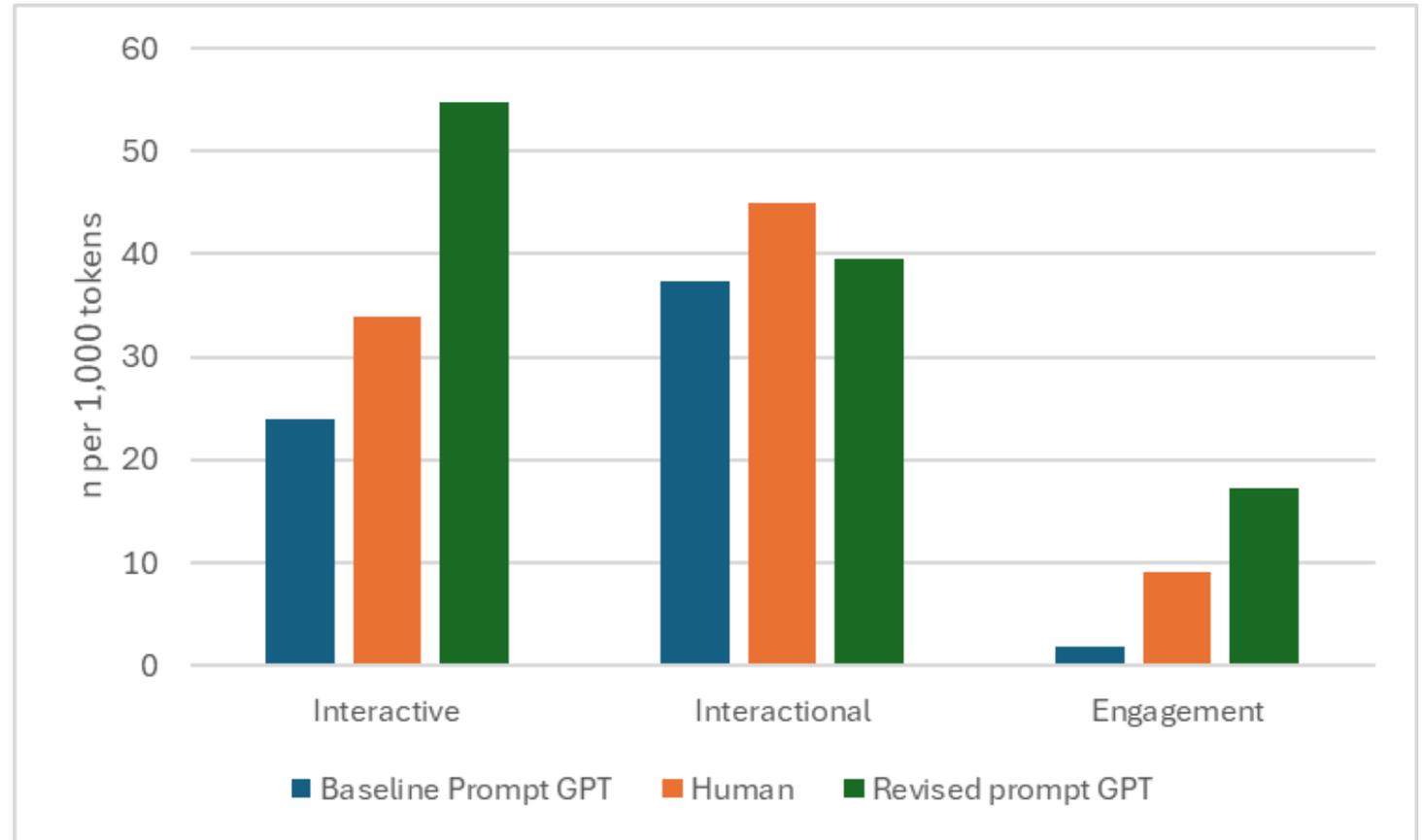


Corpus 3

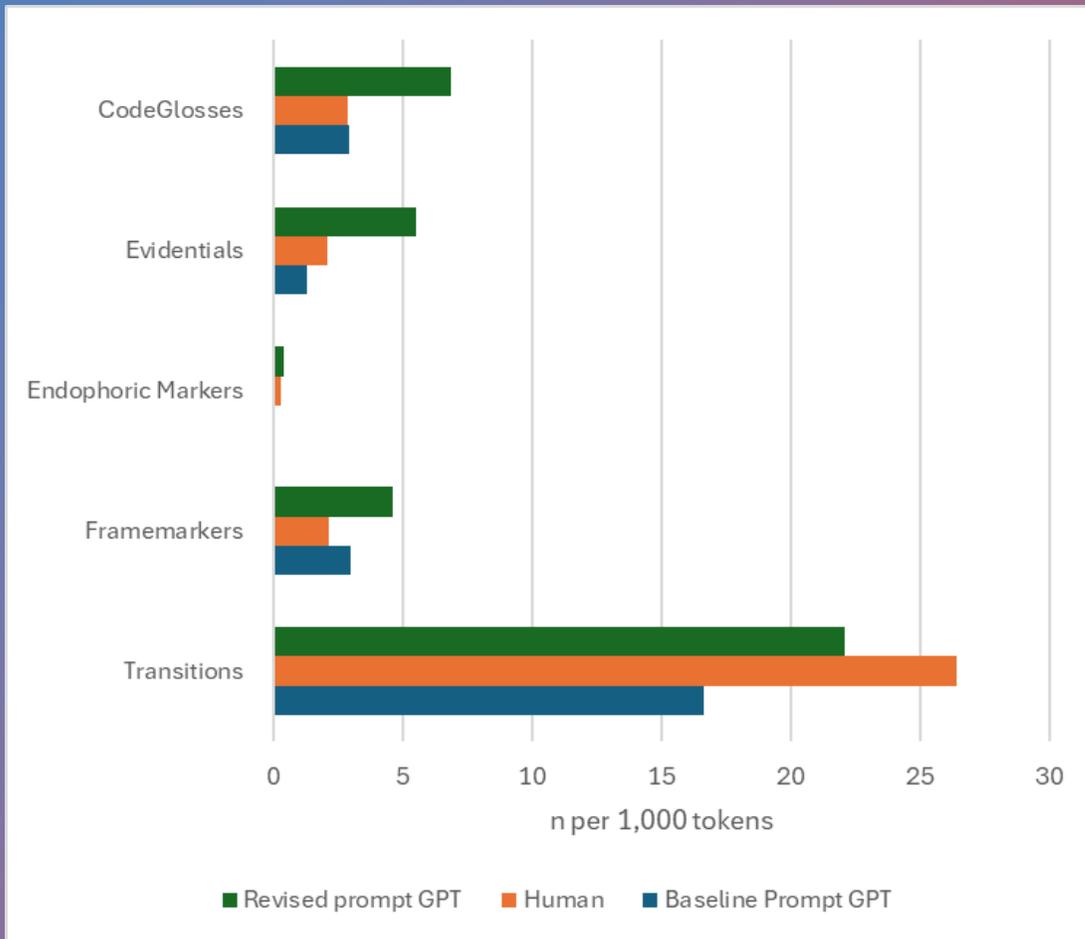
- A new AI corpus was generated using ChatGPT-4 with a revised prompt designed to minimise the metadiscourse differences identified between the human and AI texts for RQ1.
- *“You are a competent university-student writer of English texts for academic purposes. Write [5] argumentative essays with a persuasive 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.”*
- A new corpus of 145 texts was generated on the exact same topics as found for the previous LOCNESS and ChatGPT corpora, with a total of 55,371 words.

Results – baseline vs human vs revised prompt

- Comparisons with baseline AI corpus and human corpus in line with previous studies
- i.e. – human writers produce more interactive, interactional and engagement features
- BUT – with revised prompt, new AI corpus OVERproduces these features compared with humans!

