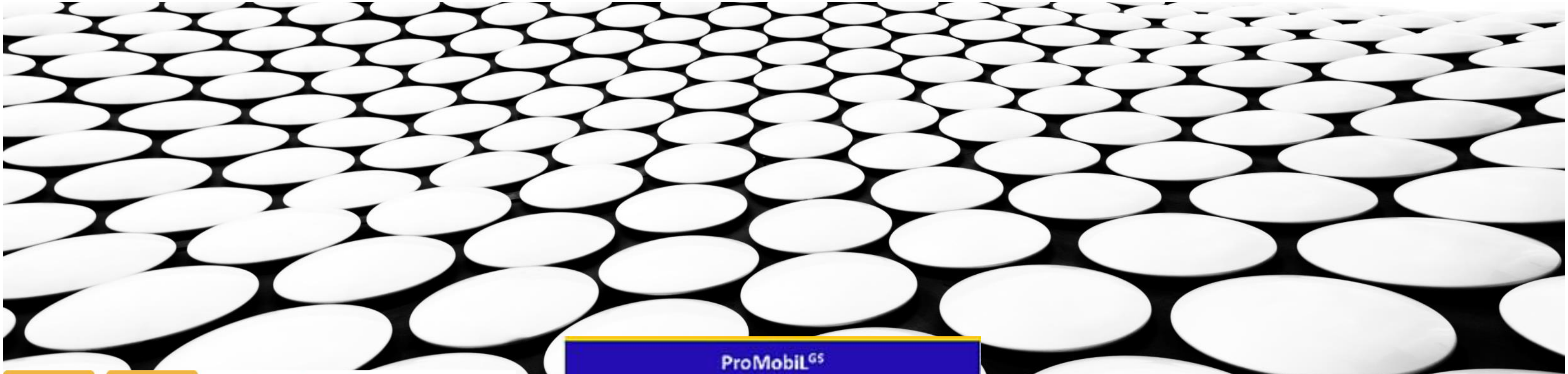
ADOPTING CHATGPT AS A WRITING BUDDY IN THE ACADEMIC L2 WRITING CLASS

CAROLA STROBL, IRYNA MENKE-BAZHUTKINA, NIKLAS ABEL AND MARIJE MICHEL



University of Antwerp
TRICS | Translation, Interpreting and Intercultural Studies

ProMobiL^{GS}
INTERNATIONAL
SYMPOSIUM 2023

university of groningen
faculty of arts

INSPIRATION AND BACKGROUND

- Translation pedagogy: AI-based MT tools since 2017 (DeepL)
  -> post-editing MT as a (new) task in the translation classroom to promote effective use of MT in translation practice (Balling et al., 2014; Chung, 2020)
- Writing pedagogy:
  - Discussion about integration of digital tools from a process- and product-oriented perspective (Oh, 2022)
  - AI-generated text takes writing support to a next level (Gayed et al., 2022)

  -> need for pedagogically sound embedding into the (L2) writing classroom to promote awareness of advantages and pitfalls of tools such as ChatGPT as "writing buddy" (Kasneci et al., 2023)
- Our approach: stimulate '**inner feedback**' (Nicol, 2021) through comparison of own text with AI-generated model

  writing > comparing > revising
- Model-based feedback: students mainly notice **vocabulary** and to a lesser extent **content** issues
  (e.g., Cánovas Guirao, 2015; Hanaoka, 2007; Kang, 2023; Mayo & Labandibar, 2017; Roothooft et al., 2022)

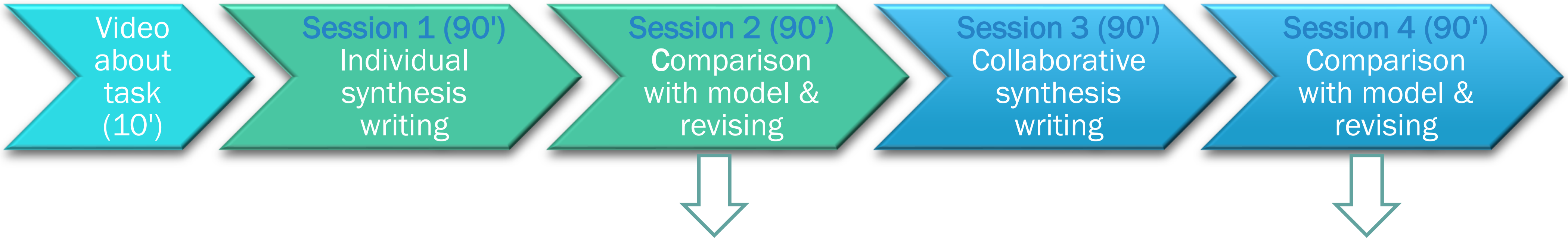
RESEARCH QUESTIONS

RQ1 What do students notice in their own output and in Chat-GPT output based on a guided comparison?

