Human-AI Interaction in Language Acquisition: Evaluating LLM as a Language Partner

Oleksandr Lytvyn
University of Vienna
Comenius University Bratislava
lytvyn3@uniba.sk

Abstract

This single-case study investigates the potential of ChatGPT's advanced voice mode as a self-guided language partner for an upperintermediate (B2.1-level) German learner, combining AI-mediated practice with the *Sicher! B2.1* textbook (a standard intermediate German coursebook [1]) across 30 structured sessions. Using a mixed-methods approach involving quantitative measures (vocabulary tracking, standardized tests, Positive and Negative Affect Schedule (PANAS) surveys) and qualitative instruments (reflective journals, word-cloud analyses, self-interviews), the study evaluated linguistic gains and affective outcomes. Results showed significant vocabulary expansion and improvements in conversational fluency, along with reduced anxiety and increased motivation attributed to ChatGPT's low-pressure environment. However, limitations including technical instability, overly optimistic feedback, and a lack of pedagogical structure highlight the tool's role as a supplementary resource rather than a standalone solution. The study underscores the value of hybrid AI-powered learning methodologies integrated with structured curricula to balance accessibility and linguistic rigor.

1 Introduction

Large Language Models (LLMs) like ChatGPT are changing the way we learn languages by offering a low-pressure, conversation-based environment. But do they truly measure up to a human tutor when it comes to improving vocabulary, listening, and speaking skills—especially for learners around the B1–B2 proficiency level? And what about the roles of motivation and anxiety in this process? These questions frame the present research.

This pilot study examines whether ChatGPT's advanced voice interface, paired with the structured exercises in *Sicher! B2.1* (an intermediate German textbook by Perlmann-Balme, Schwalb, Matussek, 2018), can effectively support learning and boost confidence for a single upper-intermediate German learner. The work builds on insights from broader research into AI-driven language learning, covering not only technical advancements but also how such tools influence emotions, motivation, and cognitive development. The study uses a combination of quantitative and qualitative approaches—including standardized tests, vocabulary tracking, and reflective journals—to gauge how ChatGPT compares to traditional tutoring. Key findings are outlined alongside a discussion of limitations, with suggestions for how LLM-based tools could be integrated or refined within regular language classes or new learning paradigms. While this is a small-scale, single-participant project, it lays the groundwork for larger studies on whether AI can eventually complement or even partially replace a human tutor in language learning.

2 Literature Review

Recent studies highlight the growing role of AIdriven tools in language education, especially conversational agents like ChatGPT. These systems integrate advanced features—such as speech recognition, contextual awareness, and multi-turn dialogue—that facilitate personalized and low-pressure language practice. In particular, chatbots have been praised for improving access and learner motivation. However, concerns remain over the scripted nature of responses and lack of cultural nuance in AI interactions. Early applications of ChatGPT in language learning report promising results; for example, Park (2023) demonstrated effective use of ChatGPT in an English learning platform, while Pratiwi et al. (2024) reviewed ChatGPT's voice conversation mode and highlighted

[1] The *Sicher! B2.1* textbook is a widely used curriculum for this German proficiency level.

both its potential for improving speaking practice and certain limitations (such as inaccuracies).

Conversational AI has shown particular promise in reducing language anxiety and supporting vocabulary development. For instance, Ji et al. (2023) found that such tools provide authentic interaction and emotional comfort, making them effective supplements to formal instruction. However, their pedagogical depth is limited, especially in providing nuanced corrective feedback. Research also underscores the importance of affective factors in language learning. Systems like AutoTutor—an intelligent tutoring system that engages learners in dialogue and even responds to learner emotions—can match the effectiveness of human tutors in specific contexts (Graesser et al., 2014). Yet, true emotional understanding remains a challenge for AI, pointing to the need for human–AI hybrid models that involve teachers for empathy and complex feedback. From a cognitive standpoint, AI tools can support gains in vocabulary, fluency, and comprehension through adaptive learning (Qiao Zhao, 2023). Nonetheless, their over-reliance on repetition may narrow the range of linguistic exposure if used in isolation. For optimal acquisition, learners benefit from diverse language input modalities— listening, reading, and interactive speaking—which purely conversational systems might not fully provide unless integrated with other resources. In summary, the literature supports the use of AI as a complementary language learning aid—boosting engagement and reducing anxiety—while emphasizing the ongoing need for structured, human-led pedagogy to ensure depth and accuracy in learning outcomes.

