Multimodal Technologies and Interaction

Article

Pitch It Right: Using Prosodic Entrainment to Improve Robot-Assisted Foreign Language Learning in School-Aged Children

Bo Molenaar [1,2], Breixo Soliño Fernández [1,*], Alessandra Polimeno [1,3], Emilia Barakova [4] and Aoju Chen [1]

[1] Department of Languages, Literature and Communication, Utrecht University, 3512 JK Utrecht, The Netherlands; bo.molenaar@ru.nl (B.M.); aapolimeno@gmail.com (A.P.); aoju.chen@uu.nl (A.C.)
[2] Centre of Language and Speech Technology, Radboud University, 6525 XZ Nijmegen, The Netherlands
[3] Department of Languages, Literature and Communication, Free University of Amsterdam, 1081 HV Amsterdam, The Netherlands
[4] Department of Industrial Design, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands; e.i.barakova@tue.nl
* Correspondence: b.solino.fernandez@gmail.com

Abstract: Robot-assisted language learning (RALL) is a promising application when employing social robots to help both children and adults acquire a language and is an increasingly widely studied area of child–robot interaction. By introducing prosodic entrainment, i.e., converging the robot's pitch with that of the learner, the present study aimed to provide new insights into RALL as a facilitative method for interactive tutoring. It is hypothesized that pitch-level entrainment by a Nao robot during a word learning task in a foreign language will result in increased learning in school-aged children. The results indicate that entrainment has no significant effect on participants' learning, contra the hypothesis. Research on the implementation of entrainment in the context of RALL is new. This study highlights constraints in currently available technologies for voice generation and methodological limitations that should be taken into account in future research.

Keywords: robot-assisted language learning; pitch entrainment; foreign language learning; robot–child interaction

check for updates

Citation: Molenaar, B.; Soliño Fernández, B.; Polimeno, A.; Barakova, E.; Chen, A. Pitch It Right: Using Prosodic Entrainment to Improve Robot-Assisted Foreign Language Learning in School-Aged Children. *Multimodal Technol. Interact.* **2021**, *5*, 76. https://doi.org/10.3390/mti5120076

Academic Editors: Sofia Serholt and Shruti Chandra

Received: 1 September 2021
Accepted: 25 November 2021
Published: 30 November 2021

1. Introduction

Over the past few decades, robots have come to assist humans in a variety of ways. Whereas, at first, this mostly concerned the automation of relatively simple tasks, it is now possible to design socially capable robots which can be used in a variety of settings requiring social interactions such as education. Other technology-based tools (e.g., intelligent tutoring systems or ITS) are already in use in classroom settings. One of their advantages is that they can provide personalized feedback in a one-to-one interaction, a method of education that has been shown to be highly effective compared to group education [1]. Robots may have an additional advantage over digital intelligent systems, namely their physical presence. When equipped with a humanoid body, social robots resemble a human teacher more closely, which has been argued to have "a strong positive effect on the students' perception of their learning experience" [2] (p. 365). In addition, humanoid robots' social presence influences children's motivation and performance in learning activities [3,4]. Randall [5] posits that robot-assisted language learning (RALL) is an improvement over 2D technologies (e.g., computers, tablets) by "increasing [learners'] motivation, interest, and engagement—and this may be important in long-term success" (p. 7:12). In a meta-analysis of the use of RALL, van den Berghe [6] points out that, theoretically, robots can be suitable language tutors. Interactive social cues of the robots, i.e., social cues that are in response to and in synchrony with the instantaneous human behavior, such as mimicking the head inclination of the human and timely social praises, can invoke higher task compliance and make the robot more liked and accepted [7]. However, in practice, successful interactions are commonly obstructed by technological constraints that cause the robot to fail to react to situations appropriately [8], including using interactive social cues.

In this study, we examined whether language learning can be influenced by making robots use speech entrainment, another interactive social cue whereby conversation partners adapt to each other in terms of verbal and non-verbal communicative behavior [9]. For example, interlocutors can entrain on verbal features, such as syntactic structures [10,11] and choice of words [12], and on non-verbal behavior, such as prosody [13] and gestures [14]. Entrainment affects how people are perceived. People who adapt their speech to match their interlocutor are generally perceived as more competent, socially attractive, and likeable [13]. As such, the implementation of entrainment may increase the acceptance of social robots because it can positively affect how the robot is perceived by humans.

