Computational Linguistics in the Netherlands Journal 15 (2026) 105-120 Submitted 01/2026; Published 06/2026

Dictionaries, DeepL, and ChatGPT: Comparing Language Tool Effects on L2 English Speaking and Learner Perceptions

Irma Schampheleer* IRMA.SCHAMPHELEER@UGENT.BE
Lieve Macken* LIEVE.MACKEN@UGENT.BE
Vanessa De Wilde* VANESSA.DEWILDE@UGENT.BE

Department of Translation, Interpreting, and Communication, Ghent University, Belgium

Abstract

In recent years, language learners have increasingly relied on digital tools such as machine translation (MT) and generative artificial intelligence (AI) tools for language support. While the effects of MT tools like DeepL on second language (L2) writing have been relatively well-documented, fewer studies have examined their influence on L2 speaking performance. Research on the impact of generative AI tools such as ChatGPT in this context is even more limited. This multi-method experimental study aims to assess how the use of MT and generative AI tools affects L2 English speaking performance, as well as learners' experiences preparing oral presentations. 30 university language students were divided into three groups and assigned either dictionaries, DeepL, or ChatGPT for language support when preparing a group presentation. The presentations were recorded, transcribed, and analysed for fluency and lexical complexity. Additionally, participants completed a two-part survey: one part reflecting on their assigned tool during the preparation process, and one evaluating their general experience with all three tools. Results showed that MT users outperformed other groups in fluency-related measures, while AI users demonstrated significantly greater lexical diversity and density than the MT group ($p = 0.04$). Lecturer evaluations significantly favoured presentations prepared with MT and AI tools over those prepared with dictionaries ($p = 0.02$). Additionally, participants using notes during the delivery demonstrated a significantly higher lexical diversity ($p = 0.025$). Survey responses indicated that perceived enjoyment and usefulness varied between groups, with generative AI tools generally rated more positively for idea generation but less so when it came to practising the presentation. Overall, participants favoured MT tools for preparing both oral and written tasks. These findings suggest that while MT tools currently offer the most consistent benefits for L2 speaking performance, generative AI tools hold promise, particularly for idea development and lexical variation.

1. Introduction

Machine translation (MT) and generative artificial intelligence (AI) tools have become increasingly integrated into language learning environments in recent years. DeepL and ChatGPT, among other popular tools, are now readily accessible and widely used by language learners across all proficiency levels (Klimova 2025). Their growing popularity raises important questions about how such technologies influence L2 learning processes and outcomes, particularly in productive skills like speaking.

While a substantial body of research has explored the use of MT tools for L2 writing (Chung and Ahn 2022, Kol et al. 2018, Lee 2020, Stapleton 2021, Tsai 2019), their impact on speaking performance remains less studied. Even fewer studies have experimentally investigated the role of generative AI tools such as ChatGPT in L2 speaking and writing. As these technologies evolve and become more sophisticated, it becomes increasingly relevant to evaluate not only how learners use them, but also how these tools impact learners' linguistic output and shape their learning experience.

This study [1] aims to address that gap by examining the effects of three types of digital language tools – dictionaries, MT (DeepL), and generative AI (ChatGPT) – on L2 English speaking

1. This study builds on research conducted as part of the author's master's thesis at Ghent University in 2025.

performance. To that end, we conducted a multi-method experimental study involving 30 university language students. The participants were divided into three groups and invited to prepare and deliver group presentations using one of the tools as their main source of linguistic support. Their presentations were recorded, transcribed, and analysed in terms of fluency, lexical complexity, lecturer evaluation, and contingent on the use of presentation aids. In addition, a two-part survey was administered to gather insights into the students' experiences using their assigned tools during the preparation process, as well as their general attitudes and preferences regarding all three tools.

The aim of this paper is twofold. First, it seeks to compare the linguistic outcomes of students who prepared their presentations with different tools (i.e. dictionaries, MT, and generative AI), in order to evaluate how each tool may impact spoken production in English as a second language. Second, it aims to better understand students' motivations for using these tools, the ways in which they integrated them into their preparation process, and their perceived usefulness and enjoyment. Through this combined lens of performance and perception, the study contributes to the growing field of research on technology-enhanced language learning and offers insights into the pedagogical implications of integrating MT and AI tools in communicative language tasks.

2. Related research

Research on digital tools in L2 learning has primarily focused on their impact on language production and learner perceptions. Four key dimensions – fluency, complexity, accuracy, and pronunciation – are widely recognised as benchmarks for evaluating L2 speaking performance (Skehan 1998, Housen et al. 2012). Fluency refers to the smoothness and speed of speech or writing, and is typically measured by the total number of words produced or the number of words produced per minute (Kellogg 1996). Complexity captures the range and sophistication of grammatical and lexical structures. Figure 1 visualises its taxonomy, alongside measures frequently used in empirical studies. Accuracy measures conformity to target norms, either by calculating the ratio of errors present in a text to a predefined unit of production (e.g., words, clauses, or sentential units), or by determining the proportion of these units that remain error-free (Lambert and Kormos 2014). Lastly, pronunciation is increasingly assessed through intelligibility and comprehensibility rather than native-like norms (Mairano et al. 2023). These constructs provide the foundation for examining how language tools influence L2 speaking performance.

2.1 MT tools

A substantial body of research has explored MT tools such as DeepL and Google Translate, particularly in writing contexts. Findings suggest that MT can enhance lexical sophistication and vocabulary profiles, with learners producing texts containing more advanced and academic words (Kol et al. 2018, Tsai 2019). However, its effects on syntactic complexity and accuracy are mixed: some studies report improvements in error reduction and text length (Chung and Ahn 2022), while others find negligible differences (Lee 2020). Learners generally perceive MT positively, citing convenience, time efficiency, and vocabulary support as key benefits, though concerns about grammatical errors, awkward phrasing, and overreliance persist.

