Using a Robot Peer to Encourage the Production of Spatial Concepts in a Second Language

Conference Paper · December 2018

DOI: 10.1145/3284432.3284433

7 authors, including:

Christopher David Wallbridge
University of Plymouth
6 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Rianne van den Berghe
Utrecht University
8 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Daniel Hernandez Garcia
University of Plymouth
16 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Junko Kanero
Sabanci University
13 PUBLICATIONS **78** CITATIONS
SEE PROFILE

Using a Robot Peer to Encourage the Production of Spatial Concepts in a Second Language

Christopher D. Wallbridge
University of Plymouth
Plymouth, UK
christopher.wallbridge@plymouth.ac.uk

Rianne van den Berghe
Utrecht University
Utrecht, Netherlands
m.a.j.vandenberghe@uu.nl

Daniel Hernández García
University of Plymouth
Plymouth, UK
daniel.hernandez@plymouth.ac.uk

Junko Kanero
Koç University
Istanbul, Turkey
jkanero@ku.edu.tr

Séverin Lemaignan
Bristol Robotics Laboratory
Bristol, UK
severin.lemaignan@brl.ac.uk

Charlotte Edmunds
University of Plymouth
Plymouth, UK
charlotte.edmunds@plymouth.ac.uk

Tony Belpaeme
Ghent University/University of Plymouth
Ghent, Belgium
tony.belpaeme@ugent.be

ABSTRACT

We conducted a study with 25 children to investigate the effectiveness of a robot measuring and encouraging production of spatial concepts in a second language compared to a human experimenter. Productive vocabulary is often not measured in second language learning, due to the difficulty of both learning and assessing productive learning gains. We hypothesized that a robot peer may help assessing productive vocabulary. Previous studies on foreign language learning have found that robots can help to reduce language anxiety, leading to improved results. In our study we found that a robot is able to reach a similar performance to the experimenter in getting children to produce, despite the person's advantages in social ability, and discuss the extent to which a robot may be suitable for this task.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Social and professional topics** → **Assistive technologies**; • **Computing methodologies** → Natural language processing; Cognitive robotics;

KEYWORDS

Robot Assisted Language Learning; Assessment; Second Language Learning

HAI '18, December 15–18, 2018, Southampton, United Kingdom

ACM ISBN 978-1-4503-5953-5/18/12…$15.00
https://doi.org/10.1145/3284432.3284433

ACM Reference Format:
Christopher D. Wallbridge, Rianne van den Berghe, Daniel Hernández García, Junko Kanero, Séverin Lemaignan, Charlotte Edmunds, and Tony Belpaeme. 2018. Using a Robot Peer to Encourage the Production of Spatial Concepts in a Second Language. In *6th International Conference on Human-Agent Interaction (HAI '18), December 15–18, 2018, Southampton, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3284432.3284433

1 INTRODUCTION

Learning the language of a new home region is vital for migrant children. It is beneficial for them to integrate with their peers, and necessary to prevent them from falling behind in school. Children need the opportunity to practice their language skills, but it may be difficult if no one at home is able to speak the language of the host region. Finding qualified teachers or tutors that know both the new language and the language of children's old homeland can also be challenging. With robots we may be able to support children's language learning needs.

When learning a second language (L2), it is difficult to master vocabulary both receptively and productively. L2 learners may find themselves capable of understanding the L2, while still struggling to produce L2 words. Indeed, previous research has shown that receptive vocabulary tends to be bigger than productive vocabulary in first language (L1) [8, 11], and that L2 learners obtain lower scores on productive tests as compared to receptive tests [14]. Thus, people are able to recognize more words than they can produce, both in their L1 and L2. This has been formalised into a hierarchy for word knowledge by Laufer et al. [9], based on knowing the words passively or actively and in being able to recognize them or recall them. The hierarchy is as follows, from easiest to most difficult: passive recognition → active recognition → passive recall → active recall. These are defined as follows:

- *Passive recognition* - The student is able to select the L1 word from a choice of words when provided the word in L2.