Several studies have found that entrainment in human–human tutoring can have a positive effect on learning in tutoring contexts: both lexical [15,16] and acoustic-prosodic [17] entrainment were found to correlate positively with learning gain. Research on entrainment in computer-mediated communication between learners of a foreign or second language (L2) and native speakers of this language provides further support, suggesting that entrainment may facilitate language learning [18]. Most relevant to the current study, [19] presented positive evidence for the effect of a social entraining robot on learning gain, compared to a nonsocial one. Kory-Westlund et al. [20] showed that entrainment in the robot combined with a backstory from the robot (about its poor speech and hearing abilities) led to more accurate retelling of the robot's stories in children. However, past studies have examined entrainment on several features at the same time, making it difficult to assess the contribution of individual entraining features. In the present study, we focus on one feature, i.e., mean pitch, and explore the relationship between entrainment on mean pitch and learning. Moreover, while previous studies of entrainment in social robots have examined learning in areas, such as mathematics [19], story-retelling [20], or games [21], entrainment has not yet been researched in the context of RALL. Past work has indicated that RALL is an effective language learning method [5,6], and that entrainment by social robots can contribute to learning for school-aged students [18]. It remains unclear whether entrainment can improve language learning assisted by robot tutors.

Against this background, we investigated whether the implementation of entrainment in a robot tutor can increase L2 learning. Acoustic-prosodic entrainment can occur in different ways (i.e., proximity—similarity between interlocuters; synchrony—sychronised changes in interlocutors; convergence—becoming more similar over time) at different levels (i.e., turn level vs. conversation level), but turn-level pitch proximity appears to be the most prosodic feature for entrainment between human interlocutors [13,22]. We thus decided to focus on turn-level pitch proximity in our study. Specifically, we have addressed the following question: does pitch-level entrainment in a robot tutor increase L2 vocabulary learning in school-aged children?

Based on the existing literature on entrainment in social robots, we hypothesize that pitch-level entrainment in a robot tutor will lead to better L2 vocabulary learning. Our prediction is that children will learn more words from a robot tutor when pitch-level entrainment is applied than when it is not applied.

2. Materials and Method

2.1. Participants

Thirty-two monolingual Dutch-speaking children aged between 8 years 10 months and 11 years 5 months took part in this study. The participants were split into two groups of 16, the experimental (entrainment) group, and the control (no entrainment) group. The mean age was 10 years 3 months for the control group and 10 years 4 months for the experimental group. Both groups consisted of 11 female and 5 male participants.

2.2. Research Design

The experiment followed a pre-test–training–post-test design [23], which allowed us to assess the effect of entrainment on learning words in the target language (i.e., English). The pre-test served to measure the participants' prior knowledge of the English words in the learning task. During the training phase, the participants completed a word-learning task with a Nao V4 humanoid robot (running on NAOqi OS, version 2.1.4.13; see Figure 1). The robot was introduced to children as Robin, which is a gender-neutral name in Dutch. By using a gender-neutral name and not dressing the robot in any way, we minimized the risk of gender-related differences in the participants' perception of Robin and in our results. In the entrainment group, Robin entrained to the mean pitch of each participant's production during word learning. In the control group, Robin did not entrain to the participant's speech. Finally, during the post-test, the participants' knowledge of the newly learned words was measured again. The difference between pre-test and post-test scores would indicate the extent to which the participants have learned new words.

Figure 1. NAO Robot v4 by Aldebaran Robotics.

2.3. Materials

2.3.1. The Word-Learning Task

The word-learning task was designed in a way that allowed the robot to entrain to each participant's pitch, while at the same time the participant had the opportunity to learn new English words. In this task, the participant received booklets containing Dutch monosyllabic nouns or verbs, which were selected and translated from PPVT-IV-EN [24]. The participant would say one Dutch word at a time from their booklets aloud, to which Robin would respond with the English translation. Each word category consisted of 10 words, of which 3 functioned as target words and 7 as filler words. Based on age markers in PPVT and feedback from the participant's English teacher, the filler words were expected to be familiar in both the participant's L1 and L2, while the target words were expected to be familiar in their L1 only.

