Singing to Remember: The Impact of AI-Generated Music and a Social Robot on L2 Retention and Recall in Children

Özde Pilli
s5257018
o.pilli@student.rug.nl

Supervisors: Paul Vogt & Suzan Dilara Scheffer

July 1, 2025

Abstract: This study investigates the impact of AI-generated music delivered through a social robot on second language (L2) vocabulary retention and recall in children. Specifically, it explores whether songs generated by a generative AI music platform, Suno, and presented via a social robot can enhance L2 learning outcomes across different age groups. Dutch-native or dominant children in two age brackets (e.g., 7-8 and 10-12 years old) participate in a single experimental condition involving interaction with a social robot that sings AI-generated songs. Pre- and post-intervention vocabulary tests are used to assess learning effectiveness, with controlled repetition and interaction settings. By comparing performance between the two age groups, the study aims to shed light on developmental differences in responsiveness to AI-supported multimodal language instruction. The findings contribute to understanding how artificial intelligence and social robotics can be effectively integrated into early bilingual education.

1 Introduction

In an increasingly globalized world, multilingual proficiency has never been more im-

portant. Children, who exhibit a remarkable capacity for language acquisition during their early developmental years, are especially well-suited for L2 learning [14]. As educational technologies continue to advance, their integration into language learning environments offers new and exciting opportunities [35], particularly through the use of artificial intelligence (AI) and social robotics [32].

This thesis explores how AI-generated music, when combined with social robot interaction, can influence the recall and retention of L2 vocabulary in children. While social robots have been studied in educational contexts, especially in language acquisition, the use of AI-generated music alongside robot interaction is a novel and largely unexplored area.

L2 learning refers to acquiring a language beyond one's native tongue. In children, this process benefits from early cognitive flexibility and exposure to rich, engaging input [14]. Vocabulary is a key starting point, as it supports later language development and is easier to measure in young learners [24]. Research shows that multimodal input, such as visuals, gestures, and music, can enhance L2 learning by making words more memorable and meaningful [5]. This aligns with theories like Krashen's Input Hypothesis, which emphasizes the importance of understandable and slightly challenging input for language acquisition [14].

Prior research on child-robot interaction highlights how robots can function as effective language tutors for young learners [32]. Focusing on key design features for robot tutors, such as peer-like interactions, *temporal*[*] and *semantic contingency*, joint attention, and gestures, has proven important for supporting language learning.

Additionally, the effectiveness of *social robots* in engaging children and providing personalized, multimodal learning experiences has been highlighted [16]. These studies underline the potential of robots to support language learning through interaction, gesture recognition, and adaptive feedback.

Other studies have explored how robots can use non-verbal cues, such as gestures, to enhance vocabulary learning [7]. It has been emphasized that gestures can significantly improve children's engagement and help them better retain new vocabulary by providing multimodal learning cues that reinforce verbal input.

Moreover, it has been examined how robots can support language learning through interactive storytelling [12]. Illustrating that robots can engage children in more complex language tasks, such as storytelling, and can adjust their prompts to encourage richer, more elaborate responses. This capacity for dynamic interaction makes robots powerful tools in fostering language production, in addition to vocabulary acquisition.

Parallel to advancements in robotics, music has long been recognized for its positive impact on children's education. Studies have shown that music enhances memory retention, aids in language development, and fosters emotional engagement [13]. Despite this, the use of AI-generated music specifi-

[*]Definitions for italicized terms can be found in the Appendix A, Glossary.

cally for language learning has not yet been thoroughly explored.

One key advantage of incorporating AI-generated songs into L2 learning is that it poses a significant advantage over human teachers. Unlike traditional methods where teachers must manually create educational songs or rely on pre-made content, AI systems can generate personalized songs on the spot, tailored to each child's learning needs. This not only makes learning more engaging but also allows for rapid customization of the content. Ensuring that each child receives the most relevant and effective learning experience possible. This level of personalization and speed is difficult to match by human tutors, offering a unique and scalable solution for educational environments.

Taken together, these studies collectively provide a foundation for the current research, demonstrating how social robots and AI can be integrated into language learning environments to enhance retention and recall in young learners. Yet, the intersection of AI-generated music and social robot interaction in this domain remains relatively unexplored, marking a significant gap this study aims to address.

Consequently, this research investigates a new approach to enhancing L2 vocabulary retention by integrating AI-generated songs into the interactions with a robot.

The research question this study seeks to answer is: **Do children aged 7-8 and 10-12 recall and retain a greater amount of English vocabulary after interacting with a social robot that uses AI-generated music, and does the amount of recalled and retained differ between the two age groups, and if so, in what direction?**

The significance of this research lies in the fact that the integration of AI-generated music and social robots for L2 acquisition in children remains largely unexplored. While the use of social robots in education has been explored, particularly in language acquisition, the combination of AI-generated music with robot interaction is a novel approach.

The upcoming parts discuss the impact of AI-generated music and social robot interaction on children's recall and retention of L2 vocabulary using an experimental design with two groups: Robot + AI-Generated Music, for the age groups 7-8 and 10-12. Pre-tests and post-tests are used to measure vocabulary retention.

The remainder of this thesis begins by describing the experimental setup, including participant characteristics, test procedures, and task design for each age group. Then the intervention components are outlined, such as the cross-linguistic similarity analysis of vocabulary items, the AI-generated song, and the robot's beat-synchronized movements. The results section presents the impact of the intervention on vocabulary recall and retention, followed by a discussion of the findings, limitations of the study, and directions for future research.

2 Methods

This section presents the methodological framework used to investigate the effect of AI-generated music and social robot interaction on children's recall and retention of English L2 vocabulary. The study employed a pre-test and post-test design to evaluate vocabulary gains after exposure to the experimental intervention.