RQ2 What do students revise in their own texts?

METHOD AND DATA

- Participants: 22 university students from U of Groningen minoring in L2 German (CEF-levels B2-C1)
- Task: S1 & S3: Synthesis writing from two popular-scientific source texts on linguistic topics of contemporary German (*Kiezdeutsch* & Anglicisms). S2 & S4: Compare with two ChatGPT models (pre-generated) + revise own texts
  Environment: Google Docs



Video about task (10')

Session 1 (90') Individual synthesis writing

Session 2 (90‘) Comparison with model & revising

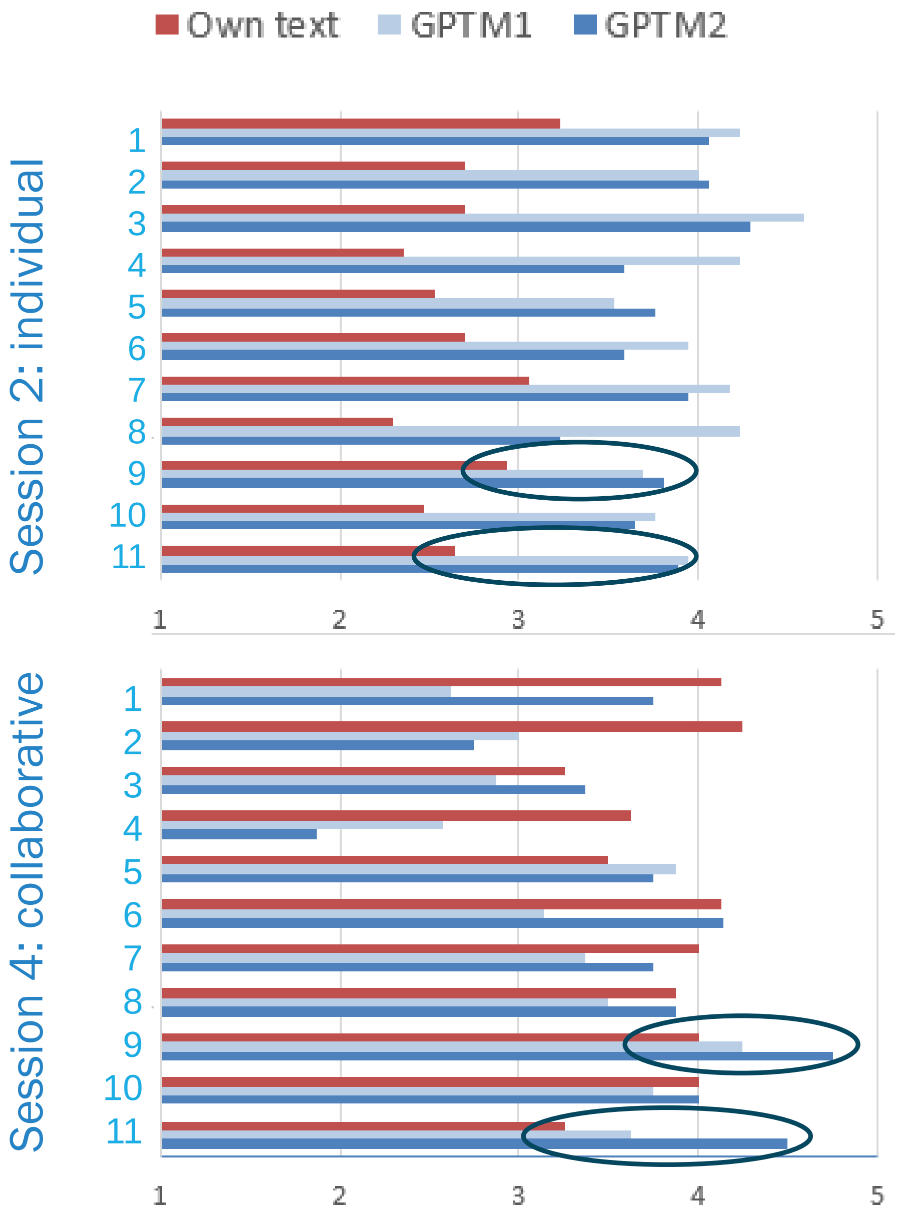
Session 3 (90') Collaborative synthesis writing

Session 4 (90‘) Comparison with model & revising

- “Noticing data” (RQ1): Guided evaluation and comparison of own text with two Chat-GPT models:
  - 11 pre-defined text quality statements (Likert-scale)
  - Free-text comments (three strong + three weak points of the models)
- “Revision data” (RQ2): Screen-recordings (Screenpresso) and audio-recordings (mobile phones)
  9 revision sessions of 6 participants (6 individual and 3 collaborative revisions) coded by three coders (Atlas-TI)

RESULTS: GUIDED COMPARISON

1. The synthesis reproduces well the content of both source texts.
2. The synthesis has a clear and logical structure.
3. The introduction summarises the theme of the synthesis.
4. The main body is divided into clear thematic paragraphs.
5. The conclusion clearly rounds off the synthesis.
6. The ideas are clearly linked.
7. The synthesis reads fluidly in one go.
8. The synthesis is reader-oriented: it explains what the reader does not know.
9. The language use overall is correct.
10. The language use overall is varied.
11. The linguistic style is appropriate for an academic synthesis.