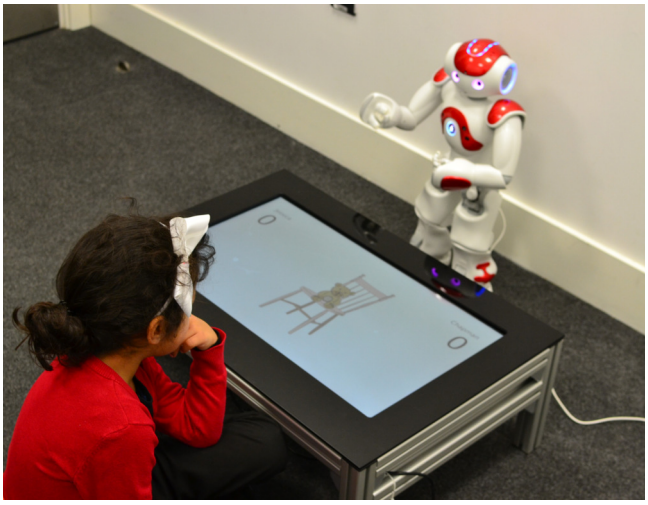

Figure 1: A child interacting with the robot in our study. The agent – in this case a robot – stands opposite from the child. An interactive table displays an image of a teddy bear and a chair. The child must use a word from a second language to describe the position of the bear in relation to the chair.

- *Active recognition* - The student is able to select the L2 word from a choice of words when provided the word in L1.
- *Passive recall* - The student is able to give the meaning of a word in L1 when provided the word in L2.
- *Active recall* - The student is able to give the L2 word when provided the word in L1.

This poses a challenge for L2 vocabulary interventions in which the trainer wants to assess the trainee's learning gains: L2 learners have difficulty learning the words productively (i.e. learning to produce foreign words), and will struggle to actively recall newly learned L2 words. There are several tests to assess an L2 learner's productive vocabulary, including assessments in which the participant has to describe pictures (e.g., the Expressive Vocabulary Test [18], the Expressive One-Word Picture Vocabulary Test [5], or the Clinical Evaluation of Language Fundamentals Test [17]), writing tests in which the learner has to fill in the blank (e.g., the Productive Vocabulary Levels Test [10]), or, for very young children, parental or teacher reports [4].

In many situations, it may not be possible to use one of these tests. For example, when the words learned concern abstract concepts, which cannot be easily depicted, it is not possible to use a picture test. If the learner is illiterate, one cannot use a writing test. Parents or teachers may struggle to report the child's L2 if they do not speak that language themselves. To further complicate the issue, producing L2 words may be intimidating for L2 learners. Even if the learner is able to produce the word, they may not produce it due to anxiety of pronouncing the word incorrectly [13].

A social robot may help overcome some of the issues described above in assessing L2 learner's vocabulary. While not being able to solve by itself the issue of vocabulary being more difficult to learn productively than receptively, a social robot may help in innovating novel ways to assess L2 vocabulary, or in reducing L2 anxiety in L2 vocabulary test settings. A robot may be less intimidating than an adult assessor, especially for young children, encouraging more speech production. This study evaluates whether school children may produce more L2 words in a productive L2 vocabulary test when playing with a social robot than with an adult. Below, we discuss relevant robot-assisted language learning (RALL) studies before detailing our study.

2 PREVIOUS WORK

RALL has been found to be effective in reducing foreign language anxiety (FLA), and teaching robots are able to improve oral skills of young students learning English as a foreign language [1]. Alemi et al. [2] performed a study using a robot teaching assistant. In the study, Persian-speaking students in Iran were taught English. A survey of the students showed that those who learned from the robot were significantly less anxious compared to the control group that did not have the robot. While a number of factors were thought to contribute to this reduction in anxiety, the authors claimed a major reason to be intentional mistakes the robot made. The mistakes not only gave the students a chance to correct the robot, but also made them less afraid of making errors of their own.

