

L2 Writing in the Age of Generative AI: Presenting a new GenAI-assisted corpus

May 11th 2026

Nathan Vandeweerd & Anna Zaffinelli

LANSPAN, University of Groningen

INTRODUCTION

THE THREAT OF GEN-AI

Chat GPT Use in Fall Quarter 2022 Finals

If you used ChatGPT to assist with your final fall quarter assignments and/or exams, to what extent did you use it?

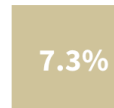
Brainstorming, outlining, and forming ideas



Answered multiple-choice questions with the help of ChatGPT



Submitted written material from ChatGPT with edits



Submitted written material from ChatGPT without edits



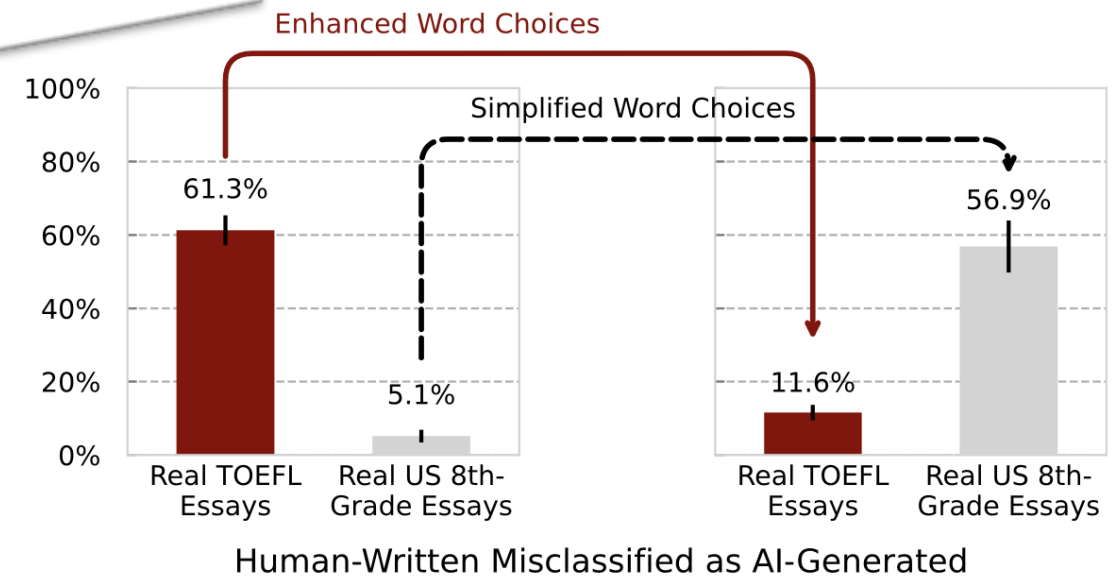
Source: [Fizz Post](#) • In an anonymous poll uploaded on Fizz, 4,497 Stanford students answered whether or not they used ChatGPT to help with their Fall final assignments.
Graphic: MARK ALLEN CU/The Stanford Daily



(Cu & Hochman, 2023)

INTRODUCTION

AUTOMATIC DETECTION



(Liang et al., 2023: 2)

INTRODUCTION

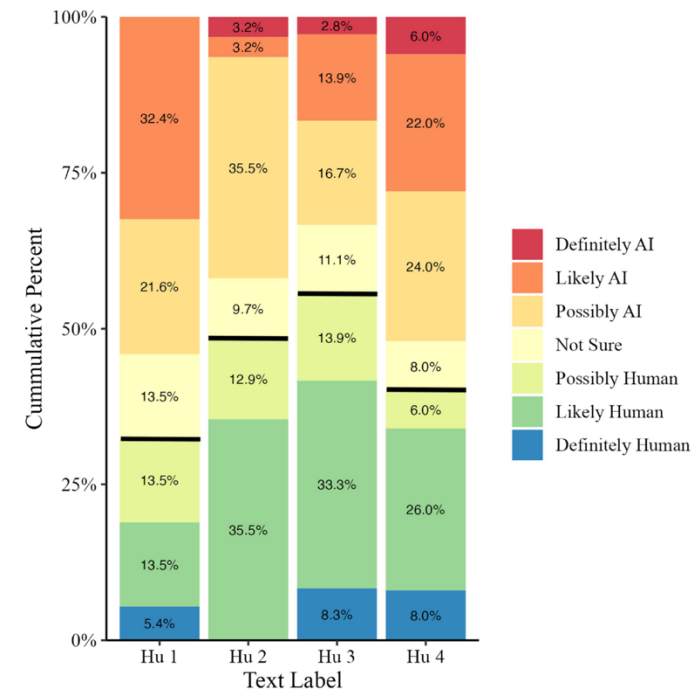
HUMAN IDENTIFICATION OF AI-GENERATED CONTENT

Laypeople (recruited from Prolific):



(Jakesch et al., 2023: 2)

Trained linguists:



(Casal & Kessler, 2023: 7)

INTRODUCTION

COMPARING AI-GENERATED CONTENT VS. L2 TEXTS

- Etaat, 2026:
 - GenAI (ChatGPT-4) essays are longer, more syntactically complex and have higher lexical diversity and more content words
 - L2 essays contain more errors
 - Similar cohesion patterns
- Yuan and Dao (forthcoming):
 - GenAI (ChatGPT-3.5) essays show more lexical sophistication
 - L2 essays more errors
 - No sig. differences in clausal complexity measures, fluency (length) measures

INTRODUCTION

COMPARING AI-GENERATED CONTENT VS. L2 TEXTS

*AI prompt:

“Write an argumentative essay of about 400 words in length for a student assignment for an English as a foreign language class, using no headings, in a single paragraph.”

