Social Robot Tutoring for Child Second Language Learning

James Kennedy, Paul Baxter, Emmanuel Senft and Tony Belpaeme
Centre for Robotics and Neural Systems
Plymouth University, Plymouth, U.K.
{james.kennedy, paul.baxter, emmanuel.senft, tony.belpaeme}@plymouth.ac.uk

***Abstract*—An increasing amount of research is being conducted to determine how a robot tutor should behave socially in educational interactions with children. Both human-human and human-robot interaction literature predicts an increase in learning with increased social availability of a tutor, where social availability has verbal and nonverbal components. Prior work has shown that greater availability in the nonverbal behaviour of a robot tutor has a positive impact on child learning. This paper presents a study with 67 children to explore how social aspects of a tutor robot's speech influences their perception of the robot and their language learning in an interaction. Children perceive the difference in social behaviour between 'low' and 'high' verbal availability conditions, and improve significantly between a pre- and a post-test in both conditions. A longer-term retention test taken the following week showed that the children had retained almost all of the information they had learnt. However, learning was not affected by which of the robot behaviours they had been exposed to. It is suggested that in this short-term interaction context, additional effort in developing social aspects of a robot's verbal behaviour may not return the desired positive impact on learning gains.**

***Index Terms*—Human-robot interaction; robot tutors; second language learning; social availability; immediacy**

I. Introduction

An increasing number of human-robot interaction (HRI) researchers are exploring the utility of robots for tutoring children [1], [2], [3]. Much of this research is centred around the social behaviour of the robot, with a view to improving learning outcomes and child responses to the robot [4], [5]. However, there are still many questions to be answered about how a robot should behave in educational interactions in order to achieve these goals [6].

Social interaction has been highlighted as a particularly important element in language learning [7], and recent research in HRI suggests that robots are able to make a positive impact on learning in such contexts [1], [8]. One aspect of social interaction which is positively correlated with learning between humans is the 'psychological availability' of an instructor [9], [10], [11]. Certain elements of 'availability' in social behaviour have been studied in HRI before [12], [13], but an explicit effort to manipulate this availability and examine the effect on child learning remains to be carried out.

Child language learning provides an ideal domain for social HRI to contribute to. In the case of language, children learn better than adults, despite the increased cognitive capacity of adults. Language learning has a 'critical period' in neurobiology [14], which means that there is a window in which it is best learned. As such, in this paper we conduct a study with children aged 8 and 9 years old. At this age, the children are still within the critical period, but have sufficient skill to read novel words without assistance.

We aim to explore how the language learning of a child can be influenced by the social behaviour of a robot tutor. This paper presents an experiment in which a robot tutor teaches children some aspects of a second language. The robot behaviour is modified to be more or less socially available through the verbal interaction it has with the child. The learning of the children is measured in the short-term (immediately after the interaction), and also the following week to check that the learned information was retained. We seek to investigate whether the intended availability of the robot is perceived by the children, and whether a more socially available robot has a positive impact on learning outcomes as predicted by the HRI and human-human interaction (HHI) literature.

II. Related Work

A. Language Learning with Robots

Social robots have proven their utility in language learning environments with improved outcomes when teaching is supplemented with robots [1], [2]. Alemi *et al.* [1] used a NAO robot in a school classroom to support a human teacher in teaching English as a foreign language. Knowledge was assessed before and after 5 lessons (one per week for 5 weeks). It was found that children in the condition with a robot learned and retained significantly more vocabulary than children who had a human teacher alone.

However, things are not as clear when the robot is interacting one-on-one with students without a human teacher present. Various experiments have sought to apply human-human learning principles to child-robot interactions in the language domain with mixed results [8], [15]. Curiosity of a robot was used to inspire reciprocal behaviour in children as the HHI literature predicts an increase in learning when children are more curious. Although the children who saw the curious robot adopted curious behaviours, their word learning did not improve any more than those children who had not seen the curious robot [4].

Some effects have been successful though: a robot with personalised story-telling complexity resulted in children using more words and more diverse words than children

who interacted with a non-personalised robot [15]. Socially supportive behaviours have also successfully been implemented in a robot which taught a novel language to children [16]. Those in the socially supportive condition scored significantly higher on a language test and in motivation measures (intrinsic and task motivation). The socially supportive condition employed many non-verbal behaviour manipulations, such as increased empathy, attention guiding, and non-verbal feedback. Whilst this is a promising result, more needs to be done to establish solid models for robot social behaviour in interactions of this nature. This paper seeks to address how the verbal social behaviour of a tutor robot affects child learning and how such behaviour might be characterised.

B. Social Behaviour and Learning

In order to maximise the potential of robots in learning contexts, it is useful to explore how they should behave socially, as many human-human studies have revealed a link between social behaviour and learning [10], [11], [14]. Social behaviour also has a great impact on how students perceive teachers [10], [17]. In turn, this influences factors such as how much students believe they have learnt, and how motivated they are to learn [11]. Therefore, it is important for students interacting with robots in educational contexts to have a positive perception of, and relationship with, the robot.