The Dutch translations of the selected words and their corresponding image from PPVT-IV-EN were printed on A6 paper cards, which were subsequently bound into booklets. For each word, two different cards were created (each featuring a different image of the word), in order to let the participant see the word multiple times, but avoid repetition of the same image. There were two sets of verb cards and two sets of noun cards, with the words appearing in a different order for each set. This resulted in four unique ordered sets of cards in total. An additional set of 5 simple words was created for practice purposes.

2.3.2. Implementation of Entrainment

During the training or the word learning task, Robin entrained with turn-level proximity to the mean pitch of each of the participant's words by selecting from a library of prerecorded answers an answer that most accurately matched the participant's pitch. This was a semi-automatic process of five steps (Figure 2), managed by an experimenter (Experimenter 1, see Figure 3) from a control panel. First, the participant's one-word utterance was recorded. The recordings were made with custom software programmed in Python [25]. Second, Experimenter 1 manually controlled the beginning and end of the recordings to ensure they spanned the entire word and the resulting recording was processed by a volume filter to clean background noise. Third, the mean pitch value of the resulting recording was obtained using Praat [26]. Fourth, the mean pitch value was rounded to a value maximally 5 Hz higher or lower (e.g., 186.36 Hz to 185 Hz) in Robin's library of English words. This library was created by generating utterances using Robin's text-to-speech engine; their mean pitch value was transformed via a Praat script [27]. This resulted in a total of 45 options for Robin to choose from for each word, ranging in mean pitch from 130 Hz to 350 Hz with intervals of 5 Hz. Finally, Robin selected the response from the library that most closely matched the pitch value obtained at step 4 (i.e., 185 Hz).

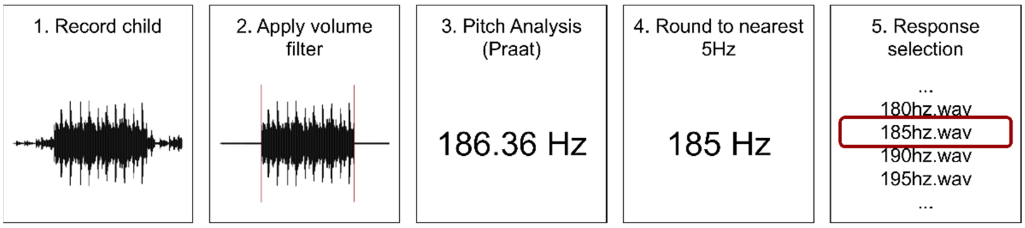

Figure 2. Overview of the semi-automatic process for the pitch entrainment.

In the control condition, Robin's mean pitch was set to 130 Hz. We opted for a fixed pitch height for the robot in the control group instead of randomly varying its pitch in order to prevent accidentally matching the robot's pitch to the participant's pitch, which would mimic the behavior of the robot in the entrainment group. This kind of accidental entrainment is very unlikely to happen with a pitch height of 130 Hz in the lower level of children's pitch range (100–700Hz) but could happen when the robot's pitch level is selected randomly.

2.4. Procedure

The participants took part in a single-session experiment with Robin. Each session consisted of seven phases. They were as follows:

Phase 1. Introduction: The participant was introduced to the robot and the task.
Phase 2. Testing (pre-test): The familiarity of the participant with the English words from the learning task was tested.
Phase 3. Practice round: The participant was prepared for the learning task.
Phase 4. Training (part 1): The first half of the word learning task was conducted.
Phase 5. Break: The robot told stories and showed the participant a few tricks (e.g., mimicking a sneeze, mimicking an orchestra conductor).
Phase 6. Training (part 2): The second half of the word learning task was conducted.
Phase 7. Testing (post-test): The participants' knowledge of the words was assessed again.

Each session started with an experimenter (Experimenter 2) leading the participant into the room, introducing themselves and the research team, and seating the participant on a pillow opposite to Robin (see Figure 3 for the test setting). The participant and Robin were seated on the floor so that their heads were at about the same height, which enabled Robin to look at the participant's face using facial recognition software. When the participant was seated, Experimenter 2 explained that they were about to play a game with Robin that would allow them to learn English words. Robin then presented themself and asked the participant introductory questions before sending the participant to do the pre-test with Experimenter 2. The pre-test was adapted from PPVT-IV-EN [24], featuring images of the 20 words from the word-learning task and recordings of these words by a female speaker of British English. Like in the original PPVT, the participant was asked to say or point at which of the four images presented to them corresponded to the word they heard. Both the visual and acoustic stimuli were presented using a PowerPoint slideshow. The participant's responses were marked on an answer sheet.