3 Methodology

This single-case (N=1) study employed a mixedmethods design to evaluate the impact of ChatGPT's voice interface on German language acquisition at the B2.1 level. The participant, an experienced language educator with intermediate (approx. B1+) German proficiency, engaged in 30 structured sessions that combined traditional study with AImediated practice. Each session included approximately 30 minutes of textbook work followed by 45 minutes of guided dialogue with ChatGPT on relevant topics (e.g. workplace culture, digital life), mirroring the content of the *Sicher! B2.1* units. This approach ensured exposure to both written and spoken language input: the textbook provided reading and structured exercises, while the ChatGPT conversations provided interactive listening and speaking practice. All sessions took place in a quiet home environment using ChatGPT's voice mode on a laptop (MacBook, 2020). Session frequency was about 3–4 sessions per week, completing the 30 sessions over an eight-week period.

To assess linguistic development and affective responses, both quantitative and qualitative tools were used. Key instruments and measures included:

- **Vocabulary tracking:** The learner's vocabulary growth was tracked through session transcripts and word logs. New lexical items encountered during each ChatGPT conversation or textbook unit were recorded, and their reuse in later sessions was noted to gauge retention. Additionally, word frequency data were visualized in periodic word clouds (see below).
- **Standardized tests:** Pre- and post-study *Sicher!* online placement tests were administered to measure overall proficiency gains, and the *Sicher! B2.1* workbook unit tests were used to evaluate performance on the textbook material after each unit.
- **PANAS surveys:** The Positive and Negative Affect Schedule (PANAS) questionnaire was given before and after each session to quantitatively measure the participant's mood and emotional state (positive affect like enthusiasm, and negative affect like nervousness) in response to the learning activities.

Qualitative data sources were likewise diverse:

- **Reflective journals:** After each session, the participant wrote a brief reflective journal entry documenting emotional state, any notable challenges or successes, and feedback on the ChatGPT interaction. These free-form entries captured personal observations about anxiety, confidence, and the perceived utility of ChatGPT's feedback in that session.
- **Word cloud analysis:** Every 10 sessions, a cumulative transcript of the ChatGPT dialogues was compiled and processed to generate a word cloud. These word clouds visualized the most frequent words and topics that had emerged, providing a snapshot of lexical diversity and prominent themes at that stage of learning. By comparing word clouds from sessions 1–10, 11–20, and 21–30, we could qualitatively observe shifts toward more advanced vocabulary and topics over time.
- **Self-interviews:** At three key intervals (after Session 10, 20, and 30), the participant conducted

a semi-structured "self-interview" as a special journal entry. This involved answering a predefined set of reflective questions about their overall progress, motivation, and any anxiety or frustration experienced up to that point. These selfinterviews provided a deeper, synthesized perspective on the learning experience at regular milestones, reducing recall bias by capturing impressions while still fresh.

Data analysis focused on vocabulary retention, test performance, and emotional trends across sessions. Quantitative data from the vocabulary logs were analyzed to count total new words learned and how many of those were retained (reused in at least three later sessions). The pre- vs. post-study test scores were compared to gauge improvement magnitude (e.g. calculating the percentage increase and considering effect size for the single participant). PANAS survey results were plotted over time to visualize changes in affect; we examined whether positive affect scores tended to be higher after each session than before, and tracked any overall shifts across the eight weeks.

