Article

LANGUAGE TEACHING RESEARCH

Language Teaching Research
1–26


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/13621688251367852
journals.sagepub.com/home/ltr

Situated L2 pronunciation instruction during small-group robot-assisted language learning activities

Teppo Jakonen
University of Turku, Finland

Derya Duran
University of Turku, Finland

Pauliina Peltonen
University of Turku, Finland

Abstract

Chatbots and other conversational agents based on speech recognition and processing technologies have been gaining ground in the field of language education. Although previous research has shown that automatic recognition of second language (L2) speech is difficult, little attention has been paid to how L2 teachers and learners interact with such technology when used as an interactional participant in classroom settings. Addressing this gap, this article provides a qualitative analysis of interactional practices of unplanned and situated pronunciation instruction as a teacher and 10- to 13-year-old young learners of L2 English complete robot-assisted language learning (RALL) activities in a primary school English-as-a-foreign-language (EFL) context in Finland. Drawing on 14 hours of video recordings, we use multimodal conversation analysis (CA) to analyse extended repair sequences that involve interactional problems related to word recognition by a social robot. Through a sequential analysis of selected data extracts, we show how the teacher and learners correct these problems by establishing a corrective focus for providing instruction on and modifying learners' word-level pronunciation, such as the quality of individual sounds or word stress. From the teacher's perspective, this consists of drawing learners' attention to pronunciation details by highlighting sounds in learners' talk and the robot's talk, using embodied conduct, and modelling a target-like word pronunciation. Our findings shed

Corresponding author:
Teppo Jakonen, School of Languages and Translation Studies, University of Turku, Arcanum Building, Arcanuminkuja 1, Turku FI-20014, Finland.
Email: teppo.jakonen@utu.fi

light on the interactional organisation of RALL activities and some of the real-life consequences of limitations in speech recognition technologies for L2 teaching and learning interactions with conversational agents. The work conducted by the teacher to convert interactional troubles into meaningful learning opportunities suggests that human agency is needed to optimally guide and mediate language learning interactions with conversational agents based on artificial intelligence (AI) and automatic speech recognition (ASR), as these agents are less capable of showing the kind of interactional and instructional adaptation that is part of human–human interaction.

Keywords
conversation analysis (CA), correction, learnables, pronunciation teaching, robot-assisted language learning, young language learners

I Introduction

Conversational applications drawing on a range of artificial intelligence (AI) technologies are rapidly becoming more prominent in language education. At the same time, research has begun to explore their impact on second language (L2) learning by analysing how teachers and learners use tools and applications such as Duolingo (e.g., Loewen et al., 2019), intelligent personal assistants such as Alexa (Dizon et al., 2022; Hsu et al., 2023), ChatGPT or other chatbots based on large language models (Han, 2024; Xiao & Zhi, 2023), and social robots (Veivo & Mutta, 2025). The emergence of these AI-based tools and the technological hype particularly around generative AI have also spurred critical debate among language teaching practitioners and researchers concerning the value and relevance of "traditional" language instruction methods as well as the role of human language teachers in the learning process. A recent example of the ongoing discussion is Kern's (2024) suggestion that language education has reached "an inflection point" that, among other things, requires a critical investigation of "what technology offers that is positive for language education [and] rethink how we organize our teaching in light of technology's affordances" (p. 516). Indeed, many language instructors are worried that now that technology can produce, assess, and correct language with a few clicks, it may lower incentives for language study and increase student cheating. These developments necessitate attention to the "pedagogy–technology interface" (Neri et al., 2002) if we wish to understand not only how language teachers and learners may optimally use the affordances of technological tools but also how they compensate for the unavoidable limitations of such tools. Rather than asking what the value of "traditional" language instruction is in new technological environments, there is a need to (also) ask what the pedagogical value of these technologies is.

In this study, we respond to Kern's (2024) research call by exploring language learning interactions with social robots. Social robots are one example of conversational agents that use a range of AI-powered language technologies and tools, such as natural language processing and automatic speech recognition (ASR). Studies of human–computer interaction have shown that ASR systems tend to struggle when processing talk by children and L2 speakers because their talk has typically not been used as data to train these systems, which may lead the ASR system to misinterpret phenomena such as

lexical disfluencies, hesitations, and code-switching (e.g., Alharbi et al., 2021; Cumbal, 2024). On the other hand, studies of social interaction show that human–human interaction is very complex yet highly systematic and organised (Sacks et al., 1974), which is why it is important to explore the situated nature and "naturalness" of human interaction with AI in real-life contexts beyond laboratory settings (see Mlynár et al., 2025; Voss & Waring, 2025). This can help us to better understand how language learning situations with ASR-based conversational agents are interactionally organised and what kind of human agency and adaptation is needed to create meaningful learning opportunities with the help of, or sometimes despite, such agents.

