Automatic Adaptation of Online Language Lessons for Robot Tutoring

Leah Perlmutter[1], Alexander Fiannaca[1], Eric Kernfeld[2], Sahil Anand[3],
Lindsey Arnold[3], and Maya Cakmak[1]

[1] Computer Science & Engineering, 185 W Stevens Way NE, Seattle WA 98195
[2] Department of Statistics, Box 354322, Seattle WA 98195
[3] Human-Centered Design & Engineering, 428 Sieg Hall, Seattle WA 98195
University of Washington

Abstract. Teaching with robots is a developing field, wherein one major challenge is creating lesson plans to be taught by a robot. We introduce a novel strategy for generating lesson material, in which we draw upon an existing corpus of electronic lesson material and develop a mapping from the original material to the robot lesson, thereby greatly reducing the time and effort required to create robot lessons. We present a system, KubiLingo, in which we implement content mapping for language lessons. With permission, we use Duolingo as the source of our content. In a study with 24 users, we demonstrate that user performance improves by a statistically similar amount with a robot lesson as with Duolingo lesson. We find that KubiLingo is more distracting and less likeable than Duolingo, indicating the need for improvements to the robot's design.

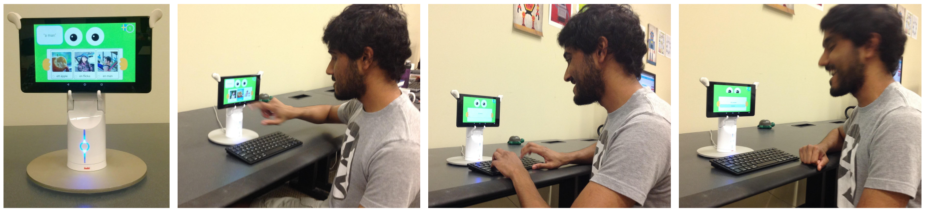

Fig. 1. KubiLingo is a social robot system that teaches languages to people

1 Introduction

The internet is a vast source of knowledge and information, but the majority of web content is in English. This means that most massive open online courses (MOOCs) [4], academic publications [21], and content in general [8] is accessible only to those privileged enough to know English. Concurrently, a great many people immigrate to countries where they do not know the language. These are just a couple situations that demand language education at a higher rate than its current availability.

In this work, we develop a tabletop robot that teaches languages to people. The role of our system is to support a human teacher with personalized supplementary lessons when the teacher is not available. We propose an embodied robot rather than a simple computerized lesson because studies show that the physical presence of a robot results in greater learning gains than when the same content is presented without a robot [18,15,14].

One bottleneck in developing robots that teach is custom content generation. It is time-consuming and requires expertise in both the subject being taught and in teaching techniques. Furthermore, it is redundant when electronic lesson content (albeit not in robot form) already exists. Our main contribution is to introduce the strategy of *content mapping*, which takes advantage of existing lesson content by adapting it to a form that the robot can teach. To the extent that source lessons follow a predictable format, our system can teach any lesson in the source corpus. As a result, we can draw on a much larger corpus of lessons.

2 Related work

Technology has long been used for teaching languages. Rosetta Stone is an example of a paid program, involving a mixture of computer-based and teleconferenced lessons [1]. Duolingo is a newer, free program offering web-based and mobile content [6].

Recently, research has been done teaching language with robots. Telepresence robots, as in the work of Kwon *et al.* [16] can be helpful when a human teacher is available but not colocated with students. Others have proposed or developed autonomous teaching robots [17,10], which can be helpful when human teachers have limited time or availability. Our robot falls into the autonomous category.

Many have demonstrated cognitive learning benefits when material is taught by embodied, autonomous robots as compared with non-embodied agents, computer-based lessons, or paper-based lessons [18,15,14]. This research indicates promise for teaching robots and helps to form the foundation of our work.

Social interaction is an important attribute of robots that teach. Saerbeck *et al.* show that a robot tutor's socially supportive behavior increases learning efficiency in students [20], while Kennedy *et al.* show that it is possible for a robot to be too social during teaching, countering learning gains [15]. This research informs our work in designing the social interactivity of our system.

Content is hard to develop for robots that teach. In many cases, researchers have created custom content for their robots to teach [17,15]. We introduce a mapping from existing content to robot content, reducing the amount of effort needed to provide content for the robot to teach. We consider our work to contribute an advancement in the area of content for robots that teach.