One concept of human social behaviour which has been positively correlated with student motivation, student achievement, and student attitudes is the 'psychological availability' of an instructor [10], [11]. This concept considers how a teacher acts towards any particular pupil (as opposed to the class as a whole, given the classroom context of many studies in this field). This availability is measured through 'immediacy' and consists of verbal and nonverbal social behaviour components [9], [18]. It should be noted that typical connotations of the word 'immediate' regarding timing do not form part of the measure. Instead, verbal immediacy includes behaviours such as whether an instructor uses personal examples in teaching, uses first names, solicits student opinions, and so on, whereas nonverbal immediacy considers the use of overt nonverbal social cues such as gaze and gestures [9], [17].

Research has been done in HRI with a view to improving the bond between children and robots through some of these means [19], although often not in the context of educational interactions. It has been found that 'off-activity talk' - dialogue with a robot which does not concern the task being completed - encourages compliance in children in a therapeutic setting [13]. Personalisation in therapeutic contexts has also been considered. Children were asked a number of questions about their preferences and the robot then mentioned these in an interaction, the children who interacted with a personalised robot enjoyed the interaction more, but subject numbers were too low for statistical comparisons [12].

Part of the social availability construct (nonverbal immediacy) has previously been used in HRI with findings in agreement with the HHI literature [20], [21], suggesting immediacy is suitable for use as a metric in HRI. This paper considers the other part of the social availability construct, verbal immediacy, to measure and motivate robot behaviour differences.

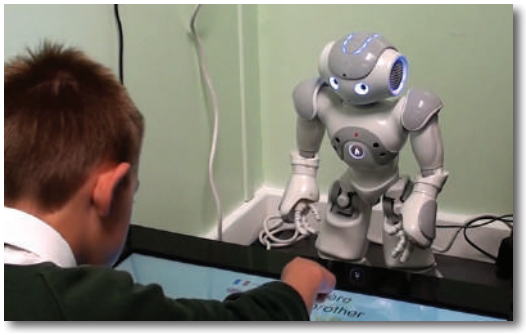

Fig. 1. A child answering a question on screen during the interaction.

III. Research Questions

Following on from previous research with humans [9] and robots [12], [13] we seek to test whether robot verbal availability has a positive impact on children's second language learning as predicted by the literature. In order to make such an assessment, it first needs to be clear that children perceive the behaviour of the robot as intended. Verbal immediacy provides a basis for measuring the children's perceptions and also for motivating differences between robot conditions. To ensure that any observed learning effects are retained and not just the product of short-term memory recall, we also aim to verify children's retention of the material outside of the short-term interaction context (as in [22]). This leads to the following hypotheses:

- H1. *Perception of robot behaviour.* Children will perceive and report differences in the robot's verbal availability (as measured through immediacy).
- H2. *Child learning.* Children will retain the language skills that they learn from the robot outside of the short-term.
- H3. *Effect of availability on learning.* A robot exhibiting more socially available verbal behaviour will lead to greater child learning gains than a robot without this behaviour.

IV. Design

French is commonly taught in English schools, so would have clear relevance for the children. However, it does not receive very much lesson time (the majority of schools offer 30-45 minutes per week at the age used in this study [23]), so there is plenty of scope to teach new concepts. As such, French was selected as the second language to teach in this study. The learning material was developed in collaboration with an academic researcher in language development, a native French speaker, and a teacher.

The structure of the lesson content was designed based on previous work in which children learnt mathematical concepts, such as [5], and a pilot study involving a human tutor and children. The aim was for the children to learn that nouns in French have a gender, that this changes which article is used

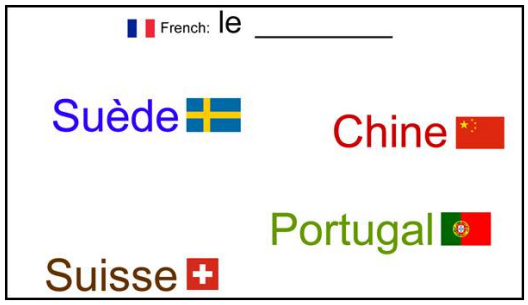

Fig. 2. Screenshot from the touchscreen showing a question. Children can touch a word, drag it to the blank space and release to answer. Here the correct answer being 'Portugal'.

('le' or 'la'), and that for some words there are patterns which can be used to help work out which article to use.

An Aldebaran NAO robot acted as a tutor, delivering all lessons through speech and moving words on a touchscreen (Fig. 1). As such, the children were exposed to both the words' pronunciation and orthography. The robot demonstrated how questions could be answered by dragging and dropping the correct answer in the blank space (see Fig. 2). The robot first explained the concept of words having a gender by using an English example (using 'waiter' for a man, and 'waitress' for a woman). Following this, it explained how the French word for 'the' could be 'le' or 'la' depending on the gender of the noun it precedes. The robot then explained rules for working out whether to use 'le' or 'la'. After explaining each rule, the child's understanding was checked (Fig. 3).