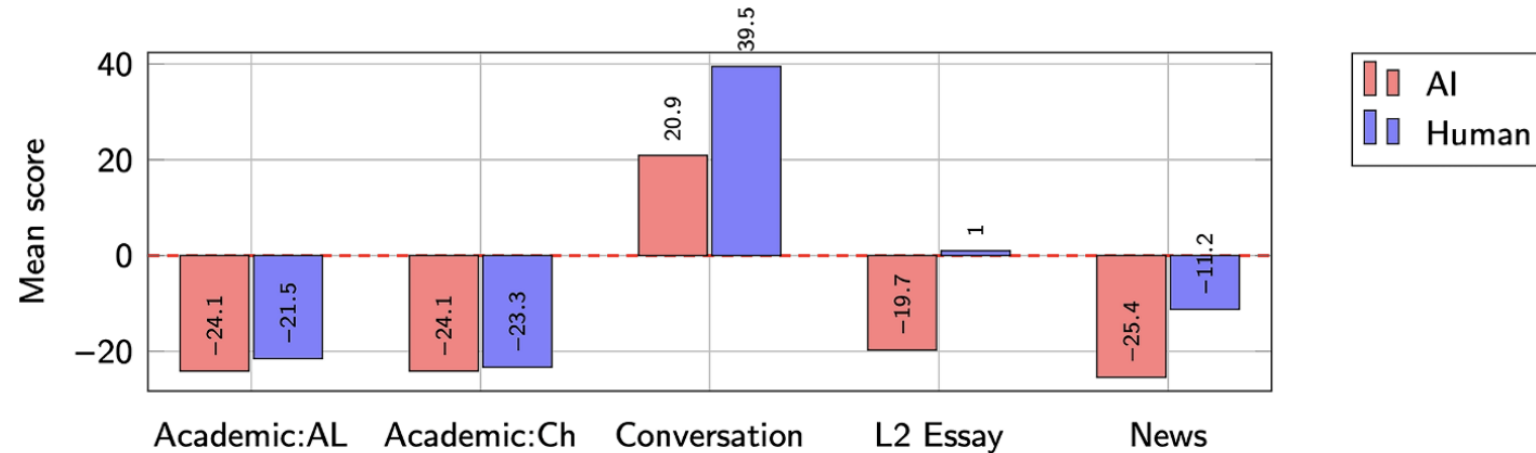


Fig. 1. Mean dimension scores for Dimension 1: Involved versus Informational Production.

(Berber Sardinha, 2023: 6)