To ensure the validity of the vocabulary assessments, a cross-linguistic similarity analysis was performed. This step was necessary to control for the possibility that children might guess the meanings of English words based on *phonological* and *semantic similarities* with words in their native language. A detailed explanation of this analysis is provided in Section 2.2.

The intervention included a custom educational song generated using Suno, a multimodal AI-based song generation platform [1]. This song was designed to incorporate the target vocabulary in a musically engaging and age-appropriate format. The process of generating the lyrics and melody is described in Sections 2.4-2.6.

To enhance the interactive experience, the robot was programmed to gesture and move in synchrony with the music. Although a brief overview of gesture use is included here, further technical details are provided in the Appendix: B Gesture Implementation.

Robot dance movements were synchronized with the beat of the AI-generated song using beat detection algorithms to ensure a natural and engaging presentation, Section 2.7.

Together, these components formed a multimodal learning experience aimed at maximizing children's engagement and supporting vocabulary acquisition in a playful and meaningful context.

2.1 Experimental Setup

2.1.1 Participants

The experiment involved two groups of children: Group 1, aged 7–8 years, and Group 2, aged 10–12 years. The groups consisted of 8 and 7 children, respectively, resulting in a total of 15 participants. Group 1 had an equal gender distribution (4 males, 4 females), while Group 2 consisted of 3 males and 4 females.

Participants were recruited through volunteer schools in the Groningen region of the Netherlands. Group 1 attended a primary school, while Group 2 attended a Sunday school. The participants from Group 2 and were enrolled in various regular schools during the week. Parental consent was obtained through ethically approved consent forms, and all data were anonymized. Socioeconomic background information was not collected.

Prior to participation, each child completed a short questionnaire (see Appendix C, Section: C.1) to determine their language background. They were asked whether they were native Dutch speakers and whether they spoke any additional languages at home. All participants reported Dutch as their first language. While three children in Group 1 and three children in Group 2 indicated bilingual proficiency, with additional languages including Syrian, Arabic, Somali, and French. It

was unknown whether any children had received prior formal English instruction. Additionally, little was known about the participants' prior English proficiency levels, as we did not have access to the curriculum details from their schools. English proficiency levels were estimated based on observations during a guest lecture conducted prior to the study.

This information can provide valuable insights regarding possible confounding variables. Cross-linguistic similarity was assessed between the languages Dutch and English, so the cross-linguistic similarity between English and one of the mentioned languages spoken by the bilingual participants could possibly influence the results.

2.1.2 Pre- and Post-Test Procedure

To assess the recall and retention in L2, both groups completed a pre-test before the intervention and a post-test afterwards. 10 vocabulary items were present in each groups pre-/post-tests. The vocabulary tests were designed to be slightly above the children's current English proficiency to better assess their ability to recall new words [33], (see Appendix C, Sections: C.2 & C.3).

For Group 1, who were beginner learners, the test was adapted from A1–A2 vocabulary lists provided by the *British Council* [4]. For Group 2, a B1-level vocabulary test was adapted from *English Grammar and Vocabulary Exercises* [9]. The test focused on descriptive vocabulary related to people for the B1 level, whereas for the A1–A2 level, kitchen-related vocabulary was used. All vocabulary items were chosen based on two criteria: minimal cross-linguistic similarity between English and Dutch to reduce the likelihood of correct guesses based on phonetic overlap, and appropriateness for the respective proficiency levels (A1–A2 for Group 1, B1 for Group 2).

2.1.3 Task Design by Age Group

For Group 1 (7–8-year-olds), a picture-pointing task was used to measure receptive vocabulary knowledge. Children listened to a spoken word and selected the corresponding image from a set. This method is developmentally appropriate for young learners as it reduces cognitive demands by avoiding reading and written responses. Instead, it focuses on receptive language skills, which typically precede productive skills [5]. Picture-pointing tasks are widely accepted in early language assessment for their engaging nature and ability to reduce test anxiety, thus promoting more accurate responses [8]. Multiple-choice formats were avoided to eliminate reading-related confounds and guessing biases [25].

For Group 2 (10–12-year-olds), a fill-in-the-gaps task was administered. This required participants to complete sentences by selecting appropriate target words from a provided word bank. This format was chosen to assess productive vocabulary skills, which are more appropriate for intermediate learners at the B1 level [24]. By offering a limited word bank, the task balanced cognitive load and open-ended recall demands, supporting lexical access and contextual understanding [18]. This method also reduces guessing compared to multiple-choice formats and provides a more

ecologically valid assessment of vocabulary knowledge [27, 30, 25].

2.1.4 Intervention Phase

The full experimental session lasted no more than 10 minutes and was conducted with groups of children. During the robot interaction, participants were divided into groups of four and seated in a half-circle approximately one meter from the robot. Immediately after the pre-test, participants interacted with a social robot, the Alpha Mini (see Figure 2.1) (provided by the University of Groningen).The robot briefly introduced itself in Dutch and described the activity before delivering the intervention: a custom AI-generated educational song introducing the target vocabulary in an engaging musical format. The intervention song was played on vloume 5 and lasted approximately 2-3 minutes. Following the song, children participated in a filler activity where the robot danced and sang a different (non-educational) song, which lasted about 3 minutes. This filler activity was identical for both groups and aimed to maintain engagement while reducing any immediate memory effects associated solely with rote repetition or test preparation. The post-test was administered immediately after the filler activity.

Both groups were tested at approximately 10 a.m. on separate days in a quiet classroom environment to ensure consistency across sessions. Care was taken to ensure that each group received identical instructions and support throughout the procedure. No additional learning materials or external aids were introduced during the session, and all interactions were standardized to control for variation in delivery. This controlled setting aimed to isolate the effects of the intervention and reduce potential confounding factors such as time of day, setting, or teacher influence.