In speaking contexts, research has focused on pronunciation, particularly the acquisition of English past tense *-ed* endings using MT's text-to-speech (TTS) and automatic speech recognition (ASR) features. Studies indicate improvements in phonological awareness and discrimination, though gains in oral production vary (He and Cardoso 2021, Khademi and Cardoso 2022). Learners also report that MT tools increase confidence and motivation during speaking tasks (Rushton 2022, Sakamoto 2022).

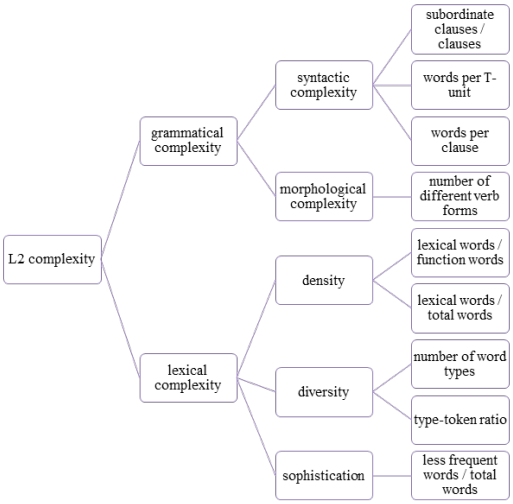

Figure 1: Simplified visualisation of the components and subcomponents of L2 complexity, inspired by Housen et al. (2012), alongside their typical measures

2.2 Generative AI tools

Compared to MT, research on generative AI tools such as ChatGPT is still emerging. Existing studies, largely situated in English as a Foreign Language (EFL) university contexts, emphasize learner perceptions rather than measurable performance outcomes (Ali et al. 2023, Huang and Mizumoto 2024, Yamaoka 2024, Tang 2025). Findings consistently highlight motivational benefits: AI tools reduce anxiety, foster autonomy, and enhance engagement, particularly in writing tasks. Learners perceive ChatGPT as helpful for idea generation, language optimisation, and drafting, though concerns about plagiarism and overreliance remain. Preliminary evidence suggests potential gains in vocabulary and grammar acquisition (Zhang and Huang 2024) and perceived improvements in speaking fluency (Laksana et al. 2024), but these claims are based on self-reports rather than controlled experiments. Overall, while generative AI shows promise as a supportive tool, empirical evidence of its impact on oral proficiency is limited.

2.3 Research Gap

Despite growing interest in digital tools for language learning, few studies have compared traditional dictionaries, MT, and generative AI in the context of L2 speaking preparation. Most research on AI tools focuses on attitudes or writing outcomes, leaving a gap in evidence-based assessments of their influence on spoken performance. The present study addresses this gap by examining how these tools affect fluency, lexical complexity, and lecturer evaluations during oral presentations, while also exploring learners' experiences and general usage patterns.

Therefore, the following research questions guide this study:

- RQ1: To what extent does the use of dictionaries, a machine translation, or generative AI tool influence L2 learners' speaking performance in terms of fluency, lexical complexity, and lecturer evaluation?

- RQ2: How do learners experience and evaluate the use of these tools when preparing for an L2 speaking task?

- RQ3: How do learners generally use and experience these tools for language support in their L2?

3. Methodology

In this experimental study, 30 L2 English students gave group presentations on a controversial topic using different language tools. These presentations were recorded and transcribed. Participants were also invited to complete a survey about their experience preparing for the presentations and their overall experience with online language tools. Both the presentations and survey responses were analysed quantitatively and qualitatively. All aspects of the methodology are discussed separately below.

3.1 Participants

This study comprised 30 participants (27 women and 3 men) between 18 and 30 years old. All participants were Dutch speakers, with some speaking additional languages at home, including Russian, Thai, Hindi, Spanish, and Arabic.

In accordance with the majority of prior studies examining the integration of language technology in education, the present study's participants were university students of EFL (Chung and Ahn 2022, Kol et al. 2018, Lee 2020, Tsai 2019, Rushton 2022, Sakamoto 2022, Yamaoka 2024, Tang 2025, Laksana et al. 2024). All participants were enrolled in the second year of the Bachelor of Arts in Applied Linguistics programme at Ghent University, specialising in English. Specifically, they were all taking the *English: Language Proficiency II* course, an EFL course focusing on writing and speaking skills. Participants were divided into three groups (hereafter called 'language tool groups') according to the day they were following the course.

3.2 Speaking task

For the first part of the study, students were asked to participate in group presentations as part of a classroom activity. Each group of three to four students selected a controversial topic from a list of 52 issues provided by their lecturer. These are three examples of the issues selected by the participants:

- Personal tech: *Does personal tech encourage asocial behaviour, and if so, what can be done to address the problem?*
- Zoos: *Zoos may be good for educational purposes but it's unnatural for animals to live in cages. Discuss.*
- Euthanasia: *An individual who is elderly or terminally ill should have the right to choose when his/her life should end. To what extent do you agree with this statement?*

Students started to prepare their presentations in class. They were allowed to continue their preparations at home. The allotted time for this task was one week. Participants were invited to document the time they had spent preparing for this task.

Each student was allocated three to four minutes of speaking time during their group presentation. The presentations consisted of three primary sections. However, the students were not instructed on how to divide the different sections among the speakers. The introduction had to include a hook to capture the audience's attention, a concise topic introduction, and a clear thesis statement. The body of the presentation then presented well-structured arguments that supported the thesis, while the conclusion summarised the key points and provided a strong closing statement.

To enhance their delivery, students were permitted to use different presentation aids. They were allowed to project slides, use the blackboard, or play audio or video materials.

3.3 Language tools

Students were divided into three groups based on the day they followed the course. Each group was allotted a different type of tool for language support during their preparation at home. They were not allowed to use the tools assigned to the other two groups. It should be noted that all students were permitted to use other (online) resources in addition to their assigned language tool to gather information about their topic and formulate their arguments. These instructions were summarised in an assignment sheet, which differed according to each language group.

The first group was instructed to employ only (online) dictionaries for language assistance. Participants in the DI-group (henceforth referred to as such) were permitted to use translation dictionaries, such as Van Dale, as well as explanatory dictionaries, such as Collins Dictionary, Oxford Dictionary, Cambridge Dictionary, and Merriam-Webster.