When looking at speaking skills, the focus can not just be on vocabulary gains, but pronunciation as well. Lee et al. [12] conducted a series of lessons to help Korean children from grades 3 to 5 (roughly 8 to 10 years old) learn English. In South Korea children start learning English from grade 3. As part of a lesson series they were given a pronunciation training with a robot, that used a lexicon that included often confused phonemes, so that the robot could correct the child's pronunciation. It was reported that the children's speaking skills improved significantly with a large effect size when measured by a teacher. As well as the improvement in speaking skills all three affective factors – interest, confidence and motivation – all improved significantly.

Instances of robots acting as care-receivers also occur in RALL. In a study by Tanaka and Matsuzoe [16], Japanese children were given the role of teaching English verbs to a NAO robot. The children had to guide the robot's arm to act out the target verbs, e.g. brushing teeth. In a comprehension post-test the children answered correctly more often with words they had taught the robot than those learnt during a regular verb-learning game. While the robot only learned from 'Direct' teaching, where the child was guiding the motion of the robot, there was a high frequency of verbal teaching using English.

We can see that there are many instances where RALL is able to assist in teaching an L2 to students. Many of these show a reduction in FLA and increase in confidence and willingness to learn in the students. In all these cases, however, they use the robot to teach, whether directly in the role of teacher or acting as a care receiver or assistant. Robots were not used in assessment, and in most cases the tests performed were aimed at measuring the comprehension of the L2 words that were being taught. We want to explore the possibility of using a robot to assess the L2 production of children. Due to the reported reductions in anxiety and increase in confidence when using a robot, we may see an increase in the amount of production.

3 STUDY DESIGN

This study was conducted at a local school with English-speaking 5- to 6-year-old children. We decided to teach spatial language, more specifically spatial prepositions, because while those concepts are more abstract than physical objects, we can still represent them using images. Spatial language itself is also particularly challenging to L2 learners as the meaning can often differ depending on context and the referent. Every morning, five children were randomly selected to participate in the study for that day and assigned a condition, balanced across gender. These five children were first given a French lesson before playing our production quiz game on an interactive table [3] individually throughout the rest of the day (Figure 1). An agent (robot or experimenter depending on our condition) is placed opposite to the child and gives instructions and encouragement to the children. The interactive table displays an image of a teddy bear and a chair. The child would have to use one of the French words taught to describe the position of the bear relative to the chair.

As well as the teacher three experimenters were involved in the study:

(1) *Lead Experimenter* - The lead experimenter acted as the interaction point for the children outside of the one to one sessions. Either the lead experimenter or the wizard was required to be in the presence of the child while outside their classroom. The lead experimenter was certified in the children's health and well being, and was there to ensure the health and safety of the children as required by the school.
(2) *Wizard Experimenter* - The wizard experimenter controlled the robot remotely via a laptop interface. The wizard experimenter was also certified in the children's health and well being, but had minimal interaction with the children so as to minimise interference during the study.
(3) *Blind Experimenter* - The blind experimenter facilitated the interactions before the main study began, provided the comprehension test and acted as the agent in the child-human condition. The blind experimenter was unaware of the purpose of the study to reduce influencing the outcome.

3.1 Hypothesis

With our study we wanted to test the following hypothesis:

H The presence of a robot will allow children to produce more spatial words verbally in an L2 than when working with a human experimenter.

3.2 Teaching

The children were taught five French words: *Nounours* (Teddy Bear), *chaise* (chair), *devant* (in front of), *sur* (on), *sous* (under). Of these, the first two were supporting words and the last three were the target words for the study. The content of the lesson was created and taught by a professional French teacher, with a goal of enabling the children to produce these words after one lesson. We decided to use a professional teacher as we did not want a robot teacher that would also influence our results. It has also been shown that human teachers can still outperform a robot teacher [7]. The lead experimenter acted as a teacher's assistant. The children were taught in groups of five. The lesson was designed to last 30 minutes.

