Multi-modal Language Learning: Explorations in Japanese Vocabulary learning

Pieter Wolfert
pieter.wolfert@donders.ru.nl
Radboud University, Donders Institute for Brain, Cognition and Behaviour
Nijmegen, The Netherlands

Lisa De Gersem
lisadegersem@icloud.me
IDLab-AIRO, Ghent University - imec
Ghent, Belgium

Ruben Janssens
ruben.janssens@ugent.be
IDLab-AIRO, Ghent University - imec
Ghent, Belgium

Tony Belpaeme
tony.belpaeme@ugent.be
IDLab-AIRO, Ghent University - imec
Ghent, Belgium

ABSTRACT

We explore robot-assisted language learning with a social robot, in which the robot teaches Japanese vocabulary. Specifically, we study if the mode of presentation of referents of nouns influences learning outcomes, and hypothesise that multimodal presentation of referents leads to improved learning outcomes. Three conditions are tested: referents are either presented as Japanese audio only, referents are visually presented, or referents are presented as actual objects that learners could pick up and manipulate. The learners were taught 4 words per condition and were distracted between the conditions with general questions related to the robot. There was a significant difference in the number of learned words between the audio-only and visual conditions, as well as between the audio-only and tactile conditions. No significant difference was found between the visual and tactile conditions.

However, from our study, it follows that both these conditions are preferred over learning through only audio.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**; • **Computer systems organization** → **Robotics**.

KEYWORDS

human-robot interaction, robot-assisted language learning, multi-modal interaction

ACM Reference Format:
Pieter Wolfert, Lisa De Gersem, Ruben Janssens, and Tony Belpaeme. 2024. Multi-modal Language Learning: Explorations in Japanese Vocabulary learning. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640685

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3640685

1 INTRODUCTION

The potential of social robots in educational settings is widely accepted, with robots having the potential to improve both cognitive and affective outcomes, especially in comparison to the social agents that are not physically present [5, 14]. A key component in HRI is multi-modal communication, with multi-modal communication covering both verbal and non-verbal communication. For example, a story-telling robot was found to be more persuasive when it combined non-verbal behaviours with its verbal behaviour [10]. Another example is the study by de Wit et al. [8], where a child-robot language tutoring setup was used to see the effect of the use of gestures during an interaction. Here, gestures positively influenced the learning abilities and the engagement of children with a social robot. However, caution is needed when implementing social behaviours, as it needs to be done right to be effective [14].

Chang et al. [7] divide robots for education into three groups: learning materials, learning companions/pets, and teaching assistants. Mubin et al. [16] in turn, define three main roles a robot can have during a learning activity: tool, peer or tutor. However, using a social robot during a learning activity brings certain challenges. Firstly, social signal processing is not yet advanced enough for all populations [5]. An example being the lack of proper automated speech recognition (ASR) for children in social robots [9]. Secondly, a proper learning experience requires personalisation towards learners [21]. Thirdly, social robots for education are meant to be used with children, which calls for a careful implementation, for example when sharing data with third parties [2]. Social robots have been used for a while for helping with second language learning (SLL), in a tutoring role [6, 12]. In [11], five methods for language teaching are identified: traditional method, direct method, audio lingual method, immersion method, and submersion method. In the traditional method, direct translations going from the native language (L1) to the target language (L2) are provided. In the direct method, all teaching is done in L2. The audio lingual method focuses on speaking and listening over reading and writing. In the immersion method, the learner is taught in L2 only, and immersed in the language. The submersion method is applicable when a learner is surrounded by L2 speakers only (for example, when the learner is in the country where L2 is exclusively spoken). All these methods will eventually offer the learner the necessary combination of both vocabulary and grammar.

In this study, we focus specifically on the learning of vocabulary, and in specific nouns. It has been shown that second language learners can better memorise words when these are combined with visual cues, which has become readily available with printed and digital learning material [15, 18]. These findings match with learning style preferences individuals have. Oxford [17] identifies four main categories for learning: auditory, visual, kinaesthetic (movement-oriented) and tactile (touch-oriented). In this paper, we explore the use of the Furhat social robot as a tutor for teaching second language vocabulary. We look at how incorporating three modalities —auditory (listening), visual (reading), and tactile (touching and doing, with real objects)— influences word learning.