After the pre-test, the participant returned to Robin, who explained the game and played a practice round with them using the practice materials. After the practice round, the participant and Robin conducted phases 4 to 6. During each training round, the participant said Dutch words from the booklet to Robin, who replied with the English translations of the words, as mentioned in Section 2.3.1. For the entrainment group, Robin's reply was entrained to the mean pitch of the participant's last utterance. Robin's voice was set to different pitch levels in each group during phase 3 and 5, using the pitchShift parameter in the NAO text-to-speech module. The parameter was set to 1.15 for the entrainment group and 1.07 for the control group. These values were selected so that Robin's pitch in the entrainment group approached the expected mean pitch of the participants, whereas Robin's pitch in the control group was closer to the 130 Hz used during the word-learning task.

After phase 6, Robin said goodbye and invited the participant to do the post-test with Experimenter 2, which concluded the session. The post-test followed the same approach as the pre-test but had different images and recordings by a different female speaker of British English. This was performed to prevent participants from simply memorizing the combinations of acoustic signal and image.

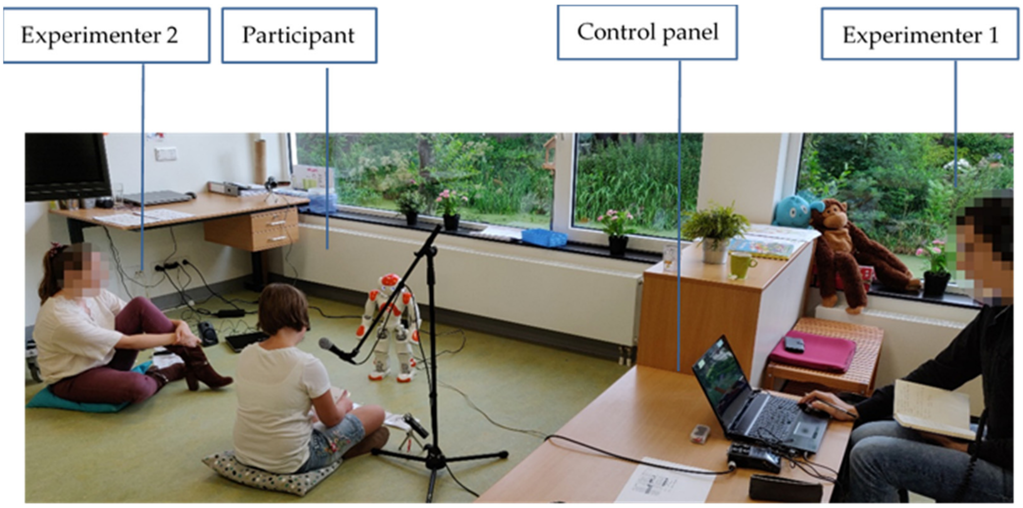

Figure 3. Overview of the test setting.

3. Results

The data of two participants were excluded from the analysis (one for prior exposure to the experiment and one for reaching the maximum score at the pre-test), leaving a total of 30 participants (15 in each group) in the final data set (Table S1). Table 1 displays the mean percentages of correctly recognized target words in decimals per phase per group. Both groups obtained a higher score in the post-test than in the pre-test. The difference between the pre- and post-test appeared to be bigger for the control group than for the entrainment group.

Table 1. Mean percentages (and standard deviation) of correct target words in decimals per test phase per group.

	Control (*N* = 15)	**Entrainment (*N* = 15)**
Pre-test	0.43 (0.50)	0.46 (0.50)
Post-test	0.75 (0.44)	0.59 (0.50)

To assess the statistical significance of these observations, a mixed-effect binary logistic regression analysis was performed using the lme4 package [28] in R [29]. Different from parametric tests, such as ANOVA, and their non-parametric counterparts, the mixed-effect model makes it possible to factor in potential influence from factors that are not controlled for, referred to as random factors. There were two fixed factors (or independent variables): the group (entrainment group vs. control group), and the test phase (pre-test vs. post-test). The random factors included participants and words. The outcome variable (or dependent variable) was the total number of correctly recognized words. Four models were built, starting with a model that only contained the random factors. Then, the fixed factors were added one by one; first came the test phase, then the group, followed by the interaction of group * test phase. The ANOVA function in R was used to compare models in order to determine the best-fit model. The best-fit model contained only a main effect of test phase ($\beta = -1.260$, SE = 0.556, $p < 0.001$), confirming that the number of words recognized by the participants was significantly higher in the post-test than in the pre-test, regardless of the group. The mix-effect logistic regression modelling thus showed that there was neither a main effect of group nor an interaction effect of group and test phase.