RESULTS: FREE COMMENTS ON THE TWO CHAT-GPT MODELS

Strong points

Language use: correct and adequate

In terms of grammar, I would never be able to write such a perfect text containing that many conjunctive and genitive constructions

It is strange that a bot would use humanlike voice, such as "Insgesamt zeigt sich, dass" [overall, we can state that]

Content: good selection

ChatGPT did a much better job than me in selecting the main information of the two source texts

Weak points

Language use: plagiarised from sources, lack of originality

Given the topic of Kiezdeutsch *as a highly creative language variety, it is a pity that ChatGPT itself does not use creative language*

Content: invented facts

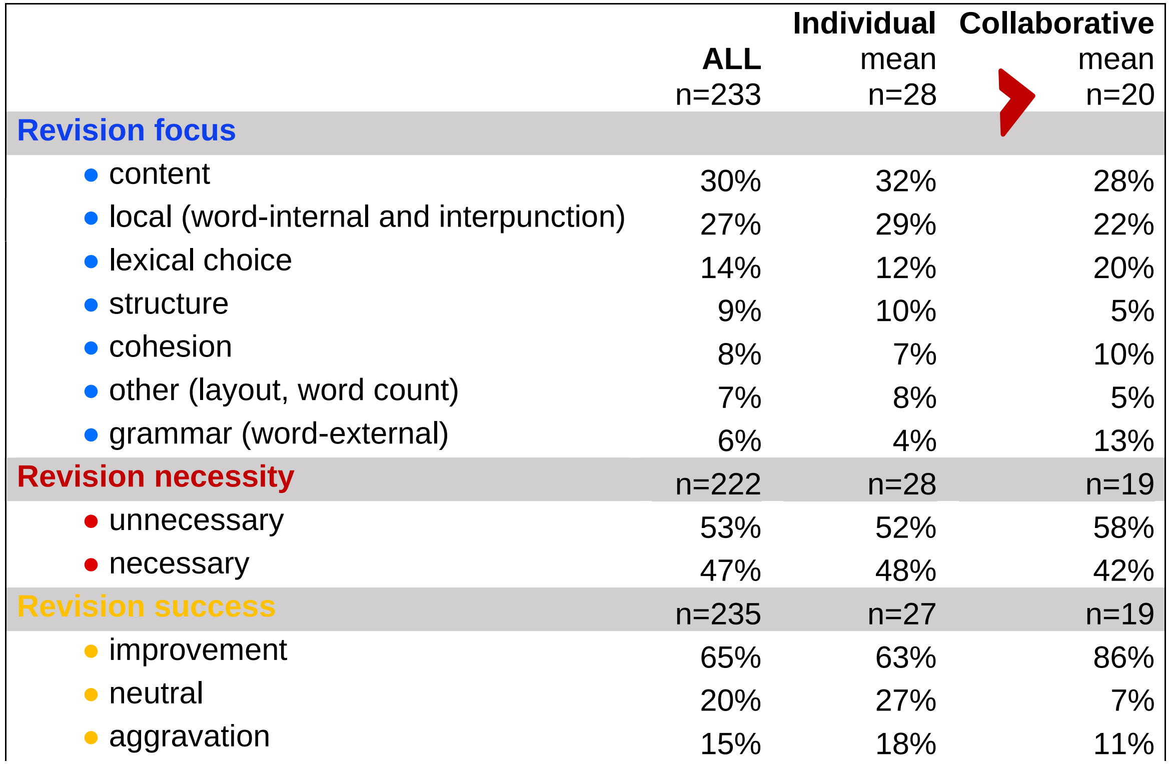
ChatGPT mentions "die Autorin", but there is no evidence of the source text being written by a female author.