Students in the MT-group were requested to only use a MT tool for language support during their preparations. Specifically, they were asked to use DeepL Translate. The decision to restrict the use of MT to a single tool was made to minimise variations in the output quality of different translation tools. We opted for DeepL as opposed to Google Translate, another widely employed MT tool, because it allows users to choose from multiple phrasing options.

Finally, the AI-group was instructed to engage with a single generative AI tool, namely ChatGPT. Once more, the decision was taken to instruct the students to use a single, designated tool, thus ensuring that the quality of the AI experience would be uniform for all participants.

Table 1 provides an overview of the three language tool groups, including the number of students and group presentations, in addition to the topics selected by each group for their presentations. It should be noted that exchange students were excluded from the recordings to allow for a homogeneous set of participants in terms of L1. This is why one of the groups only has a single participant.

	DI-group	**MT-group**	**AI-group**
Number of students	9	7	14
Number of group presentations within the language tool groups	3	3	4
Topics chosen by each group, along with the number of students in that group (exchange students excluded)	1. Personal tech (3) 2. Gap years (3) 3. Zoos (3)	1. Zoos (4) 2. Euthanasia (1) 3. TikTok (2)	1. Obesity (4) 2. Beggars (3) 3. Zoos (4) 4. Abortion (3)

Table 1: Overview of the numbers of students in each language tool group, and the topics they selected for their group presentations

3.4 Recordings and transcriptions

During their presentations, participants were invited to wear a lavalier microphone. This initially caused some unease and fiddling among the participants, but was considered to be the least intrusive way of recording their speech. To distinguish participants within groups, the first and last sentences of each group member were written down.

Subsequently, the presentations were automatically transcribed using Whisper. Afterwards, they were manually corrected by the researcher. For the purpose of the current study, we focused on speed fluency, and therefore, pauses were not included in the transcriptions. Nevertheless, disfluencies were included to ensure an accurate representation of the number of words and, consequently, the speech rate. For instance, when one of the participants said the following: "[...] but it can also... it can also give the student a new interest", this was not simply transcribed as "but it can also give the

student a new interest", as that would distort the number of words and therefore also the speech rate. Smaller disfluencies, such as 'often ... often' and 'and ... and', were also manually included in the transcriptions.

3.5 Survey

After their presentations, the participants were sent a link to an online survey. Most participants completed the questionnaire on the day of the recordings, with a smaller number doing so up to a week later. The objective of this Qualtrics survey was to ascertain two aspects regarding the participants' experiences with (online) language tools. Therefore, the survey was divided into two sections. The survey questions are based on the surveys conducted in previous studies (Chung and Ahn 2022, Niño 2020).

The first section of the survey was designed to assess the participants' experience of using their designated tool (i.e. dictionaries, DeepL, or ChatGPT) during the preparation of their presentation assignments. Consequently, the questions in this section differed for each language tool group. In the second section of the survey, students were invited to share their overall experience with online dictionaries, MT tools, and generative AI tools. All students therefore filled in the same questions.

Finally, the survey incorporated a series of demographic questions. The objective of this last section was to ascertain whether there were considerable differences in age, gender, or linguistic background.

3.6 Analysis

Given that these group presentations were not only the subject of this research paper but also primarily a classroom activity, the students were provided with oral feedback immediately after their presentations. In addition, the lecturer provided a global performance evaluation on a scale of 1 to 20. It is important to note that these marks were not disclosed to the students, nor did they have an impact on their final mark for the course – the marks were exclusively shared with the researcher. It is noteworthy that the lecturer had known the students for a minimum of eight weeks before the presentations. Consequently, the marks may be perceived as slightly biased, though the lecturer has extensive experience in grading students' speaking skills.

The analysis of both the presentation data and the survey results were conducted using Microsoft Excel. Descriptive statistics were calculated to summarise speaking performance in terms of fluency (presentation duration, word count, and speech rate) and lexical complexity. Lexical complexity was operationalised through measures of lexical density, lexical diversity (type-token ratio), and lexical sophistication. This last category refers to the proportions of words belonging to four distinct frequency levels: the list of the 1000 most frequent word families (K1), the second 1000 most frequent word families (K2), the Academic Word List (AWL), and words that do not appear in the other lists (off-list). These lexical measures were computed using LexTutor (www.lextutor.ca/vp/eng/), following the approach of earlier studies (Tsai 2019, Kol et al. 2018). Due to time constraints and because all participants were intermediate L2 English speakers, no analyses were performed on grammatical accuracy or complexity.

To identify statistically significant differences between the three language tool groups, Kruskal-Wallis tests were carried out for each quantitative variable. Pivot tables in Excel facilitated the organisation and comparison of descriptive data, and the results were visualised using bar charts and box plots to clearly illustrate group differences.

In addition to the quantitative analysis, a limited qualitative analysis was performed on the open-ended survey responses. Both the prompts used by the AI-participants and the advantages and disadvantages of the assigned language tools, as described by the participants, were categorised into thematic groups.

4. Results

In this section, we will first discuss the participants' speaking performance results (4.1) in terms of fluency, lecturer evaluation, lexical complexity, and presentation aid use. Then, we will report on the survey findings. Section 4.2 reports on the students' preparation process with their assigned tool and their reflections on their specific tool use. In 4.3, we will discuss their general experience with all three tools.

4.1 Speaking performance

4.1.1 Fluency

All participants were instructed to speak for a duration of three to four minutes during their group presentations. They were not encouraged to learn a script of a particular number of words by heart; however, the transcriptions allowed for the calculation of the word count of each participant's speaking part. Consequently, the speech rate could also be calculated and subsequently compared across participants and language tool groups.

The mean duration of speech by each participant was 3 minutes and 45 seconds, with an average word count of 469.8. This equates to a speech rate of 124.25 words per minute. This rate is notably lower than the average speaking speed in British English, which is reported to be 198 words per minute (Wang 2021).