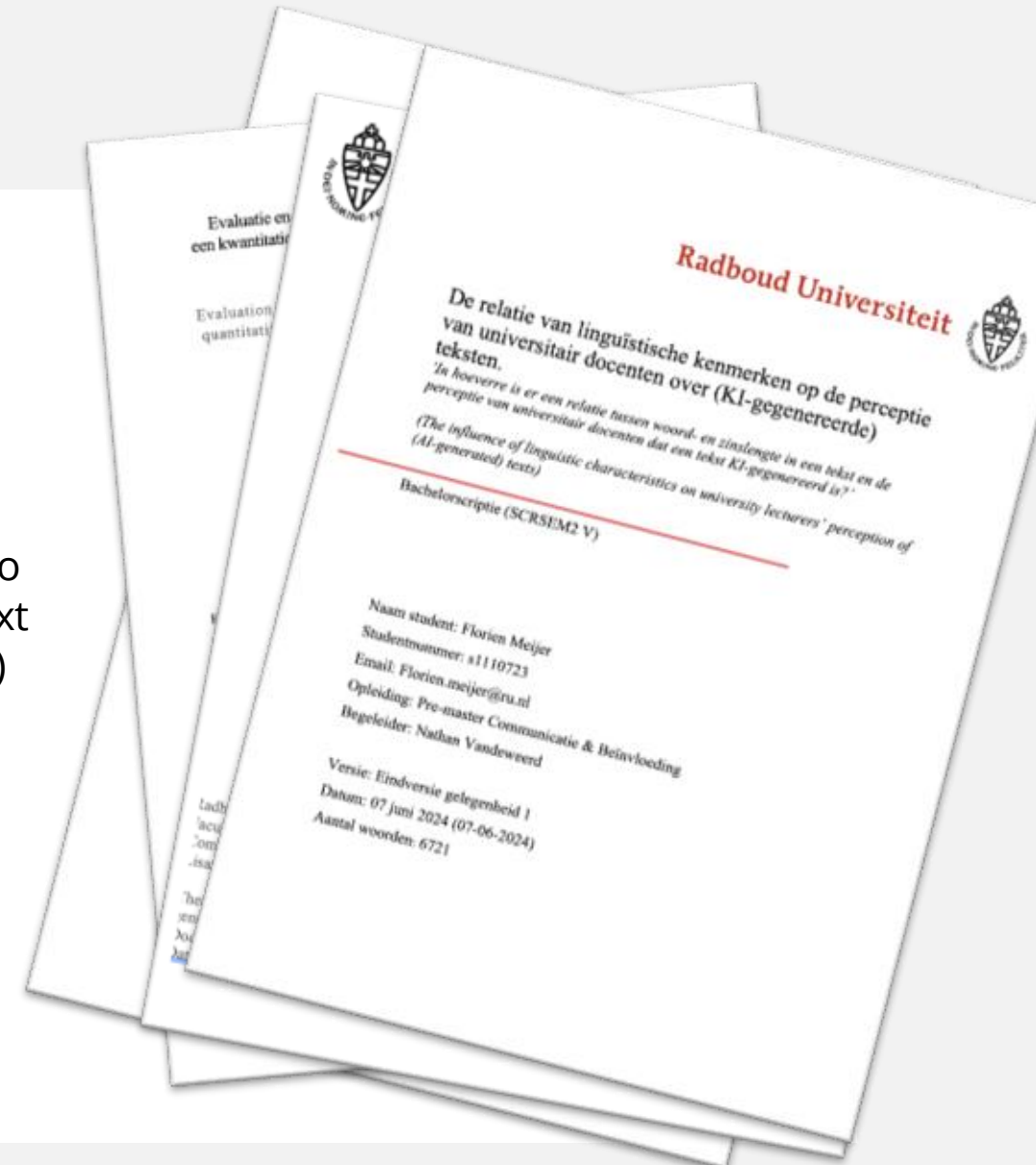
INTRODUCTION

L2 DUTCH

BA Thesis project (2023-2024)

- 30 Argumentative Texts from the Leendercorpus Nederlands (Degand & Perrez, 2004)
 - 30 texts generated by ChatGPT-3.5 using the same essay topics using Chain of Thought prompting
 - All texts rated by a sample of university lecturers ($n = 30$) who each rated at least 10 texts on a 5-point Likert scale ("This text was written by AI", 1 = Written by a human, 5 = Written by AI)
- Spelling mistakes per 100 words $\rho = -0.64^{**}$
 - Incorrect word order $\rho = -0.67^{**}$
 - Average sentence length $\rho = 0.66^{***}$
 - Proximity of temporal connective makers (cohesion) $\rho = -0.11$ (n.s.) but: $L2 > AI^{**}$
 - Overlap in lemmas between sentences (cohesion) $\rho = 0.12$ (n.s.) but $L2 > AI^{**}$
 - Average word length $\rho = 0.59^{***}$
 - MTLD (lexical diversity) $\rho = 0.39^{**}$
 - Logged Word Frequency $\rho = -0.69^{***}$

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$



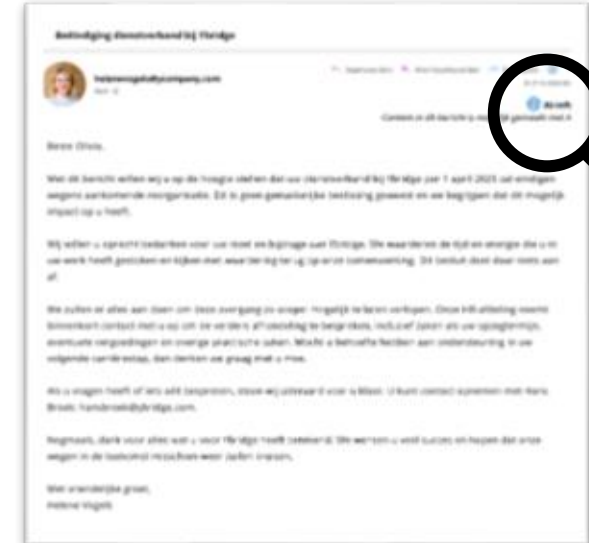
INTRODUCTION

AN INTERESTING OBSERVATION...