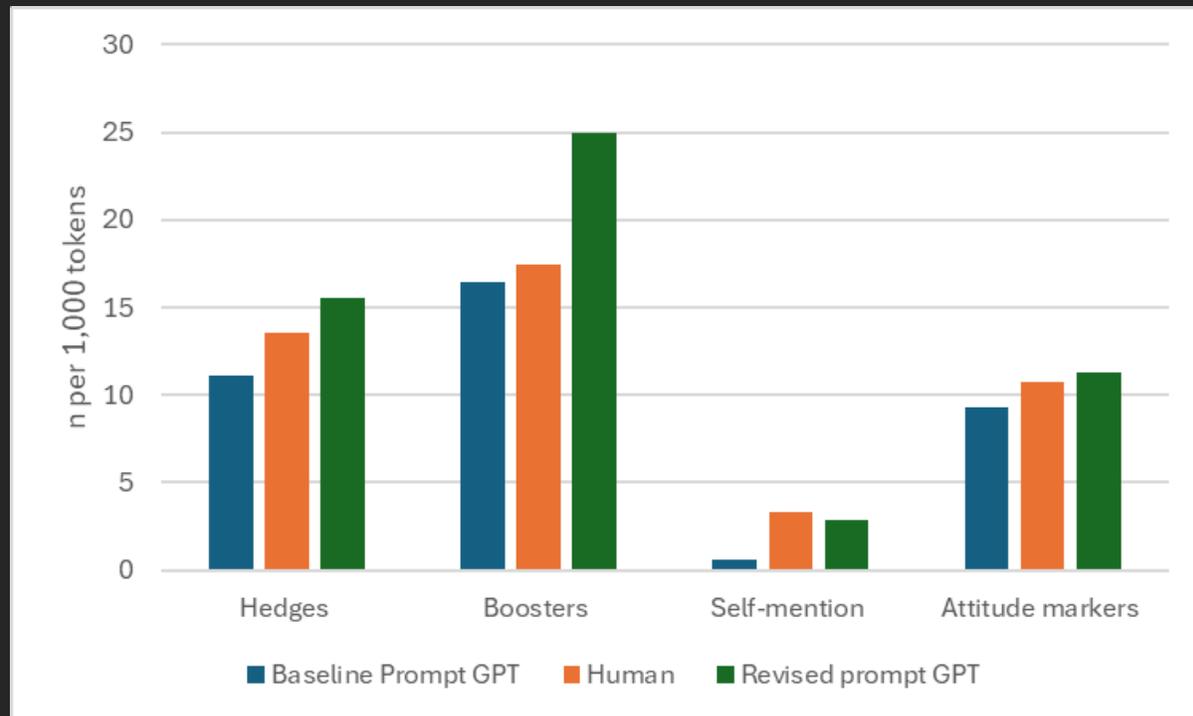


Results – Interactive MD



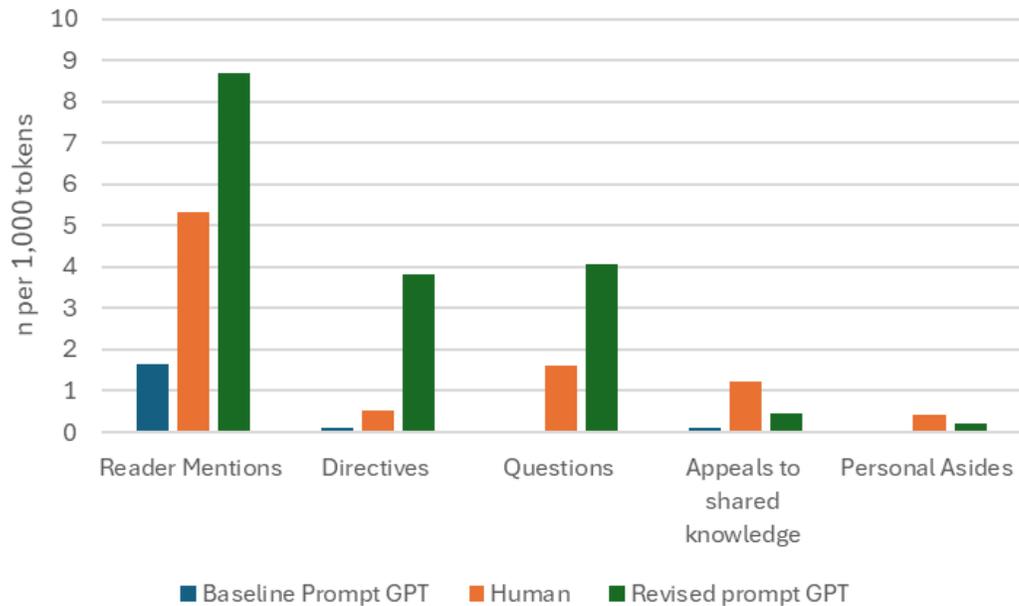
- We asked for code glosses, evidentials, frame markers and transitions – we got them!
- GPT’s logical connective use can be selectively reshaped by targeted prompting, enabling it to adopt more complex contrastive and conditional relations
- Explicit prompting can eliminate earlier differences in endophoric marking by inducing GPT to adopt conventional discourse-referential formulas e.g. ‘as discussed earlier...’, ‘as we have already discussed’, and ‘as mentioned earlier’