Figure 2.1: Alpha Mini Robot

A visual overview of the experimental procedure can be found in Figure 2.2, which outlines the sequence from pre-test to intervention and post-test.

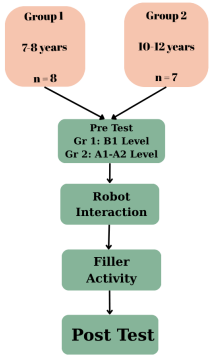

Figure 2.2: Experimental Procedure

2.2 Cross-Linguistic Similarity Analysis

The list of words within the pre- and post-tests was subjected to a cross-linguistic similarity analysis using a system pre-trained with multilingual word embeddings. This was done to assess the possible *orthographic and semantic overlaps* of each English test vocabulary with its Dutch equivalent.

In cross-linguistic vocabulary acquisition studies, cognates, words that share similar forms and meanings across languages due to common etymology (e.g., English "orange" and Dutch "oranje"), can significantly distort L2 vocabulary assessment results [28]. When learners encounter *cognates*, they may correctly guess meanings based on L1 knowledge rather than demonstrating actual L2 recall, leading to inflated performance scores that do not reflect true learning gains. By identifying and minimizing such cross-linguistically similar items, we aimed to ensure that the test items required actual learning rather than inference from L1 knowledge.

The procedure used for the cross-linguistic similarity analysis:

1. **Word Selection:** Vocabulary items from two proficiency levels were selected. Each English word had a corresponding Dutch translation based on test materials.

2. **Embedding Model:** We used the multilingual version of *fastText* word embeddings [2], which represent words from different languages in a shared vector space. These embeddings capture both orthographic (visual form) and semantic (meaning) similarities based on distributional patterns learned from large multilingual corpora.

3. **Similarity Computation:** Each English-Dutch word pair was embedded into a 300-dimensional vector space. Cosine similarity was computed between each pair to quantify their cross-linguistic similarity, using the formula:

$$\text{similarity}(A, B) = \frac{A \cdot B}{|A||B|}$$

   Where A and B are the embedding vectors for the English and Dutch words, respectively.

Similarity values range from -1 to 1, where values closer to 1 indicate higher cross-linguistic similarity (potentially problematic cognates or near-cognates), and values near 0

indicate minimal similarity. To minimize the potential for cross-linguistic transfer effects, only word pairs with similarity scores below 0.2 were retained. This conservative threshold was chosen to exclude potential cognates and semantically transparent items that could be easily guessed based on L1 knowledge. Negative values were retained as they indicate dissimilar word pairs, which aligns with our goal of testing actual L2 vocabulary acquisition rather than cross-linguistic inference.

The resulting values for the two tests can be observed below:

English Word	Dutch Translation	Similarity Score
sink	gootsteen	-0.01
bowl	kom	-0.07
table	tafel	0.09
cupboard	kast	0.04
cup	beker	0.04
knife	mes	-0.03
chair	stoel	0.11
cooker	fornuis	0.02
plate	bord	-0.04
spoon	lepel	-0.08

Figure 2.3: Cross-linguistic Similarity Scores for A1-A2 Level Vocabulary Test Items

English Word	Dutch Translation	Similarity Score
amusing	grappig	-0.07
careless	zorgeloos	0.01
cautious	voorzichtig	-0.04
cheerful	vrolijk	0.04
greedy	hebberig	-0.04
healthy	gezond	0.05
polite	beleefd	0.03
slim	slank	0.16
tall	lang	0.01
wealthy	rijk	-0.00

Figure 2.4: Cross-linguistic Similarity Scores for B1 Level Vocabulary Test Items

As shown in Figures 2.3 and 2.4, all similarity scores for both final test sets fall below the 0.2 threshold, confirming that the selected vocabulary items have minimal cross-linguistic similarity and are unlikely to be correctly guessed based on Dutch knowledge alone.

2.3 Gesture Implementation During Robot Introduction

At the start of the experiment, the Alpha Mini robot used a combination of beat and iconic gestures during its self-introduction and explanation of the activity. Beat gestures were synchronized with prosodically stressed words to enhance speech rhythm and engagement, while iconic gestures were triggered by specific semantic content (e.g., waving for greetings, pointing for self-references).

Beat gestures were timed to coincide with stressed syllables, while iconic gestures were mapped to specific keywords in the robot's speech.

All gestures were designed to be subtle and proportional to the robot's size to maintain naturalness and avoid distracting from the verbal content. The multimodal approach aimed to support communicative effectiveness and increase perceived social presence during the robot's initial interaction with the children. See the Appendix for a detailed explanation of the implementation.

2.4 AI-Generated Educational Song

To generate the educational songs used in this experiment, the Suno app (version v4) was employed. Suno is a state-of-the-art AI music generation platform capable of producing complete musical compositions from text prompts, including lyrics, melodies, harmonies, and instrumental arrangements [1]. Version v4 was selected for this study due to its improved coherence and stylistic flexibility, which are essential for creating engaging and developmentally appropriate content for young second-language learners.