3 System

3.1 Hardware and software

Our hardware consists of a Nexus 7 tablet mounted on a Kubi base, as seen in Fig. 1. Designed by Revolve Robotics [7], Kubi is a telepresence platform which holds the tablet and is actuated with two degrees of freedom: pan and tilt.

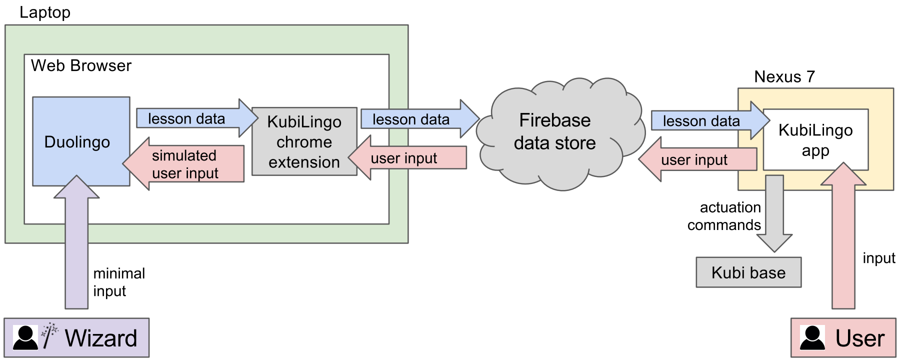

Fig. 2. System architecture diagram

Our robot system, KubiLingo, runs in an Android app. The robot character, named “Kubi,” has an animated face with large eyes, illustrated in Fig. 3. It shows emotions by speaking, animating the eyes, and actuating the base. Kubi also has virtual, animated hands, and displays lesson material on a virtual card held in its hands.

As the backend for our lesson content, we use Duolingo, an online language teaching platform designed to teach any language to speakers of any other language. Duolingo’s content is crowdsourced, so there is a sustainable way to generate more content, resulting in a large and growing corpus of lesson material. Duolingo has given us permission to use their content as described below.

Our system wraps Duolingo, automatically adapting Duolingo’s content to be rendered on the robot and transmitting user input back to Duolingo, as illustrated in Fig. 2. To do this, we built a browser extension (“KubiLingo chrome extension”) that runs on Chrome while Chrome runs a Duolingo lesson. It parses lesson data from the DOM and sends it to the robot in real-time via Firebase, a cloud-based service [2]. It also receives user input data from the robot and simulates that user input in the browser. The flow of data is fully automated except the confirmation step, in which the user verbally confirms their answer to each prompt. A wizard-of-oz operator listens for the user’s verbal confirmation and clicks the confirmation button in the web browser.

3.2 Visual design and animation

We chose a vibrant color scheme similar to Duolingo’s to make the learning experience fun and playful. We designed Kubi to resemble Duo, the owl mascot of Duolingo, by giving it green and orange coloring. Related works inspired some other features. Ribiero *et al.* applied Disney’s animation principles to help users understand a robot’s emotions [19]. Accordingly, we designed Kubi’s facial expressions using these principles. Cuijpers *et al.* found that users perceive robots with idle motions as more alive and empathic [12], so we implemented idle motions in the form of random head movements and periodic eye blinking.

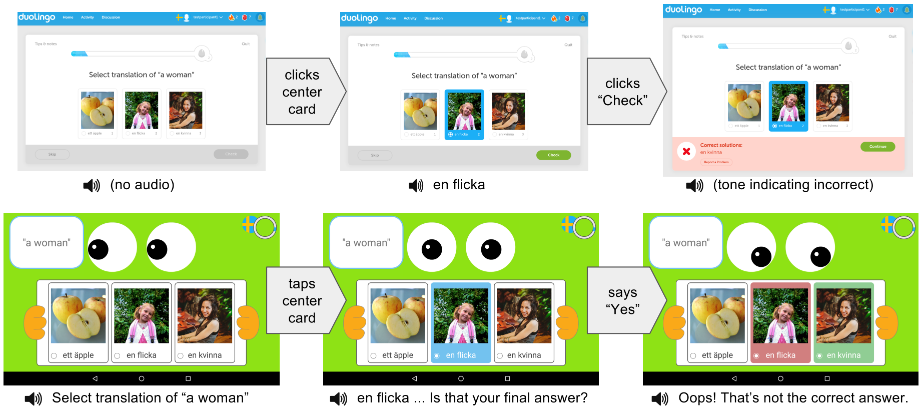
duolingo
Select translation of "a woman"
ett äpple
en flicka
en kvinna
clicks center card
clicks “Check”
Correct solution:
en kvinna
Continue
(no audio)
en flicka
(tone indicating incorrect)
"a woman"
taps center card
says “Yes”
Select translation of “a woman”
en flicka ... Is that your final answer?
Oops! That's not the correct answer.