Results – Interactional MD



- The degree of difference in the use of AI-human self-mention has reduced considerably, each in line with the revised prompt.
- However, the use of boosters (previously non-significant) has now significantly increased in GPT-produced texts
- This highlights the difficulty of achieving balanced, human-like stance through prompt control alone

Results - Engagement



- Sharp increase in reader mentions, directives and questions in GPT texts as a result of the revised prompt as compared with the baseline, now far surpassing those of human produced texts.
- The use of appeals to shared knowledge and personal asides is now more in line with human produced texts.
- Therefore, engagement in GPT writing is highly prompt-sensitive, with targeted instructions capable of driving reader-oriented features beyond natural human levels

Summary – implications for research

- Taken together, these findings demonstrate that metadiscourse patterns in GPT-generated academic writing are highly prompt-sensitive rather than intrinsically human-like or human-unlike.
- More fundamentally, the findings call into question the construct validity of treating “AI metadiscourse” as a stable object of analysis.
- If the distribution and realisation of metadiscourse features can be so substantially altered through relatively minor changes in prompt design, then what is being measured in such studies is no longer a stable property of AI writing, but a prompt-contingent textual artefact.
- In this sense, metadiscourse in AI-generated texts constitutes a moving target whose linguistic profile is co-produced by the model and the instructions it receives

Summary 2 – implications for pedagogy

- If AI-generated academic texts can display radically different profiles of stance, engagement, and textual organisation depending on how they are prompted, then such texts cannot be treated as stable or reliable models of academic writing for learners.
- Students who rely on AI output as exemplars of “good academic style” may inadvertently acquire distorted patterns of metadiscourse which diverge from disciplinary norms of calibrated commitment and reader-oriented argumentation.

Summary 3 – implications for pedagogy

- Pedagogical practice can therefore shift from asking whether students should use AI, to how they can be trained to evaluate, revise, and rhetorically reshape AI-generated drafts, particularly with respect to hedging, evidentiality, self-mention, and engagement strategies.
- Developing learners' awareness of prompt design and its rhetorical consequences may also form part of emerging "AI literacy", enabling students to recognise that academic voice is not mechanically produced but strategically constructed through socially situated choices about how knowledge claims are presented and how readers are positioned.

What about using corpora to teach? Is DDL dead?

We all know and love DDL!

DDL 'works' – as proved by so many secondary reviews (e.g. Vyatkina, Boulton, Lee et al, Lusta et al, etc).

So many wonderful OA DDL resources and DDL-friendly tools (Antconc, CorpusMate, SKELL, BYU suite, Le Foll, Gablasova, Pinto et al...)

But problems persist (Crosthwaite & Boulton, *coming out one day*)

- Improved (albeit mostly directly untested) definitions of the theoretical underpinnings of DDL.
- A lack of studies focusing on the impact of DDL at the cognitive level or actual improvements to learner autonomy
- A general lack of information on the training provided in the use of corpus tools and integration of DDL in many DDL studies
- Applied linguists, not teachers as main stakeholders (Crosthwaite & Sanhueza, forthcoming)
- Overreliance on concordancing at the expense of other data types (Crosthwaite & Schweinberger, forthcoming)
- A need for DDL studies with learners other than tertiary students, especially for English (Crosthwaite, 2019)
- A lack of methodological rigour in data collection, analysis, instruments and methods preventing replication, often coupled with poor reporting practices

DDL vs Generative AI

The continued case for corpora over GenAI includes:

- **Knowing the data** - we know – exactly - the domain of texts from which the corpus data is derived
- **Authenticity** - that is, language data actually produced by humans – should be seen as a more reliable indicator of real language-in-use.
- **Replicability** – we 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.
- **Safety** - lack of clarity about how GenAI tools and companies use user data
- **Active vs. passive learning** – The L in DDL requires much intention on the part of the learner to succeed

But - Death to concordancing (for DDL)? (Crosthwaite & Schweinberger, 2025)



Overreliance on concordancing at the expense of other software, data and corpus types (Crosthwaite & Schweinberger, forthcoming)



“concordancing itself is hard work, requiring a relatively high degree of reading comprehension depending on the corpus data involved, the ability to sort/manipulate surrounding context to ‘see’ the patterns in the data. And - in fairness - concordancing itself might not be the most exciting classroom activity for many learners”