Motivated by the need to increase understanding about interactions involving human participants and ASR-based conversational agents, our aim in this article is to shed new light on practices of L2 teaching and learning in naturalistic settings in which technology is not just a tool but an interactional participant. We present a qualitative study that explores interactional practices of pronunciation instruction that occur between a teacher and 10- to 13-year-old L2 users of English as they work in small groups with a social robot programmed to complete robot-assisted language learning (RALL) activities. Drawing on the methodological perspective of multimodal conversation analysis (CA), we focus our analysis on sequences involving interactional problems, which the participants observably attribute to a learner's (deviant) pronunciation. Our analysis is guided by the following research question:

> RQ: How do human participants turn interactional troubles related to a social robot's word recognition into L2 pronunciation teaching and learning opportunities?

In what follows, we will first review existing literature on ASR technologies in pronunciation teaching (section 2) and introduce our data and methods. We will then present our findings by analysing five selected episodes of RALL interaction (section 4), discuss these findings in relation to earlier studies (section 5), and explain our conclusions and the main contributions of the study (section 6).

II Theoretical background

In this section, we discuss relevant literature for our study from two perspectives: the use of ASR-based technologies, including social robots, in the context of pronunciation learning (section on ASR systems in computer-assisted pronunciation teaching and RALL) and of L2 pronunciation instruction, with a focus on young learners (section on pronunciation teaching practices in L2 learning).

1 ASR systems in computer-assisted pronunciation teaching and RALL

The use of technology for pronunciation teaching and feedback has been previously investigated under the heading of computer assisted pronunciation teaching (CAPT; see e.g., O'Brien et al., 2018). While many researchers have called for more research in CAPT (Derwing, 2017), especially due to the rapid advances in technology,

even in relatively recent overviews on the topic (e.g., Chun & Jiang, 2022), emerging technologies such as social robots are extremely rarely mentioned as potential sources of pronunciation feedback, highlighting our study's novel angle on the topic. Existing CAPT studies have typically investigated the benefits of different kinds of ASR-based programs for L2 pronunciation learning using experimental designs (Liakin et al., 2015; McCrocklin, 2019). These studies have demonstrated that ASR-based systems can help learners practice and improve their pronunciation skills (Inceoglu et al., 2020; Neri et al., 2008; Tejedor-García et al., 2021) as well as enhance their interest in using them (Shadiev et al., 2019). Liakin et al. (2015), for example, found that mobile ASR-based pronunciation training had a greater impact than human instruction involving a teacher on the development of L2 French segmental features. Similarly, in a 15-week-long classroom study, García et al. (2020) showed that a group of L2 Spanish learners trained through ASR-based pronunciation programs improved more on certain phonemes than an instructor-led pronunciation training group.

As one application area of ASR technology, social robots are gaining momentum in second language acquisition (SLA) research as a potential way to facilitate language development. One justification of social robots is that they could offer opportunities for language practice that, in addition to talk, involves a broader range of nonverbal resources, such as facial expressions, gestures, and body posture, compared to voice-only conversational agents. According to RALL research reviews, social robots have been found to bring about at least short-term language learning gains and affective benefits in the form of lowering learners' anxiety as well as increasing their motivation to learn and willingness to interact (e.g., H. Lee & Lee, 2022; Randall, 2019; van den Berghe et al., 2019). RALL technologies are developing rapidly, but existing state-of-the art reviews suggest that currently, a feasible mode of RALL implementation is to supplement human instruction rather than replace it (Randall, 2019). Robots can nevertheless provide learners with opportunities for language production, repetition, and interaction with an embodied conversational partner.

While an increasing number of studies have explored how social robots can be used to teach speaking skills to both school-aged children and adults (e.g., Cumbal, 2024; Iio et al., 2019; S. Lee et al., 2011), only a handful of studies have investigated the specific topic of robot-assisted L2 pronunciation learning. Iio et al. (2019), for example, developed a robot-assisted system to enhance English conversations of Japanese adult learners, which focused on the complexity, accuracy, and fluency of the speech of those learners. The authors found that the participants had improved their English conversational skills by the end of the experiment. In and Han (2015) investigated the use of synthesised speech for RALL and found the range in speech melody in synthesised speech to be much lower than native speaker utterances. That is, it did not serve as useful learning stimuli for language learning because learners even corrected their speech in the wrong direction. As a rare example of an L2 pronunciation-focused study involving a social robot, Krisdityawan et al. (2022) employed a social robot as a teacher to instruct Japanese adult learners' English pronunciation and prosody. The learning materials—29 target English words to improve vowel and consonant sounds, and two short passages with patterns of intonation and phrasing—were provided in slides and a NAO robot voiced the instructions to the learners. It led learners to produce each target word,

instructed them on the correct pronunciation, and guided them to read the texts. This led the learners to improve the intelligibility of their pronunciation and to adopt a more target-like prosody, highlighting the potential of social robots for facilitating pronunciation development.