 **AI-info**
Content in dit bericht is mogelijk gemaakt met AI

BA/MA Thesis projects 2025-2026

- Perception of AI-generated texts (e.g., empathy, trust, credibility)
- Manipulation check comparing “AI/human generated texts” (**only difference is the presence/absence of an “AI label”**)
 - Either... no sig. difference between conditions (Deuzeman, 2025)
 - Or... even when difference is significant, the non-AI labelled texts still often have higher than average AI-ratings (on a 7-point scale)
 - Mean_{AI-label} = 4.45 (SD = 1.00) vs. **M_{no-label} = 3.89** (SD = 1.57) (Croonen, 2025; Deuzeman, 2025; Hettinga, 2025; Linssen, 2025; Otte, 2025; van de Kuit, 2025; Oude Brunink, 2025)
 - Mean_{AI-label} = 5.73 (SD = 0.93) vs. **M_{no-label} = 4.07** (SD = 1.35) (Akay, 2025)



SOME GENERAL OBSERVATIONS

- Both humans and automated tools are not very good at distinguishing between (L2) writers and AI output
- Humans sometimes rely on “incorrect heuristics”
- Baseline level of suspicion for AI-content (even when texts are human-written)

...So what is it that makes an (L2) text feel *authentic*?

INTRODUCTION

A SIMILAR CONSTRUCT



B2



C1/C2

- ✓ Similar lexical complexity
- ✓ Similar syntactic complexity
- ✓ Differences in word-combinations

E.g.,
délinquance + juvenile ($PMI = 11.62$)
'juvenile delinquency'

(Vandeweerd et al., 2021, 2023a, 2023b)

*“[...] the ability of the native speaker routinely to convey his meaning by an expression that **is not only grammatical but also nativelike**...how he selects a sentence that is **natural and idiomatic** from among the range of grammatically correct paraphrases”*

“[...] the ability of the speaker routinely to convey his meaning by an expression that **is not only grammatical but also *humanlike***...how he selects a sentence that is **natural and *authentic*** from among the range of grammatically correct paraphrases”

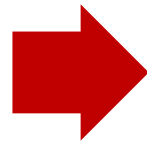
INTRODUCTION

THE AI AUTHENTICITY PROJECT

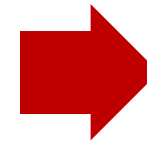
Which are the defining features of L2 writing that lead readers to perceive it as human-authored?



Perceptions of
authenticity in L2
writing



Perceived linguistic
features vs. actual
patterns



Perceptions of
authenticity after
changes in linguistic
features

INTRODUCTION

THE AI AUTHENTICITY PROJECT

Which are the defining features of L2 writing that lead readers to perceive it as human-authored?



Perceptions of
authenticity in L2
writing



Perceived linguistic
features vs. actual
patterns



Perceptions of
authenticity after
changes in linguistic
features

INTRODUCTION

INTERVIEWS AND FOCUS GROUPS WITH EDUCATORS

Which linguistic features do educators associate with human L2 and GenAI writing?

- Participants presented with **portfolio of human L2 and AI-assisted texts**
- Interviews and focus groups with 7 educators teaching 1st year BA courses (EMI)
- Content and thematic analysis of written transcripts
- Main findings:
 - Difficulty distinguishing AI and L2 texts (and judgements generally inaccurate)
 - No clear consensus on what constitutes “authentic” writing
 - AI texts: advanced vocab, complex syntax, clear structure
 - L2 texts: looser structure, personal tone, distinct voice

(Zaffinelli et al., forthcoming)

TOWARDS A NEW CORPUS OF AI-ASSISTED L2 WRITING

MOTIVATIONS

- Traditional L2 corpora predate GenAI rise (e.g., ICLE)
- Previous studies have focused on only one writing task (cf. Berber Sardinha, 2024)
- Often no matched human-only control condition
- Process data usually incomplete
- **AI/human operationalized as a binary variable**

AI-ASSISTED L2 CORPUS **CORPUS COMPILATION**

SAMPLE POPULATION

- **1st year English-taught Bachelor's** programme
- Taal- en Letterkunde (Engels)

EXPERIMENTAL DESIGN

- Mixed within and between subject design
- 2 conditions: **AI-assisted** vs. **unaided writing**
- 2 task types: **narrative** vs. **persuasive essay**
- 2 writing sessions
- **Writing process data:** Inputlog (Leijten & van Waes, 2013) and OBS Studio

AI-ASSISTED L2 CORPUS

CORPUS COMPILATION

WRITING TASKS

- Topic control: language
- **Argumentative** essay on the topic of English Medium Instruction (EMI) (Source material provided)
- A personal **narrative** about experience with language

Universiteit
Antwerpen

METADATA

- **LexTALE** score
- **Learner metadata** (based on the core metadata scheme for learner corpora; Paquot et al., 2024)
 - Language background
 - Learners' exposure to English L2
- **GenAI questionnaire:**
 - GenAI acceptance,
 - perceived extent of GenAI usage
 - perceived type of GenAI usage

AI-ASSISTED L2 CORPUS

A CONTROLLED AI-ENVIRONMENT

- Participants interact with ChatGPT-5.2 within the Qualtrics survey (dedicated API key)
- Ensures comparability and replicability across participants (no previous history)
- Saves the conversation history

ChatGPT Chat

Waiting for Response.

Hi

2 user what should I definitely include in a narrative essay about my experience with language?

0 user Okay, I have to write a text of 400 words. I have to talk about a true life story about my experience with languages. I want to talk about my brother who moves to Ameirica when I was 4 years old. Thats when I started learning English and how I develepoded my love for languages

2 user how would one translate 'Algemeen nederlands' in Dutch?