2 RELATED WORK

Language learning with robots (RALL) goes back almost 20 years. One of the first studies, taking place in Japan, used two identical robots. These were placed outside a classroom on a corridor. In total, 119 first-grade and 109 sixth-grade students interacted with the robots. Learners were recognised using RFID tags such that the robot could personalise its interaction. The robots knew 50 English words and could reply using a limited set of basic utterances. Over a period two weeks learners significantly improved their knowledge of English vocabulary. A follow-up study looked at what would happen when the robots would adapt and personalise their behaviour directed at the children. They found that personalised behaviour led children to see the robot as a friend [13]. A 2014 study made use of the NAO robot to teach English as a second language [1]. They found that children in the RALL condition were able to learn and retain vocabulary better in comparison to children that were taught vocabulary using traditional methods. These studies show the potential of robots that (co)teach the vocabulary of a second language.

A 2017 study compared second language word learning with a tablet against learning with real objects [20]. The participants, young Dutch-native speaking children with a mean age of 5 years, were assigned to either a tablet or physical object condition. The children were told a story with a total of six target words (in English, i.e. 'heavy', 'light', 'full', 'empty', 'in front of', and 'behind'). The rationale behind these words was that children would benefit from the physical interactions with objects that were associated with one of these six words. Children had to repeat the target words and perform simple actions with objects. Against expectation, no significant differences between word learning on a tablet or with physical objects were found. The authors hypothesised that the concepts to which the words refer would already be known by the children in their native language, and that the physical manipulation of objects did not contribute to L2 learning. However, another study by Bara and Kaminski [3] that compared vocabulary learning in Rwandan children (in the range from five to 10 years), showed that children who were learning L2 vocabulary with physical objects had a higher memorization rate than when children learned with pictures. Key differences between the studies were a higher number of participants, and an older target group (mean age of 88.8 months versus 60.6 months for the tablet study), and a larger number of learned words. These contradicting findings call for further research.

Figure 1: Example dialogue interaction in condition 3. On the left, a robot is visualized, with a participant on the rights.

3 METHODS

We use a within-subject design to teach 12 Japanese words to adult learners. This section describes the procedure and study materials used.

3.1 Study design

The three conditions are explained in more detail below:

(1) Condition 1: auditory – This condition uses speech only and is implemented as a simple two-way interaction between the learner and robot. To overcome the problem of unnatural switches between voices of the English and Japanese text-to-speech (TTS) engines, the Japanese lexicon is always placed at the end of each sentence, e.g. *"What is the English translation for いちご?"*.
(2) Condition 2: visual – In addition to the spoken interaction from condition 1, an image and word is shown on a tablet positioned next to Furhat.
(3) Condition 3: tactile – In addition to the spoken interaction from condition 1, the robot now invites learners to pick and show objects to the robot.
An example of the interaction in condition 3 can be found in Figure 1.

3.2 Procedure

Participants provided informed consent before the study[1]. The participants were told that they will see three lessons on Japanese vocabulary, each using a different teaching style. Each lesson was followed by a short quiz to assess recall. Between each lesson, to distract the participant, a number of questions were asked: "Did you like this way of learning a new language? Is the robot loud and clear enough?", "What could be better? Would you like more repetitions?", and "Do you think that you learned a lot? Could you focus on on the task at hand?". The order in which the conditions were shown was randomized and balanced between participants.

Four Japanese words were offered per condition, as pilot studies showed that more words could lead to an unacceptable cognitive load and disappointing learning outcomes. In addition, a longer experiment was deemed tedious by the participants. The Japanese vocabulary came from four categories: fruit, electronics, utensils,

[1]The study complies with the guidelines of Ghent University in regard to ethics and data collection

Table 1: Japanese vocabulary per condition

Auditory	Visual	Tactile
Strawberry いちご - *Ichigo*	Watermelon スイカ - *Suika*	Apple 林檎 - *Ringo*
Washer 洗濯機 - *Sentakki*	Fridge 冷蔵庫 - *Reizōko*	Telephone 電話 - *Denwa*
Kitchen knife 包丁 - *Hocho*	Glass コップ - *Koppu*	Spoon 匙 - *Saji*
Scissors はさみ - *Hasami*	Book 読書 - *Dokusho*	Toothbrush 歯ブラシ - *Haburashi*

and other. Table 1 shows the lexicon per learning condition used in randomized order.