As RALL research is still very much emerging, more research-based evidence is needed to identify favourable conditions and best practices for using social robots in language education. So far, RALL practices and processes have been mainly studied through experimental methodologies, controlled task interactions, and quantitative analyses (e.g., Amioka et al., 2023; Iio et al., 2019; Krisdityawan et al., 2022), similarly to how research has approached other kinds of conversational agents using ASR technologies. While experimental and quantitative studies have provided valuable information on the effects of ASR-based training on pronunciation development, there is a need to complement this with qualitative, fine-grained explorations of actual instances of teaching involving ASR-based tools, such as the social robot examined in the present study. Such studies can help us obtain a clearer picture of the "pedagogy–technology interface" (Neri et al., 2002) of ASR-based tools, including their affordances and limitations in L2 classroom contexts and the potentially very different ways in which teachers and learners can relate to and interact with such tools in naturalistic settings. In the field of RALL, this could provide insight into not only how instructional patterns of interaction between a robot and learners may differ from teacher–learner interaction, but also whether robots may be able to support language teachers to produce pedagogically meaningful instructional actions. As Ward (2025) argues, there is a "fantasy" of authenticity around conversational agents, which needs to be critically examined. The present study is one step towards this direction in its aim to explore situated pronunciation instruction as the interplay of human and technological resources during small-group RALL activities in classroom-based settings involving young learners.

2 Pronunciation teaching practices in L2 learning

The impact of pronunciation instruction on the development of L2 pronunciation skills has been widely examined in SLA research, with results generally suggesting a positive effect for instruction (for a meta-analysis, see J. Lee et al., 2015). To explain why pronunciation instruction may be effective, Sardegna and McGregor (2022) conducted a systematic literature review of studies reporting on teaching experiments published between 2015 and 2020. Based on closer scrutiny of 15 studies, they found that all the studies included explicit instruction and complemented it with some essential instructional approaches, such as awareness-raising, perceptual training, oral production practice, and/or corrective feedback. Overall, the bulk of research on pronunciation instruction has focused on adult learners, such as university students, while studies with younger beginner learners are rarer (see also e.g., O'Brien et al., 2018). The focus on adult learners is often reflected in the instructional approaches documented in previous literature, some of which may not be entirely applicable to young L2 learners. However, some recent experimental studies conducted in laboratory conditions (e.g., Immonen et al., 2022) and classrooms (e.g., Baills & Prieto, 2023; Gómez Lacabex & Gallardo-del-Puerto, 2014; Gómez Lacabex et al., 2022) have addressed young learners'

pronunciation and highlighted aspects that are suitable for supporting pronunciation development in this age group, such as repetition and embodiment (both relevant also to the present study). Furthermore, Tergujeff (2022, p. 244) has argued that playful activities may be most suitable for young learners and, especially for children who do not yet know how to read, mimicking and listening might be particularly relevant activities (see also Kirkova-Naskova, 2019). Repetition and embodied pronunciation practice are also central to the activity analysed in the present study, in which young learners interact with a social robot.

Previous studies have also shown that L2 teaching activities with a focus on other language skills, such as vocabulary practice (e.g., Duran & Kääntä, 2023; Gordon, 2022) or reading aloud (Kääntä, 2017), may also bring about an emergent focus on pronunciation. Typically, such a focus is established through a range of corrective feedback practices provided by the teacher, a peer learner, or the learners themselves in various forms (Ellis, 2009). As Sardegna and McGregor (2022, p. 118) aptly summarise, "L2 learners are often lost in their pronunciation journey; corrective feedback can be a compass guiding them in the direction of their pronunciation goal destination" (see also e.g., Saito & Lyster, 2012). In line with the overall focus on adult learners in pronunciation instruction research, previous studies on corrective feedback in the context of L2 pronunciation have focused, for instance, on pre-service and in-service teachers' beliefs and practices (e.g., Baker & Burri, 2016; Lintunen et al., 2023), but rarely on young learners. In addition, some experimental studies have examined pronunciation features along with other aspects of L2 production as part of a broader focus on interactional feedback aimed at improving comprehensibility ("negotiation for comprehensibility," Saito & Akiyama, 2017; on intelligibility and comprehensibility as the key targets in L2 pronunciation, see e.g., Galante & Piccardo, 2022).