2.5 Custom Lyric Development

Although Suno can generate lyrics automatically, custom-written lyrics were used for this experiment. Preliminary trials revealed that Suno's automatically generated lyrics were often too complex in vocabulary and syntax for children at A1–A2 and B1 *CEFR* language levels, featuring abstract concepts and advanced grammatical structures unsuitable for beginner and intermediate learners. To ensure pedagogical appropriateness, all lyrics were manually crafted using a controlled vocabulary that directly aligned with the target words featured in the pre- and post-tests. The lyric development process involved several stages:

1. **Initial ideation:** ChatGPT-4 Turbo was used to generate preliminary ideas for choruses and verses incorporating the target vocabulary items.
2. **Simplification:** Generated content was simplified to remove complex sentence structures, abstract metaphors, and idiomatic expressions inappropriate for the target proficiency levels.
3. **Manual composition:** Final lyrics were adjusted manually, emphasizing simplicity, repetition, and contextual clarity to reinforce target vocabulary through short, grammatically accessible constructions.
4. **Pronunciation optimization:** Lyrics were refined to ensure clear pronunciation when rendered by Suno's vocal synthesis, as complex structures often resulted in unclear or distorted audio output.

Each target vocabulary item appeared multiple times throughout the songs to maximize exposure and reinforce learning through repetition.

2.6 Musical Style Selection

To align with children's musical preferences, songs were generated using the style prompts "Pop", "Fun," and "Dance." This selection was based on established research on musical preferences in the target age groups. Research indicates that children begin showing clear preferences for pop music around age 8, with earlier ages (5-7 years) showing broader genre acceptance [11]. This trend is supported by Szabó et al. [31], who surveyed over 1,100 students aged 9-19 and found that accessible pop music was the dominant preference among

children aged 9-12, largely due to exposure through school and family environments.

The selected style tags were intended to enhance song appeal and increase the likelihood of sustained attention, emotional engagement, and motivation during the vocabulary learning activity. Two separate songs were created: one incorporating the A1-A2 vocabulary for Group 1, and another incorporating the B1 vocabulary for Group 2.

Two versions of the song were generated, one featuring a female voice and the other a male voice. To maintain fairness across groups, the female-voiced version was chosen, as the male-voiced version typically conveyed a more downbeat tone.

2.7 Synchronized Robot Dance

To enable the robot to dance in synchrony with the AI-generated music, beat detection techniques were employed to extract rhythmic cues from the audio and map them to pre-defined dance routines. This ensures that the robot's movements are timed to align with the beat of the song, rather than being arbitrarily timed.

The synchronization process is performed using the *Librosa* library, which is a powerful Python toolkit for music and audio analysis [20]. The process begins by loading the target audio file, which returns both the audio time series and the sampling rate. The tempo is then estimated in beats per minute and detects the positions of beats within the audio, which is represented as frame indices. These beat frames are subsequently converted to time values (in seconds), providing a sequence of timestamps that correspond to the rhythmic pulse of the song.

During execution, these beat times are used to synchronize predefined dance movements with the music. The robot's dance routines consist of simple head and arm movements, chosen specifically to avoid distracting the child during the interaction. Robot's dance routines, defined as sequences of joint movements, are scheduled to start in alignment with specific beats, ensuring that the performance is rhythmically synchronized with the song. This method allows the robot to "dance" in time with music based on precise beat detection rather than arbitrary timing.These timestamps are saved for later use. Figure 2.5 shows an outline of the process.

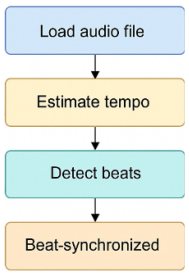

Figure 2.5: Beat Analysis & Synchronized Moves

3 Results

This section presents the intervention outcomes for two age groups: 7–8 years (Group 1, n = 8) and 10–12 years (Group 2, n = 7). Paired-samples t-tests served as the primary analysis, with Wilcoxon signed-rank tests and bootstrap confidence intervals providing additional validation.

The Figure 3.1 displays the boxplots comparing pre- and post-test distributions for both groups. Group 1 shows minimal change with greater variability compared to Group 2. While Group 2 demonstrates a clear upward shift in scores with the median increasing from 5.0 to 7.0 points.

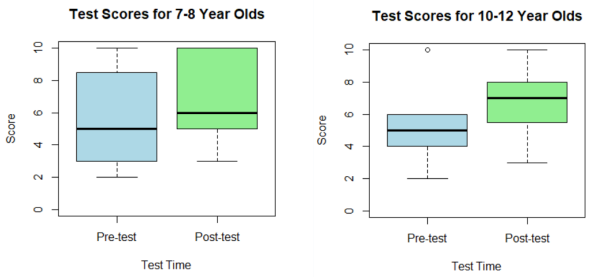

Figure 3.1: Box Plots for Pre- and Post-Test Scores of Group 1 and Group 2

Figure 3.2 below presents the complete descriptive statistics for both groups. Group 1 demonstrated minimal change between pre-test (M = 6.12, SD = 3.22) and post-test (M = 6.87, SD = 2.80) phases. Group 2 showed improvement from pre-test (M = 5.2, SD = 2.56) to post-test (M = 6.0, SD = 2.29), with the median increasing from 5.0 to 7.0 points.

Group	Test Phase	n	Mean (SD)	Median	IQR	Range
1	Pre-test	8	6.12 (3.22)	5.0	3.0–8.5	2–10
1	Post-test	8	6.87 (2.80)	6.0	5.0–10.0	3–10
2	Pre-test	7	5.2 (2.56)	5.0	4.0–6.0	2–10
2	Post-test	7	6.0 (2.29)	7.0	5.5–8.0	3–10

Vocabulary scores range from 0–10 points.

Figure 3.2: Descriptive Statistics

Primary Statistical Analysis: Paired-samples t-tests revealed no significant improvement for Group 1: mean difference = 0.75 points, 95% CI [–0.65, 2.15], $t(7) = 1.27$, $p = .244$, Cohen's $d = 0.24$, 95% CI [–0.17, 0.65]. For Group 2, results indicated statistically significant improvement: mean difference = 1.43 points, 95% CI [0.25, 2.61], $t(6) = 2.97$, $p = .025$, Cohen's $d = 0.58$, 95% CI [0.12, 1.03].