The robot runs both the lesson and the evaluation. All lessons had a similar structure, with the only variation being due to the used modality. During a lesson, the robot tutor first mentioned an English word, provide its Japanese translation and ask the learner to repeat it. In the visual condition, a picture and the learned object's English name were shown on a tablet next to the robot. For the tactile condition, the robot asked the learner to pick up and show the object. The robot tested the participants' recall using a Japanese to English translation task. For this, the learners replied by using speech (auditory), the tablet (visual) or through showing an object (tactile). After the experiment finished, a debriefing followed. Participants were asked to fill in a questionnaire to get insights in their perceived language skills and were asked to fill out a Godspeed questionnaire [4].

3.3 Participants

Native Dutch speaking Adult participants were recruited through social media. The experiment took place on campus at the Ghent University over a period of two weeks, and participants were allowed to book a 30 minute time-slot convenient to them. Participants gave themselves on average a 4.4 (SD=0.54) out of 5 on proficiency in English. On average, participants gave themselves a 3.1 (SD=1.00) out of 5 on the ability of learning a new language. Of the 43 participants, 9 stated that they knew some Japanese. We excluded these participants from our study. 34 people participated in the study ($F = 12, M = 22$, mean age = 28.32 (SD = 11.96)).

3.4 Materials

The Furhat from *Furhat Robotics* [2] was chosen as platform. It can show human-like facial expressions on its projected face and has advanced conversational capabilities. It supports multi-language conversations, an essential feature for language tutoring, and provides access to several TTS and speech-to-text (STT) engines. In addition, it has a built-in camera (1080p RGB 135° diagonal field of view) and an in-built speaker.

[2]https://furhatrobotics.com/

Figure 2: Setup as used in the experiment

For the visual condition, we used a consumer grade tablet that displayed a website that was linked to Furhat. The setup used during the experiments is depicted in Figure 2.

The visual object recognition, necessary in the tactile condition, used the YOLOv5 network (pre-trained on MS COCO dataset) to recognise objects in camera frames. Processing was done remotely, and camera images and processing results were communicated to the robot using a TCP socket connection. To trade off speed and accuracy, the medium YOLOv5 network was chosen and to further reduce the latency, the model ran on a machine connected to the local network and was only applied every ten frames. In addition, the MS COCO classes were filtered on the objects that were used during the interaction to limit wrong detections

3.5 Measurements

We compared the effect of the three learning conditions on Japanese word recall and on the learning experience. The objective measure is the number of Japanese words a subject has learned per condition. This was followed by the Godspeed questionnaire for future reference.

4 RESULTS

A repeated-measures ANOVA was conducted to examine the impact of three distinct learning conditions on the number of learned words. The analysis revealed a statistically significant difference in the number of learned words across conditions ($F(2, 78) = 5.174$, $p < 0.001$). While there was no significant overall effect of the order of presentation ($F(5, 78) = 1.981$, $p = 0.09068$), the interaction between condition and order was also not significant ($F(10, 78) = 0.830$, $p = 0.60159$). These findings suggest that the observed differences in learning outcomes across conditions are consistent across various orders of presentation.