Fig. 3. The chronological progression of a prompt on Duolingo (top) and KubiLingo (bottom). User actions are described in the gray arrows.

3.3 Content mapping

With the differences between computer and robot in mind, we created a mapping to automatically convert Duolingo lessons into robot lessons.

The basic unit of interaction in Duolingo is a *prompt*, in which some material (text, audio, pictures) is delivered to the user and the user is expected to reply by typing or clicking. Below we describe the different conceptual parts of a prompt, and how they are delivered in Duolingo and on the robot.

The *directive* is a description of what to do in this prompt, e.g. “Select translation of ‘a woman’.” Duolingo displays the directive as text; Kubi speaks the directive and displays key words (e.g. “a woman”) in a speech bubble.

The *body* is the main part of the prompt, e.g. selectable flash cards. Both Duolingo and KubiLingo display the body centrally; KubiLingo displays it on a virtual card in the robot’s hands. A Duolingo prompt requires clicking or typing input. In KubiLingo, clicking is replaced with tapping and typing is done using a bluetooth keyboard.

A *hint* is extra data related to a word. In Duolingo, the user can click an underlined word for a hint. A popup appears just below the cursor showing translations of the clicked word. In KubiLingo, we underline words with available hints, and when the user taps a word, Kubi shows the hint in a speech bubble.

Checking is when the user indicates they are finished providing input and wish to check their answer. Duolingo provides a “Check” button for the user to click. Kubi asks, “Is that your final answer?” The wizard listens for an affirmative response from the user, and clicks “Check” in the Duolingo interface.

Feedback is displayed after the user submits a response, and indicates whether the the response was correct. Feedback can display alternate answers or corrections to mistakes. Duolingo displays feedback just below the prompt body; KubiLingo displays it overlaid with the body. In addition, Kubi provides verbal feedback and emotes a reaction. For example, if the response was incorrect, Kubi might lower its head, show sad eyes, and say "Sorry, that's not the right answer." KubiLingo has a variety of positive and negative responses, and randomly selects an appropriate one in the feedback stage of each prompt.

When the user is finished with the feedback, they *advance* to the next prompt. For this purpose, Duolingo provides a "Continue" button to click. In KubiLingo, the wizard waits for Kubi to present the feedback, then clicks "Continue".

Progress indicates the number of prompts the user has completed in the current lesson. A visual indicator advances when the user answers correctly and recedes when the user answers incorrectly. Duolingo displays a progress bar above the body; KubiLingo displays a progress ring in the upper right corner.

We implemented the mapping for five different types of prompt: *Select*, in which the user selects a flash card matching the spoken word, *Translate*, in which the user types the translation of a word or phrase, *Name*, in which the user types the noun shown in three pictures, *Listen*, in which the user types a spoken phrase, and *Judge*, in which the user selects one or more options from a list of choices. These five prompt types account for all the material taught in the lessons we used for our system evaluation. Fig. 3 illustrates the chronological progression of a Select Prompt on both Duolingo and KubiLingo.

4 Evaluation

Hypotheses and conditions We conducted a user study to compare KubiLingo with Duolingo. We hypothesized:

- **H1 (Performance hypothesis)** User performance from pre-test to post-test will improve more with a robot lesson than with a screen lesson.
- **H2 (Preference hypothesis)** Users will subjectively rate the robot higher than the screen.

H1 is supported by previous work showing that embodied robots have cognitive learning benefits [18,15,14]. Proving H1 alone, however, would not sufficiently show that KubiLingo is a better learning platform. It would be possible to be more effective but less appealing, in which case users may abandon it, learning less overall. We formulated H2 to test whether KubiLingo is more appealing.

Another goal of our study, not covered by our hypotheses, was to gather feedback for the next design iteration of KubiLingo. We wanted to measure the quality of KubiLingo's user experience and learn which features impacted it.

To test our hypotheses, we designed a study with two conditions: **Screen lesson**, a Duolingo lesson taken on a laptop; and **Robot lesson**, a lesson taken from KubiLingo. In our crossover study design, each participant took one of each type of lesson, in two consecutive sessions. We counterbalanced the order, creating two arms of the study, **Screen-first (S)** and **Robot-first (R)**.