The findings per language tool group indicate that participants who had used MT for their presentations exhibited the highest word count, duration, and speech rate. In contrast, the DI-group demonstrated the lowest word count, the shortest duration, and the slowest average speech rate.

The box plot below (Figure 2) visualises the speech rate results per group. However, a Kruskal-Wallis test showed that the differences in word count, duration, and speech rate across the three groups were not significant ($p = 0.250$, $p = 0.671$, and $p = 0.724$, respectively).

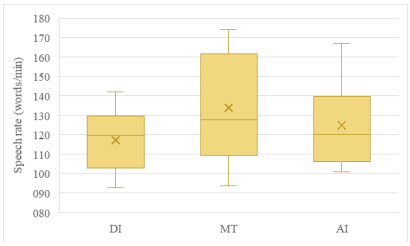

Figure 2: Speech rate (words/min) dispersion per language tool group

4.1.2 Lecturer evaluation

The mean mark awarded by the students' lecturer was 14.33 out of 20. The MT-group demonstrated the highest average score of 15.29. Students who had prepared their presentations with ChatGPT received an average score of 14.71. The DI-group demonstrated the poorest performance according to the lecturer, with an average mark of 13.00 out of 20. A Kruskal-Wallis test revealed that these results are statistically significant ($p = 0.020$). Pairwise comparisons of these results indicated that the differences in the evaluation are significant between the DI- and AI-group ($p = 0.005$), and between the DI- and MT-group ($p = 0.009$).

4.1.3 Lexical complexity

The transcriptions of the presentations were subsequently processed through the VocabProfiler by Lextutor to allow for a comparison of their lexical density and diversity (type-token ratio). While the average type-token ratio for the MT and DI groups is equivalent (0.41), that for the AI group is slightly higher at 0.45. The Kruskal-Wallis test suggested no statistically significant difference between the groups ($p = 0.179$).

Furthermore, the AI-group demonstrated the highest lexical density score (0.49). The average lexical density of the MT-group was 0.48, and the DI-group presentations were the least lexically dense with a score of 0.46. However, the results regarding lexical density were not statistically significant ($p = 0.336$).

A third component of lexical complexity according to Housen et al. (2012) is lexical sophistication. The results demonstrate that the presentations by the DI-group exhibit the highest average proportion of K1 words. In contrast, the MT-group and AI-group presentations contain a lower proportion of K1 words, which is counterbalanced with higher proportions of the other categories. No significant differences across all three groups were found according to the Kruskal-Wallis test ($p = 0.088$). Additionally, the findings indicate that the MT-group demonstrates the highest proportion of off-list words, while the AI-group exhibits the highest percentage of academic vocabulary, though these results were not found to be significant either ($p = 0.069$ and $p = 0.090$, respectively). Averages of all aforementioned results can be found in Table 2.

Language tool	**Word count**	**Duration**	**Speech rate**	**Mark**	**Type-token ratio**	**Lexical density**	**K1 Words (%)**	**K2 words (%)**	**AWL words** (%)	**Off-List words (%)**
DI	413.44	03:34	117.26	13.00	0.41	0.46	86.14	4.64	3.36	5.85
MT	562.43	04:04	133.92	15.29	0.41	0.48	80.65	4.59	4.31	10.45
AI	459.71	03:43	124.81	14.71	0.45	0.49	80.79	5.01	6.11	8.09
Total	**469.80**	**03:45**	**124.67**	**14.33**	**0.43**	**0.48**	**82.37**	**4.80**	**4.86**	**7.97**
p-value	0.250	0.671	0.724	0.020	0.179	0.336	0.088	0.542	0.090	0.069

Table 2: Average presentation marks, fluency, and lexical complexity depending on language tool group

4.1.4 Presentation aid use

As mentioned in section 3.2, participants were allowed to use different presentation aids. Participants could opt for any of the following options: projecting slides, utilising the blackboard, or playing audio or video materials. However, the only aid that was chosen were slides.

In total, 21 out of the 30 participants (or 6 out of the 10 groups) made use of slides during their presentations. In each of the three language tool groups, there was one group presentation without slides. Table 3 shows the average results of the parameters discussed in sections 4.1 to 4.3, depending on the use of slides. Participants who employed slides spoke for an average of 7 seconds longer. Furthermore, their use of academic vocabulary was marginally lower, whilst their use of off-list words was somewhat higher. However, a Mann Whitney U-test identified no significant disparities among these variables when comparing participants who employed slides with those who did not (see last row of Table 3).

During the course of the presentations, the researcher observed that some participants were consulting notes. The subjects had not received any instructions as to whether this was permitted, but the question arises as to whether this has an effect on their delivery. Table 4 shows the average

Use of slides	**Duration**	**Word count**	**Speech rate**	**Mark**	**Type-token ratio**	**Lexical density**	**K1 words (%)**	**K2 words (%)**	**AWL words (%)**	**Off-List words (%)**
No	03:40	474.11	126.25	14.67	0.44	0.48	82.04	4.96	5.59	7.40
Yes	03:47	467.95	124.00	14.19	0.43	0.48	82.51	4.73	4.55	8.21
Total	**03:45**	**469.80**	**124.67**	**14.33**	**0.43**	**0.48**	**82.37**	**4.80**	**4.86**	**7.97**
p-value	0.904	0.904	0.973	0.657	0.940	0.973	0.994	0.656	0.597	0.551

Table 3: Average presentation fluency, marks, and lexical complexity depending on the use of slides

results regarding fluency, evaluation, and lexical complexity, contingent on the use of notes. The findings indicate that participants who presented without the use of notes spoke for an average of 19 seconds longer, and at an average speech rate of 14 words per minute faster than the participants who did use notes. In contrast, participants using notes during their presentation demonstrated an elevated type-token ratio (0.46 vs. 0.41, $p = 0.025$) and lexical density (0.49 vs. 0.47). The vocabulary profiles (i.e. the proportion of K1, K2, AWL, and off-list words) also present slightly more pronounced disparities than participants who did and did not use slides during their presentations. However, these differences were not statistically significant.