Studies e.g., Mishan (2004) note the “density and apparent complexity of the concordance” (p. 222)



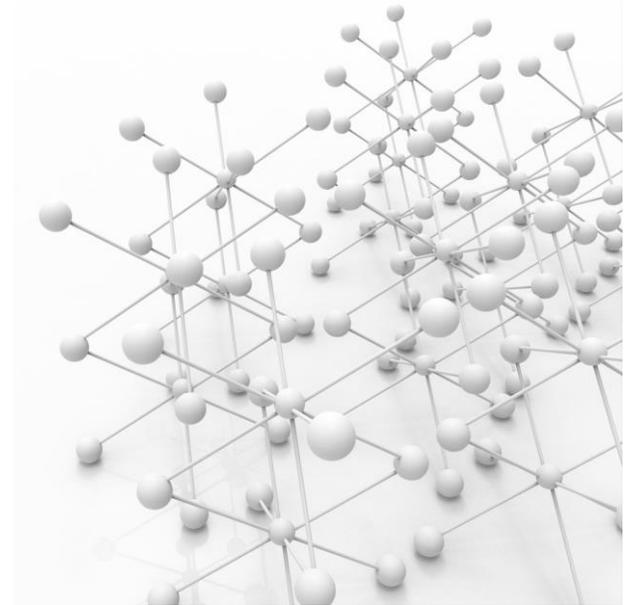
Students vary in their preference for concordancing over more statistical forms of information provided via corpora (e.g. wordlists, collocation statistics), with students in the sciences preferring the latter over students in the humanities (Crosthwaite et. al., 2019).



Therefore - and as much as it might pain us to say it - it may be time to drop concordancing as the central feature underpinning DDL pedagogy in favour of presenting other forms of corpus data such as collocates, frequency visualization, etc. at least in the initial stages of DDL training

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
- GenAI can help boost DDL by:
 - Lowering required technical knowledge
 - Simplified user experience
 - Opportunities for **differentiation** of output
 - Data size scaling
 - Prior user input as next input
 - Translation of findings to teaching materials





English-Corpora.org now offers something entirely new: the ability to combine the depth and reliability of corpus data with the analytic power of Large Language Models (LLMs) like GPT, Gemini, Claude, Perplexity, Llama, Mistral, and DeepSeek.

With just one click, the corpus can send collocates, frequency patterns, phrase lists, or concordance lines to an LLM — which will instantly group, explain, and interpret the data. These AI-powered insights appear directly in the interface, alongside the original corpus results.

The result? Faster understanding of patterns, clearer semantic groupings, and deeper insight into how language works — for language learners and researchers alike. The corpus data remains front and center, but now with the option of an intelligent assistant working behind the scenes to guide your analysis.

Corpora first. AI second. Insights faster.

Integrating GenAI into DDL - tools

- Antconc – the perennial mainstay of much DDL research, has already introduced ChatAI into its latest version
- Atsushi Mizumoto has already embedded GenAI into many of his online DDL resources
- Mark Davies has already incorporated GenAI into the BYU/English Corpora suite (<https://www.english-corpora.org/ai-llms/>) with accompanying video tutorials of how and why.

Integrating GenAI into DDL: Lessons (Ma, Lee & Crosthwaite, 2026 - Advancing English Language Teaching – A Practical Guide to Corpus-Based Language Pedagogy

- Task 2: Consult ChatGPT [Pair work]
Work in pairs. Take turns reading your dialogues in Task 1 and see if you have different answers. Then, ask *ChatGPT* on *Poe* about the definitions of “see”, “look” and “watch”, and write down some keywords or sentences.

To consult ChatGPT, you can provide prompts like the follows:

1. Can you explain the differences among “see”, “look” and “watch” in words for primary school students?
2. Can you use simple words to explain the differences among “see”, “watch” and “look at”?
3. Explain the verbs: “see”, “look at” and “watch” for P6 students.
4. ...

After consulting ChatGPT, can you summarise the differences among these three verbs?

Suggested answers:

See: When you *see* something, you use your eyes to notice without trying.

Look: When you *look* at something, you direct your eyesight to it intentionally.

Watch: When you *watch* something, you pay attention and keep looking for a while.

Follow the steps, read the underlined concordance lines, and then try to define the verb “see” with the help of the teacher.