Sensitivity Analyses: Bootstrap confidence intervals (10,000 resamples) supported the primary findings: Group 1 showed mean difference = 0.75, 95% CI [–0.36, 1.75]; Group 2 showed mean difference = 1.43, 95% CI [0.57, 2.29]. Wilcoxon signed-rank tests yielded consistent results: Group 1, $V = 10.5$, $p = .375$; Group 2, $V = 0$, $p = .057$.

Analysis	Group 1 (7–8 years)	Group 2 (10–12 years)
Paired t-test	$t(7) = 1.27$, $p = .24$	$t(6) = 2.97$, $p = .025^*$
Mean Difference [95% CI]	0.75 [−0.65, 2.15]	1.43 [0.25, 2.61]*
Cohen's d [95% CI]	0.24 [−0.17, 0.65] (Trivial)	0.58 [0.12, 1.03]* (Moderate)
Bootstrap Mean [95% CI]	0.75 [−0.36, 1.75]	1.43 [0.57, 2.29]*
Wilcoxon Test	$V = 10.5$, $p = .38$	$V = 0$, $p = .057$ (Trend)

Figure 3.3: Overall View Of The Statistical Results

In summary, the intervention had a statistically significant and practically meaningful effect for older children (10–12 years), with consistent results across the statistical methods used. The younger group (7–8 years) did not show a reliable improvement.

4 Discussion

This study was designed to explore whether a combination of an AI-generated song and a social robot could support the learning of

L2 vocabulary in children. In relation to the research question, both age groups displayed a positive trend in post-test scores compared to pre-test scores, suggesting that the intervention had some beneficial effect on vocabulary retention. However, only the older group (10–12 years) showed a statistically significant improvement, indicating that age may moderate the effectiveness of this multimodal learning approach. Also displaying no drastic differences in the pre- and post-test scores of only Dutch speaking and bilingual participants.

One possible explanation for the age-related difference is the developmental variation in cognitive capacities such as working memory, attention span, and metalinguistic awareness. This may be an aiding factor in vocabulary acquisition during auditory tasks [10]. Explaining the somewhat higher post-test scores observed in the older group. This aligns with Krashen's Input Hypothesis, which posits that learners acquire language most effectively when exposed to input slightly above their current level. The older group may have been more capable of processing and retaining such input, especially when delivered in an engaging, musical format.

It is also likely that the older participants had greater familiarity with formal testing, allowing them to focus more effectively during the post-test. In contrast, the younger group may have found the testing environment more cognitively demanding, or the tasks may have been developmentally misaligned despite efforts to tailor them to age-appropriate formats. These findings suggest that intervention success depends not only on content but also on the match between task format and developmental stage.

The use of music in L2 learning has been shown to aid memory through structured repetition, rhythm, and emotional engagement [22, 26]. While the AI-generated songs in this study were customized for vocabulary alignment and age appropriateness, it remains unclear whether the musical style or structure optimally supported learning for both age groups. Children's engagement with the song likely varied depending on their personal musical preferences, emotional response, or familiarity with similar music genres.

Additionally, the presence of a social robot may have amplified the learning experience by enhancing perceived social presence, which is known to improve engagement and memory [32]. However, because the music and robot were presented simultaneously, this study cannot isolate whether learning gains were due to the musical component, the robot, or their combined effect.

Moreover, while the AI-generated song was designed to be engaging and educational, it remains uncertain whether this specific prompt formulation or musical style aligned well with the children's preferences. Children's enjoyment, motivation, and willingness to engage likely play a significant role in vocabulary acquisition, and future studies could explore which types of musical prompts or narrative formats children respond to most positively. This could include varying rhythm, repetition, visual supports, or emotional tone.

Finally, while the intervention showed

promise, individual differences such as prior English exposure, language aptitude, and comfort with technology could have influenced the results. These factors were not formally assessed but may have contributed to performance variability across participants.

5 Limitations

This study has several limitations that should be acknowledged. The most significant is the small sample size (N = 15), which limits statistical power and the generalizability of the findings. While some group-level trends were observed, they should be interpreted cautiously given the potential for variability due to individual differences.

Second, the experimental design combined the robot and the AI-generated song into a single intervention without control conditions, making it impossible to isolate the specific contributions of either component. This limits our ability to determine whether learning gains were primarily driven by musical input, social interaction, or their combination.

Third, the two age groups were exposed to different songs and vocabulary test items, introducing a confound that prevents direct comparison between them. Although efforts were made to match task difficulty to age-appropriate levels, this difference in materials may have influenced group-level outcomes.

Fourth, while the AI-generated song was customized for vocabulary alignment, its musical structure and delivery style were not formally evaluated for developmental suitability or learner preference. It is unknown whether the selected musical genre, tempo, or vocal style was equally effective across age groups.

Fifth, children's prior exposure to English and individual language proficiency were estimated informally and not directly measured. This lack of baseline control may have introduced variance unrelated to the intervention. Notably, the boxplots revealed greater variability in Group 1's scores compared to Group 2, suggesting that the younger group had more heterogeneous English proficiency levels. This could reflect differences in informal exposure, home language environment, or readiness for vocabulary learning tasks.

Finally, key affective and behavioral variables, such as engagement, enjoyment, and motivation, were not systematically measured. These factors likely mediate the effectiveness of music- and robot-based learning and could help explain individual differences in learning outcomes.