The teacher started the lesson by introducing the children to the support words. At all stages the children were encouraged to repeat any French words they heard. The children were taught a song that used the three target words and hand gestures to go along with them. After singing, the children would position themselves relative to the chair based on the words announced by the teacher. The children were then each given a teddy bear and repeated the process with the bear. The children then played a game of 'Telephone'. In this game one child was first given one of the target words, and each child would whisper the word to the next child down the line until the last child. The last child would announce to the rest of the group the word they heard. The game was repeated several times with the children re-organised into a different order so that the announcing child changed each time. This was followed by a game of 'Corners'. In each corner of the lesson area, a teddy was placed in a position relative to a chair that referred to one of the target words. The children were then encouraged to sing and move around until the teacher would stop them, and say one of the target words. The children then had to move to the relevant corner and say the word three times. Variants of this game were then played in teams with the chairs lined up, and then individually. Finally each child was told to say one of the target words and then go stand by the correct chair. The lesson wrapped up with one more repetition of the song they had been taught near the beginning.

During the interaction we also established any prior knowledge in the target language. They were split into the following categories:

(1) *No Exposure* - The children have not been exposed to any French, other than potentially those used in popular culture e.g. *C'est la vie*.
(2) *Beginner* - The child has potentially received some lessons in French and knows simple phrases that do not include our target words e.g. *Je m'appelle John*.
(3) *Intermediate* - The child has knowledge of French, including our target words.
(4) *Advanced* - The child has an intricate knowledge of French, and is able to produce words with a high capability or are fluent.

Children of intermediate or advanced knowledge were excluded from the data analysis. 25 children took part in our study of which three were excluded from the analysis of results, leaving 22 children.

3.3 Individual Interactions

Upon completing another familiarity task and a 10 minute activity with the robot–that required the child to describe the position of objects to the robot in English–a comprehension test was administered by a blind experimenter who was unaware of the purpose of the study (Figure 2). This served as a small refresher of what the children had learned earlier in the day, as well as allows us to establish a baseline for the efficacy of the lesson. For the comprehension test there were 6 sheets with 3 images each (representing the 3 target words), placed on the left, in the centre or on the right. Together, the 6 sheets covered all possible permutations of the 3 target words (*devant*, *sur*, *sous*) with each of the 3 positions. The images were similar but not the same as the ones used for the production quiz questions. For each sheet the experimenter asked the child to point at the picture that matches the statement (see below). If the

Figure 2: A child being administered the comprehension test before moving onto the main production quiz.

Figure 3: The 'wizard' experimenter was positioned behind the child to minimise interaction between them.

child pointed to the wrong picture they were allowed to try again until they pointed to the correct image. We repeated each target word twice to account for guessing and to ensure they weren't just picking based on location on the question sheet. The statements and their order were the same for every child:

(1) Le nounours est sous la chaise.
(2) Le nounours est devant la chaise.
(3) Le nounours est sur la chaise.
(4) Le nounours est devant la chaise.
(5) Le nounours est sur la chaise.
(6) Le nounours est sous la chaise.

The child then played the production quiz with either the robot or the blind experimenter based on the group they were in (child-robot or child-human). In both conditions, the production quiz was displayed on the sandtray. The robot was controlled through a Wizard-of-Oz interface, with the 'wizard' sat behind the child, out of sight, so as to minimise effects on the child (Figure 3). The rules of the game were explained by the agent (blind experimenter or robot). The child was sat in front of the sandtray upon which the production quiz game was displayed. The agent sat opposite the child. The sandtray displayed an image of the teddy bear in a position relative to the chair, and the agent or child must answer "Où est le nounours?" (Where is the teddy bear?). The agent was to give the answer in the form "sur/sous/devant la chaise", but any answer given by the child that included one of the target words 'sur', 'sous' or 'devant' was accepted. Each correct answer scored a point. If either the question was answered correctly or both the child and the agent answered incorrectly then the production quiz moved onto the next question. If the child did not answer after a short period then the agent would give encouragement in proceeding levels:

(1) Encourage the child to guess e.g. "Just have a guess".
(2) Targeted encouragement, such as asking them to remember the lesson from the morning.
(3) The agent will attempt the question.
    - If the child was ahead on points then the agent (adult/robot) would answer correctly so as to keep up an appearance of a challenging opponent in the game.
    - If the child was level or behind the agent (adult/robot) then the agent would answer incorrectly to demonstrate a willingness to answer even if wrong.