During the lessons the robot would explain a rule and then use the screen to show an example. The rules used were taken from online French language learning guides[1,2] and were verified by a French native speaker. The rules were as follows: 1) 'le' is used for male people, and 'la' is used for female people, 2) 'la' is used for countries ending in 'e', 3) 'la' is used for fruit or vegetables ending in 'e'. Whilst these are recognised techniques for people learning a second language, it should be made clear that it is unlikely that a native speaker would learn in this way, and that there are a limited number of exceptions to rules 2 and 3 (but these were avoided in the lesson content here). We do not seek to determine the best teaching strategy for the concept, but the effect that robot behaviour has on any learning.

Questions were designed to get progressively more complex as the interaction progressed. To start with, English translations and pictorial representations of the words were provided alongside the French. At this stage, the child was only required to select the article 'le' or 'la' to add to the word. Towards the end of the interaction, all English translations were removed so that only the French and the pictures remained. The question structure was also changed in later stages: the child was required to match a noun to the article (Fig. 2), which requires them to assess several nouns for each question, rather than just one as in the earlier questions.

[1]http://goo.gl/JPjmPO
[2]https://goo.gl/WY37z5

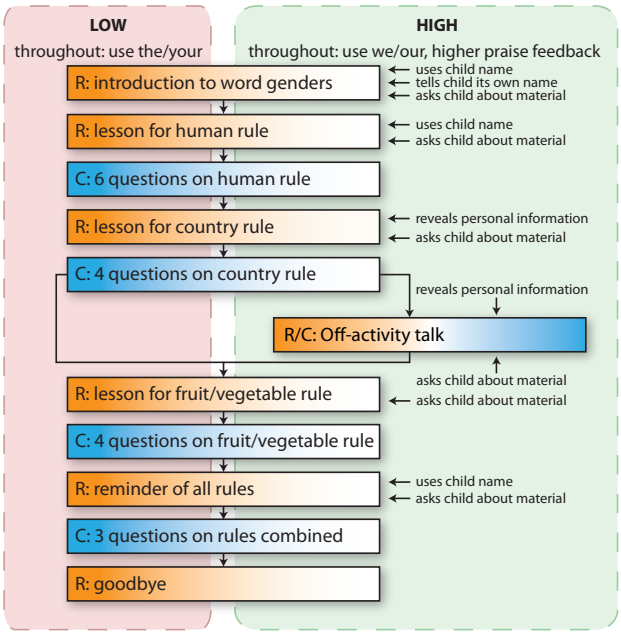

Fig. 3. Structure of the task. *R* refers to robot explanation sections and *C* refers to child question answering sections. The robot dictates the structure of the interaction through speech and by presenting questions on the touchscreen, informing the child of when it is their turn answer questions on the screen. The HIGH condition includes many manipulations in the verbal behaviour to make it more 'available'.

All feedback was provided verbally by the robot; no feedback was shown on the screen. When providing feedback, the robot's TTS would switch to French so that the child could hear the correct pronunciation. The robot was autonomous throughout, except for some short vocal phrases in one condition, which were triggered by the experimenter (see Section V-C).

V. Evaluation

A. Participants

A total of 67 children were included in the study after exclusions due to technical issues (1 child) or absence from school during one of the two visiting periods (7 children). All children were native English speakers and were from the same year group (with three class teachers) from a primary school in the U.K. (average age M=8.8, SD=0.4; 30M, 37F). Only one child was fluent in another language (this language was not used in this study). Children were distributed randomly between groups whilst balancing for gender and class teacher. All children had parental/guardian permission and gave their consent to take part in the study.

B. Measures

Learning was measured through pre-, post- and retention tests, which can be seen online[3]. These tests sought to examine various aspects of the children's learning, including their

[3]http://goo.gl/hrIQEe

vocabulary acquisition, and their ability to apply each of the 3 rules in isolation and combination with each other. The test consisted of 12 questions: 3 vocabulary-based (1 French-English and 2 English-French), 2 about humans (rule 1), 2 about countries (rule 2), 3 about fruits and vegetables (rule 3), and 2 combined all three rules. Each question had 4 multiple choice answers and used the same formats as questions on the touchscreen. The majority of the test questions used words that the children had not seen in the learning material in order to ensure generalised learning was taking place, rather than memorisation of specific instances; exceptions are discussed in Section VII. The pre-, post- and retention-tests were all the same as this was necessary to account for children's prior knowledge (they had learnt some French vocabulary in school before), and to accurately measure their recall. The children were not given any feedback on their tests at any stage.

The child's perception of the robot was measured through a questionnaire combining verbal immediacy and nonverbal immediacy items. This 23 question questionnaire was completed on paper and was multiple choice. The verbal immediacy and nonverbal immediacy items were based on those used in [10], but were modified such that the language could be understood by children. The final questionnaire used can be seen online[4]. Verbal immediacy includes aspects of behaviour such as personalisation, off-activity talk, and student opinion solicitation. Nonverbal immediacy covers overt social behaviours, such as whether gestures are used, whether the robot looks at the child, and so on.