Subsequent Tukey's Honestly Significant Difference (HSD) post hoc tests were employed for multiple comparisons. These tests demonstrated significant differences in the mean number of learned

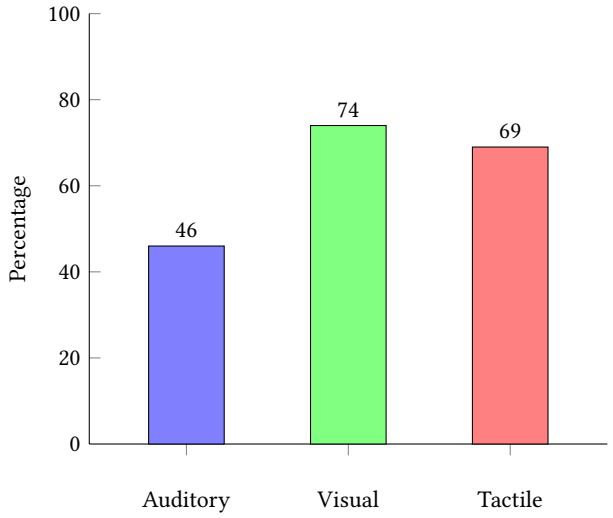

Figure 3: Bar chart showing the average percentage of remembered words per condition.

words between condition 1 (auditory) and condition 2 (visual) ($p < 0.001$, 95% C.I. = [0.57, 1.66]) and between condition 1 (auditory) and condition 3 (tactile) ($p < 0.001$, 95% C.I. = [0.38, 1.48]). However, there was no statistically significant difference observed between condition 2 (visual) and condition 3 (tactile) ($p = 0.69$).

4.1 Survey Results

We had participants filling in the Godspeed Questionnaire after the session. The robot scored on average a 3 (SD=1) on anthropomorphism (out of 5). It scored a 3 (SD=1) on animacy (out of 5). It scored a 4 out of 5 (SD=1) on like-ability. On perceived intelligence it scored a 3 (out of 5) (SD=1) as well as on perceived safety.

We asked participants which condition they preferred. Most participants preferred the visual condition (condition 2) (with 50%). This was followed with the tactile condition (with 40.5%). The auditory condition was least preferred with 9.5%.

5 DISCUSSION

While we expected the tactile condition to lead to better word learning than the visual condition, as reported in earlier work by Bara and Kaminski [3], we did not find this. However, we did find differences between those conditions and the auditory condition, suggesting that adding the visual modality already brings benefits which are not improved upon by adding a tactile/physical modality.

Participants mentioned the lack of feedback and the lack of repetition. We chose to not use repeated exposure for practical reasons, but agree that repetition or the ability to request repeated demonstration of a word would be needed in a practical L2 tutoring application.

Participants also mentioned the lack of feedback on pronunciation. In future work, we would like to implement feedback in phonetic writing and how well the pronunciation was per word. This could help participants in understanding how well they did, and which words they must practice more. In our scripted interaction, we did not consider for pauses between the different interactions. However, this time could be needed for participants to process the presented information. These pauses could be implemented after each word, so that participants have more time for memorization.

During the design of the experiment, the medium YOLOv5 object detection model was chosen instead of a larger one based on speed requirements and the decision to run it in a local network, rather than on a remote GPU server, to limit additional delays and to retain the fluency of the interaction. However, this choice caused a higher error rate on the object detection side, which potentially could break the interaction flow. The primary issue was the numerous false positives for the telephone class and the high number of false negatives for the spoon class.

To improve object detection, one could take a larger version of the object detection network to reduce the amount of detection errors. The bad placement and quality of our microphone that was part of the setup resulted in participants needing to repeat themselves multiple times. This mainly occurred with short words and silent speech, which was often not captured by the microphone. On particular occasions, the microphone was placed close to the robot, which subsequently suffered from noise pollution by the robot's ventilation system. Another issue is that the TTS and STT engines are not equipped for grading the participants' speech, which is important for L2 learning. It also creates the need for a Wizard-of-Oz approach, which goes against the idea of an autonomous social robot for second language tutoring. The lack of a multilingual TTS makes that a participant hears two different voices, as often the same voice is not available for multiple languages. As the perceived personality also relies on the voice of a person, and also for a social robot, this could create an uncanny effect and could harm learning.

We acknowledge the potential influence of our study design, wherein different words were utilized across conditions, which may have introduced a subtle degree of variability in our results. Additionally, the large age range of our participants might have had an effect on our results. However, we believe that sharing these findings remains valuable to the research community. Transparency about this aspect of our design allows for an open dialogue about its implications.