Send Reset

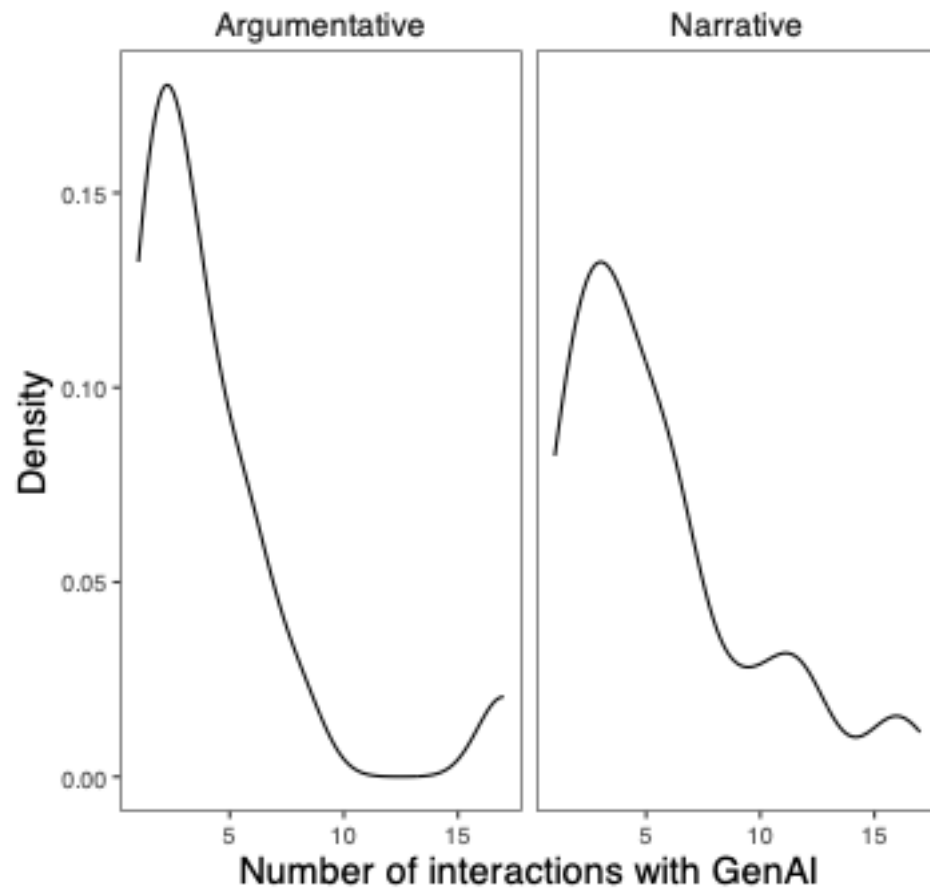
AI-ASSISTED L2 CORPUS

DATA COLLECTED TO DATE

- Collected data from 100 participants
 - 80 consented to use of text and survey data
 - 67 consented to use of Inputlog
 - 10 consented to the use of OBS Studio
 - Final sample of **64 participants** with usable data
- 48 females, 10 males, 6 other/prefer not to answer
- Age 17-39 ($M = 19.69$, $SD = 3.12$)
- 89% L1 Dutch ($n = 57$);
- Self-assessed proficiency B1 ($n = 2$), B2 ($n = 28$), C1 ($n = 29$), C2 ($n = 6$)
- LexTALE score (64.62-100/100, $M = 87.52$, $SD = 8.66$)

Condition	Arg.	Nar.	Total
Unaided	33	33	66
AI-Assisted	31	31	62
Total	64	64	128

INITIAL FINDINGS: INTERACTION WITH GENAI



- Many ($n = 27$) in AI condition did not interact at all with GenAI!
- Range of 1-17 interactions with ChatGPT ($M = 4.78$, $SD = 3.92$)