C. Conditions and Robot Behaviour

In order to address the hypotheses in Section III, three conditions were devised: 1) a robot with high verbal availability (HIGH, n=20), 2) a robot with low verbal availability (LOW, n=20), 3) a control with no robot and just a pre- and retention test (CTRL, n=27). The robot with low verbal availability doesn't have the verbal behaviours which lead to being considered available as measured by verbal immediacy (Fig. 3). The control condition is used to verify that there are no learning effects due to exposure to the test material.

In both robot conditions, the nonverbal behaviour was kept constant. The behaviour used was designed to be of high nonverbal immediacy, with the robot's gaze randomly moving in the direction of the child, gestures during speech, a slight lean forward of the body, and slight motor noise in the arms to give the impression of being relaxed. The perception of this behaviour as being of high nonverbal immediacy is verified through the questionnaire after the interaction (as described in Section V-B).

The speech of the robot was kept the same in both conditions outside of the experimental manipulations as described below. This ensures that the lesson content is largely unchanged between conditions, although the experimental manipulations require some language adjustments, these should not impact on the coherence or intelligibility of the lessons.

Verbal immediacy can be used to measure aspects of availability of an instructor, so the verbal immediacy questionnaire [9] was used to create the robot conditions with different availability levels. In order to generate the behaviour for the conditions, we therefore applied all of the verbal immediacy questionnaire items possible to the speech for the HIGH condition, and did not apply them for the LOW condition. The following differences were present in the HIGH condition robot behaviour, but not in the LOW condition[5]:

1) use the child's name (3 times)
2) tell the child its name
3) reveal personal information about itself (twice in addition to its name)
4) ask the child how they felt about the material (e.g. "does everything make sense to you so far?" 6 times)
5) ask the child about their hobbies and continue the discussion for 2 or 3 speech turns
6) use "we/our" work (as opposed to "the/your", throughout)
7) provide higher praise feedback (e.g. "You're doing really well! That was right", as opposed to simply "That was right" in the LOW condition)

Two items of the verbal immediacy questionnaire were not manipulated: humour and feedback provision. Humour was considered to be inappropriate to add given the context of the interaction and difficulties in selecting a comment that would be universally funny. Whether or not feedback was provided was not manipulated between conditions as in this context, the only way of getting feedback was from the robot and missing feedback here would confound any findings related to learning.

To compensate for unreliable speech recognition, a Wizard-of-Oz intervention was used in the HIGH condition to let the robot reply 'that's great' after the children answered a question from the robot about their understanding of the material (children always said they had understood the lesson), and to trigger pre-scripted phrases at the appropriate time for the discussion about the child's hobby.

D. Procedure

The interactions took place on the school premises in a quiet working space familiar to the children. The child sat across from an Aldebaran NAO with a 27 inch touchscreen placed horizontally between them (Fig. 1). Two video cameras were used to record the interactions. One experimenter sat behind and to the side of the child, out of their view (Fig. 4). The time children spent interacting with the robot was on average M=11min 26s (SD=1min 11s).

The experimenter spent a full week in the school, plus one day the following week. On the first Monday of the visit, pre-tests were delivered to all children in their main classrooms. These were completed under the supervision of the experimenter and the class teacher to make sure that children completed them individually. Throughout the week those children interacting with the robot would be taken out of class individually, take part in the interaction, and then

[4]http://goo.gl/UoL5QM

[5]Please also refer to the video figure for this publication

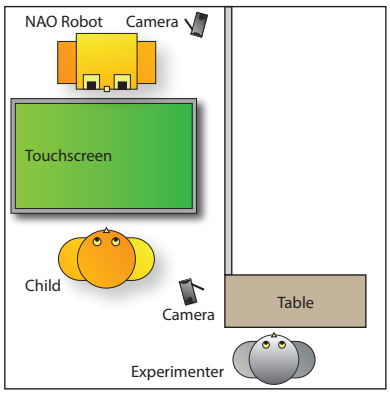

Fig. 4. Schematic overview of the interactions being investigated in this paper. The child and the Aldebaran NAO robot sit across a touchscreen from one another. An experimenter sits behind and out of view of the child. Two video cameras record the interaction. Figure not to scale.

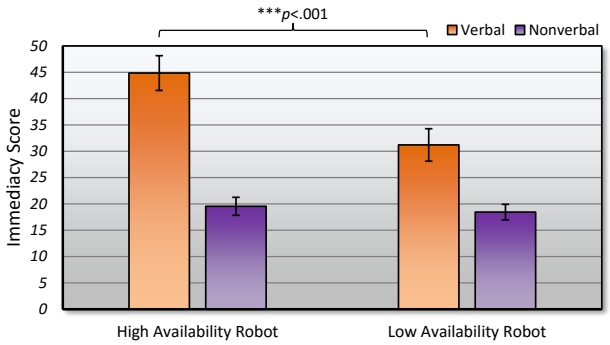

Fig. 5. Verbal and nonverbal immediacy scores for the high availability (HIGH) and low availability robot (LOW) conditions. The HIGH condition is perceived to have significantly higher verbal immediacy while having the same nonverbal immediacy, showing that the children perceive it as more available. *Error bars* show 95% CI.

complete the post-interaction test and questionnaire on paper, to the side of where the experimenter had been sitting (so they can no longer see the robot or touchscreen). The robot condition was switched between each interaction to ensure a balance throughout the week.