4. Discussion

This paper presents a novel experimental study investigating the effect of speech entrainment in child–robot interaction in RALL. As expected, the results show a significant difference between the pre- and post-test, indicating that the participants learned new words during the word learning task. This finding was in accordance with previous RALL research [30–32]. Notably, we found no interaction between the test phase (pre-test or post-test) and the group (control or entrainment). There is, thus, no evidence that entrainment in the robot can lead to a larger learning effect than otherwise. Hence, the hypothesis that pitch-level entrainment in a robot tutor will lead to better L2 vocabulary learning is not supported by the current data.

Research on the effect of robot–child speech entrainment in the context of second language learning is still sparse because implementing interactive cues as speech entrainment on currently available robots is technically difficult. While [8] showed that interactive social cues that are in synchrony with the instantaneous human behavior, such as mimicking the head inclination of the human and timely social praises, and invoked higher task compliance and liking of the robot, the social cues used in the current study are more challenging to implement and notice. While the head movement of a robot is clearly noticeable and maybe even more noticeable than when it is performed by a human due to the accompanying noise, the prosody entrainment may be less noticeable, especially for children that are engaged in a task of remembering words in a foreign language. Another study [33] previously compared the learning performance of children when a NAO robot had a robotic voice or a human voice with natural prosody. In this case, the children clearly noticed the difference, since they showed higher engagement with the robot that had prosody in its voice, i.e., the robot with the human voice. In the current study, however, the difference between different mean pitch values is more subtle than the difference between robotic prosody and natural speech prosody. In addition, the change in the voice prosody of a NAO robot is restricted and not comparable with that of a human voice. Future research may consider including a brief questionnaire on the extent to which participants notice entrainment in mean pitch on the part of the robot and testing the effect of noticing entrainment or not on learning.

Some methodological limitations that have been encountered in the current study should also be taken into account in future work. First, it is possible that our word learning task was not difficult enough for the participants, both in terms of word complexity and number of items. This may have left limited room for the participants to improve their scores from the pre-test to the post-test. Originally, the ratio of target versus filler words was 6:14, but two of the target words had to be discarded as filler words because most of the participants turned out to be familiar with them, contra to the teacher's expectations. To compensate, one filler word was recoded as a target word because not many participants were familiar with it, decreasing the ratio of target versus filler words to 5:15. In future research, it would be recommendable to conduct a vocabulary test in advance to establish children's knowledge of words and then select a sufficient number of words from the words unknown to children for the word-learning task. Furthermore, although the sample size of our study was comparable to the standard in past published studies (e.g., 23 participants per group in [19] and 19 participance per group in [21]), it was modest. It is desirable to use a large sample size in future studies. Moreover, the current analysis focused on learning outcomes after one learning session. It is, however, possible that certain interventions in child–robot interactions may have an immediate effect on learning experience, such as learner engagement, but need more time to eventually also impact learning outcomes. It might, thus, be fruitful to extend the measurement of learning to the degree of learner engagement. Engagement has been found to correlate with learning in previous research (for social robots specifically) [34,35] and with affection for and interest towards social robots among children with autism spectrum disorder [33]. These findings suggest that engagement may improve the learning experience and, thus, the motivation to learn, even if it does not immediately improve performance measures. In the same vein, which include multiple training sessions are needed to find out whether speech entrainment can lead to better learning outcomes over a longer learning period.

5. Conclusions

This study investigated whether the implementation of prosodic entrainment in a Nao robot tutor led to improved performance in learning L2 English words by monolingual Dutch 8–11-year-old children. We have found no evidence for an effect of prosodic entrainment in the robot. However, future research is called for, where the methodological limitations in the current study and the issues in implementing prosodic entrainment in robots can be circumvented and the measurement of learning can be extended to learning experience, such as engagement.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/mti5120076/s1, Table S1: Molenaar_et_al.social_robotics_MTI_TableS1.