Qualitative data from journals and interviews were analyzed using thematic coding (following a reflexive thematic analysis approach in the spirit of Braun & Clarke, 2006). Recurring themes related to anxiety reduction, confidence gains, feedback quality, and self-regulation strategies were identified in the participant's narratives. The word cloud outputs were also interpreted for semantic shifts, indicating whether conversation topics and vocabulary became more complex as the sessions progressed. Throughout the study, reflexivity logs were maintained to mitigate researcher-participant bias (since the researcher was also the learner), and standard ethical considerations (informed consent, data privacy) were observed. The overall design aimed to replicate a realistic self-guided learning scenario where an AI tool supplements traditional study materials, in order to explore the feasibility of such hybrid learning in practice.

4 Findings

4.1 Vocabulary Development

Throughout the 30 sessions, the learner demonstrated a clear expansion of vocabulary. Lexical tracking revealed especially strong gains in noun usage and retention compared to verbs and adjectives, suggesting a bias toward concrete terminology during AI interactions. In other words, the participant tended to acquire and repeatedly use many new nouns (often topic-specific terms introduced by the chatbot or textbook), whereas fewer new verbs and descriptive words were picked up. This pattern is consistent with early stages of vocabulary development, where concrete nouns are learned before more abstract vocabulary.

A comparative analysis of vocabulary sources showed that the ChatGPT conversations yielded a larger pool of new words than the textbook activities. Across the study, the AI-mediated sessions exposed the learner to approximately 775 word tokens, covering 134 unique new words, whereas the textbook study introduced about 319 word tokens with 84 unique new words. In total, combining both sources, about 218 distinct new words were learned (some overlap existed between the AI and textbook, but the majority of new terms were exclusive to one

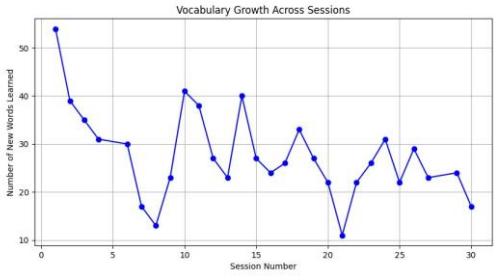

Figure 1: Vocabulary Growth Across Sessions.

source). These results support ChatGPT's potential as an expansive lexical resource: the conversational practice not only reinforced words from the textbook but also introduced additional vocabulary beyond the curriculum. Despite the greater volume of new words from ChatGPT, it was observed that many of these were concrete nouns tied to the conversation topics; comparatively fewer abstract or technical terms were introduced until later sessions (when the conversations became more complex). This finding suggests that while LLM-driven dialogues can greatly broaden vocabulary exposure, they may need guided prompts or supplemental materials to ensure a balance of word types and linguistic structures.

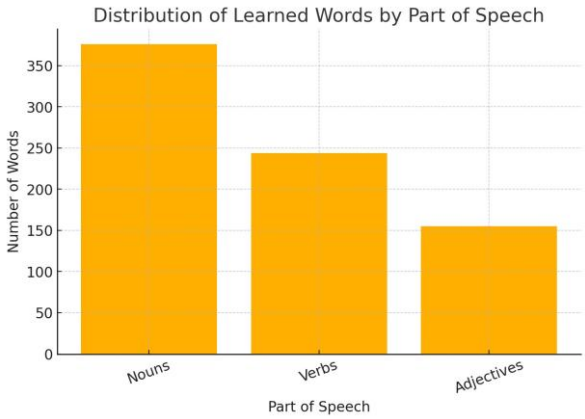

Figure 2: Distribution of Learned Words by Part of Speech.