If the child still did not have a guess after all stages then the game proceeded as if they had answered incorrectly. The agent began the production quiz after explaining how to play by answering the first question correctly. There were nine subsequent questions which we expected the child to answer, three for each target word.

4 RESULTS

4.1 Participants

25 children took part in our study of which three were excluded from our analysis of results leaving us with 22 children. 11 Children were in the Human Condition (4 Female) and 11 in the Robot Condition (6 Female). There were 11 5 year olds (6 Female) and 11 6 year olds (4 Female). Of these children two had an L1 other than English (1 Female), but their English level was high enough to still participate.

4.2 Comprehension

We scored the comprehension test by taking the maximum attempts per question (3) and subtracting the number of attempts they took to get the correct answer. This meant each question was scored between 0 and 2, giving a maximum possible score of 12 on the comprehension test. The mean score for the comprehension test was 8.5 (SD=1.92). In the Human condition the children averaged 8.27 (SD=2.20) at the comprehension test while in the Robot condition the children averaged 8.72 (SD=1.68). Using a Welch Two Sample t-test, no significant difference between the two conditions was found ($t = 0.55$, $df = 18.72$ $p = 0.59$). This shows that the groups between our two conditions were roughly equal in ability before beginning the

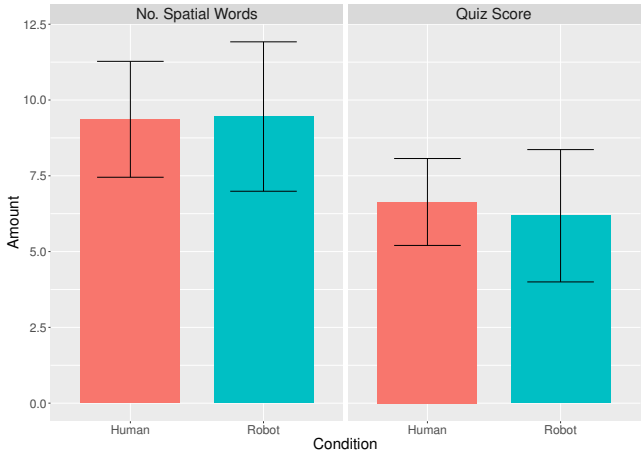

Figure 4: Analysis of L2 spatial words used during the production quiz. Left: spatial words used without additional prompting to attempt the question; right: number of correct words said by the children during the production quiz. In both cases no significant difference was found between the robot and adult conditions. Error bars are showing the standard deviation.

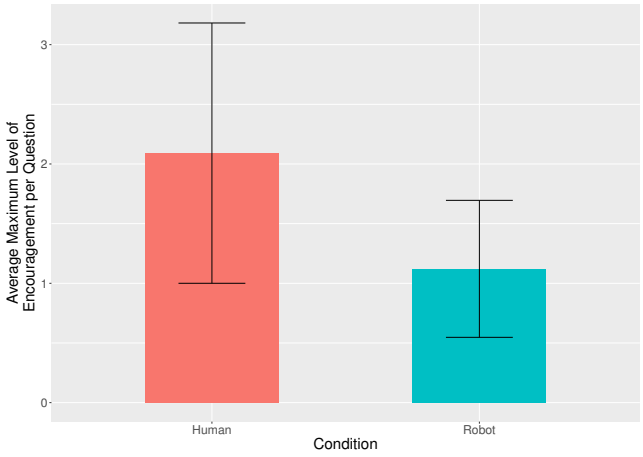

Figure 5: Analysis between participants of the average maximum level of encouragement reached across conditions. A significant difference is seen between the two conditions, Human and Robot. Error bars are showing the standard deviation.