To prevent carryover effects, we taught Swedish in Session 1 and Dutch in Session 2. We did not counterbalance languages (which would create four study arms) because the two study arms mentioned above were sufficient to isolate the effects necessary to test our hypotheses. Each "lesson" of our study consisted of three Duolingo lessons: "Basics" lessons 1-3 for the language being taught (L2).

Participants We recruited participants from a university community. Those interested completed an eligibility survey, and those qualified were sorted into the two study arms using stratified randomization [13]. We excluded participants understanding Swedish, Norwegian, Danish, Icelandic, Dutch, German, or Flemish and stratified those knowing “a little bit” of those languages. We also stratified on bilingual ability and age at which English was learned. No participant reported visual or hearing impairments that would block perception of visual or audio lesson material. Participants were assumed to be proficient in English. We had a total of 24 participants, 12 in each study arm.

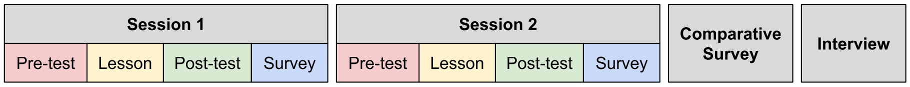

Fig. 4. Parts of the user study

Procedure When they arrived in the study room, participants signed a consent form and the facilitator started video and audio recording. The wizard sat on the other side of the room with a laptop, and was introduced as “tech support”.

The participant was seated at a laptop for a pre-test. Next, they completed the first language lesson with either laptop or robot, followed by a post-test and subjective survey. Then they completed the second session – pre-test, lesson, post-test, and survey. After both sessions, they completed a survey comparing the two. Then the facilitator provided a debrief, interviewed them for feedback on the robot, and offered compensation. Total duration was about 45 minutes.

5 Findings

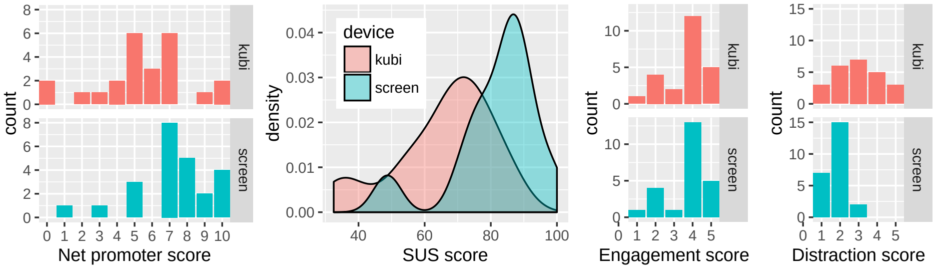

Fig. 5. System characterization results. Net promoter score is out of 10. SUS is out of 100, as described in [11]. Engagement and distraction are on a 5-point Likert scale.

System characterization To characterize usability, likeability, and engagement of robot and screen users completed a survey after each session. The survey included questions for System Usability Scale (SUS) [11], Net Promoter Score [5], engagement, and distraction. Afterwards, the facilitator interviewed the user

	Net promoter			SUS			Engagement			Distraction		
	mean	SD	SE	mean	SD	SE	mean	SD	SE	mean	SD	SE
screen	7.21	2.21	0.45	81.74	12.82	2.67	3.71	1.12	0.23	1.79	0.59	0.12
kubi	5.50	2.57	0.52	65.73	14.95	3.05	3.67	1.13	0.23	2.96	1.23	0.25

Table 1. Mean, standard deviation, and standard error of the mean for Fig. 5. "SE" denotes standard error of the mean.

for feedback on which features were most engaging and distracting, and how to improve the robot. Fig. 5 shows quantitative results. We used paired t-tests to test for a difference between robot and screen scores in each category.

Users found KubiLingo to be engaging, rating it an average of 3.67, 95% CI [3.16, 4.18] on a 5 point Likert scale. Duolingo's engagement score is similar: 3.71, 95% CI [3.20, 4.22] (p=0.90). Despite KubiLingo's engaging ratings, many participants found it to be distracting as well. Not everyone found the robot distracting – the standard deviation of 1.23 indicates a wide spread of user opinions. Duolingo's distraction ratings were significantly lower (p=0.001). Comments are shown in Table 2.

An analysis of many studies' SUS ratings by Sauro finds an average score of 68 [3]. At 65.73, 95% CI [59.02, 72.44], KubiLingo's score is near the average. Using the semantic scale from Bangor *et al.* [9], we can assign semantic interpretations to the SUS ratings. KubiLingo's SUS ratings are between "OK" and "Good", but closer to "Good". In comparison, we measured Duolingo's SUS ratings significantly higher (p<0.001), attaining better than "Good", but not quite "Excellent". Past Duolingo use had no significant effect on SUS.