4.2 Word Cloud Analysis

Progression in lexical complexity and topic breadth was visualized through word clouds generated at three key intervals in the intervention. Each word cloud highlights the most frequently used words in the ChatGPT dialogues for that period, providing insight into the dominant themes and vocabulary focus. The evolution of these word clouds reflected the shifting content of the sessions:

- **Sessions 1–10:** Focus on personal growth and basic everyday communication. (Common words in the first word cloud included personal pronouns and simple terms related to daily life and personal experiences, aligning with introductory conversational practice.)
- **Sessions 11–20:** Shift toward workplace and social themes. (The second word cloud showed frequent terms related to work, technology, and social interactions, indicating that topics had broadened to professional and societal contexts as the textbook units progressed.)
- **Sessions 21–30:** Emergence of academic and abstract vocabulary. (By the final sessions, the word cloud featured more complex, abstract terms – including academic language and nuanced vocabulary – reflecting the higher-level discussions and the more advanced textbook content covered at the end.)

This progression from concrete to more abstract vocabulary indicates that the AI-assisted conversations scaled in complexity alongside the textbook curriculum. Early sessions were dominated by simple, personal-topic vocabulary, but as the learner's confidence and the material difficulty grew, the conversations naturally incorporated more sophisticated lexicon. The word cloud analysis thus visually confirmed an increase in lexical diversity over time. It also helped identify which types of words were being reinforced: for example, the prominence of nounbased topics in early sessions, and the later inclusion of more verbs and adjectives as discussion themes expanded. Overall, the word clouds provided an intuitive way to track vocabulary development and ensured that the learner was indeed encountering increasingly varied language input throughout the study.

Figure 3: Word Cloud for Sessions 1-10.

Figure 4: Word Cloud for Sessions 11-20.

Figure 5: Word Cloud for Sessions 21-30.

4.3 Test Performance

Standardized testing results showed a significant improvement in formal language proficiency by the end of the study. The participant's score on the *Sicher!* placement test rose from 56% (pre-test) to 92% (post-test), indicating a substantial gain in overall competence at the B2 level. Similarly, the ongoing assessments using the *Sicher! B2.1* workbook unit exams reflected consistently high performance throughout the intervention: the

learner scored well on each unit's exercises and quizzes, demonstrating solid grasp of the textbook material. There was a slight decline in scores on the final few workbook units, but this was attributed to the rising difficulty of those later units (rather than a loss of knowledge). It should be noted that no long-term follow-up test was conducted beyond the immediate post-study exam, so retention of these gains over time was not evaluated. In other words, this study did not measure whether the participant's improved test performance would sustain after weeks or months without practice. Future research would need to include a delayed post-test to assess long-term knowledge retention or any post-study decline.

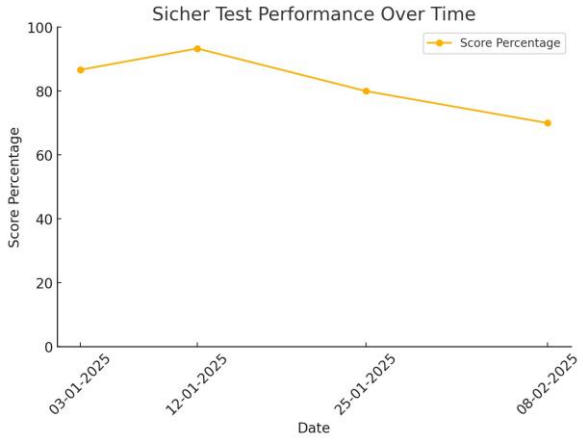

Figure 6: Test Performance Over Time.

4.4 Affective Outcomes (PANAS)

The affective measures revealed positive trends in the learner's emotional state over the course of each session and across the program. According to the PANAS surveys administered before and after sessions, positive affect (PA) scores consistently rose from pre-session to post-session. In practice, the participant tended to start a session with moderate enthusiasm or confidence, and finish the session reporting significantly higher positive feelings such as enjoyment, interest, or accomplishment. This indicates that engaging with ChatGPT in German conversation boosted the learner's engagement and enjoyment by the end of each practice session. Concurrently, negative affect (NA) scores decreased over time. The learner's reported anxiety, nervousness, or stress levels were higher before starting practice and then notably lower after interacting with the AI. Moreover, a gradual downward trend in baseline negative affect was observed as the sessions progressed week by week: the participant became less anxious overall about speaking German as they accumulated more practice in the low- pressure chatbot environment. These affective outcomes suggest that the ChatGPT voice partner had a confidence-building effect. The ability to practice speaking without fear of harsh judgment, and to receive patient responses, likely reduced language anxiety. The repeated pattern of rising PA and falling NA around each session underscores that the learner not only enjoyed the AI sessions but also grew more comfortable and less stressed about using German, which is a crucial component for successful language acquisition.