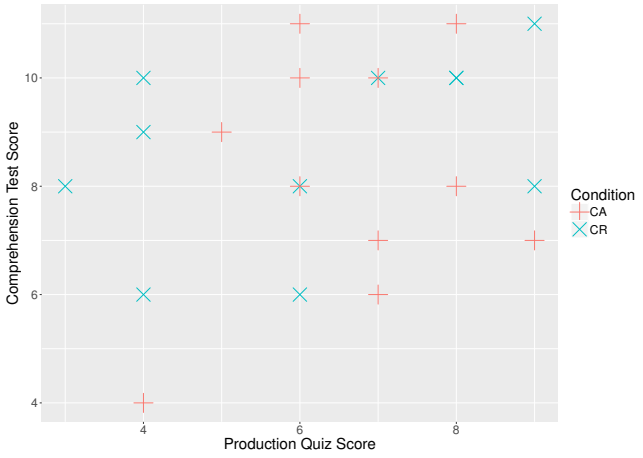

Figure 6: A comparison between the score in the production quiz and the score on the comprehension test. No significant correlation was found.

production quiz. The scores remained consistent throughout the test, with no learning effect seen when the first half and the second half of the comprehension test were compared (first half: mean=4.5, SD=1.26; second half: mean=4 SD=0.93; t=1.50, df = 38.51, p=0.14).

4.3 Production

Children in the child-human condition scored M=6.64 (SD=1.43) out of 9 on the production quiz and M=6.18 (SD=2.18) in the child-robot condition. Using a Welch Two Sample t-test no significant difference between the two conditions was found (t=-0.58, df =17.27, p=0.57).

We also analysed the total number of spatial vocabulary used in L2 (Figure 4). Due to a break in protocol, children were sometimes prompted to attempt a question again instead of moving on in the production quiz. As such our analysis is on words used without being prompted for an additional attempt. In the Robot condition, the children averaged M=9.45 (SD=2.46) spatial words, compared to M=9.36 (SD=1.91) in the Human condition. Using a Welch Two Sample t-test no significant difference was found (t=0.10, df=18.4, p=0.92).

Finally we analysed the amount and level of encouragement given (see levels in Section 3.3). While encoding encouragement given to the children we added a fourth level for analysis of the results:

(4) Encouragement is given that changes or disrupts the task, e.g. telling the child that the current question is the same as a previous one.

The mean amount of encouragement given was M=12.36 (SD=7.46) in the Human condition and M=13.09 (SD=7.78) in the Robot condition. No significant difference was found between the conditions (p=0.83). However we see a significant difference in the average maximum level of encouragement per question across the two conditions (Robot: M=1.12, SD=0.57. Adult: M=2.09, SD=1.09, p=0.02). This is strongly influenced by the amount of level 4 encouragement given by the adult, of which we see 33 instances across 10 children. We see a significant difference between the average amount of level 4 encouragement given per child between the amount given in the first half of the study compared to the second showing an increase in deviation from the protocol over time (First Half: M=1.25, SD=.0.88. Second Half: M=4.25, SD=2.64, p=0.04).

4.4 Comprehension and Production

The data we collected also provided us with an opportunity to test the predictions of Laufer et al. [9], a key foundation for our research.

By looking at the children's scores on comprehension (passive recognition) and production (active recall) we should be able to see evidence of a hierarchy, where comprehension is required for production.

Across both conditions the children had an average score on the production quiz of 6.41 (SD=1.82) out of 9 and is significantly above chance (p=0.03). A positive but non-significant correlation was found between the comprehension test score and their production quiz score (Pearson's r=0.29, p=0.19). The lack of a significant correlation suggests that abilities in comprehension and production are not directly related.

We marked a child as having achieved comprehension on a particular word if they required less than four attempts across the two relevant questions in the comprehension test. For example if we were looking at whether a child could comprehend the word 'sur' we would look at the number of attempts they took for questions three and five. If a child takes two attempts on question three and one attempt on question five their total number of attempts for 'sur' would be three. We would mark this child as being able to comprehend 'sur'. We marked a child as being able to produce a word if they scored at least two points in the production quiz on the three relevant questions. Using Guttman's Coefficient of Reproducibility (reported in Table 1), we were unable to find a hierarchy. A hierarchy would show that comprehension is needed for production. Guttman's Coefficient measures whether such a hierarchy exists based on the number of deviations from that hierarchy. A coefficient of over 0.9 is expected to display such a hierarchy.