Use of notes	**Duration**	**Word count**	**Speech rate**	**Mark**	**Type-token ratio**	**Lexical density**	**K1 words (%)**	**K2 words (%)**	**AWL words (%)**	**Off-List words (%)**
No	03:53	502.47	130.71	14.47	0.41	0.47	83.20	4.65	4.08	8.07
Yes	03:34	427.08	116.77	14.15	0.46	0.49	81.28	5.00	5.88	7.83
Total	**03:45**	**469.80**	**124.67**	**14.33**	**0.43**	**0.48**	**82.37**	**4.80**	**4.86**	**7.97**
p-value	0.791	0.255	0.134	0.934	0.025	0.655	0.678	0.386	0.107	0.847

Table 4: Average presentation fluency, marks, and lexical complexity depending on the use of notes

4.2 Preparation process and reflections on tool use

In this section, we will discuss the first part of the survey results, in which participants reflected on their experience of preparing presentations using their assigned tools. We will compare how much time they spent preparing, how they engaged with their tool, and their attitude towards it.

4.2.1 Preparation time

Participants were asked to approximately measure, in minutes, the amount of time they allocated to the preparation of their presentations at home. The mean duration for this activity was 116.67 minutes. In contrast, students who were instructed to use dictionaries required an average of 85 minutes. The MT-group spent an average of 90 minutes on the task. The AI-group allocated the most time to this activity, with an average of 150.36 minutes. The differences in preparation time were not shown to be significant ($p = 0.680$). The box plot (Figure 3) below shows the dispersion of the results per language tool group.

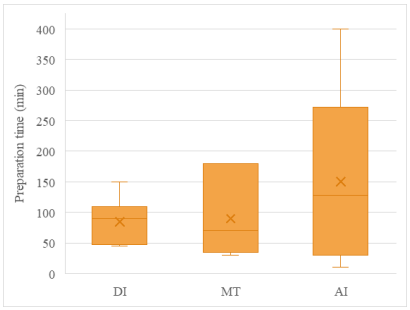

Figure 3: Preparation time (min) dispersion per language tool group

4.2.2 Language tool use

After indicating how much time they had spent preparing their presentations, participants were asked how they had used the (type of) tool they had been assigned. Naturally, questions differed slightly for each language tool group.

Participants in the DI-group mostly relied on translation dictionaries, such as Van Dale and mijnwoordenboek.nl. However, explanatory dictionaries were also used by about half of the group. Most participants used more than one dictionary, often combining a translation dictionary with an explanatory one. While most of them used dictionaries to look up words (7 out of 9 participants), a third also consulted them to look up phrases or for pronunciation purposes.

The MT-group primarily used DeepL for vocabulary lookup, with 6 out of 7 participants reporting this as their main use, while fewer used it for phrases, clauses, or sentences, and none for paragraphs. DeepL was considered most beneficial during the composition stage of their group presentations. Participants mainly valued the tool to look up vocabulary, but also used it to verify pronunciation, spelling, and grammar, though these functions were less common. Most accepted DeepL's output 61–80% of the time, rejecting it mainly when the contextual meaning differed from their intent, and occasionally due to grammatical errors or uncertainty. One participant highlighted DeepL's unique feature of offering alternative translations.

Students in the AI-group primarily used ChatGPT for three purposes: linguistic support, content development, and presentation preparation, with prompts ranging from translation queries to generating full drafts. Table 5 presents a few examples. The tool was most valued for idea generation (13 out of 14 participants), followed by answering questions and drafting presentation parts. While nearly two-thirds found it most helpful during brainstorming, five participants highlighted its usefulness when writing, and none during practice. Acceptance rates were high, with most students accepting ChatGPT's output 61–80% of the time and six accepting it even more frequently. Rejections were mainly due to mismatched contextual meaning, uncertainty, overly formal tone, or concerns about credibility. Compared to the MT-group, which relied on DeepL mainly during writing, the AI-group emphasized idea generation as an advantage ($p = 0.025$), while acceptance rates between groups were similar overall.

4.2.3 Attitudes towards assigned tool

At the end of the first part of the survey, students rated their enjoyment and perceived usefulness of their assigned tools (dictionaries, DeepL, or ChatGPT) for this assignment on a five-star scale.

Both the DI-group and AI-group reported high enjoyment, with an average of four stars, while the MT-group rated DeepL slightly lower at 3.57 stars; however, these differences were not statistically significant ($p = 0.599$).

In terms of usefulness, the DI-group and AI-group again scored similarly high (4.33 and 4.36 stars), with all participants in these groups giving four or five stars. In contrast, the MT-group

Linguistic support	Content development	Presentation support
"How do you translate 'kant-en-klaar'?" "Can you say 'what's more is that ...'? Or does it sound clunky? " "Is grocery store British English?"	"Can you give me some examples of wildlife sanctuaries around the world?" "Could you give me some examples of physical consequences of being denied an abortion?"	"Make a presentation about 20 minutes long about why or why not should we give money to beggars on the street."

Table 5: Examples of prompts by participants in the AI-group divided into three categories

again rated DeepL less favorably, with scores ranging from one to four stars and a mean of 3.29, a difference that proved statistically significant ($p = 0.039$).

4.3 General experience with language tools

The second part of the survey was designed to gauge all participants' general experience with dictionaries, MT, and generative AI tools. This section will address how the students in this study usually employ the three types of tools and what their attitudes towards their use are.

4.3.1 Language tool use

All participants reported prior experience with dictionaries, most commonly Van Dale and mijnwoordenboek.nl, while Cambridge and Oxford were the most frequently used monolingual dictionaries. A few students mentioned other resources such as Longman, Thesaurus, and Reverso Context, though the latter indicates some confusion about what constitutes a dictionary. Dictionaries were primarily used to look up words, followed by expressions and phrases. About one-third of participants consulted them for pronunciation, and a small number used them to verify spelling or find synonyms and antonyms.