As a relatively recent development in research on L2 pronunciation instruction and corrective feedback, some experimental and classroom-based studies have explored the embodied nature of pronunciation instruction and correction. In particular, these studies have shed light on how resources such as gestures can facilitate pronunciation teaching and learning. For instance, in a training experiment involving 28 Catalan children, Baills and Prieto (2023) demonstrated that rhythmic clapping facilitated pronunciation learning among young learners of French. Similar observations on the facilitative role of gestures and other multimodal resources have been made in interactional studies. For example, gestures can make suprasegmental features of pronunciation (e.g., stress, rhythm) visible, and thus help L2 learners recognise errors in their production and correct them accordingly (e.g., Nguyen, 2016; Smotrova, 2017). For suprasegmental features, language instructors may, for example, use upward body movements (Smotrova, 2017) to visualise and embody word stress and the rhythmic patterns of a target language. Nguyen (2016) also showed that teachers' pointing gestures and demonstrations of how and where sounds are produced can help students recognise and practice target sounds better. On the other hand, close multimodal analyses have also suggested that rather than deploying a single resource such as a gesture, instructing pronunciation in interaction involves configuring multiple resources, including talk, gestures, and facial expressions (Duran & Kääntä, 2023). Overall, these and other multimodal studies point to a need to attend to embodied practices in pronunciation-focused instructional interaction, but, to

the best of our knowledge, studies have not yet explored how the use of conversational agents such as social robots may configure such practices.

III Data and method

The data used in this study come from a regional initiative in which municipalities in a bilingual (Swedish and Finnish) area in Finland had received external funding to try out a social robot in language teaching. The municipalities recruited a teacher with previous experience in RALL implementation to visit local schools and introduce a NAO6 robot to students who were learning English as a foreign language (EFL). Released in 2018 by the French company Aldebaran, NAO6 is an autonomous and programmable humanoid robot that is used in education for a broad range of purposes, including not only language teaching but also science, technology, engineering, and mathematics (STEM) contexts. The focal RALL activities were administered to small groups (each including two to four students) in a separate room during the students' regular EFL lessons. A team of two researchers from the University of Turku video-recorded altogether 14 hr of such activities in four rural Swedish-speaking schools in Finland in 2019. Apart from recording, the researchers did not influence the design or implementation of these activities.

The video corpus includes 111 participants, who completed the activities with a NAO6 robot in 42 small groups that were formed from the schools' existing EFL classes. At the time of data collection, the students, aged between 10 and 13, had been learning EFL for 1.5–3.5 years; none of them had earlier experience talking to a social robot. Permission for video-recording the activities was secured from the local school board and headteachers, and informed consent was provided by the students' parents. The recorded learning activities took approximately 15–20 min and focused on vocabulary (family words or colours).

In our dataset, the NAO6 robot is integrated with the Elias language learning application (https://www.eliasrobot.com/). NAO6 is a humanoid robot with a torso, arms, legs, and facial features. It has voice recognition, speech synthesiser, gaze recognition, and the ability to move, which enable it to react verbally and in an embodied manner to human participants. Developed in Finland, Elias is a commercial application that includes pre-programmed short conversational tasks in a range of target languages, but teachers can also use it to program their own tasks. The Elias app can be operated on different voice–user interfaces, including mobile phones, tablets, and social robots, as is the case in our study. NAO6 was connected to a laptop computer on which the participants used the Elias app and controlled their progress in the learning task, for example, moving between task elements and sometimes repeating problematic words and utterances. The laptop would also show images related to the focal vocabulary to guide learners. As the selected voice–user interface, NAO6 uttered all the spoken language that was a part of the Elias app tasks and related to the human participants in an embodied manner, such as by nodding, shifting posture, and blinking lights.

In this study, we analyse word repetition sequences, which occurred in an early phase of the overall activity. In this phase, the robot uttered a family word (uncle, father, sister, etc.), and the learners took turns to repeat it. After this, the robot would ratify the repeated word as "correct" by repeating the word once more (most of the time), beeping, and/or

flashing eyes. Conversely, the robot would react by nodding or readjusting its body to ambiguous utterances that its ASR system recognised as sound but not as the target utterance. At a later stage of the word repetition activity, learners would be prompted to utter a word by an image depicting one of the focal words shown on the laptop screen, without the robot initiating the utterance. As the orchestrator of the activity and with previous experience with the NAO6 robot and the Elias app gained through the RALL initiative, the teacher knew what the robot's different ways of reacting meant. However, for the learners, who interacted with a robot for the first time, these responses were initially less transparent during the activity.

Although this part of the task was not primarily intended to be a pronunciation exercise in the form of a drill, we noticed that it involved recurrent interactional troubles between the students and the robot. This was particularly apparent in the sequential context between a student's response and the robot's ratification of correctness, in which case it would often lead to multiple repetitions of the focal utterance (see also Veivo & Mutta, 2025). Many of these troubles seemed to indicate limitations in the robot's voice recognition capabilities. Analysing how the teacher and students addressed these troubles, we became interested in cases in which either the teacher or a student specifically treated the responding student's pronunciation as a potential reason behind the trouble, for example, by instructing how the word should be pronounced or by modifying their own pronunciation. We then formed a collection of 40 such sequences in which participants seemed to work on an unplanned pronunciation-focused "learnable," and investigated their practices of correcting, modelling, and modifying pronunciation in these sequences. For this article, we have selected five representative cases from that collection to illustrate recurring aspects of such emergent pronunciation instruction and the breadth of the teacher's practices to direct a student's attention to pronunciation detail in the focal RALL context.