	Sur	Sous	Devant
No. Deviations	5	3	4
Guttman's Coefficient λ_4	0.11	0.57	0.56

Table 1: Table detailing the number of deviations from the expected hierarchy and the Guttman's Coefficient of reproducibility. In the case of all three words, we fail to meet the reliability expectation of 0.9

5 DISCUSSION

5.1 Effectiveness of the robot to support L2 production

While this study does not show statistical improvement to a child's ability to produce by using a robot over a person, it does show a similar performance in this task, with no significant difference between the two conditions being found. It may still be desirable to use a robot to allow standardization and automation of assessment. With a minimal amount of support being provided by an agent, only a narrow set of phrases can be given – otherwise the nature of the task could be changed from production. This can make interactions very repetitive for the assessor. Though the scores were higher than expected it still proved to be a challenging task for the children. With the minimal amount of support available to an experimenter it could be emotionally stressful to be unable to intervene when a child is finding the task difficult.

The scores from the production quiz are higher than we expected. From the literature we expected L2 production to be difficult for the children, and our expert tutor believed that it would take two to three sessions for most children to produce at all. The observed prowess of the children may be partially explained by the design of the lessons, directly aimed at encouraging the children to produce the target words for this study. It should be noted that most productions were only single words. Only two children produced any of the support words (*nounours* – teddy bear, and *chaise* – chair).

Several factors may contribute to the high performance of the experimenter. Even within the context of a limited set of responses a person is able to provide much better cues and encouragement based on reading the child. These kind of social skills are still a gold standard to which robotics researchers strive. Though this experiment was conducted using a 'wizard', their position and the time delay in actions for the robot prevented this fine grained social interaction. Some of the cues provided by the experimenter were not programmed into the robot but should be added into its repertoire

(1) *Direct phonetic cues* - Giving part of the word e.g. the starting s.
(2) *Indirect phonetic cues* - Giving clues to the word about how it sounds e.g. "It's the one with a strange sound in it"
(3) *Rhythmic cues* - Giving the syllables of the word e.g. "Duh-dum". This may work well for the small target vocabulary, like ours, where this could refer to a single word, but may be less effective in larger vocabularies.
(4) *Gestural cues* - Movements with the hands that mimic gestures used by the teacher in the lesson.

Despite the more limited social skills of this implementation of the robot, it still achieved a similar performance level to a person. This may be the expected reduction of anxiety, that previous research has shown, balancing the limited social behaviours.

However we also saw a large amount of encouragement given to the children by the blind experimenter that was outside of the original protocol, that could be deemed to have affected the scores of the children in an undesirable way. While in the first half the amount of these encouragements by the experimenter remained low, there was a sharp increase in the latter half. This could be caused by forgetting the protocol over the days of the study or just growing more lax in its use, or even the emotional stress that is put on a person by the children's difficulties.

The presence of a wizard in the room may also have been a contributing factor. The presence of a person, even when not in view, may have prevented the robot from reducing anxiety as much as it could have done, as the child might be aware someone else is listening in. We minimized the affect of the wizard by ensuring there was no reason for them to interact with the children either before the study. Analysis of the videos showed that the majority of children never turned towards the wizard at any point during the study, and focused on the robot. So we believe the impact of the wizard's presence was minimal.

Finally, it must be noted that the school where we performed the study cultivated a much friendlier relationship between adults in the school and the students than is typically seen. This may have made the children feel more comfortable and confident in the presence of our experimenter, reducing anxiety. Future work will

focus on broadening this study to multiple schools to see whether our results can be replicated in different settings.

5.2 Relative difficulty of comprehension versus production

The lack of correlation shown between the production quiz score and the number of attempts on the comprehension test (Figure 6) shows that there was no direct relation between comprehension and production vocabularies. However when we look at the possibility of a hierarchy from comprehension to production we do not find evidence to support a hierarchy. This could have had several causes. While we were hoping to find support within our data, we were not directly testing for this hierarchy. Laufer et al. [9] looked at students 16 years and older at high school and university who had been studying their L2 as part of a national curriculum for between 6 to 9 years. Ours is based on a single lesson focused entirely on being able to say the target words. The younger children in our study may also have been more receptive to learning words productively, as they are still increasing their phonological vocabulary. These skills have been shown to have a correlation with word vocabulary [6]. These factors could account for an increase in deviations from the previously established hierarchy.