On the Monday of the following week the experimenter returned to deliver the retention test to the children under the same conditions as the pre-test. Children in the control group therefore completed a pre-test and a retention test without any teaching input. The children had not been informed that they would be tested again on the material that they had covered with the robot. After each class had completed the retention test, the experimenter gave an overview of the study and a presentation of social robots to all children. This meant that all children understood the study and had the opportunity to interact with the robot.

VI. Results

A. Perception of the Robot

To address H1 (that children will perceive differences in the verbal availability of the robot), the results of the post-interaction questionnaire were analysed. The questionnaire is broken down into the several parts which measure different constructs, as described in Section V-B. The manipulations were conducted on the verbal immediacy element of the questionnaire, where a higher verbal immediacy score would indicate a higher perception of verbal availability. An unpaired t-test reveals a significant difference between the average verbal immediacy measure for the LOW condition (M=31.2, 95% CI [28.1,34.3]) and the HIGH condition (M=44.9, 95% CI [41.6,48.2]); $t(38)$=6.322, $p<.001$. This confirms H1; children could indeed perceive the difference between the conditions (despite not having seen the other condition for comparison).

Nonverbal immediacy scores were also compared; the difference between the nonverbal immediacy score in the LOW condition (M=18.5, 95% CI [17.0,19.9]) was not found to be significantly different to that of the HIGH condition (M=19.6, 95% CI [17.8,21.3]); $t(38)$=1.020, p=.314 (Fig. 5). This provides some validation for the control of nonverbal behaviour between the conditions.

B. Learning Gains

Learning gains are measured through scores on the tests conducted before the interaction (pre-test), immediately after the interaction (post-test), and 3-7 days after the interaction (retention test). Questions on the tests are equally weighted, so scores are out of a maximum of 12. Before analysis of the two robot conditions can be conducted there are some potential confounds which must be eliminated as factors: the differences in time between the interaction and retention test, and the impact of exposure to the test (as the same test is used).

It could be expected that children who interacted with the robot at a time closer to the retention test would outperform those who interacted with the robot earlier in the visit. To explore whether this was a factor, the day on which the interaction took place was correlated with the difference between the post-test and the retention test. The correlation is weak and non-significant; $r(36)$=-.079, p=.637, indicating that the time from interaction to retention test can be eliminated as a factor. We would suggest that the absolute number of days does not make a difference to the retention, but the number of days out of school during this period is more important, which was constant for all children (a weekend of 2 days).

The control condition is used to verify whether exposure to the test makes a difference to the findings. It would not be expected that there would be a difference as the children are given no feedback on the tests at any stage, but the control condition allows verification. For children in the control condition, the pre-test score (M=3.96, 95% CI [3.26,4.66]) and retention test score (M=3.89, 95% CI [3.28,4.49]) can be considered equivalent. Two one-sided t-tests (TOST) [24] with a 1 point threshold confirm the test scores are equivalent at the $p<.05$ level: $t(52)=-2.061$, p=.022/$t(52)$=2.391, p=.010. This indicates that exposure to the test is not a confounding factor.

A repeated measures ANOVA was used to explore H2 (that children will retain their learning) and H3 (that the robot condition will affect learning); Fig. 6 and Table I show

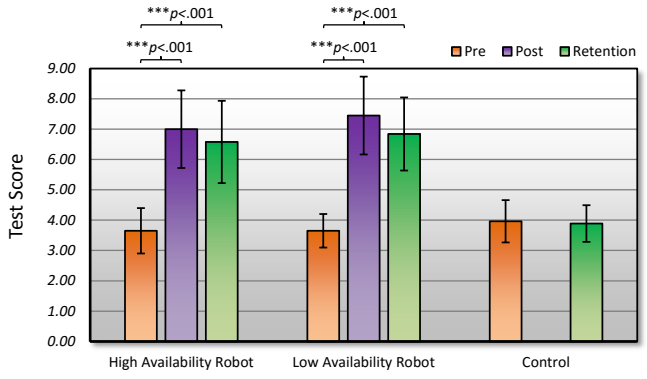

Fig. 6. Pre-test, post-test and retention test scores by condition (chance score=3; maximum score=12). Children learn a significant amount from the robot between pre- and post-tests; this gain is sustained to the retention test. *Error bars* show 95% CI.

the results for test scores by condition. Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated, $\chi^2(2)=1.873$, $p=.392$. No significant interaction was found between test and condition; Wilk's Lambda=.998, $F(2,35)=0.04$, $p=.963$. A main effect was found for test, Wilk's Lambda=.391, $F(2,35)=27.21$, $p<.001$, but not for condition; $F(1,36)=0.08$, $p=.774$. Bonferroni pairwise comparisons find that there is a significant difference between pre-test and post-test, and pre-test and retention test scores (all $p<.001$), but no difference between post-test and retention test ($p=1.00$).