CONCLUSIONS

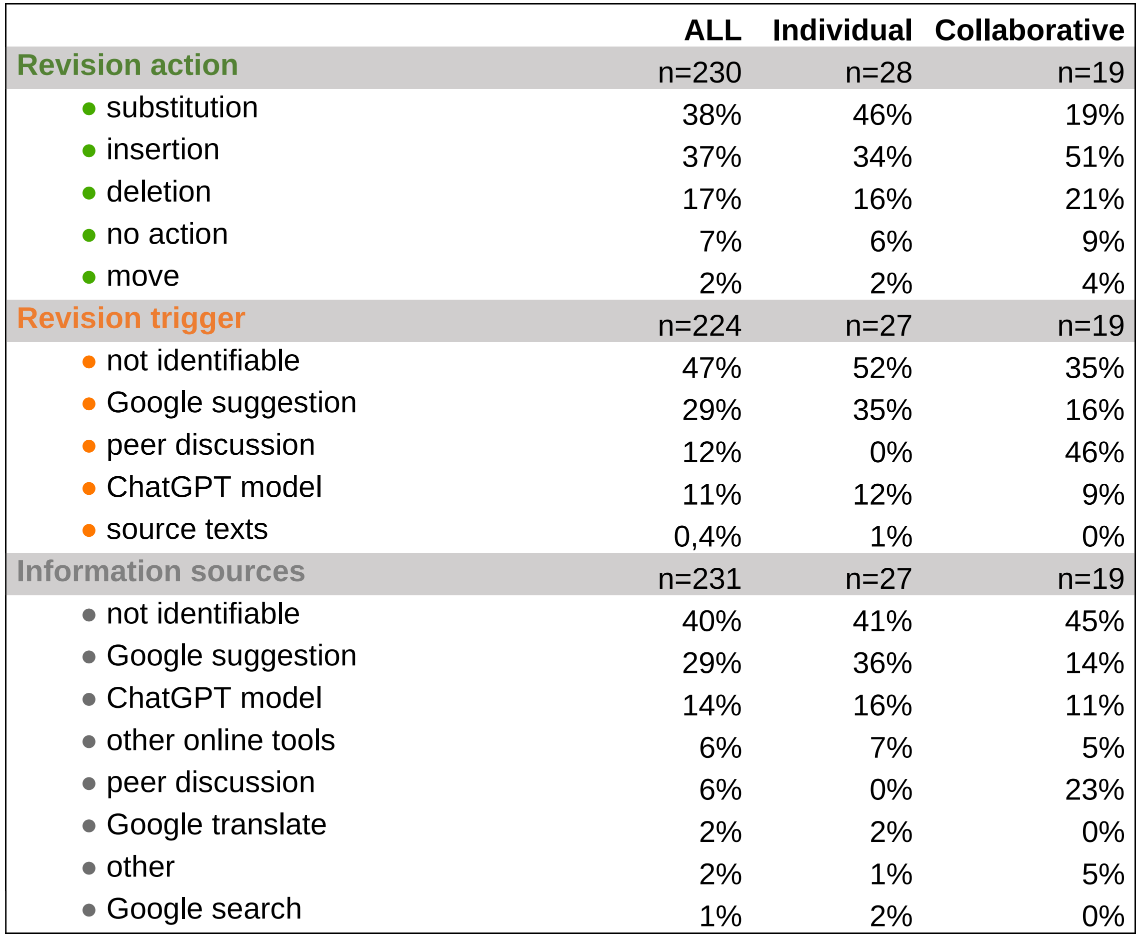
RQ1 What do students notice in their own output and in Chat-GPT output based on a guided comparison?

- Students rated their own output consistently low in terms of linguistic accuracy and appropriate writing style in comparison with ChatGPT-output.
- In terms of content, students rated ChatGPT-output high, but also noticed problems with trustworthiness of information (Ranalli, 2021: “calibrated trust”).
- Overall, students´ confidence with their own text quality compared with Chat-GPT output grew during the intervention.

RESULTS: REVISION BEHAVIOUR OF SIX FOCUS PARTICIPANTS



	ALL	Individual	Collaborative
		mean	mean
	n=233	n=28	n=20
Revision focus			
● content	30%	32%	28%
● local (word-internal and interpunction)	27%	29%	22%
● lexical choice	14%	12%	20%
● structure	9%	10%	5%
● cohesion	8%	7%	10%
● other (layout, word count)	7%	8%	5%
● grammar (word-external)	6%	4%	13%
Revision necessity	n=222	n=28	n=19
● unnecessary	53%	52%	58%
● necessary	47%	48%	42%
Revision success	n=235	n=27	n=19
● improvement	65%	63%	86%
● neutral	20%	27%	7%
● aggravation	15%	18%	11%



	ALL	Individual	Collaborative
Revision action	n=230	n=28	n=19
● substitution	38%	46%	19%
● insertion	37%	34%	51%
● deletion	17%	16%	21%
● no action	7%	6%	9%
● move	2%	2%	4%
Revision trigger	n=224	n=27	n=19
● not identifiable	47%	52%	35%
● Google suggestion	29%	35%	16%
● peer discussion	12%	0%	46%
● ChatGPT model	11%	12%	9%
● source texts	0,4%	1%	0%
Information sources	n=231	n=27	n=19
● not identifiable	40%	41%	45%
● Google suggestion	29%	36%	14%
● ChatGPT model	14%	16%	11%
● other online tools	6%	7%	5%
● peer discussion	6%	0%	23%
● Google translate	2%	2%	0%
● other	2%	1%	5%
● Google search	1%	2%	0%