Author Contributions: Conceptualization: B.M., B.S.F., A.C. and E.B.; software: B.S.F.; experiment conduction: B.M. and B.S.F.; data analysis: A.P., B.M. and B.S.F.; writing—original draft preparation: A.P., B.M. and B.S.F.; writing—review and editing: A.C. and E.B.; supervision: A.C.; funding acquisition: A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a Westerdijk Talent grant from the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek).

Institutional Review Board Statement: The study was conducted as the bachelor thesis project of B.M. and the master thesis project of B.S.F. under the supervision of A.C. according to the guidelines of the Declaration of Helsinki. Ethical review and approval were waived for this study, because it was not required by the institution where this study was conducted.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data reported in this paper are added as Supplementary Materials.

Acknowledgments: We would like to thank the UiL OTS labs at Utrecht University for technical support, Judith Brinksma for her assistance as Experimenter 2, Nelleke Jansen and Judith Brinksma for recording the words. A big thank-you goes to Kees de Korte and the teaching staff of the

Noachschool in Schoonrewoerd for providing testing space, helping us find participants and hosting the testing team. We would also like to extend our warm thanks to the children who participated and their parents.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bloom, B.S. The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educ. Res.* **1984**, *13*, 4–16. [CrossRef]
2. Lester, J.C.; Converse, S.A.; Kahler, S.E.; Barlow, S.T.; Stone, B.A.; Bhogal, R.S. The persona effect: Affective impact of animated pedagogical agents. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Atlanta, GA, USA, 22–27 March 1997; pp. 359–366. [CrossRef]
3. Riether, N.; Hegel, F.; Wrede, B.; Horstmann, G. Social facilitation with social robots? In Proceedings of the 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boston, MA, USA, 5–8 March 2012; p. 41. [CrossRef]
4. Song, H.; Barakova, E.; Ham, J.; Markopoulos, P. Personalizing HRI in Musical Instrument Practicing: The Influence of Robot Roles (Evaluative versus Non-evaluative) on the Child's Motivation for Children in Different Learning Stages. *Front. Robot. AI* **2021**, *8*, 282. [CrossRef] [PubMed]
5. Randall, N. A Survey of Robot-Assisted Language Learning (RALL). *ACM Trans. Hum. Robot Interact.* **2019**, *9*, 7:1–7:36. [CrossRef]
6. van den Berghe, M.A.J. Social Robots as Second-Language Tutors for Young Children: Challenges and Opportunities. Doctoral Dissertation, Universiteit Utrecht, Utrecht, The Netherlands, 2019.
7. Ghazali, A.S.; Ham, J.; Barakova, E.; Markopoulos, P. Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance. *Adv. Robot.* **2019**, *33*, 325–337. [CrossRef]
8. Belpaeme, T.; Kennedy, J.; Ramachandran, A.; Scassellati, B.; Tanaka, F. Social robots for education: A review. *Sci. Robot.* **2018**, *3*, eaat5954. [CrossRef] [PubMed]
9. Giles, H.; Coupland, J.; Coupland, N. (Eds.) *Contexts of Accommodation: Developments in Applied Sociolinguistics*; Editions de la Maison des Sciences de l'Homme; Cambridge University Press: Cambridge, UK, 1991. [CrossRef]
10. Reitter, D.; Keller, F.; Moore, J.D. Computational modelling of structural priming in dialogue. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York, NY, USA, 4–9 June 2006; pp. 121–124. [CrossRef]
11. Reitter, D.; Keller, F.; Moore, J.D. A Computational Cognitive Model of Syntactic Priming. *Cogn. Sci.* **2011**, *35*, 587–637. [CrossRef]
12. Brennan, S.E.; Clark, H.H. Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* **1996**, *22*, 1482–1493. [CrossRef] [PubMed]
13. Levitan, R.; Hirschberg, J. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In Proceedings of the 12th Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011.
14. Benuš, Š. Social aspects of entrainment in spoken interaction. *Cogn. Comput.* **2014**, *6*, 802–813. [CrossRef]
15. Mitchell, C.; Boyer, K.; Lester, J. From strangers to partners: Examining convergence within a longitudinal study of task-oriented dialogue. In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Seoul, Korea, 5–6 July 2012; pp. 94–98.
16. Ward, A.; Litman, D.J. Dialog Convergence and Learning. In Proceedings of the Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, Amsterdam, The Netherlands, 8 June 2007; pp. 262–269.
17. Thomason, J.; Nguyen, H.V.; Litman, D. Prosodic entrainment and tutoring dialogue success. In Proceedings of the International Conference on Artificial Intelligence in Education, Memphis, TN, USA, 9–13 July 2013; pp. 750–753.
18. Michel, M.; Cappellini, M. Alignment During Synchronous Video Versus Written Chat L2 Interactions: A Methodological Exploration. *Annu. Rev. Appl. Linguist.* **2019**, *39*, 189–216. [CrossRef]
19. Lubold, N.; Walker, E.; Pon-Barry, H.; Ogan, A. Automated Pitch Convergence Improves Learning in a Social, Teachable Robot for Middle School Mathematics. *Artif. Intell. Educ.* **2018**, *10947*, 282–296. [CrossRef]
20. Kory-Westlund, J.M.; Breazeal, C. Exploring the Effects of a Social Robot's Speech Entrainment and Backstory on Young Children's Emotion, Rapport, Relationship, and Learning. *Front. Robot. AI* **2019**, *6*, 54. [CrossRef] [PubMed]
21. Sadoughi, N.; Pereira, A.; Jain, R.; Leite, I.; Lehman, J.F. Creating Prosodic Synchrony for a Robot Co-player in a Speech-controlled Game for Children. In Proceedings of the 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Vienna, Austria, 6–9 March 2017; pp. 91–99. [CrossRef]
22. Xia, Z.; Levitan, R.; Hirschberg, J. Prosodic Entrainment in Mandarin and English: A Cross-Linguistic Comparison. *Proc. Speech Prosody* **2014**, *5*, 65–69. [CrossRef]
23. Podesva, R.J.; Sharma, D. *Research Methods in Linguistics*; Cambridge University Press: Cambridge, UK, 2013.
24. Dunn, L.M.; Dunn, D.M. *Peabody Picture Vocabulary Test IV*; American Guidance Service: Circle Pines, MN, USA, 2007.
25. Soliño Fernández, B. 2019-Speech-Entrainment (Version 1.1.1). 2021. Available online: https://zenodo.org/record/5330715#.YaWLe7oRWHt (accessed on 29 August 2021). [CrossRef]