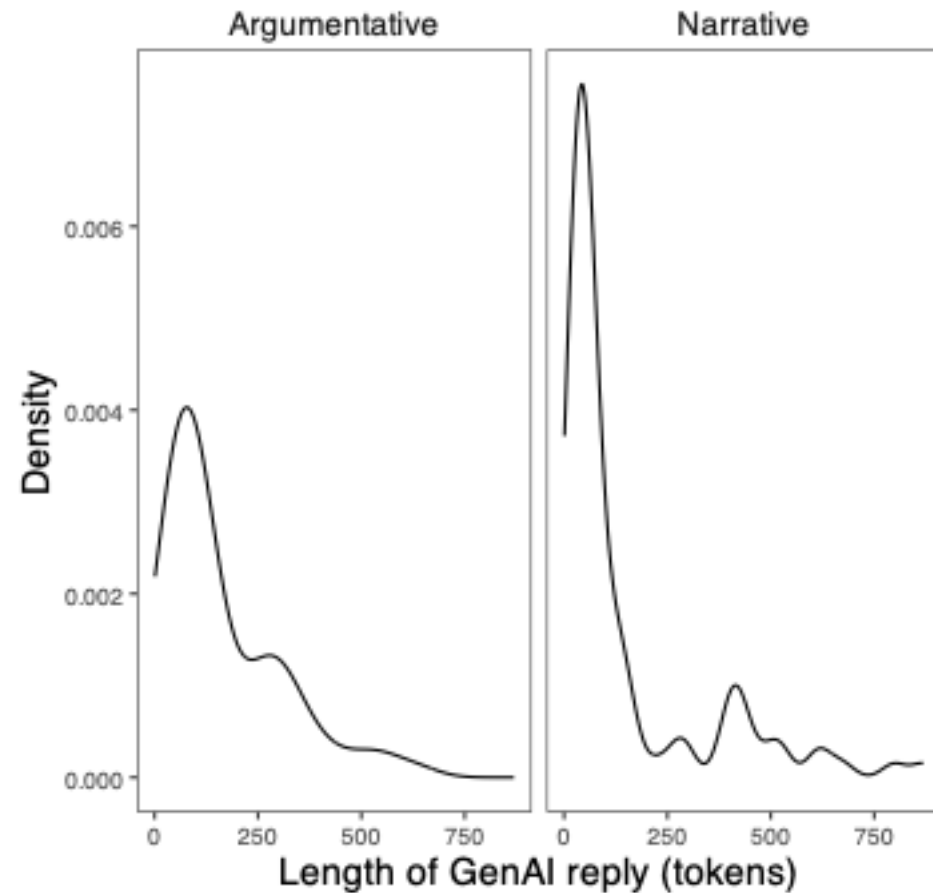
WHY DID PARTICIPANTS CHOOSE NOT TO USE GEN-AI?

PP043: I believe it does not add much value to the writing. However, it can be useful as a revision tool when used properly. Also, I do have a general aversion against AI. I believe **one can learn the skills to write better texts with more depth than a machine can do**. Everything the machine can do, you can do yourself, or with the help of other people.

PP077: Because **I'm too scared that using GenAI will lead to my texts lacking personality** and I prefer having a text that is absolutely mine than to one that does not feel like it is.

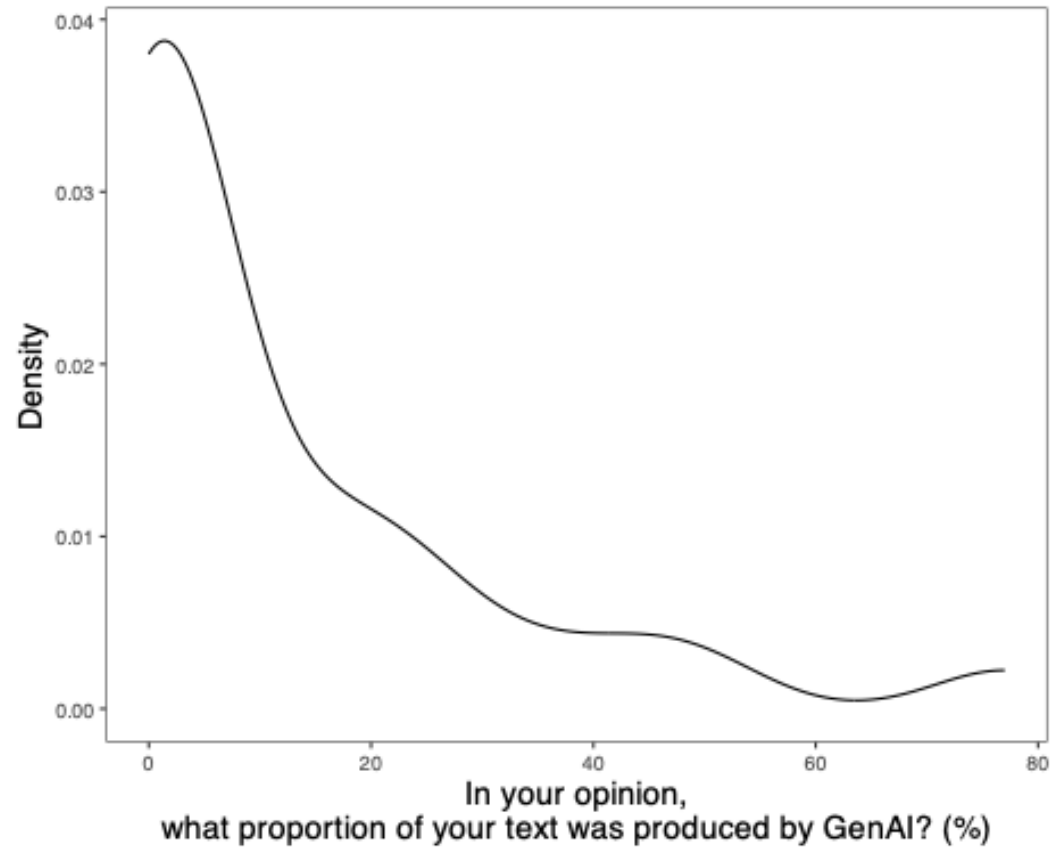
PP067: Because **I think it is important to develop the skill myself**, instead of relying on GenAI.