We take a qualitative approach and analyse the data from the perspective of ethnomethodological conversation analysis (EMCA), grounding our analytical claims in participants' observable orientation to interactional conduct, which constitutes emic evidence in the form of so-called next-turn proof procedure (Sacks et al., 1974). We have transcribed the interactional data according to standard CA conventions developed by Jefferson (2004), reflecting features of interaction (i.e., intonation, silences, overlaps) with as much detail as possible (see Appendix). In addition, the extracts include an International Phonetic Alphabet (IPA) transcription of the focal words (right-hand column) to make the pronunciation detail to which participants themselves oriented also accessible to the reader in as close to renditions as possible for recordings made in non-laboratory contexts. All potentially identifying person and place references have been changed to maintain anonymity.

IV Data analysis

In this section, we first illustrate how participants establish a corrective focus on a student's word-level pronunciation in moments of problems in child–robot interaction, illustrating different multimodal practices for drawing attention to pronunciation details (section on *Instructional practices for highlighting pronunciation detail*). We show how

participants make sense of the interactional and embodied responses of the robot to verify whether the robot "hears" and "understands" learner talk. Throughout the analysis, we pay particular attention to the relationship between the teacher's corrective instructional actions and learners' subsequent pronunciation modifications. In the section on *Complications in following pronunciation instructions*, we analyse two more complex instances in which, despite the learner's pronunciation modification in response to the teacher's corrective action, the interactional trouble remains, largely due to ASR failures.

1 Instructional practices for highlighting pronunciation detail

Extract 1 illustrates recurrent aspects of the interactional organisation of the repetition task, including participants' orientation to the interactional functions of the robot's nodding and the relevance of pronunciation for task accomplishment. It shows a case in which the teacher treats a student's (S2) production of the target word ("aunt") as problematic by providing a multimodal instruction that combines talk and a hand gesture to correct the student's pronunciation. The extract starts as the teacher has turned the robot to face S2 (see Figure 1.1 in Extract 1), whose turn is to complete the word repetition sequence.

Extract 1. Identifying problematic phonemes

		Transcript of interaction	IPA transcript (focal word)
01	R	aunt, (.) beep beep	ænt
02		(0.9)	
03	S2	a:nd, #FIG 1.1	ænd

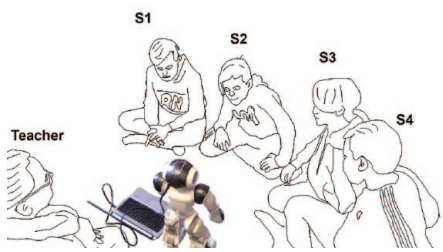

Figure 1.1. S2 talks to the robot.

04		(2.0)	
05	T	°kom närmare° ((gestures closer)) *"Come closer"*	
06		(1.2) ((S2 comes closer to robot))	
07	S2	a:nd,	ænd
08		(0.7)	
09	R	((nods))	
10		(0.4)	
11	T	kommer lite tee där. ((points up)) #FIG 1.2 *"There is a little 'T' there"*	

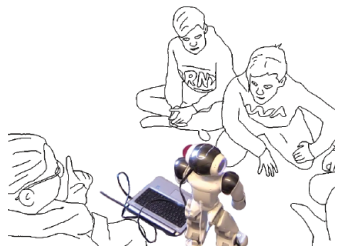

Figure 1.2. The teacher's hand gesture (Line 11).

12		(0.9)	
13	S4	a[unt]	ænt
14	S3	[aunt]	ænt
15		(0.5) ((S2 keeps gazing at robot))	
16	S2	aunt,	ænt
17		(0.5)	
18	R	((nods))	
19		(1.1)	
20	S2	aunt,	ænt
21		(0.7)	
22	R	((nods))	
23		(0.4) ((S2 smiles))	
24	S4	.hhh (när man) nickar ↑mhm ((nods repeatedly)) *"(When one) nods ↑mhm"*	
25		(0.6)	
26	S2	au:n°t°	æ:nt
27		(0.9)	
28	R	aunt, (.) beep beep ((nods, eyes flash))	
29	T	mhmh ((moves robot to face S1))	

As S2 responds to the robot's prompt in Line 3, he produces the target word "aunt" with elongation and using the voiced phoneme (/d/) instead of the voiceless phoneme (/t/) as the word's final sound, making it sound more or less like the word "and." The robot does not react to the word visibly or audibly during the silence that follows, which the teacher treats as a signal that it has not registered S2's talk, as evidenced by her instruction for S2 to come closer (Line 5). S2 gets closer to the robot and pronounces his second attempt in the same way as the first one. This time, the robot responds with a nod (Line 9).