RESULTS: REVISION BEHAVIOUR OF SIX FOCUS PARTICIPANTS

		Individual	Collaborative
	ALL	mean	mean
	n=233	n=28	n=20
Revision focus			
● content	30%	32%	28%
● local (word-internal and interpunction)	27%	29%	22%
● lexical choice	14%	12%	20%
● structure	9%	10%	5%
● cohesion	8%	7%	10%
● other (layout, word count)	7%	8%	5%
● grammar (word-external)	6%	4%	13%
Revision necessity	n=222	n=28	n=19
● unnecessary	53%	52%	58%
● necessary	47%	48%	42%
Revision success	n=235	n=27	n=19
● improvement	65%	63%	86%
● neutral	20%	27%	7%
● aggravation	15%	18%	11%

	ALL	Individual	Collaborative
Revision action	n=230	n=28	n=19
● substitution	38%	46%	19%
● insertion	37%	34%	51%
● deletion	17%	16%	21%
● no action	7%	6%	9%
● move	2%	2%	4%
Revision trigger	n=224	n=27	n=19
● not identifiable	47%	52%	35%
● Google suggestion	29%	35%	16%
● peer discussion	12%	0%	46%
● ChatGPT model	11%	12%	9%
● source texts	0,4%	1%	0%
Information sources	n=231	n=27	n=19
● not identifiable	40%	41%	45%
● Google suggestion	29%	36%	14%
● ChatGPT model	14%	16%	11%
● other online tools	6%	7%	5%
● peer discussion	6%	0%	23%
● Google translate	2%	2%	0%
● other	2%	1%	5%
● Google search	1%	2%	0%

RESULTS: REVISION BEHAVIOUR OF SIX FOCUS PARTICIPANTS

		Individual	Collaborative
	ALL	mean	mean
	n=233	n=28	n=20
Revision focus			
● content	30%	32%	28%
● local (word-internal and interpunction)	27%	29%	22%
● lexical choice	14%	12%	20%
● structure	9%	10%	5%
● cohesion	8%	7%	10%
● other (layout, word count)	7%	8%	5%
● grammar (word-external)	6%	4%	13%
Revision necessity	n=222	n=28	n=19
● unnecessary	53%	52%	58%
● necessary	47%	48%	42%
Revision success	n=235	n=27	n=19
● improvement	65%	63%	86%
● neutral	20%	27%	7%
● aggravation	15%	18%	11%

	ALL	Individual	Collaborative
Revision action	n=230	n=28	n=19
● substitution	38%	46%	19%
● insertion	37%	34%	51%
● deletion	17%	16%	21%
● no action	7%	6%	9%
● move	2%	2%	4%
Revision trigger	n=224	n=27	n=19
● not identifiable	47%	52%	35%
● Google suggestion	29%	35%	16%
● peer discussion	12%	0%	46%
● ChatGPT model	11%	12%	9%
● source texts	0,4%	1%	0%
Information sources	n=231	n=27	n=19
● not identifiable	40%	41%	45%
● Google suggestion	29%	36%	14%
● ChatGPT model	14%	16%	11%
● other online tools	6%	7%	5%
● peer discussion	6%	0%	23%
● Google translate	2%	2%	0%
● other	2%	1%	5%
● Google search	1%	2%	0%

RESULTS: REVISION BEHAVIOUR OF SIX FOCUS PARTICIPANTS

		Individual	Collaborative
	ALL	mean	mean
	n=233	n=28	n=20
Revision focus			
● content	30%	32%	28%
● local (word-internal and interpunction)	27%	29%	22%
● lexical choice	14%	12%	20%
● structure	9%	10%	5%
● cohesion	8%	7%	10%
● other (layout, word count)	7%	8%	5%
● grammar (word-external)	6%	4%	13%
Revision necessity	n=222	n=28	n=19
● unnecessary	53%	52%	58%
● necessary	47%	48%	42%
Revision success	n=235	n=27	n=19
● improvement	65%	63%	86%
● neutral	20%	27%	7%
● aggravation	15%	18%	11%

	ALL	Individual	Collaborative
Revision action	n=230	n=28	n=19
● substitution	38%	46%	19%
● insertion	37%	34%	51%
● deletion	17%	16%	21%
● no action	7%	6%	9%
● move	2%	2%	4%
Revision trigger	n=224	n=27	n=19
● not identifiable	47%	52%	35%
● Google suggestion	29%	35%	16%
● peer discussion	12%	0%	46%
● ChatGPT model	11%	12%	9%
● source texts	0,4%	1%	0%
Information sources	n=231	n=27	n=19
● not identifiable	40%	41%	45%
● Google suggestion	29%	36%	14%
● ChatGPT model	14%	16%	11%
● other online tools	6%	7%	5%
● peer discussion	6%	0%	23%
● Google translate	2%	2%	0%
● other	2%	1%	5%
● Google search	1%	2%	0%