INITIAL FINDINGS: INTERACTION WITH GENAI

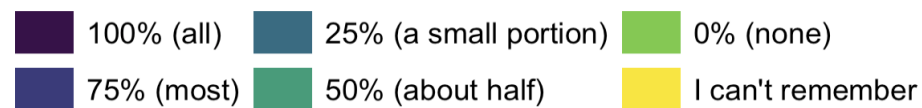
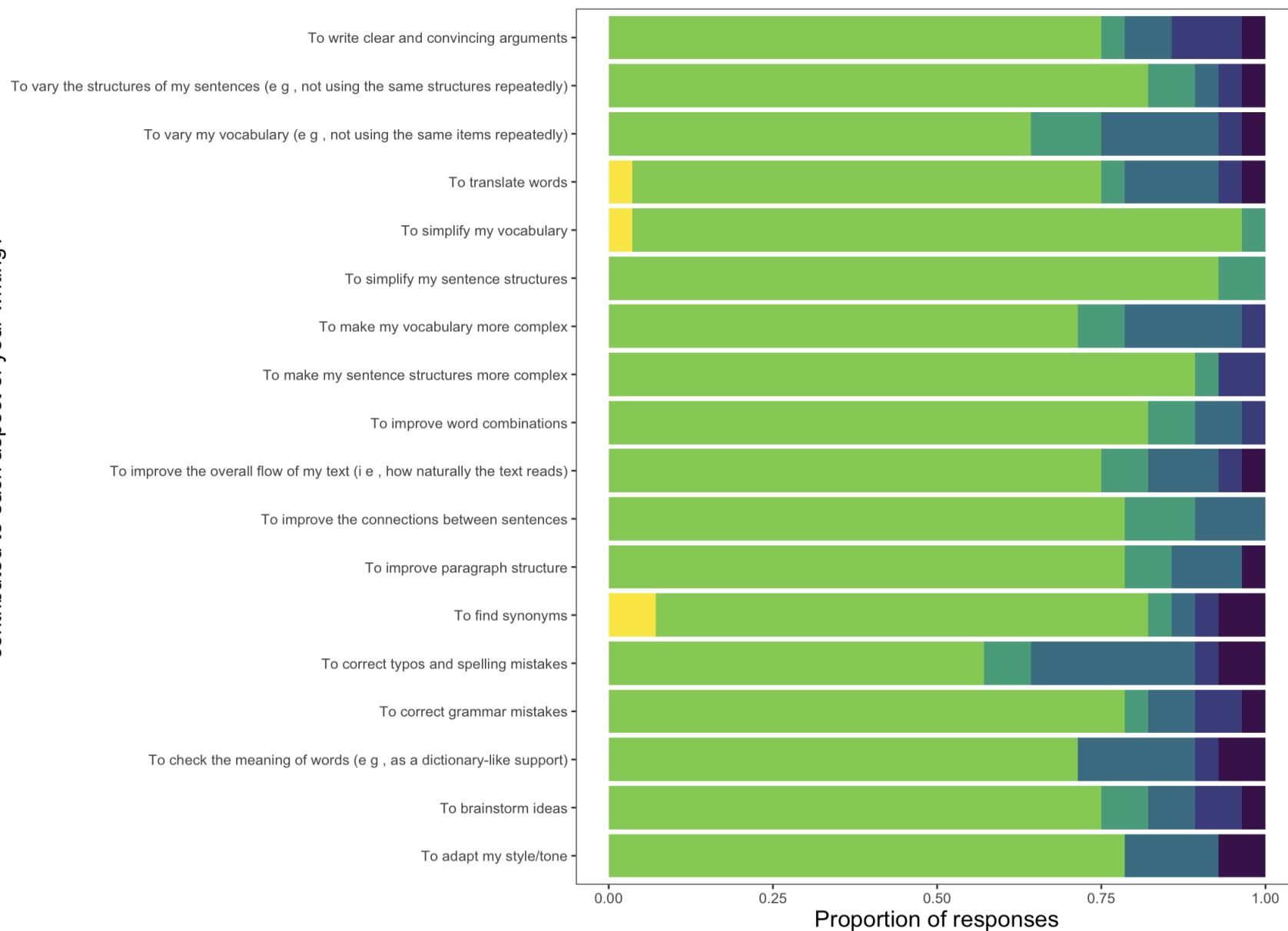


- Range of 1-869 token responses from ChatGPT ($M=151.43$, $SD = 172.83$)

USE OF GEN-AI DURING SESSION 1 (ACCORDING TO PARTICIPANTS)