These results support H2, as children learn between the pre- and post-tests, and retain their learning in the retention test. Further support for H2 can be gained through Weber & Popova paired-samples equivalency tests [25] which show that the post and retention test scores are equivalent in both the HIGH ($t(18)=0.67$, $p=.022$) and LOW ($t(18)=0.73$, $p=.025$) conditions, with Cohen's $d=.50$. Whilst this is an 'intermediate' effect size for demonstrating equivalency, it should be noted that the sample size is relatively small on a per-condition basis, leading to a higher variation in scores, which raises the level at which equivalency can be shown. Combined, these findings provide evidence in support of H2 as the children learn a significant amount from the pre-test to the post-test, and the post-test and retention test scores can be considered largely equivalent, demonstrating their retention of the learning.

The ANOVA results do not support H3 (that higher availability will lead to greater learning) as no significant effect was found for robot condition. Nor can a significant difference be seen between the improvement in the LOW condition ($M=3.80$, 95% CI [2.55,5.05]) and the HIGH condition ($M=3.35$, 95% CI [1.78,4.92]); $t(38)=0.470$, $p=.641$. The drop in score from post-test to retention test can also be considered equivalent between conditions; using a Weber & Popova independent-samples equivalance test, $t(36)=0.07$, $p=.004$ with Cohen's $d=.50$. Therefore, Hypothesis H3 must be rejected as there are no significant differences observed between conditions in terms of learning.

Based on the rules taught to the children, one could suggest that learning a very simple rule of: "if the word ends in an 'e', then use *la*, otherwise use *le*" may be adopted as a 'shortcut' and could account for the learning differences. This would then have nothing to do with learning aspects of language, but be a basic memory phenomenon. This had been anticipated in the study design, so later questions in the learning material made sure to challenge this approach by including several words ending in 'e' as possible answers, but with those words relating to humans of male gender (therefore requiring 'le', rather than 'la' and violating the shortcut rule). Additionally, a question in the tests used adopted this approach, with several words ending in an 'e', but not all being feminine. This was done to verify whether the shortcut rule had been adopted, or whether the children had really learnt the material as it had been taught, with the ability to discriminate between different types of words. If the children had only learnt the shortcut rule then they would answer this verification question incorrectly, however, it was answered correctly above the average level for the rest of the questions in the test (63% for the verification question, versus 60% for the other questions). This provides some evidence that the children learnt intricacies of the language that was presented to them; further evidence in support of this will be provided in Section VII.

TABLE I
TEST SCORE RESULTS BY CONDITION.

Condition	Pre-Test *M* [95% CI]	Post-Test *M* [95% CI]	Retention Test *M* [95% CI]
CTRL	3.96 [3.26, 4.66]	N/A	3.89 [3.28, 4.49]
LOW	3.65 [3.10, 4.20]	7.45 [6.17, 8.73]	6.84 [5.64, 8.05]
HIGH	3.65 [2.90, 4.40]	7.00 [5.72, 8.28]	6.58 [5.22, 7.94]

VII. Discussion

The results show that the children perceived the verbal availability of the robot conditions as intended, which confirms that the behaviour was designed appropriately to address the research hypotheses. The nonverbal behaviour was kept constant between the two conditions, and this was reflected in the children's questionnaire responses. The children in both robot conditions exhibited significant learning gains between the pre-test and post-test, as well as between the pre-test and retention test, with equivalent scores in the retention test and the post-test. This is a positive result, as it would have been plausible that the children would quickly forget what the robot had taught them once the interaction was over, especially as the children were not aware that they would be re-tested, and so had little motivation to attempt to actively try and retain the information.

The tests which the children had to complete were designed to be challenging. Each answer had four options with no obviously incorrect answers, so the likelihood of a guess being correct would be chance (25%). It was found that children scored slightly above this on the pre-tests as they had done a small amount of French before, so scored closer to 4 than the 3 that would be expected with random guessing. This significantly improved to over 7 out of 12 in the post-tests.

Given the difficulty of the tests and the relatively short time the child interacts with the robot learning and practising the material, this is an impressive increase. Indeed, only 6 of the 40 children who interacted with the robot did not improve from pre-test to post-test. Learning of 'le' or 'la' as the article choice could have contributed to part of the increase in scores, however if children had learnt the choice to be le/la then the chance score would go up by 1.5 points from pre-test (chance = 3) to post-test (chance = 4.5). The children actually improve by an average of 3.6 (95% CI [2.6,4.5]), suggesting learning beyond any improvement due to the higher chance score.

Despite the children being able to perceive the difference in verbal aspects of availability between the two robot conditions (measured through verbal immediacy), no significant difference was observed in learning in either the post- or retention-test. This finding is surprising given the positive correlation between verbal immediacy and learning in human studies [9], [11]. Previous work has found that nonverbal aspects of availability can lead to additional learning above that gained through just exposure [20]. The work here explored whether verbal aspects of availability would have a similar positive effect on learning, but they did not.

Aspects of the behaviour manipulated here, such as personalisation [12] and off-activity talk [13], have been studied before in HRI with promising results. However, these studies had too few subjects to make conclusions about learning [12], or did not assess learning [13]. In contrast to [13], we do find here that the children perceive differences between the conditions, but in our study the questionnaire is targeted towards specifically measuring the perception of the behaviours which were manipulated, rather than assessing an overall feeling towards the robot. It is possible that despite children perceiving differences in the availability of the robot, this did not translate into any difference in feeling towards the robot. If the relationship the child feels towards the robot is no different between conditions then this may go some way to explaining the lack of difference in learning.