6 Future Work

To better understand the mechanisms behind the observed learning gains, future studies should adopt a controlled experimental design that separates the effects of music and robot interaction. Specifically, three experimental conditions are recommended: (1) robot + AI-generated song, (2) robot + verbal vocabulary explanation (no song), and (3) speaker + AI-generated song (no robot). This would allow researchers to isolate the individual and combined effects of musical input and social presence on L2 vocabulary acquisition.

Second, the role of specific musical fea-

tures, such as rhythm, repetition, emotional tone, and melodic structure, should be systematically investigated. These elements may influence memory encoding and learner engagement in distinct ways. It is also likely that children's musical preferences and familiarity with certain genres affect their level of attention and recall. Future studies could compare different musical styles, emotional tones, or delivery formats to identify what combinations are most effective for different age groups or learner profiles.

Third, long-term studies are needed to assess the durability of the learning gains. Retention tests administered days or weeks after the intervention would provide insights into whether the observed benefits are sustained over time. Additionally, repeated exposure across multiple sessions could help determine whether continued engagement with the robot and musical content enhances consolidation.

Finally, integrating qualitative measures, such as interviews, enjoyment ratings, or observational coding of engagement, could provide richer insights into how children experience and respond to these interventions. These methods would help capture emotional and motivational factors that are not reflected in vocabulary test scores but are likely critical for learning success.

7 Conclusion

This study provides preliminary evidence that the combination of a social robot with an AI-generated song can support vocabulary learning in children. Improvements in post-test performance suggest some educational benefit, particularly for older learners. While the effect sizes were modest, the results highlight the potential of integrating multimodal, interactive technologies into early language instruction.

By combining musical input with embodied interaction, the intervention reflects a growing trend in educational technology toward engaging learning experiences. The robot provided social presence and gesture-based cues, while the AI-generated music offered repetition, rhythm, and contextual embedding of vocabulary. These findings contribute to the expanding field of technology-enhanced L2 learning and demonstrate the feasibility of using AI-driven tools in child-centered pedagogical settings.

Despite its exploratory nature, the study points toward scalable, adaptable approaches to language instruction, particularly when interventions are developmentally tailored. Future iterations of this approach can further optimize personalization, feedback, and learner engagement based on age, proficiency, and preference.

Overall, to answer the research question, the findings suggest that both age groups, children aged 7–8 and 10–12—demonstrated improved English vocabulary recall and retention after interacting with the social robot enhanced by AI-generated music. However, the older group showed greater gains in post-test performance, indicating that age may influence the effectiveness of such interventions, with older children benefiting more from the multimodal input.

References

[1] Suno AI. *Suno: AI Music Generation Platform*. https://www.suno.ai. 2024.

[2] Piotr Bojanowski et al. "Enriching word vectors with subword information". In: *Transactions of the association for computational linguistics* 5 (2017), pp. 135–146.

[3] Rishi Bommasani et al. "On the Opportunities and Risks of Foundation Models". In: *arXiv preprint arXiv:2108.07258*. 2021.

[4] British Council. *A1–A2 Vocabulary*. Accessed: 2025-05-04. n.d. URL: https://learnenglishteens.britishcouncil.org/vocabulary/a1-a2-vocabulary.

[5] Lynne Cameron. *Teaching languages to young learners*. Cambridge university press, 2001.

[6] Justine Cassell, David McNeill, and Karl-Eric McCullough. "Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information". In: *Pragmatics and Cognition* 9.2 (2001), pp. 217–237.

[7] Jan De Wit et al. "The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies". In: *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 2018, pp. 50–58.

[8] Lloyd M Dunn and Leota M Dunn. "Peabody picture vocabulary test–". In: (1965).

[9] English Practice. *English Grammar and Vocabulary Exercises*. https://www.english-practice.at. Accessed: 2025-05-27. n.d.

[10] Susan E Gathercole. "Nonword repetition and word learning: The nature of the relationship". In: *Applied psycholinguistics* 27.4 (2006), pp. 513–543.

[11] Heiner Gembris and Gabriele Schellberg. "Musical preferences of elementary school children". In: *Proceedings of the 5th Triennial ESCOM Conference*. Vol. 8. 2003, p. 13.

[12] Nicole Goossens, Rian Aarts, and Paul Vogt. "Storytelling with a social robot". In: *Robots for Learning R4L* (2019).

[13] Erin E Hannon and Sandra E Trehub. "Tuning in to musical rhythms: Infants learn more readily than adults". In: *Proceedings of the National Academy of Sciences* 102.35 (2005), pp. 12639–12643.

[14] Ruyun Hu. "The age factor in second language learning". In: *Theory and practice in language studies* 6.11 (2016), pp. 2164–2168.

[15] Cheng Huang and Bilge Mutlu. "Robotic nonverbal behavior improves task performance in human-robot collaboration". In: *Proceedings of the ACM/IEEE International Conference*

on Human-Robot Interaction (HRI). 2011, pp. 57–64.

[16] Junko Kanero et al. "Social robots for early language learning: Current evidence and future directions". In: *Child Development Perspectives* 12.3 (2018), pp. 146–151.

[17] Emiel Krahmer and Marc Swerts. "Adding gesture to speech can improve information recall". In: *Human Factors* 49.3 (2007), pp. 515–524.

[18] Batia Laufer and Zohreh Goldstein. "Measuring and explaining the lexical threshold of reading comprehension: A case for semantically related words". In: *Language Testing* 16.1 (1999), pp. 33–52.

[19] Weiting Li, Chieh-Yang Lee, and Eduard Hovy. "A Survey of Emotion Recognition Using Transformer-based Models". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2022.

[20] Brian McFee et al. "librosa: Audio and music signal analysis in python". In: *Proceedings of the 14th python in science conference*. Vol. 8. 2015, pp. 18–25.

[21] David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

[22] Suzanne L Medina. "The Effects of Music upon Second Language Vocabulary Acquisition." In: (1990).