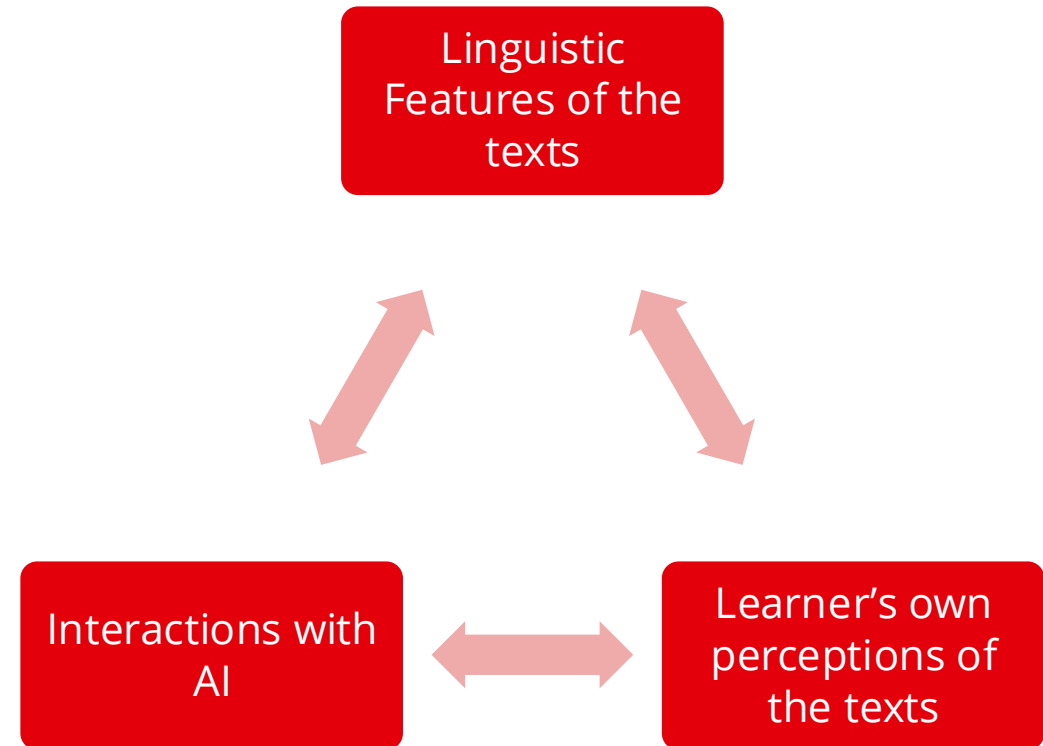


How much do you feel that GenAI contributed to each aspect of your writing?



NEXT STEPS: TRIANGULATING THE DATA

- Authenticity in L2 writing cannot be located in the product or process alone but must be investigated jointly across process and product data.
- Integrate lexico-grammar features and writing process data using multidimensional analysis.



NEXT STEPS: TENTATIVE METHODOLOGY

PRINCIPAL COMPONENT ANALYSIS

- Lexico-grammatical features (Biber Tagger)
- Writing process variables from Inputlog (Leijten & van Waes, 2013)
 - o Pausing behavior
 - o Fluency
 - o Revision behavior
 - o Source use
- New dimensions?

NEXT STEPS: TENTATIVE METHODOLOGY

PRINCIPAL COMPONENT ANALYSIS

- Lexico-grammatical features (Biber Tagger)
- Writing process variables from Inputlog (Leijten & van Waes, 2013)
 - o Pausing behavior
 - o Fluency
 - o Revision behavior
 - o Source use
- New dimensions?

STATISTICAL MODELING

```
model_dim1 <- lmer(dim1_score ~  
condition * task_type + lexTALE +  
(1|participantID), data = dataset)
```

WRAPPING UP

METHODOLOGICAL AND ETHICAL CONSIDERATIONS

- Replicable use of genAI
- Ethical considerations
 - Use of AI (in general)
 - Assignment to conditions
- GenAI use is not a binary variable

NEXT STEPS: DATA SHARING

Aim: Open data to be shared with the research community (hopefully mid 2027)

- L2 and AI-assisted texts
- Chatlogs
- Metadata questionnaire
- GenAI questionnaire
- LexTALE results
- Inputlog data

WRAPPING UP

SOME CONCLUSIONS

- AI-use is not a binary variable: Wide degree of variation in terms of amount and type of AI-use by L2 learners
- Importance of including both product, process and perception data
- Creation of new AI-assisted corpora involve new methodological and ethical challenges
- This is an ongoing project and we are open for collaboration!

Get in touch here:

genaisis.corpus@ru.nl

Q & Eh?

SELECTED REFERENCES

- Berber Sardinha, T. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1). <https://doi.org/10.1016/j.acorp.2023.100083>
- Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, 2(3), 100068. <https://doi.org/10.1016/j.rmal.2023.100068>
- Etaat, F. (2026). Exploring linguistic fingerprints in human and AI-generated texts: An NLP-based approach in second language writing. *Ampersand*, 16. <https://doi.org/10.1016/j.amper.2026.100258>
- Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11). <https://doi.org/10.1073/pnas.2208839120>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- Yuan, S., & Dao, P. (2026). *ESL tertiary learners' engagement in ChatGPT-assisted second language writing*. Routledge. <https://doi.org/10.17863/CAM.128762>