RESULTS: REVISION BEHAVIOUR OF SIX FOCUS PARTICIPANTS

	ALL n=233	Individual mean n=28	Collaborative mean n=20
Revision focus			
● content	30%	32%	28%
● local (word-internal and interpunction)	27%	29%	22%
● lexical choice	14%	12%	20%
● structure	9%	10%	5%
● cohesion	8%	7%	10%
● other (layout, word count)	7%	8%	5%
● grammar (word-external)	6%	4%	13%
Revision necessity	n=222	n=28	n=19
● unnecessary	53%	52%	58%
● necessary	47%	48%	42%
Revision success	n=235	n=27	n=19
● improvement	65%	63%	86%
● neutral	20%	27%	7%
● aggravation	15%	18%	11%

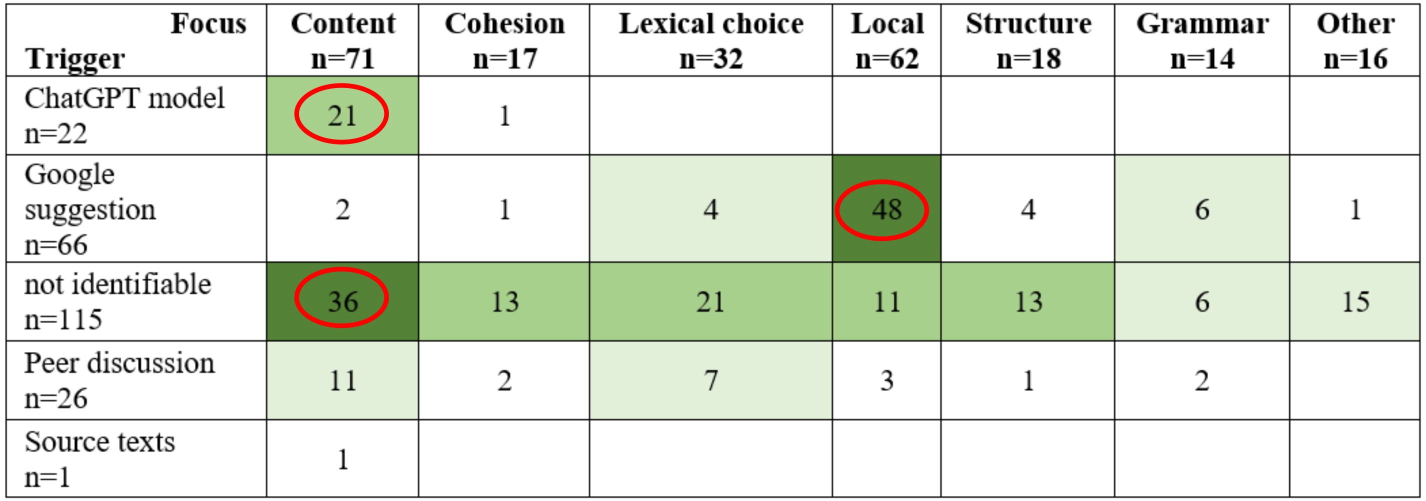
	ALL	Individual	Collaborative
Revision action	n=230	n=28	n=19
● substitution	38%	46%	19%
● insertion	37%	34%	51%
● deletion	17%	16%	21%
● no action	7%	6%	9%
● move	2%	2%	4%
Revision trigger	n=224	n=27	n=19
● not identifiable	47%	52%	35%
● Google suggestion	29%	35%	16%
● peer discussion	12%	0%	46%
● ChatGPT model	11%	12%	9%
● source texts	0,4%	1%	0%
Information sources	n=231	n=27	n=19
● not identifiable	40%	41%	45%
● Google suggestion	29%	36%	14%
● ChatGPT model	14%	16%	11%
● other online tools	6%	7%	5%
● peer discussion	6%	0%	23%
● Google translate	2%	2%	0%
● other	2%	1%	5%
● Google search	1%	2%	0%