[23] Louis-Philippe Morency et al. "Head gestures for perceptual interfaces: The role of context in improving recognition". In: *Artificial Intelligence* 171.8-9 (2007), pp. 568–585.

[24] I.S.P. Nation. *Learning Vocabulary in Another Language*. Cambridge University Press, 2001.

[25] Anne-Catherine Nicolay and Martine Poncelet. "Cognitive abilities underlying second-language vocabulary acquisition in an early second-language immersion education context: A longitudinal study". In: *Journal of experimental child psychology* 115.4 (2013), pp. 655–671.

[26] Kelli R Paquette and Sue A Rieg. "Using music to support the literacy development of young English language learners". In: *Early Childhood Education Journal* 36 (2008), pp. 227–232.

[27] John Read. *Assessing Vocabulary*. Cambridge University Press, 2000.

[28] Håkan Ringbom. *Cross-linguistic similarity in foreign language learning*. Vol. 21. Multilingual Matters, 2006.

[29] Maja Salem et al. "Generation and evaluation of communicative robot gesture". In: *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2012, pp. 61–68.

[30] Norbert Schmitt. *Vocabulary in Language Teaching*. Cambridge University Press, 2000.

[31] Norbert Szabó et al. “Musical Preferences among Students Aged 9–19: A Study on Musical Genres and Styles”. In: *Education Sciences* 14.3 (2024), p. 290.

[32] Paul Vogt et al. “Child-robot interactions for second language tutoring to preschool children”. In: *Frontiers in human neuroscience* 11 (2017), p. 73.

[33] Lev Semenovich Vygotsky and Michael Cole. *Mind in society: Development of higher psychological processes*. Harvard university press, 1978.

[34] Petra Wagner, Zofia Malisz, and Stefan Kopp. “Gesture and speech in interaction: An overview”. In: *Speech Communication* 57 (2014), pp. 209–232.

[35] Darrell M West. “Mobile learning: Transforming education, engaging students, and improving outcomes”. In: *Brookings Policy Report* 9.7 (2013), pp. 1–7.

[36] David Wilkins. “Adam kendon (2004). gesture: Visible action as utterance”. In: *Gesture* 6.1 (2006), pp. 119–144.

A Glossary

- **Temporal Contingency:** Temporal contingency refers to the timely and responsive interaction between the robot and the child. This concept is critical in ensuring that the child remains engaged with the robot, as the interaction feels natural when the robot responds quickly to the child's actions or speech.

- **Semantic Contingency:** Semantic contingency emphasizes the importance of the robot's responses being aligned with the child's focus of attention. It ensures that the robot's answers are relevant to what the child is currently engaging with, whether verbal or non-verbal.

- **Social Robots:** Social robots are machines designed to interact with humans in a way that mimics human-to-human interaction. In this study, social robots are used to engage children in second language (L2) learning activities, providing interactive, personalized feedback.

- **Beat Gestures:** Small, rhythmic hand movements that co-occur with speech, typically marking prosodic stress or the beat of an utterance. They do not carry semantic content themselves but emphasize spoken words.

- **Iconic Gestures:** Gestures that depict or represent the form or movement of an object, person, or action being described in speech. They typically carry semantic content related to the spoken word.

- **Phonological Similarity:** The degree to which words or sounds share similar pronunciations.

- **CEFR:** Common European Framework, a standard used to describe language proficiency levels.

- **Cross-linguistic similarity analysis:** An examination of how words in different languages may resemble each other in form or meaning, potentially influencing language learning or word recognition.

  Orthographic Overlap: The degree to which two words share similar spelling patterns or letter sequences. High orthographic overlap means the words look similar in written form, which can influence word recognition, especially in language learning or bilingual contexts. Example: "house" (English) and "haus" (German) have high orthographic overlap.

- **Semantic Overlap:** The extent to which two words share similar meanings or refer to related concepts. High semantic overlap means the words are conceptually similar, which can aid comprehension and recall across languages. Example: "mother" (English) and "madre" (Spanish) have high semantic overlap because they refer to the same concept.

- **Cognates:** Words in two languages that look and mean the same because they come from the same original word. Example: “animal” in English and “animal” in Spanish.

B Gesture Implementation

B.1 Beat Gestures

At the start of each group the robot introduces the activity with a brief explanation. During this, it uses two types of gestures, *beat* and *iconic* gestures, to make the interaction look as natural as possible. The design of the beat gestures aimed to create arms and combined arm-head movements that closely resemble human-like beat gestures. To maintain a natural appearance, the gestures were deliberately kept subtle and proportional to the robot’s size, avoiding exaggerated motions that might appear unnatural. Instead, the movements were designed to be small, rhythmic, and smooth, enhancing the expressiveness of the robot while maintaining coherence with its speech. This approach ensures that the gestures effectively support verbal communication without overwhelming or disrupting the interaction. Research has shown that beat gestures consisting of simple rhythmic movements aligned with speech prosody, enhance listener comprehension, emphasize speech rhythm, and increase speaker engagement [21, 17]. In the context of robots, well-timed, human-like gestures improve perceived naturalness, social presence, and communicative effectiveness [29, 15].

Stress words in speech are identified to determine beat gesture placement using a **hybrid approach: LLM-based selection and part-of-speech (POS) tagging**. Beat gestures typically align with content words such as; nouns, verbs, adjectives, and adverbs, as these often carry prosodic stress and semantic load [21, 34]. However, speakers also emphasize emotionally salient or contextually meaningful words, which are not strictly limited to grammatical categories [6, 23]. Since such emphasis varies by speaker and situation, we introduced controlled randomness using a large language model (Chat-GPT 3.5), which selects keywords based on semantic and emotional relevance [3, 19]. POS tagging filters out stopwords and function words, and a 7-word buffer between consecutive gestures prevents overuse, making the Alpha Mini robot look more natural. Additionally, iconic gestures were given priority and they override beat gestures when they co-occur.