Step 1: Access COCA. Select “SEARCH” in the top menu, then fill in “see” in the blank.

Step 2: Click “Sections”, select “FICTION” for column 1 and “IGNORE” for column 2. Then click “Find matching strings”.

Step 3: Click on “SEE” to view more information.

Step 4: Read the selected concordance lines of *see* and complete the following activities.

10 sample concordance lines to show students:

1	behind the latticed shutters, not even able to peek her head far enough to <u>see</u> the drops of water land on the dirt courtyard below,
2	If she had turned and returned to the dancehall, pretending she didn't <u>see</u> Rosa, if she had dragged her friend back inside and found the men and...
3	as she could and then turned away, as if she hoped she would never <u>see</u> Fidelia again. When she turned back Fidelia was gone. Elmo had been...
4	Elmo dropped Fidelia at his sister's feet and Fidelia was so frightened to <u>see</u> Rosa again—her eyes ruined plums, her lips a smear -- that she grabbed for
5	got the <u>sweeetest</u> deal going on out there. <u>Woosss</u> , you got <u>ta see</u> it to believe it, “Dub told Jock a few days earlier.
6	to block out the sun and send a chill down her spine. Shade can <u>see</u> the source of the shadow in gaps between the trees.
7	Then, “Okay. So, are we staying? Will we see dragons? I mean, I know it would be super dangerous.
8	“I don't ever <u>see</u> them,” True says. She doesn't add that if she did,
9	do more damage to them than to the dragon. # “I can still <u>see</u> it”, he shouts. “Let's get a move on!”
10	says, two hours later back at the convenience store. “Did you <u>see</u> her?” True asks.

Activity 1:

Circle the noun or pronoun after see in the above sentences and find more examples.

See	
somebody	something
Rosa	it
Fidelia	...
...	

CorpusChat (Cheung & Crosthwaite, 2025, Computer-Assisted Language Learning)

The screenshot displays the CorpusChat interface. At the top left is the CorpusChat logo, and at the top right is the user name 'CAES Super Admin'. A vertical sidebar on the left contains navigation icons for Chat, Chatbots, Craft, Corpus, Admin, and Logout. The main content area is titled 'Chatbot Gallery' and features a 'Favorite Chatbots' section with two entries: 'Chatbot for Humanitie...' and 'Chatbot for BAWE', both using 'gpt-4o-mini' and having a 'Chat Now' button. Below this is a 'Specific Corpus' button. The main gallery has 'Tags' for 'ALL' and 'CAES', and a 'Humanities' filter button. It contains five chatbot cards: 1. 'Chatbot for Humanities and Arts' (starred) with a description of the CHAT corpus (656,484 words) and buttons for 'View Details' and 'Chat Now'. 2. 'Chatbot for BAWE' (starred) with a description of the British Academic Written English Corpus and buttons for 'View Details' and 'Chat Now'. 3. 'Professional Dentistry Corpus (...)' (starred) with a description of the Corpus of Dentistry Research Articles and buttons for 'View Details' and 'Chat Now'. 4. 'Corpus of Social Sciences (CSS)' (starred) with a description of the CSS corpus (nearly 1.4 million words) and buttons for 'View Details' and 'Chat Now'. 5. 'Science Corpus' (starred) with a description of the Science Corpus (1 million words) and buttons for 'View Details' and 'Chat Now'. Additional filter buttons are present: 'British Academic Written Corpus' (orange), 'Dentistry' (blue), 'Social Sciences' (blue), and 'Science' (blue).

Next project – Recursive Language Models (and the BAWE) – Comparing human, baseline AI and AI+human RLM

Recursive Language Models (RLMs) represent an emerging, advanced GenAI methodology designed to overcome the limitations of finite context windows and "context rot" (forgetting information) in large language models.