RESULTS: REVISION BEHAVIOUR OF SIX FOCUS PARTICIPANTS

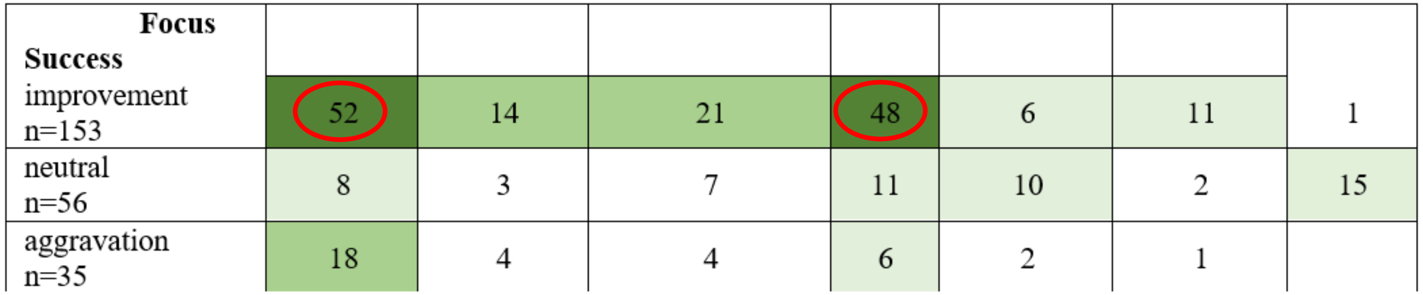
	ALL	Individual mean	Collaborative mean
	n=233	n=28	n=20
Revision focus			
● content	30%	32%	28%
● local (word-internal and interpunction)	27%	29%	22%
● lexical choice	14%	12%	20%
● structure	9%	10%	5%
● cohesion	8%	7%	10%
● other (layout, word count)	7%	8%	5%
● grammar (word-external)	6%	4%	13%
Revision necessity	n=222	n=28	n=19
● unnecessary	53%	52%	58%
● necessary	47%	48%	42%
Revision success	n=235	n=27	n=19
● improvement	65%	63%	86%
● neutral	20%	27%	7%
● aggravation	15%	18%	11%

	ALL	Individual	Collaborative
Revision action	n=230	n=28	n=19
● substitution	38%	46%	19%
● insertion	37%	34%	51%
● deletion	17%	16%	21%
● no action	7%	6%	9%
● move	2%	2%	4%
Revision trigger	n=224	n=27	n=19
● not identifiable	47%	52%	35%
● Google suggestion	29%	35%	16%
● peer discussion	12%	0%	46%
● ChatGPT model	11%	12%	9%
● source texts	0,4%	1%	0%
Information sources	n=231	n=27	n=19
● not identifiable	40%	41%	45%
● Google suggestion	29%	36%	14%
● ChatGPT model	14%	16%	11%
● other online tools	6%	7%	5%
● peer discussion	6%	0%	23%
● Google translate	2%	2%	0%
● other	2%	1%	5%
● Google search	1%	2%	0%

CO-OCCURRENCE ANALYSIS: REVISION FOCUS, TRIGGER, AND SUCCESS



Focus Trigger	Content n=71	Cohesion n=17	Lexical choice n=32	Local n=62	Structure n=18	Grammar n=14	Other n=16
ChatGPT model n=22	21	1					
Google suggestion n=66	2	1	4	48	4	6	1
not identifiable n=115	36	13	21	11	13	6	15
Peer discussion n=26	11	2	7	3	1	2	
Source texts n=1	1						



Focus Success							
improvement n=153	52	14	21	48	6	11	1
neutral n=56	8	3	7	11	10	2	15
aggravation n=35	18	4	4	6	2	1	

CONCLUSIONS

RQ1 What do students notice in their own output and in Chat-GPT output based on a guided comparison?

- Students rated their own output consistently low in terms of linguistic accuracy and appropriate writing style in comparison with ChatGPT-output.
- In terms of content, students rated ChatGPT-output high, but also noticed problems with trustworthiness of information (Ranalli 2021: “calibrated trust”).
- Overall, students´ confidence with their own text quality compared with Chat-GPT output grew during the intervention.

RQ2 What do students revise in their own texts?

- Overall, students revised more in the first (individual) session than in the second (collaborative) session.
- Revision focus is on content (frequently induced by the models) and on local issues (mostly induced by automated Google-suggestions), followed by vocabulary in the third place (⇔ previous literature on model-based revision) → Students skillfully draw on their resources for text optimalisation.
- More than half of the revisions are unnecessary (“overrevisions”), however often lead to text improvement.
- High number of unidentified revision triggers underlines the suitability of the task sequence (writing > comparing > revising) to stimulate “inner feedback” (Nicol, 2021) loops.

MY FAVOURITE QUOTE OF A COLLABORATIVE REVISION SESSION

Ann* [referring to a model]: I like this sentence. Should we just copy-paste it into our text or try to rephrase it?

Jos*: Just copy-paste it! If ChatGPT can do this, we also can.

*pseudonyms

JCDecaux
Et ça,
il sait le faire
ChatGPT ?
Restons
naturels.
LES FERMIERS DE
Loué
COOPÉRATIVE D'ÉLEVEURS
DEPUIS 1958

THANK YOU!