The interpretation of the robot character could have been influenced by the TTS voice used by the robot, which would switch when the language changed. These voices were clearly different and this could have impacted how the children perceived the robot. However, the children have no prior experience with the robot, so they may have accepted this as part of the robot's behaviour. As the voices are clearly different, they may also have interpreted this not to be part of the robot's character, but to be the robot playing back other media (akin to a teacher playing recorded French). It is not possible to determine how the children perceived this switch in voice from the data collected, but perceptions of voice switching of multi-lingual robots could be worth explicitly exploring in future work.

Another factor which may have influenced the learning results is novelty. Novelty is often an issue for HRI studies [26], [27], and it possibly played a role here as the children interact just once with the robot for a brief period of time. Verbal immediacy has been found to consist of four factors, including 'individual friendliness' [10]. Even if the children were to bond more strongly with the high availability robot because of increased friendliness, the short interaction time might not be enough for differences in the relationship to manifest into learning outcomes. Furthermore, it could be that the behaviour of the more available robot cancels out its own benefits by being so novel as to distract from the learning material. For example, when the robot is conducting off-activity talk during the interaction, this is time when the children are not focussing on the learning task and are possibly forgetting information they have learnt. This doesn't mean that off-activity talk should be avoided for fear of distraction, but that it might only be appropriate in longer, or repeated interactions where novelty is less of an issue. We would hypothesise that given a longer interaction timescale, the learning benefits predicted by the literature of greater availability [9], [11] would be observed as the novelty wears off [2], [26].

In the HHI literature, a lower correlation between verbal immediacy and learning has been found when compared to nonverbal immediacy and learning [11]. Nonverbal immediacy has previously been found to make a difference to learning in HRI [20], [21]. This could suggest that verbal behaviour may not be as important for learning (at least in short-term interactions) as overt nonverbal behaviour. It has also been found in humans that the impact of immediacy behaviours is enhanced in line with increases in class size [9]. It could be that the effect of verbal immediacy is simply too far reduced when placed in a one-to-one tutoring context as in this study, rather than the larger classroom setting. The availability of the robot would be experienced to some extent in both conditions simply through the nature of the one-to-one interaction.

One interesting finding from the data collected which was not hypothesised was the ability of the children to acquire vocabulary despite the learning material not explicitly requiring them to do so. Three questions of the test were vocabulary based: two requiring translation from English to French, and one French to English. Two of these questions referred to words which the children would have seen on screen and heard the robot say (as they were answers to questions in the learning material). The remaining question was about a word which they would have seen on screen, but the robot did not say (as it was not a correct answer). It is suggested that the two words which were answers in the learning material were more likely to be recalled as the children would have looked at the word for longer and the robot would have said the word. However, a significant increase was found for all 3 of the questions independently, and a repeated measures ANOVA found a significant increase for the average score (out of 3) of children who correctly translated the words from pre-test to post-test, and from pre-test to retention test. Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated, $\chi^2(2)=0.661$, $p=.719$. No significant interaction was found between test and condition; Wilk's Lambda=.968, $F(2,35)=0.58$, $p=.565$. A main effect was found for test, Wilk's Lambda=.595, $F(2,35)=11.94$, $p<.001$, but not for condition; $F(1,36)=0.14$, $p=.710$. Post-hoc Bonferroni pairwise

comparisons find that there is a significant difference between pre-test (M=0.8, 95% CI [0.6,1.0]) and post-test (M=1.6, 95% CI [1.3,1.9]), and pre-test and retention test (M=1.4, 95% CI [1.1,1.7]) scores ($p<.001$ and $p=.001$, respectively), but no difference between post-test and retention test scores ($p=.883$).

It is of course possible that the children remembered the words from the pre-test and made an effort to learn these words when they were presented on screen, but this seems unlikely given the time (up to 4 days) between many of the pre-tests and the interactions, and the sheer number of words they were exposed to in the learning content (over 40). For a child to concentrate on learning 3 words from the pre-test, days after having seen it, when being taught a different aspect of language would seem to be highly improbable. As such, this is a promising finding with robots that confirms data from human-human literature whereby children of this age will acquire language through exposure in social interactions [14].

VIII. Conclusion

Children perceived the relative social availability of the two robot conditions as intended in the design. This confirms that the manipulations made were appropriate to address the question of whether an increase in verbal aspects of availability would lead to an increase in learning. As expected, the children did learn elements of a second language from the robot. This was measured immediately after the interaction and also some days later. The retention test scores were slightly lower than the pre-test scores, but can be considered statistically equivalent. However, surprisingly there was a lack of any significant difference between conditions in the immediate post-test score, or the longer-term retention test score. Literature from human-human interaction studies [9], [11] and human-robot interaction studies [12], [13] would predict an increase in robot verbal availability to lead to an increase in learning, but this was not found. These findings suggest that in this short-term dyadic interaction context, additional effort in developing social aspects of a robot's verbal behaviour may not return the desired positive impact on learning gains.