Finally, future work should take account of the cognitive load and engagement of a participant, as both matter a lot in terms of learning ability. However, keeping track of this might make the use of other equipment necessary (e.g. to measure a participant's stress response) [19].

6 CONCLUSION

In this paper, we used the robotic head *Furhat* to implement an autonomous L2 tutor for Japanese lexicon. We implemented three well-known learning modalities – auditory, visual and kinesthetic – by extending Furhat's functionality with a graphical user interface and providing it with awareness of the objects in its field of view.

Our autonomous robot tutor's measured academic achievement and enjoyment were best when employing the visual and kinesthetic learning modalities. Results indicate that it is vital to explore the addition of visual representations and tactile stimuli in language tutoring setting to optimize personalized learning experiences.

REFERENCES

[1] Minoo Alemi, Ali Meghdari, and Maryam Ghazisaedy. 2014. Employing humanoid robots for teaching English language in Iranian junior high-schools. *International Journal of Humanoid Robotics* 11, 03 (2014), 1450022.

[2] Minoo Alemi, Ali Meghdari, and Maryam Ghazisaedy. 2015. The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics* 7, 4 (2015), 523–535.

[3] Florence Bara and Gwenael Kaminski. 2019. Holding a real object during encoding helps the learning of foreign vocabulary. *Acta Psychologica* 196 (2019), 26–32.

[4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[5] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018), eaat5954.

[6] Tony Belpaeme, Paul Vogt, Rianne Van den Berghe, Kirsten Bergmann, Tilbe Göksun, Mirjam De Haas, Junko Kanero, James Kennedy, Aylin C Küntay, Ora Oudgenoeg-Paz, et al. 2018. Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics* 10, 3 (2018), 325–341.

[7] Chih-Wei Chang, Jih-Hsien Lee, Po-Yao Chao, Chin-Yeh Wang, and Gwo-Dong Chen. 2010. Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Journal of Educational Technology & Society* 13, 2 (2010), 13–24.

[8] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2018. The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 50–58.

[9] Olov Engwall, José Lopes, and Ronald Cumbal. 2022. Is a Wizard-of-Oz Required for Robot-Led Conversation Practice in a Second Language? *International Journal of Social Robotics* (2022), 1–19.

[10] Jaap Ham, Raymond H Cuijpers, and John-John Cabibihan. 2015. Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics* 7, 4 (2015), 479–487.

[11] Trevor A Harley. 2013. *The psychology of language: From data to theory*. Psychology press.

[12] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction* 19, 1-2 (2004), 61–84.

[13] Takayuki Kanda, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. 2007. A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on robotics* 23, 5 (2007), 962–971.

[14] James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics* 7, 2 (2015), 293–308.

[15] Farzad Mashhadi and Golnaz Jamalifar. 2015. Second language vocabulary learning through visual and textual representation. *Procedia-Social and Behavioral Sciences* 192 (2015), 298–307.

[16] Omar Mubin, Catherine J Stevens, Suleman Shahid, Abdullah Al Mahmud, and Jian-Jie Dong. 2013. A review of the applicability of robots in education. *Journal of Technology in Education and Learning* 1, 209-0015 (2013), 13.

[17] Rebecca L Oxford. 2003. Language learning styles and strategies: An overview. Gala Oxford.

[18] Karim Sadeghi and Bahareh Farzizadeh. 2013. The effect of visually-supported vocabulary instruction on beginner EFL learners' vocabulary gain. *Mextesol Journal* 37, 1 (2013), 1–12.

[19] Alexander Skulmowski and Günter Daniel Rey. 2017. Measuring cognitive load in embodied learning settings. *Frontiers in psychology* 8 (2017), 1191.

[20] MAJ Vlaar, Josje Verhagen, Ora Oudgenoeg-Paz, and PPM Leseman. 2017. Comparing L2 Word Learning through a Tablet or Real Objects: What Benefits Learning Most?

[21] Hansol Woo, Gerald K LeTendre, Trang Pham-Shouse, and Yuhan Xiong. 2021. The use of social robots in classrooms: A review of field-based studies. *Educational Research Review* 33 (2021), 100388.