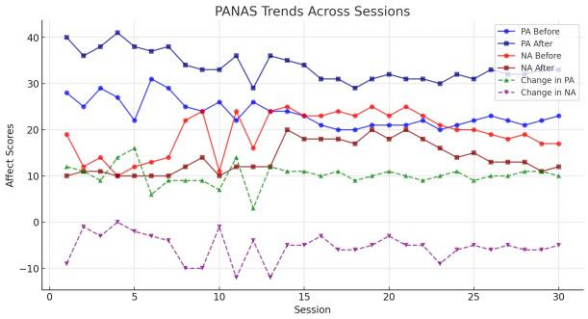

Figure 7: PANAS Trends Across Sessions.

4.5 Qualitative Insights

Qualitative data from the reflective journals and selfinterviews provided deeper insight into the learner's subjective experience and highlighted both strengths and limitations of using ChatGPT as a language partner. Over the 30 sessions, the participant's journal entries documented growing confidence in speaking and a steadily reduced fear of judgment. Initially, the learner noted nervousness about making mistakes in German, but as sessions went on, they frequently mentioned feeling more at ease conversing with the AI. The non-judgmental, patient nature of ChatGPT's responses was repeatedly cited as a positive factor that made practicing less intimidating. The learner also expressed appreciation for ChatGPT's supportive tone—the AI often provided encouragement and gentle praise, which helped maintain the learner's motivation. These comments align with the quantitative PANAS results, reinforcing that the AI partner created a comfortable learning atmosphere that bolstered the learner's self-confidence and willingness to speak.

However, the journals and interviews also pointed out some limitations and areas of concern. A recurring theme was that ChatGPT tended to give overly optimistic feedback and rarely offered corrections unless explicitly asked. The participant observed that the AI would often respond with polite affirmations ("That's great!" or a correct rephrasing

of the learner's sentence) even when errors were made, rather than directly correcting mistakes. This lack of systematic error correction meant that certain grammatical or pronunciation issues might not have been adequately addressed by the AI. The learner felt that more structured correction mechanisms would be beneficial – for example, having the AI occasionally point out mistakes or provide gentle corrections in real time, to more closely mimic a human tutor's guidance. Additionally, there were a few technical disruptions during the voice conversations (such as speech recognition errors or occasional crashes of the application). These interruptions, while not frequent, did break the flow of practice on a few occasions and were noted as a frustration. They suggest a need for improved stability in the voice integration of such AI systems.

In summary, the qualitative feedback indicates that ChatGPT in voice mode can serve as a comfortable and engaging conversation partner, helping to reduce anxiety and build speaking fluency. The learner's overall experience was positive, with clear motivational benefits. Yet, the findings also underline that ChatGPT is not a complete substitute for a human instructor. Important pedagogical functions like error correction and adaptive feedback were limited in the AI's responses, implying that human oversight or enhanced AI training is needed to address those gaps. Despite these limitations, this case study demonstrates the promise of integrating an LLM-based agent into language learning. The AI provided plentiful practice opportunities and exposure to new vocabulary in a way that kept the learner motivated. Going forward, a hybrid approach—using ChatGPT alongside a structured curriculum (such as *Sicher!*) and under teacher guidance—may strike the best balance between the AI's strengths (availability, patience, breadth of topics) and the human expertise required for thorough language instruction.

Acknowledgments

This research was made possible with the support of the University of Vienna through a mobility research project. Special thanks go to Univ.-Prof. Dr. Susanne Maria Reiterer for her expert supervision, continuous support, and valuable insights throughout the study. The author also extends gratitude to Yuta Watanabe for the inspiring discussions and his advice in the practical organization of the experiment.