The teacher's pronunciation instruction comes in Line 11, using Swedish. Its design combines a mitigated metalinguistic comment (*kommer lite tee där*, "there's a little 'T' there") that draws attention to the quality of the word's final sound in S2's pronunciation, and a pragmatic hand gesture pointing upwards with the right hand's index finger (see Figure 1.2 in Extract 1), as if to seek attention and emphasise the focus of her instructional turn. The timing of the instruction in the repetition sequence is noteworthy: it comes only after the robot has nodded for the first time, although S2 produced the word with a voiced rather than voiceless phoneme already in Line 3. This indicates that the teacher orients to the robot nod as a signal that the robot has indeed "heard" talk but not the target word needed to complete the repetition sequence. The nod, therefore, provides a signal for the participants to consider whether some element of the student's pronunciation needs adjustment to make it recognisable as the target word to the robot. Conversely, here (and in many other examples in our data), the teacher orients to the robot's lack of nodding (as in during the silence in Line 4) as a signal of nonhearing and tends to address

that before offering pronunciation instruction. In this sense, the nod is treated as a prerequisite for pronunciation instruction.

The teacher's instruction also elicits the target word from S3 and S4, who are waiting for their turn to talk to the robot (Lines 13–14). The way they almost overemphasise the word final sound (/t/) of "aunt" not only demonstrates their understanding of the teacher's instruction but also provides a concrete pronunciation model for S2. In response, S2 changes the word's final sound in his subsequent attempt (Line 16), but it is met with yet another robot nod. Eventually, it takes him two more attempts (Lines 20 and 26) before the robot recognises S2's utterance as the target word and closes the sequence with word repetition, a jubilatory beep, and flashing eyes (Line 28). On each attempt, S2 slightly modifies his word delivery from the previous occasion, as if trying out different alternatives. In Line 20, he stresses the beginning of the word, and, in Line 26, his whole utterance is considerably louder than previously, apart from the word's final phoneme. Although the teacher no longer intervenes, these changes show that S2 uses the lack of robot ratification as negative evidence of his pronunciation.

In summary, Extract 1 shows how human participants take it as their task to modify their talk so that it is recognisable to the robot. Through their corrective instructional actions, the teacher and S2's peers transform a problem in the progression of the activity into a pronunciation-related learnable for S2. In this sense, their interactional work can be seen to display and promote an awareness of the importance of pronunciation in human–robot interaction. Similarly, in Extract 2, the focal learner (S1) treats the robot's nod (and lack of word ratification) as negative feedback on her pronunciation. In contrast to the previous extract, she modifies the way she utters the target word "cousin" first without any teacher intervention. However, that changes her utterance closer to how the corresponding word is pronounced in the student's first language, Swedish (*kusin*), at which point, the teacher intervenes. Unlike in the previous extract, Extract 2 shows the participants completing the word memorisation phase of the task cycle. When the teacher presents the robot to S1 in Line 1, she thus sees the same picture on the laptop as 5 min earlier, when the robot first taught the word "cousin" to the students.

Extract 2. Modelling word stress

		Transcript of interaction	IPA transcript (focal word)
01		(5.7) ((T moves robot to face S1))	
02	S1	cousin,#FIG 2.1	ˈkʌsin

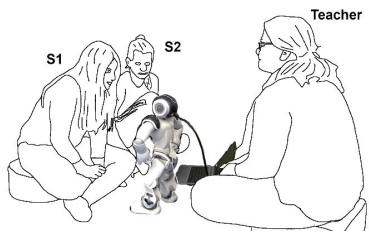

Figure 2.1. S1 leans in to talk to the robot

03		(0.4)	
04	R	((nods))	
05		(4.5)	
06	S1	kusin,	kuˈsi:n

```
07      (0.7)
08  R   ((nods))
09      (1.8) ((S1 nods and smiles))
10  T   cousin hh ((nods and smiles))#FIG 2.2                    'kʌsin
```

Figure 2.2. T nods and utters target word

```
11      (1.9)
12  S1  cousin,                                                  'kʌsin
13      (0.5)
14  R   ((nods))
15      (1.1)
16  S2  >ska jag< ta next, ((points at laptop))
        "Shall I take next"
17      (0.7)
18  T   mh- (0.4) >låt henne prova ännu<= ((points at S1))
                  "Let her try again"
19  S1  =cousin,                                                 'kʌsin
20      (0.8)
21  R   cousin (.) [blip ↑yay:::  ]                              'kʌzɪn
22  T              [(°yes°) (.) så] (.) bra?
                              "So (.) good?"
```

S1 pronounces the word "cousin" in two different ways in Lines 2 and 6, modifying her pronunciation after the robot reacts to her first attempt with a mere nod (Line 4) and no signal of ratification. Interestingly, S1's first attempt is close to the target word pronunciation, apart from the quality of the sibilant (voiceless alveolar /s/ vs. voiced alveolar /z/) and the second vowel. In contrast, her second attempt resembles the way the corresponding word is pronounced in Swedish (*kusin*), having a markedly different quality in the first vowel and stress on the second syllable instead of word-initial stress. The way S1 reacts to the robot's negative feedback (nod) in Line 4 thus takes her pronunciation further away from what could be seen as target-like pronunciation.