IX. Acknowledgements

This work is funded by the EU FP7 DREAM project (grant 611391), H2020 L2TOR project (grant 688014), and SoCEM, Plymouth University, U.K. Thanks goes to Dr. Caroline Floccia who provided valuable feedback on the study design.

References

[1] M. Alemi *et al.*, "Employing Humanoid Robots for Teaching English Language in Iranian Junior High-Schools," *Int. Journal of Humanoid Robotics*, vol. 11, no. 3, 2014.

[2] T. Kanda *et al.*, "Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial," *Human-Computer Interaction*, vol. 19, no. 1, pp. 61–84, 2004.

[3] E. Short *et al.*, "How to Train Your DragonBot: Socially Assistive Robots for Teaching Children About Nutrition Through Play," in *Proc. of the 23rd IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2014, pp. 924–929.

[4] G. Gordon *et al.*, "Can Children Catch Curiosity from a Social Robot?" in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI*. ACM, 2015, pp. 91–98.

[5] J. Kennedy *et al.*, "The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning," in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI*. ACM, 2015, pp. 67–74.

[6] ——, "Can Less be More? The Impact of Robot Social Behaviour on Human Learning," in *Proc. of the 4th Int. Symp. on New Frontiers in HRI at AISB 2015*, 2015.

[7] P. K. Kuhl, "Cracking the speech code: How infants learn language," *Acoustical Science and Technology*, vol. 28, no. 2, pp. 71–83, 2007.

[8] J. Herberg *et al.*, "Robot watchfulness hinders learning performance," in *Proc. of the 24th IEEE Int. Symp. on Robot and Human Interactive Communication*, 2015.

[9] J. Gorham, "The relationship between verbal teacher immediacy behaviors and student learning," *Communication education*, vol. 37, no. 1, pp. 40–53, 1988.

[10] J. H. Wilson and L. Locker Jr, "Immediacy scale represents four factors: Nonverbal and verbal components predict student outcomes," *The Journal of Classroom Interaction*, vol. 42, no. 2, pp. 4–10, 2007.

[11] P. L. Witt *et al.*, "A Meta-Analytical Review of the Relationship Between Teacher Immediacy and Student Learning," *Communication Monographs*, vol. 71, no. 2, pp. 184–207, 2004.

[12] O. A. Blanson Henkemans *et al.*, "Using a robot to personalise health education for children with diabetes type 1: A pilot study," *Patient Education and Counseling*, vol. 92, no. 2, pp. 174–181, 2013.

[13] I. Kruijff-Korbayova *et al.*, "Effects of Off-Activity Talk in Human-Robot Interaction with Diabetic Children," in *The 23rd IEEE Int. Symp. on Robot and Human Interactive Communication*, 2014, pp. 649–654.

[14] P. K. Kuhl, "Brain mechanisms in early language acquisition," *Neuron*, vol. 67, no. 5, pp. 713–727, 2010.

[15] J. K. Westlund and C. Breazeal, "The interplay of robot language level with children's language learning during storytelling," in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI Extended Abstracts*. ACM, 2015, pp. 65–66.

[16] M. Saerbeck *et al.*, "Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, 2010, pp. 1613–1622.

[17] V. P. Richmond *et al.*, "Development of the Nonverbal Immediacy Scale (NIS): Measures of Self- and Other-Perceived Nonverbal Immediacy," *Communication Quarterly*, vol. 51, no. 4, pp. 504–517, 2003.

[18] A. Mehrabian, "Some Referents and Measures of Nonverbal Behavior," *Behavior Research Methods & Instrumentation*, vol. 1, no. 6, pp. 203–207, 1968.

[19] T. Belpaeme *et al.*, "Multimodal Child-Robot Interaction: Building Social Bonds," *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 33–53, 2012.

[20] J. Kennedy *et al.*, "Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions," in *Proc. of the Int. Conf. on Social Robotics*, 2015, pp. 327–336.

[21] D. Szafir and B. Mutlu, "Pay Attention!: Designing Adaptive Agents that Monitor and Improve User Engagement," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, 2012, pp. 11–20.

[22] F. Tanaka and S. Matsuzoe, "Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 78–95, 2012.

[23] K. Board and T. Tinsley, *Language Trends 2014/15: The state of language learning in primary and secondary schools in England*. CfBT Education Trust, 2015.

[24] D. J. Schuirmann, "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability," *Journal of pharmacokinetics and biopharmaceutics*, vol. 15, no. 6, pp. 657–680, 1987.

[25] R. Weber and L. Popova, "Testing equivalence in communication research: Theory and application," *Communication Methods and Measures*, vol. 6, no. 3, pp. 190–213, 2012.

[26] I. Leite *et al.*, "Social robots for long-term interaction: A survey," *Int. Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[27] J. Sung *et al.*, "Robots in the wild: understanding long-term use," in *Proc. of the 4th ACM/IEEE Int. Conf. on HRI*. IEEE, 2009, pp. 45–52.