Every participant had used MT tools before, with DeepL being the most popular, followed by Google Translate and Reverso Translate. These tools were predominantly employed for written production, and to a lesser extent for reading comprehension and oral production. The main purposes included retrieving vocabulary and expressions, expanding vocabulary, and verifying spelling, while some students used them to check contextual appropriateness or post-edit translations. Convenience and time-saving were the primary motivations for using MT tools, along with reviewing their own work. A smaller group cited lack of confidence in their second language or the independence these tools provide.

Nearly all participants had prior experience with generative AI tools, most commonly ChatGPT, while Copilot and QuillBot were rarely used. These tools were primarily applied to written production and, to a lesser extent, reading comprehension and oral tasks. Students mainly used AI to gather ideas, answer questions, generate drafts, and receive feedback, with occasional use for summarizing or finding relevant articles. The dominant motivation for using AI was its convenience and time-saving nature, followed by its ability to help review work. A few participants mentioned using AI for clarifying topics, providing structure, and explaining grammar in foreign languages.

4.3.2 Attitudes towards language tools

After examining how students typically use the three tools for language tasks, this section explores their attitudes towards them. It considers perceived enjoyment, usefulness, and overall preferences for speaking and writing tasks.

Participants generally enjoyed using all three types of language tools, with MT tools rated highest (average 4.33 out of 5), followed by generative AI tools (3.97) and dictionaries (3.90). While overall differences were not statistically significant, MT tools were perceived as significantly more enjoyable than dictionaries ($p = 0.033$). Enjoyment ratings varied slightly across groups: the AI-group rated generative AI tools most positively (4.36), while the DI-group gave them the lowest score (3.44). Dictionaries received their highest rating from the MT-group (4.29) and lowest from the AI-group (3.64).

Dictionaries were considered the most useful overall (average 4.43), followed by MT tools (4.13) and generative AI tools (3.97). Although all tools were rated positively, generative AI tools were found significantly less useful than dictionaries ($p = 0.035$). Across groups, usefulness ratings for dictionaries and MT tools were consistently high, while generative AI tools showed greater variability, particularly among the DI-group.

For speaking tasks, generative AI tools were most frequently ranked first (14 participants), followed by MT tools (10) and dictionaries (6). However, when weighted averages were calculated (where the most preferred tool and is given a weight of 3, and the least preferred a score of 1), MT tools emerged as the overall preferred option (2.17), indicating they were a consistent middle-ground choice, while generative AI tools were more polarising. For writing tasks, MT tools were clearly preferred, receiving 16 first-place votes and the highest weighted average (2.47). Dictionaries ranked second overall (1.93), and generative AI tools were least preferred (1.60) on average, while they received some first-place rankings.

5. Discussion

This study aimed to investigate the impact of different language tools – dictionaries, MT, and generative AI – on L2 English learners' oral task performance and their perceptions of them. By combining performance-based measures with survey data, we aimed to gain insight into both the linguistic outcomes of tool use and learners' attitudes towards these technologies. The following section discusses the answers derived from the study's findings.

Speaking performance

The analysis of the presentation recordings and transcriptions revealed that the use of MT tools resulted in the highest levels of fluency, with the MT-group outperforming both the AI- and DI-groups in terms of total word count ($p = 0.250$), speech rate ($p = 0.724$), and presentation duration ($p = 0.671$). Generative AI tools followed closely, while the dictionary group consistently produced the least fluent presentations. It should be noted that the average speech rate of all presentations was notably lower than the average in British English (Wang 2021).

In terms of lexical complexity, generative AI users produced presentations with the highest lexical diversity and density, with statistically significant differences compared to the MT-group for type-token ratio ($p = 0.04$). In contrast, the dictionary group produced language with a higher proportion of high-frequency (K1) words ($p = 0.088$), while MT users used significantly more off-list words than the dictionary group ($p = 0.008$). This is in line with findings from previous studies (Kol et al. 2018, Tsai 2019). However, it could be due to the fact that, proportionally, there were more members of the MT-group discussing the theme of zoos and therefore using names of zoos both in Belgium and abroad, that are considered off-list words. ChatGPT users demonstrated greater use of AWL terms ($p = 0.090$), indicating that their presentations contained a more academic register. Notably, the use of notes during presentations negatively affected fluency ($p = 0.134$) while slightly increasing lexical density ($p = 0.655$) and diversity ($p = 0.025$). This pattern may indicate that students who relied on notes had prepared a more scripted, vocabulary-rich text in advance, but consulting these notes during delivery disrupted their flow and slowed their speech.

Lecturer evaluations reflected the fluency trend, with both the MT- and AI-groups receiving significantly higher scores than the DI-group ($p = 0.020$). However, the lecturer was aware of which students had used which tools, which may have affected the scores.

Preparation process and reflections on tool use

Survey data from the first part of the questionnaire showed variation in preparation processes across the three groups. ChatGPT users reported the longest preparation times ($p = 0.680$), though there was considerable dispersion within this group. In contrast to the more language-focused use of dictionaries and DeepL, the AI-group made effective use of ChatGPT's wider functionalities, employing it not only for linguistic support but also for idea generation, answering content-related questions, and drafting their presentation parts. This may explain the extensive overall preparation times in this language tool group: while ChatGPT may make certain tasks easier, it also encourages learners to explore more options and ask follow-up questions. The large variation in preparation times may be due to differences in prior experience of using the tool to prepare a presentation in an L2. It would have been insightful to have surveyed the students about this before the task.

Acceptance rates for output were similar across the MT- and AI-groups, with most participants accepting between 61 and 80% of the generated content. When rejected, it was often because the output did not align with the intended contextual meaning.

In terms of perceived enjoyment and usefulness, dictionaries and ChatGPT received higher average ratings than DeepL. The difference in perceived usefulness was statistically significant ($p = 0.039$), with the MT-group rating their tool lower than the other two. The MT-group's moderate satisfaction ratings correspond with findings from previous research (Rushton 2022, Sakamoto 2022). In contrast, the higher satisfaction reported by the DI- and AI-groups is not directly supported by earlier studies.