Both S1 and the teacher orient to the robot's second nod (Line 8): S1 reciprocates the nod and smiles, and the teacher models the pronunciation of "cousin" (Line 10). The teacher looks at S1 and utters the word at the same time as she nods (see Figure 2.2 in Extract 2). The timing of the downward movement of her head is synchronised with the first syllable of the word, and the upward movement with the second syllable. Such a multimodal synchronisation of talk and head movement gesture thus uses the body to highlight the word's initial stress in the English word "cousin" and create a contrast with the corresponding Swedish word (in which, stress falls on the second syllable). In

addition to word stress, the teacher's first vowel is also different from S1's prior attempt (Line 6) and, therefore, hearable as a model for vowel quality.[1] S1 modifies her pronunciation along the lines of the teacher's instruction. Her subsequent attempt (Line 12) is not yet ratified by the robot, but the one following it (and S2's suggestion to skip the word [Line 16]) eventually achieves task completion (Line 19). The accepted pronunciation is very similar, if not identical, to both her initial attempt (Line 2) and the one following the teacher's instruction (Line 12).

The previous extracts illustrate instructional practices in our data collection for verbalising what the learner should correct in their pronunciation (Extract 1) and modelling the target utterance for the learner (Extract 2). Another way to model target utterances in our data collection involves relistening the prompted word by making the robot repeat it once more. Extract 3 exemplifies how the robot's repetition of the prompted word is used as a secondary pronunciation model after peer instruction does not prove to be successful. Extract 3 shows the beginning of a longer sequence that involves more problems with the target word beyond an individual sound (to which we will come back in section). The extract also differs from the previous ones in that the student who is talking to the robot (S2) interrupts the repetition sequence. The way she initiates repair without repeating the target utterance (Line 3) after the robot's initial prompt suggests that she may have trouble recognising the target word ("aunt").

Extract 3. Co-constructing pronunciation instructions with technological resources

		Transcript of interaction	IPA transcript (focal word)
01	R	be-beep aunt (.) be-beep beep	ænt
02		(2.0)	
03	S2	va ↑sa han. ((frowns and looks at T))#FIG 3.1 *"What did he say?"*	

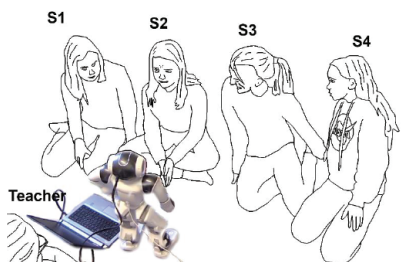

Figure 3.1. S2 frowns after the robot's prompt.

04	S3	aunt	ænt
05	T	mm-hm	
06		(0.9)	
07	S2	va(hh)?= ((smiles))#FIG 3.2 *"What?"*	

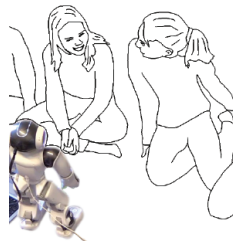

Figure 3.2. S2 smiles and initiates repair.

08	S3	=aunt ((looks at S2))	ælt
09		(1.0)	
10	S2	aunt ((leans to robot))	ælt
11		(0.5)	
12	S3	hhh [hehehe]	
13	S1	[hehehe]	
14	T	[nä:] utan, *"No but"*	
15		(2.2) ((T replays word on laptop, robot nods))	
16	R	aunt, be-beep beep	ænt
17		(1.2)	
18	S2	and,	ænd
19		(0.6)	
20	S3	hhh [°hehe]hehehe° .nh	
21	R	[((nods))]	
22	S1	[((puts hands in front of mouth, smiles))	
23	T	det är lite tokigt. ((continues instruction)) *"It is a bit crazy"*	

S2 addresses her repair initiation (Line 3) to the teacher by looking at her and produces it with a noticeable frown (see Figure 3.1 in Extract 3). The frown and the falling turn-final intonation mark the repair initiation, conveying puzzlement with word recognition, which suggests the problem is beyond not hearing. Even though S3 is a nonaddressed party, she repeats the word for S2 (Line 4) and thereby provides a target-like pronunciation model, which the teacher also confirms (Line 5). However, S3's model does not enable S2 to solve the problem but instead leads her to reinitiate repair in Line 7 (Figure 3.2 in Extract 3). S2's emphatic intonation and smiling show a sense of surprise or disbelief in what S3 has just said, which indicates she may be having trouble recognising rather than articulating the word.