26. Boersma, P.; Weenink, D. Praat: Doing Phonetics by Computer (Version 6.0). 2019. Available online: http://www.praat.org/ (accessed on 16 April 2019).
27. Weise, A. Cuny-Gc-Entrainment (Commit: ddc93b82f8352074eb23a839a583dea1c1d54d49). 2016. Available online: https://github.com/andreas-weise/cuny-gc-entrainment/ (accessed on 4 July 2019).
28. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
29. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021. Available online: https://www.R-project.org/ (accessed on 4 November 2020).
30. Alemi, M.; Meghdari, A.; Ghazisaedy, M. Employing Humanoid Robots for Teaching English Language in Iranian Junior High-Schools. *Int. J. Hum. Robot.* **2014**, *11*, 1450022. [CrossRef]
31. Eimler, S.; von der Pütten, A.; Schächtle, U.; Carstens, L.; Krämer, N. Following the White Rabbit—A Robot Rabbit as Vocabulary Trainer for Beginners of English. In Proceedings of the 6th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering in Work and Learning, Life and Leisure, Klagenfurt, Austria, 4–5 November 2010; Volume 6389, pp. 322–339. [CrossRef]
32. Mazzoni, E.; Benvenuti, M. A Robot-Partner for Preschool Children Learning English Using Socio-Cognitive Conflict. *J. Educ. Technol. Soc.* **2015**, *18*, 474–485.
33. van Straten, C.L.; Smeekens, I.; Barakova, E.; Glennon, J.; Buitelaar, J.; Chen, A. Effects of robots' intonation and bodily appearance on robot-mediated communicative treatment outcomes for children with autism spectrum disorder. *Pers. Ubiquitous Comput.* **2018**, *22*, 379–390. [CrossRef]
34. Carini, R.M.; Kuh, G.D.; Klein, S.P. Student Engagement and Student Learning: Testing the Linkages. *Res. High. Educ.* **2006**, *47*, 1–32. [CrossRef]
35. de Wit, J.; Schodde, T.; Willemsen, B.; Bergmann, K.; de Haas, M.; Kopp, S.; Krahmer, E.; Vogt, P. The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, Chicago, IL, USA, 5–8 March 2018; pp. 50–58.