KubiLingo's Net Promoter Scores were significantly lower than Duolingo's (p=0.01).

	most engaging	most distracting	interview comments
kubi	gesture (9) positive feedback (4) robot's eyes (4)	confirmation step (10) robot's movement (9)	speed it up (7) confirmation annoying (6) meaningless movements (4) gaze direction inapropriate (3)
screen	pictures (6) pronouncing words (5) sound effects (5)	(nothing) (9) lesson structure (4)	n/a

Table 2. Summary of the most common user feedback. Number of users in parentheses.

Performance evaluation To test the performance hypothesis (H1), we administered a pre-test before each lesson and a post-test after. We designed the tests to measure short-term memory of the material. Each test had two pages of 16 questions each which required typing the translation of a phrase. Page 1 had English-to-L2 translation and page 2 had L2-to-English translation. The pre-test and post-test had the same questions with order randomized on each page. We measured performance by subtracting post-test score from pre-test score.

We modeled the data using a generalized estimating equations (GEE) linear regression with an unstructured working correlation matrix [22]. GEE accounts for dependence between repeated measurements of the same person and allows us to estimate the effect of KubiLingo vs. screen. Our model also accounts for variability caused by Swedish versus Dutch and Session 1 versus Session 2.

Performance difference between KubiLingo and Screen was not significant (p=0.50). The effect size had a mean of -0.78 (SE=1.49, 95% CI [-3.04, 1.47]), meaning that users improved by about 1 question more with KubiLingo than with Screen. This result neither supports nor contradicts our performance hypothesis (H1).

	mean	lower	upper	p-value	most cited reasons
natural	0.79	0.22	1.36	0.01	computer more familiar (7) computer pacing more natural (7) robot confirmation not natural (7)
fun	-0.29	-0.94	0.36	0.36	robot's movement (8) robot's novelty (5) robot's eyes (4) robot's encouragement (4)
prefer	1.08	0.64	1.53	<0.001	screen faster or more efficient (8)

Table 3. Preference evaluation ratings and reasons. Scale is from -2 (strong robot) to 2 (strong screen). Data includes mean, upper and lower ends of confidence interval, and p-value from two-sided t-test of H2: mean=0.

Preference evaluation To test the preference hypothesis (H2), we administered a survey at the end, comparing the devices used in both sessions. The survey contained three questions, "Which system felt more natural", "Which system was more fun or entertaining", and "Which system would you prefer to use next time you are trying to learn a language". Users responded on a 5-point Likert scale and wrote an explanation for their choice. Choices were randomly flipped left-to-right to account for left-preference.

Users reported that the screen felt more natural and that they would prefer to use the screen to learn a language in the future. Results for which system was more fun/entertaining were inconclusive (p=0.36) but the mean is slightly in favor of KubiLingo. Past Duolingo use had no significant effect. These results showing preference for the computer contradict our preference hypothesis (H2).

6 Discussion

The results show the KubiLingo system has average usability and that users prefer Duolingo. KubiLingo didn't match Duolingo's usability score; however, Duolingo is a professionally designed product that sets a high bar. KubiLingo's

SUS rating indicates that there is room for improvement, and the user feedback summarized in Table 2 will inform revisions in Kubi's movement pattern and interaction cadence.

The results do not conclusively show that KubiLingo or screen has greater learning gains. This means we haven't made the lessons detectably worse by teaching them with KubiLingo. KubiLingo's effectiveness is on par with Duolingo's. Our system has a room for improvement in usability, and we speculate that the requisite usability improvements will increase likeability, reduce distraction, and lead to learning gains, potentially exceeding Duolingo's effectiveness.

In future work, we hope to iterate on KubiLingo's design and test it in different contexts. One new context would be as a conversation partner for language learners. To maintain KubiLingo's independence from custom hand-built content, we could harness web-based chat bots. Conversation would be a good way for users to gain experience and boost their confidence with spoken language. It also offers more opportunities for linguistic production, an important skill in language learning. We believe that Kubi would be a better conversation partner than a computer because of its embodiment and social agency.

We also hope to test KubiLingo with children. We had kids in mind when we designed the system, but did this study with adults because it is easier to get adult participants. A number of our users thought that KubiLingo would be better suited for a younger audience. It is also likely that children would rate KubiLingo differently in terms of subjective preference.