6 CONCLUSION

We hypothesized that a robot could surpass human performance in encouraging the production of spatial language: this hypothesis is not supported by our study; however, the robot and the facilitator's performance were very similar, with no significant difference between the two conditions being found. This was despite the greater social ability of the human experimenter. This may be explained by the previous research that shows that robots can make people less anxious in foreign language learning scenarios. Future work expanding the robot's social ability may improve the robot's ability to assess and support a student's learning.

Measuring the production skills of a child at this level is a repetitive and lengthy task. An autonomous robot that is able to measure the production level of a child could be used as a tool to alleviate these factors, enabling more accurate data collection for both research and assessment purposes. Currently we are planning on expanding this work to more schools while increasing the social skills of the robot.

7 ACKNOWLEDGEMENTS

This work was supported by the EU H2020 L2TOR project (grant 688014). The authors would also like to thank the teacher, who wished to remain anonymous, who provided the French lessons for the children. All statistics and graphs were obtained using R [15].

REFERENCES

[1] Minoo Alemi. 2016. General Impacts of Integrating Advanced and Modern Technologies on Teaching English as a Foreign Language. *International Journal on Integrating Technology in Education* 5, 1 (2016), 13–26.
[2] M Alemi, A Meghdari, and M Ghazisaedy. 2015. The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics* (2015), 1–13.
[3] Paul Baxter, Rachel Wood, and Tony Belpaeme. 2012. A touchscreen-based'sandtray'to facilitate, mediate and contextualise human-robot social interaction. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 105–106.
[4] Larry Fenson, Philip S. Dale, Steven Reznick, Donna J. Thal, Elizabeth Bates, Jeff Hartung, Stephen J. Pethick, and Judy Reilly. 1993. *The MacArthur Communicative Development Inventories: UserâĂŹs guide and technical manual*. San Diego, CA: Singular Publishing.
[5] Morrison F. Gardner. 1990. *Expressive One-Word Picture Vocabulary Test - Revised*. Novato, CA: Academic Therapy.
[6] Susan E Gathercole and Alan D Baddeley. 1989. Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of memory and language* 28, 2 (1989), 200–213.
[7] James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme. 2016. Heart vs hard drive: children learn more from a human tutor than a social robot. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 451–452.
[8] Batia Laufer. 1998. The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics* 19, 2 (1998), 255–271.
[9] Batia Laufer and Zahava Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning* 54, 3 (2004), 399–436.
[10] Batia Laufer and Paul Nation. 1999. A vocabulary-size test of controlled productive ability. *Language Testing* 16, 1 (1999), 33–51.
[11] Batia Laufer and T. Sima Paribakht. 1998. The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning* 48, 3 (1998), 365–391.
[12] Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, and Munsang Kim. 2011. On the effectiveness of robot-assisted language learning. *ReCALL* 23, 01 (2011), 25–58.
[13] Didier Maillat. 2010. The pragmatics of L2 in CLIL. Language use and language learning in CLIL classrooms. *Language Use and Language Learning in CLIL Classrooms* (2010), 39–58.
[14] Jan-Arjen Mondria and Boukje Wiersma. 2004. Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In *Vocabulary in a Second Language: Selection, Acquisition and Testing*, Paul Bogaards and Batia Laufer (Eds.). John Benjamins Publishers, 79–100.
[15] R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[16] Fumihide Tanaka and Shizuko Matsuzoe. 2012. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction* 1, 1 (2012).
[17] E. H. Wiig, W. Secord, and E. Semel. 1992. *CELF-Preschool: Clinical Evaluation of Language Fundamentals - Preschool*. New York: Psychological Corp.
[18] Kathleen Williams. 1997. *Expressive Vocabulary Test*. Minnesota: American Guidance Service.