Carola Strobl

Assistant professor in Applied Linguistics and Translation

University of Antwerp

Department of Applied Linguistics, Translation and Interpreting

Spokesperson research group TricS – Translation, Interpreting and Intercultural Studies / https://twitter.com/TricS_research

carola.strobl@uantwerpen.be
https://www.uantwerpen.be/en/staff/carola-strobl/

University of Antwerp
TRICS | Translation, Interpreting and Intercultural Studies

Strobl et al. (forthcoming): Adopting ChatGPT as a writing buddy in the advanced L2 writing class. *Technology in Language Teaching & Learning.*

Iryna Menke-Bazhutkina
German language teacher and coordinator

Niklas Abel

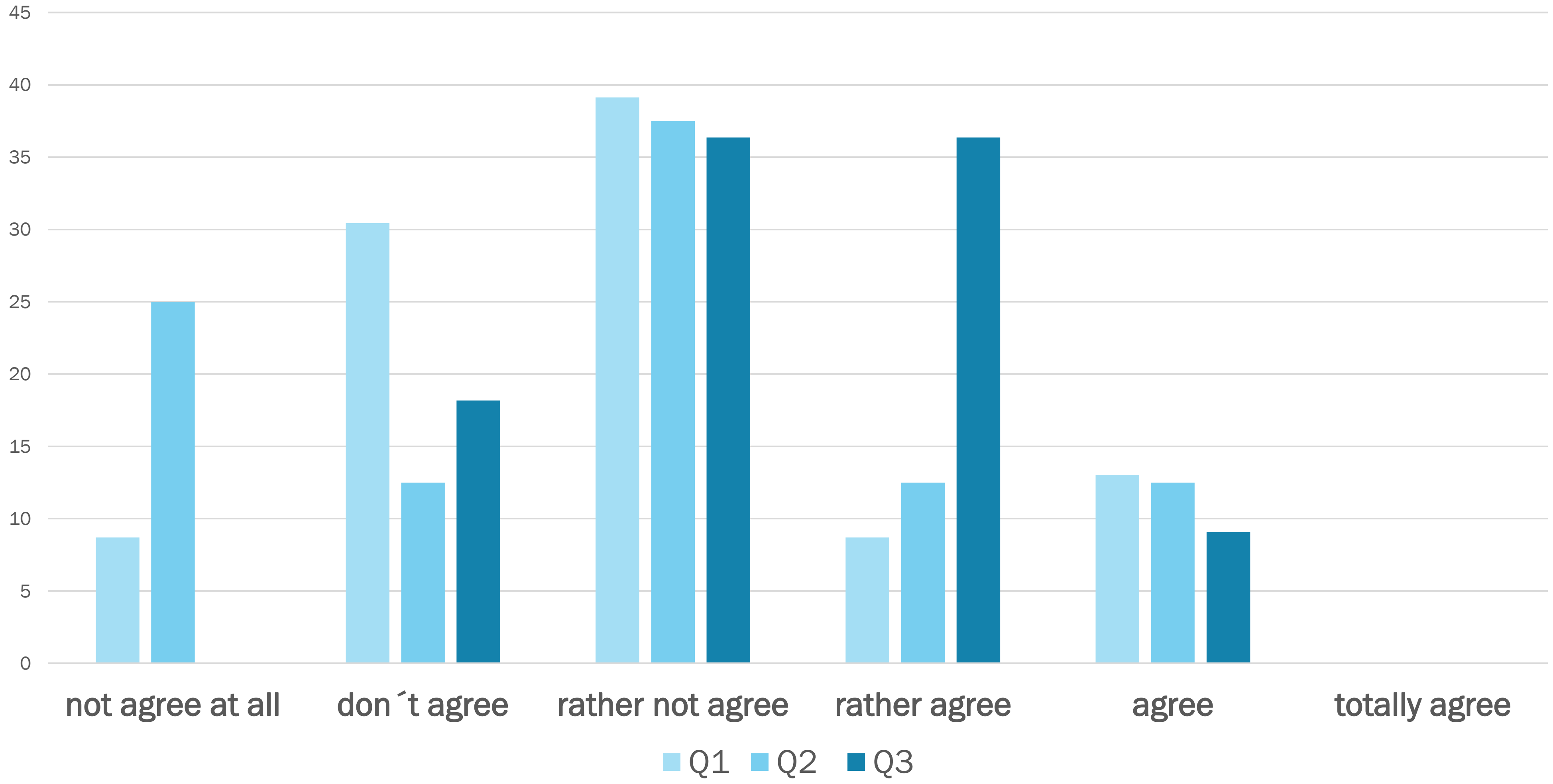
German language teacher

Marije Michel

Full Professor - Chair of Language Learning

university of groningen / faculty of arts

“Tools such as ChatGPT will make human writing redundant in the future”


not agree at all
don´t agree
rather not agree
rather agree
agree
totally agree
Q1
Q2
Q3