- Core Principles of the RLM Method
 - **Recursive Decomposition:** The model breaks complex, long-horizon tasks into smaller sub-problems.
 - **Active Agentic Behavior:** Instead of reading everything at once, the model actively queries the input data, using tools to search (grep/retrieve) and analyze.
 - **Self-Correction:** The model can call itself recursively to verify intermediate results and refine its output.
 - **Externalized Memory:** Context is treated as an external file or database rather than entirely occupying the model's active working memory (context window).
- Key Benefits
 - **Handling Massive Context:** RLMs have demonstrated the ability to process over 10 million tokens, far exceeding standard context windows.
 - **Higher Accuracy & Less "Rot":** By breaking down information, the model suffers less from "context rot" or forgetting, outperforming standard, large-context models on long-input tasks.
 - **Lower Costs:** Rather than paying for the entire 10M-token context to be re-processed, users pay for high-value, targeted reasoning tokens, reducing API costs.
 - **Improved Interpretability:** The "reasoning tree" produced by recursive calls makes it easier to trace how the model reached its conclusion.
- RLM vs. Standard RAG (Retrieval-Augmented Generation)
 - **Standard RAG:** Simply retrieves top-k chunks and sends them to the model, which might still be overwhelmed or lose context.
 - **RLM:** Acts as an agent that can iteratively search and refine its findings, making it more suitable for complex, multi-step, or long-form reasoning, such as in legal analysis, software development, or auditing.

Next project 2: Machine-driven prompting vs human prompting

Next project 3: Vibe coding CorpusMate updates with LLM support in Claude Sonnet

The screenshot displays the CorpusMate BAWE Edition interface. At the top, it shows corpus statistics: Files: 2761, Tokens: 6,640,198, Types: 158,431, and With meta: 2,761. The search bar contains the query "how is 'however' used in social science essays?". The interface is divided into several sections:

- Left Sidebar:** Contains filters for Corpus Files (158,431 types indexed), Disciplinary Group (AH, LS, PS, SS), Discipline (All disciplines), Genre Family (Essay, Research report, Critique, Case study, Methodology recount, Literature survey, Explanation, Narrative recount, Proposal, Exercise), Level (1, 2, 3, 4), Writer L1 (All L1s), Grade (D, M), and View (KWIC).
- Search Results Summary:** Shows "Searching for: however" with a frequency of 11,856 hits. It includes sections for Frequency & Spread, Typical Usage Pattern, Key Collocates (BY MI), and BAWE Context.
- Concordance View:** Displays 11 hits for "however" in a KWIC format. Each hit shows the surrounding text, the word "however" highlighted in yellow, and the source (SS Sociology - essay - L1).

Line	Text	Word	Text	Source
1	racism is an integral and internalised part of our society,	however	I will not be looking at individual or specific cases	SS Sociology - essay - L1
2	and that certain races were superior to others (Miles, 1989).	However	throughout the nineteenth century racism was rejected as a scientific	SS Sociology - essay - L1
3	it has left a legacy of racism in contemporary society.	However	this is not the only reason for the continuation of	SS Sociology - essay - L1
4	racialism has been imbedded into the economy since before capitalism.	However	since the emergence of capitalism these racist ideas have been	SS Sociology - essay - L1
5	that ethnic minorities provide this army (Procter, Lecture notes, 2004).	However	by constructing ethnic minorities as a reserve army of labour	SS Sociology - essay - L1
6	are needed as a 'reserve army of labour' (Miles, 1989).	However	politics, and the policies of British political parties over the	SS Sociology - essay - L1
7	for reliability and validity in order to produce realistic statistics,	however	there are many problems with official statistics in this aspect,	SS Sociology - essay - L1
8	of statistics in the end. Cicourel also supports this idea;	however	he takes it one step further by saying that these	SS Sociology - essay - L1
9	involved, and the different ways in which people apply concepts.	However	according to the radical approach there are more limitations with	SS Sociology - essay - L1
10	the servants, who were working class and held little power.	However,	despite these limitations official statistics do have some uses. According	SS Sociology - essay - L1
11	are not produced using legal requirements can also be useful,	however.	Bulmer claims that statisticians often go to many lengths to	SS Sociology - essay - L1

Summary

Integrating GenAI into DDL requires careful consideration of the affordances and weaknesses of both if meaningful integration is to occur

Why compete? Why gatekeep DDL to concordances? Hasn't the ship sailed?

We aren't proposing a replacement of corpus linguistics, authentic texts, active learning, constructivism, usage-based learning, sociocultural DDL.

We are, however, proposing to make DDL *better, faster, smarter and more accessible...*

Is that what we all want as teachers, learners and practitioners?

We hope that what we have demonstrated today can go some way in doing that.



Thanks for listening



Peter Crosthwaite
(p.cros@uq.edu.au)



Send a mail for any
papers mentioned in
this talk, or the slides.



Hope to hear from you
soon!