General experience with language tools

The second part of the survey provided broader insight into participants' general experience with all three language tools. Across all tools, written production was the primary skill for which these technologies were reported to be applied. MT tools, such as DeepL and Google Translate, were predominantly used to look up vocabulary and expressions, which is in line with previous research regarding MT-supported writing (Kol et al. 2018, Lee 2020, Tsai 2019, Chung and Ahn 2022). Furthermore, they were valued for their time-saving properties, convenience (as in e.g. Chung and Ahn), and ability to review one's work. Generative AI tools, especially ChatGPT, were reported to be commonly used to generate ideas for argumentation and to answer questions, again appreciated for their efficiency and ability to review work.

Attitudinal measures showed that MT tools were rated as significantly more enjoyable to use than dictionaries ($p = 0.033$), with generative AI tools receiving slightly more mixed reactions, particularly among DI- and MT-group members. In terms of perceived usefulness, dictionaries and MT tools were rated positively overall, though generative AI tools were evaluated less favourably ($p = 0.091$), especially by participants in the DI-group ($p = 0.035$). These positive MT ratings are consistent with prior research (Chung and Ahn 2022, Tsai 2019, Rushton 2022, Sakamoto 2022), though no studies to date have compared learners' attitudes toward MT, dictionaries, and generative AI within a single study.

Lastly, participants were asked to rank the three tools in terms of their preference for preparing speaking and writing tasks. For speaking assignments, generative AI tools were most frequently ranked first, followed by MT tools and dictionaries. However, when considering the weighted averages of these rankings, a different pattern emerged: MT tools were most preferred, followed by generative AI and dictionaries. This suggests that while AI tools attracted some highly positive rankings, MT tools were more consistently preferred overall. One possible explanation is that students were mindful of both the strengths and potential drawbacks of generative AI, whereas they are more familiar and confident using MT tools, which have been available for longer. However, additional research into learner attitudes is required to explore this further.

For writing assignments, MT tools were the clear preference, ranked first by a majority of participants and achieving the highest weighted average score. Dictionaries were the second choice, while generative AI tools were least preferred for writing tasks. This pattern highlights a divide be-

tween tool preferences for productive language tasks, with MT tools occupying a consistently strong position across both modalities.

At the same time, this study faced several limitations. The small, unevenly distributed yet authentic sample, short-term nature of the task, variety of presentation topics and the artificiality introduced by the recording setting (e.g., wearing microphones, researcher presence) may have influenced the results. Furthermore, the language proficiency of the participants was not tested in advance, while it is likely to have influenced their speaking performance. Additionally, the division of the group was partly based on the other L2 that the students were studying at university, which may be a confounding factor. Furthermore, language tool use was measured through a survey, i.e. through self-report. Consequently, the researchers had no control over nor insight into the actual use of the assigned tool. Finally, the lecturer evaluations were global rather than categorised by aspects such as fluency, pronunciation, or argument structure, and the analyses did not include measures of grammatical complexity or accuracy. The limitations described here can be attributed primarily to the time constraints and contextual limitations inherent to the nature of a master's thesis.

However, these limitations can be addressed and improved upon in future research: it should build on these findings by conducting longitudinal studies that examine the sustained impact of MT and AI tool use on both writing and speaking performance across different proficiency levels. Incorporating more detailed evaluation criteria, pre- and post-test language proficiency evaluations and including learner reflections over a longer period could yield deeper insights into how digital tools shape L2 learning experiences and outcomes.

6. Conclusion

This explorative study set out to research the impact of different language support tools – dictionaries, MT, and generative AI – on L2 speaking performance, as well as learners' experiences and preferences regarding these tools in an authentic classroom context. The following section revisits the research questions

RQ1: ***To what extent does the use of dictionaries, a machine translation, or generative AI tool influence L2 learners' speaking performance in terms of fluency, lexical complexity, and lecturer evaluation?*** The findings showed that both MT and AI-supported groups outperformed the dictionary group on several speaking performance measures. DeepL users displayed the highest fluency (in terms of duration, word count, and speech rate), while ChatGPT users demonstrated the most complex vocabulary, as evidenced by a higher type-token ratio and greater use of academic vocabulary. Lecturer evaluations also favoured MT and AI-groups over the dictionary group, though it would have been insightful to break these evaluations down by subcategories such as pronunciation, accuracy, content quality, structure, etc.

RQ2: ***How do learners experience and evaluate the use of these tools when preparing for an L2 speaking task?*** The survey results revealed clear patterns in tool use and learner attitudes. MT was primarily valued for vocabulary and expression lookups during the writing stage, while AI was particularly helpful for idea generation and answering questions. Interestingly, AI users reported the longest preparation times, though this group also exhibited the most variation in individual preparation durations. In terms of learner attitudes, both the dictionary and ChatGPT groups rated their assigned tool as more enjoyable and useful than the MT group, suggesting that enjoyment and perceived usefulness may not align directly with performance outcomes.

RQ3: ***How do learners generally use and experience these tools for language support in their L2?*** Students reported primarily using digital language tools for written production. MT tools are valued to look up vocabulary and expressions, while AI tools were reported to be useful to generate ideas for argumentation and to answer questions. Both were appreciated for their time-saving properties and option to review work. Across the board, participants preferred MT tools for writing tasks and speaking preparation. However, they expressed slightly more mixed feelings about generative AI tools, particularly participants from the dictionary and MT groups.

These findings point to valuable pedagogical implications. A blended approach to tool use, tailored to task type (speaking versus writing) and learner preference, may enhance language learning outcomes. Additionally, explicit training in tool literacy – including how to critically evaluate and appropriately integrate digital tools into language production tasks – would empower learners to make informed decisions and avoid potential pitfalls such as overreliance.

References

Ali, Jamal Kaid Mohammed, Muayad Abdulhalim Ahmad Shamsan, Taha Ahmed Hezam, and Ahmed A. Q. Mohammed (2023), Impact of ChatGPT on Learning Motivation: Teachers and Students' Voices, *Journal of English Studies in Arabia Felix* **2** (1), pp. 41–49. https://journals.arafa.org/index.php/jesaf/article/view/51.