LLM-based selection prompts the model to identify key words based on semantic relevance and emotional weight.

Prompt = “Identify the MOST IMPORTANT words that should be emphasized with a small arm or head movement in this text: {text}. Select at most 1 word

per 9 words. Focus on words that carry key meaning or emotion. Do NOT emphasize common nouns, generic verbs, or function words. Return only a comma-separated list of their positions in the text starting from 0."

Once the stress words and their corresponding indices are identified, beat gestures are assigned from a predefined set of motion patterns. To introduce variation, different beat gestures are randomly shuffled, cyclically dis- tributed across stress words.

During execution, the speech and gesture sequences are run in parallel. The robot first initiates its default stance before dynamically adjusting its movements in real time. After each gesture, the robot returns to a neutral position to maintain fluidity and prevent abrupt transitions.

This structured approach ensures that the Alpha Mini robot produces synchronized beat gestures that enhance speech expressiveness while maintaining natural, human-like movements.

Iconic Gestures

The system incorporates a set of iconic gestures to reinforce the semantic content of spoken utterances, enabling the robot to express meaning not just through speech but also through coordinated visual signals. Iconic gestures are defined as those that visually represent attributes of the objects or actions they refer to, such as shape, direction, or movement [21]. They serve a critical role in multimodal communication, especially in human-robot interaction (HRI), where they can enhance comprehension, engagement, and naturalness [6].

In the current system, specific verbal triggers are mapped to gesture templates designed to resemble natural human gestures. These mappings include:

Greetings (e.g., "hello," "hi," "bye"): Initiate a waving motion, distinct from the default robotic wave, with smoother joint articulation and a shorter amplitude to maintain subtlety.

Self-references (e.g., "I," "me," "mine"): Trigger a pointing gesture toward the robot's chest or torso.

Direct references to the listener (e.g., "you," "your"): Cause the robot to extend its arm outward to point toward the presumed listener's direction.

This selection is informed by findings in gesture studies showing that pointing and waving are universally recognized gestures in early language acquisition and interpersonal communication [36]. They are also frequently used in robotic applications due to their clarity and cultural consistency [29].

To synchronize gestures with verbal content, each associated word is assigned a gesture onset time aligned with the beginning of its phonetic realization. Gesture templates are translated into motion sequences consisting of joint po-

sitions and time-stamped frames, enabling the robot to execute movements that are temporally coordinated with speech. This alignment enhances coherence, avoids unnatural delays, and supports the semantic saliency of spoken words.

Moreover, prioritization rules ensure that iconic gestures override beat gestures when both coincide, reflecting their greater semantic load and communicative intent. By combining these rules with smooth, size-proportional movements, the system aims to strike a balance between expressiveness and naturalness, minimizing the risk of uncanny or distracting motion patterns.

C Vocabulary Tests and Questionnaire

C.1 Questionnaire

- Is Dutch your native language?
- If not, what is your native language?
- How many languages do you speak, and what are they?

C.2 Group 1: A1–A2 Vocabulary Pre/Post-Test

Vocabulary: Kitchen

Do these exercises to help you learn words for things you find in the kitchen.

1. Check your vocabulary: picture matching

Write the correct word in the box below the picture.

sink	bowl	table	cupboard	cup
knife	chair	cooker	plate	spoon

			
			

		

C.3 Group 2: B1 Vocabulary Pre/Post-Test

B1	Describing People	TOP006

Complete the sentences with an adjective from the box.

amusing - careless – cautious – cheerful - greedy - healthy – polite – slim - tall - wealthy

1. My girlfriend likes to do sports and eat fresh vegetables. She doesn't smoke so she's a very _____ person.
2. _____ people always want more and more. They are never satisfied with what they have.
3. Since she has gone on a diet and lost 10 kg she has become a _____ young lady.
4. Nobody is more _____ than my little brother. He never picks up anything and throws all his belongings around.
5. Dan and Benny are very _____ when they cross the street. They are afraid of being hit by passing cars.
6. Nothing seems to make grandfather sad. He is such a _____ person and smiles all the time.
7. Jack is not _____ enough to be a basketball player. He's only 1.50 meters.
8. We taught our children to be _____ and always say "please" and "thank you".
9. My uncle likes to tell jokes and entertain people. He's very _____.
10. She has two cars, a large house and always wears the most expensive clothes. She seems to be very _____.

C.4 Group 1 Song Lyrics

This is the kitchen song!

A bowl holds the soup Put it on the table, nice and neat. A cup is small, we drink some tea, And the cupboard keeps things tidy.

A knife can cut some bread or cheese, A chair is used to sit A cooker is used to cook our food. A plate to place our food, And a spoon helps us eat soup.

The sink is where we wash plates, And this is the kitchen song!

C.5 Group 2 Song Lyrics

Cheerful is feeling happy inside.

You smile and laugh and eyes open wide.

Careless is when you don't take care.

You drop your stuff and leave it there.

Cautious means you take it slow.

You stop and think before you go.

Amusing is something that makes you laugh.

Like silly faces and jokes.

Polite kids say "please" and "thank you."

They're kind and nice. Greedy means you want too much.

You keep it all, not sharing much.

Healthy kids eat fruits, not fries

They play, they run, and feel just right.

Slim is when you are thin and healthy

You're quick and light .

Tall is standing high and straight.

You reach up far, it feels so great!

Wealthy means you own a lot.

With toys and games, very rich