7 Conclusion

In this paper, we described a novel approach to creating lessons for robots that teach: *content mapping*, in which existing content for electronic lessons is automatically converted to a form that an embodied robot can teach, thereby greatly expanding the amount of content available for limited effort. We presented an embodied robot that teaches language to people and uses content mapping to construct lessons. We tested our system in a user study to compare its performance with Duolingo, the source of the lesson content. Our system had equivalent learning results to those of Duolingo but users found Duolingo more likeable. We interpret the likeability result as a tribute to Duolingo's excellent user experience and a sign that our system has room for design improvements. Independently of this result, we believe that our contribution of automatic content mapping has great potential for near-term application in the development of robots that teach.

8 Acknowledgements

The authors thank Vivek Paramasivam, Joanna Bailet, Amy Tang, Evan Blajev, and Daniel Zhu for facilitating user studies and Burr Settles for enabling our use of Duolingo. This work was funded in part by a grant from RAS-SIGHT.

References

1. About Rosetta Stone. http://www.rosettastone.com/about
2. Firebase. https://www.firebase.com/
3. Measuring Usability with the System Usability Scale (SUS): MeasuringU. http://www.measuringu.com/sus.php
4. MOOCs Directory. http://moocs.co/Home_Page.html
5. The Net Promoter Score. https://www.netpromoter.com/know/
6. Where can I use Duolingo? http://support.duolingo.com/hc/en-us/articles/204829260-Where-can-I-use-Duolingo-
7. Why Kubi. https://www.revolverobotics.com/
8. Usage Statistics and Market Share of Content Languages for Websites, June 2016. https://w3techs.com/technologies/overview/content_language/all (Jun 2016)
9. Bangor, A., Kortum, P., Miller, J.: Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. J. Usability Studies 4(3), 114–123 (May 2009)
10. Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E.E., Kopp, S., Bergmann, K., Leseman, P., Küntay, A.C., Göksun, T., Pandey, A.K., Gelin, R., Koudelkova, P., Deblieck, T.: L2TOR - Second Language Tutoring using Social Robots. In: WONDER 2015. Paris, France (2015)
11. Brooke, J.: SUS: a 'quick and dirty' usability scale. In: Usability Evaluation In Industry, pp. 189–194. CRC Press (Jun 1996)
12. Cuijpers, R.H., Knops, M.A.M.H.: Motions of Robots Matter! The Social Effects of Idle and Meaningful Motions. In: Tapus, A., André, E., Martin, J.C., Ferland, F., Ammi, M. (eds.) Social Robotics. pp. 174–183. Lecture Notes in Computer Science, Springer International Publishing (Oct 2015)
13. Friedman, L.M., Furberg, C., DeMets, D.L., Reboussin, D.M., Granger, C.B.: Fundamentals of clinical trials, vol. 4. Springer (2010)
14. Han, J., Jo, M., Park, S., Kim, S.: The educational use of home robots for children. In: ROMAN 2005. pp. 378–383 (Aug 2005)
15. Kennedy, J., Baxter, P.: The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In: HRI 2015. vol. 2015 (2015)
16. Kwon, O.H., Koo, S.Y., Kim, Y.G., Kwon, D.S.: Telepresence robot system for English tutoring. In: ARSO 2010. pp. 152–155 (Oct 2010)
17. Lee, S., Noh, H., Lee, J., Lee, K., Lee, G.G., Sagong, S., Kim, M.: On the effectiveness of Robot-Assisted Language Learning. ReCALL 23(01), 25–58 (Jan 2011)
18. Leyzberg, D., Spaulding, S., Toneva, M., Scassellati, B.: The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. In: CogSci 2012. pp. 1882–1887. Cognitive Science Society (2012)
19. Ribeiro, T., Paiva, A.: The Illusion of Robotic Life: Principles and Practices of Animation for Robots. In: HRI 2012. pp. 383–390. HRI '12, ACM, New York, NY, USA (2012)
20. Saerbeck, M., Schut, T., Bartneck, C., Janse, M.D.: Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor. In: CHI 2010. pp. 1613–1622. CHI '10, ACM, New York, NY, USA (2010)
21. van Weijen, D.: The Language of (Future) Scientific Communication. https://www.researchtrends.com/issue-31-november-2012/the-language-of-future-scientific-communication/ (Nov 2012)
22. Zeger, S.L., Liang, K.Y.: Longitudinal data analysis for discrete and continuous outcomes. Biometrics pp. 121–130 (1986)