Chung, Eun Seon and Soojin Ahn (2022), The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres, *Computer Assisted Language Learning* **35** (9), pp. 2239–2264. https://www.tandfonline.com/doi/full/10.1080/09588221.2020.1871029.

He, Yue and Walcir Cardoso (2021), Can online translators and their speech capabilities help English learners improve their pronunciation?, *in* Zoghlami, Naouel, Cédric Brudermann, Cédric Sarré, Muriel Grosbois, Linda Bradley, and Sylvie Thouësny, editors, *CALL and professionalisation: short papers from EUROCALL 2021*, Research-publishing.net, pp. 126–131. Google-Books-ID: enhUEAAAQBAJ. https://doi.org/10.14705/rpnet.2021.54.1320.

Housen, Alex, Folkert Kuiken, and Ineke Vedder, editors (2012), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, Vol. 32 of *Language Learning & Language Teaching*, John Benjamins Publishing Company, Amsterdam. http://www.jbe-platform.com/content/books/9789027273260.

Huang, Jerry and Atsushi Mizumoto (2024), Examining the effect of generative AI on students' motivation and writing self-efficacy, *Digital Applied Linguistics* **1**, pp. 102324. https://www.castledown.com/journals/dal/article/view/dal.v1.102324.

Kellogg, Ronald T. (1996), A Model of Working Memory in Writing, *in* Levy, C. Michael and Sarah Ransdell, editors, *The science of writing: theories, methods, individual differences, and application*, Routledge, New York, pp. 57–71.

Khademi, Hamidreza and Walcir Cardoso (2022), Learning L2 pronunciation with Google Translate, *in* Arnbjörnsdóttir, Birna, Branislav Bédi, Linda Bradley, Kolbrún Fririksdóttir, Hólmfríur Gararsdóttir, Sylvie Thouësny, and Matthew James Whelpton, editors, *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022*, Research-publishing.net, pp. 228–233. https://research-publishing.net/manuscript?10.14705/rpnet.2022.61.1463.

Klimova, Blanka (2025), Use of machine translation in foreign language education, *Cogent Arts Humanit.*, Informa UK Limited.

Kol, Sara, Miriam Schcolnik, and Elana Spector-Cohen (2018), Google Translate in Academic Writing Courses?, *The EuroCALL Review* **26** (2), pp. 50. https://polipapers.upv.es/index.php/eurocall/article/view/10140.

Laksana, I Putu Yoga, Putu Dyah Hudiananingsih, Ketut Suciani, Putu Tika Virginiya, and Luh Gede Eka Wahyuni (2024), Exploring the Possible Use of Generative Artificial Intelligence in Supporting Students' Speaking Performance.

Lambert, Craig and Judit Kormos (2014), Complexity, Accuracy, and Fluency in Task-based L2 Research: Toward More Developmentally Based Measures of Second Language Acquisition, *Applied Linguistics* **35** (5), pp. 607–614. https://academic.oup.com/applij/article-lookup/doi/10.1093/applin/amu047.

Lee, Sangmin-Michelle (2020), The impact of using machine translation on EFL students' writing, *Computer Assisted Language Learning* **33** (3), pp. 157–175. https://www.tandfonline.com/doi/full/10.1080/09588221.2018.1553186.

Mairano, Paolo, Fabián Santiago, and Leonardo Contreras Roa (2023), Can L2 Pronunciation Be Evaluated without Reference to a Native Model? Pillai Scores for the Intrinsic Evaluation of L2 Vowels, *Languages* **8** (4), pp. 280. https://www.mdpi.com/2226-471X/8/4/280.

Niño, Ana (2020), Exploring the use of online machine translation for independent language learning, *Research in Learning Technology*. https://journal.alt.ac.uk/index.php/rlt/article/view/2402.

Rushton, Andy (2022), Motivating Japanese University EFL Learners to Produce Longer Speaking Turns with Web Based Machine Translation: A Pilot Study, *Kobe Kaisei Jogakuin University Research Bulletin* (66), pp. 73–81.

Sakamoto, Kiyo (2022), Japanese university students' reflections on machine translation used as part of an English presentation activity, (53), pp. 33–45.

Skehan, Peter (1998), *A cognitive approach to language learning*, Oxford applied linguistics, 1. publ., 6. impr ed., Oxford University Press, Oxford. https://www.google.be/books/edition/A_Cognitive_Approach_to_Language_Learnin/Yzdl3pW0Yf4C?hl=nlgbpv=0.

Stapleton, Paul (2021), Using Google Translate as a Tool to Improve L2 Writing: A Case Study of Primary-Level Writing in Hong Kong, *International Journal of Computer-Assisted Language Learning and Teaching* **11** (3), pp. 92–98. http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJCALLT.2021070106.

Tang, Xiaoyi (2025), L2 Writing with AI: Perceptions and Engagement of EFL Learners in China, *English Language Teaching* **18** (2), pp. 68. https://ccsenet.org/journal/index.php/elt/article/view/0/51222.

Tsai, Shu-Chiao (2019), Using google translate in EFL drafts: a preliminary investigation, *Computer Assisted Language Learning* **32** (5-6), pp. 510–526. https://www.tandfonline.com/doi/full/10.1080/09588221.2018.1527361.

Wang, Li (2021), British English-Speaking Speed 2020, *Academic Journal of Humanities & Social Sciences*. https://francis-press.com/papers/4225.

Yamaoka, Kanako (2024), ChatGPT's Motivational Effects on Japanese University EFL Learners: A Qualitative Analysis, *International Journal of TESOL Studies*. https://www.tesolunion.org/journal/details/info/6MjQ4dLjhl/ChatGPT's-Motivational-Effects-on-Japanese-University-EFL-Learners:-A-Qualitative-Analysis.

Zhang, Zhihui and Xiaomeng Huang (2024), The impact of chatbots based on large language models on second language vocabulary acquisition, *Heliyon* **10** (3), pp. e25370. https://linkinghub.elsevier.com/retrieve/pii/S2405844024014014.