References

[1] AI-ENABLED LANGUAGE SPEAKING COACHING FOR DUAL LANGUAGE LEARNERS. (2019). *IADIS International Journal on WWW/Internet, 17*(1). https://doi.org/10.33965/ijwi_2019171105

[2] Akhiat, M. (n.d.). *Second Language Acquisition in the Era of Technology and Artificial Intelligence: Exploring New Frontiers.*

[3] Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using Chatbots as AI Conversational Partners in Language Learning. *Applied Sciences, 12*(17), 8427. https://doi.org/10.3390/app12178427

[4] Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by Communicating in Natural Language with Conversational Agents. *Current Directions in Psychological Science, 23*(5), 374–380. https://doi.org/10.1177/0963721414540680

[5] Ji, H., Han, I., & Ko, Y. (2023). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education, 55*(1), 48–63. https://doi.org/10.1080/15391523.2022.2142873

[6] Jian, M. J. K. O. (2023). Personalized learning through AI. *Advances in Engineering Innovation, 5*(1), 16–19. https://doi.org/10.54254/2977-3903/5/2023039

[7] Kim, W.-H., & Kim, J.-H. (2020). Individualized AI Tutor Based on Developmental Learning Networks. *IEEE Access, 8*, 27927–27937. https://doi.org/10.1109/ACCESS.2020.2972167

[8] Lemon, O. (2022). Conversational AI for multiagent communication in natural language: Research directions at the Interaction Lab. *AI Communications, 35*(4), 295–308. https://doi.org/10.3233/AIC-220147

[9] Mageira, K., Pittou, D., Papasalouros, A., Kotis, K., Zangogianni, P., & Daradoumis, A. (2022). Educational AI Chatbots for Content and Language Integrated Learning. *Applied Sciences, 12*(7), 3239. https://doi.org/10.3390/app12073239

[10] Park, H.-Y. (2023). Application of ChatGPT for an English Learning Platform. *STEM Journal, 24*(3), 30–48. https://doi.org/10.16875/stem.2023.24.3.30

[11] Perlmann-Balme, M., Schwalb, S., & Matussek, M. (2018). *Sicher! B2.1 – Kursbuch und Arbeitsbuch.* Mu¨nchen, Germany: Hueber Verlag.

[12] Pratiwi, N., Efendy, A. G., Rini, H. C., & Ahmed, N. A. (2024). Speaking Practice using ChatGPT's Voice Conversation: A Review on Potentials and Concerns. *Journal of Language Intelligence and Culture, 6*(1), 59–72. https://doi.org/10.35719/jlic.v6i1.149

[13] Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: Illuminating the impact on speaking skills and selfregulation in Chinese EFL context. *Frontiers in Psychology, 14*, 1255594. https://doi.org/10.3389/fpsyg.2023.1255594

[14] Sato, M. (2014). Exploring the construct of interactional oral fluency: Second Language Acquisition and Language Testing approaches. *System, 45*, 79–91. https://doi.org/10.1016/j.system.2014.05.004

[15] Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070.

[16] Wei, Z. (n.d.). *ChatGPT Integrated with Voice Assistant as Learning Oral Chat-based Constructive Communication to Improve Communicative Competence for EFL Learners.* arXiv preprint. https://doi.org/10.48550/arXiv.2311.00718

[17] Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages, 8*(3), 212. https://doi.org/10.3390/languages8030212

[18] Zhang, Z., Huang, X. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon, 10*(3), e025370. https://doi.org/10.1016/j. heliyon.2024.e25370

[19] Ziaja, R. (2024). Korzy´sci i granice wykorzystania sztucznej inteligencji na lekcjach jezyka niemieckiego na przyk ladzie ChatGPT. *Ne-ofilolog, 62*(2), 521–540. https://doi.org/10.14746/n.2024.62.2.11