S3 responds to S2 by repeating the target word (Line 8) but does so with a different pronunciation from her first time (Line 4). The quality of the consonant is somewhat closer to /l/ than the target sound /n/. S2 turns towards the robot and repeats a similar utterance to it (Line 10), showing that she treats S3's turn as a pronunciation model. However, S2's utterance is treated as an insufficient model of the target word: both students around her (S3 and S1) burst into laughter, and the robot does not immediately react to S2's utterance by nodding, suggesting it did not register talk. Moreover, the teacher begins a no-prefaced turn (Line 14) as she reaches for the laptop. Her verbalisation projects a corrective action, but instead of completing the turn herself, she replays the word on the laptop so that the robot repeat the original prompt and provide the model

pronunciation on the teacher's behalf (Line 16). The result is a pronunciation instruction that is co-constructed by the teacher, the robot, and S3.

S2 changes her pronunciation after the teacher's and the robot's pronunciation instruction so that it now contains a clearer "n" sound in the middle of the word. However, at the same time, the word's final sound also changes from a voiceless (/t/) to a voiced phoneme (/d/), similar to the problem in Extract 1. Although the word is now closer to the target utterance, it is still problematic in the sense that it is hearable as another word ("and"). Indeed, S3 (Line 20), the robot (Line 21), S1 (Line 22), and the teacher (Line 23) treat S2's utterance as an insufficient pronunciation, and the pronunciation instruction continues (see Extract 5).

2 Complications in following pronunciation instructions

In Extracts 1 to 3, the teacher's instructional actions are taken up by students in the form of modifying their pronunciation of the target word in subsequent utterances addressed to the robot. In this section, we will discuss two more complex instances (Extracts 4 and 5) in which the student remains "stuck" in the word repetition task cycle even after the teacher's instruction. Generally speaking, many such cases in our dataset reflect problems in the robot's voice recognition capabilities rather than the students' pronunciation. Nevertheless, as Extracts 4 and 5 demonstrate, the participants continue to investigate these repeated "failed" attempts at saying a word as evidence of potential pronunciation problems.

Extract 4 illustrates a case in which a student (S1) does not seem to take up the teacher-modelled stress pattern. That, together with the way he times some of his turns to the robot, may be among the reasons why it takes a long time for the robot to recognise his utterance as the target word ("sister"). Before the extract begins, S1 has already repeated the word four times without receiving a nod from the robot, which the teacher treats as a sign that the robot has not "heard" S1's turns, as indicated by the teacher's encouragement to S1 to produce the word more loudly in Line 1. After the robot finally nods for the first time in the repetition cycle (Line 4), the teacher provides her first pronunciation-focused corrective action (Line 6; similar timing as in Extract 1), but it still takes seven repetitions by S1 to complete the cycle.

Extract 4. Providing multiple instructions

		Transcript of interaction	IPA transcript (focal word)
01	T	**å högre röst** *"And louder voice"*	
02	S1	**siste:r,**	**sɪsˈtə:**
03		**(0.8)**	
04	R	**((nods))**	
05		**(0.5)**	
06	T	**s:ister. ((nods))** #FIG 4.1	**ˈs:ɪstə**

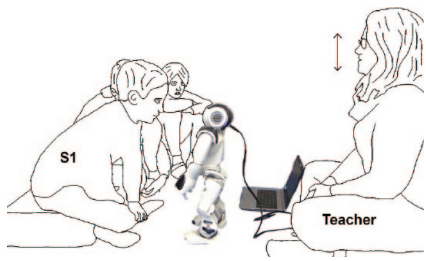

Figure 4.1. T nods to emphasise word-initial stress (Line 6)

07		(2.2) ((robot adjusts posture))	
08	S1	>siste:r,< ((robot adjusts posture))	sɪsˈtə:
09	R	((nods))	
10	T	°en gång till° *"One more time"*	
11	S1	siste:r,	sɪsˈtə:
12		(5.7) ((robot adjusts posture))	
13	T	å repetera en gång, *"And repeat one more time"*	
14		(0.9) ((T replays word on laptop))	
15	R	sister. [beep] [beep] ((nods))	ˈsɪstə
16	S1	[s-] [s]ister,	sɪsˈtə:
17	R	((nods))	
18		(1.4)	
19	S1	siste:r,	sɪsˈtə:
20		(2.2) ((robot adjusts posture))	
21	S1	siste:r,	sɪsˈtə:
22		(2.9) ((robot adjusts posture))	
23	T	prova ett lite till. *"Try a little more"*	
24	S1	sist-er.	sɪst.ˈə
25		(0.5)	
26	R	((nods))	
27		(0.4)	
28	T	sister, ((head down))	ˈsɪstə
29		(0.5)	
30	S1	sister,	sɪsˈtə:
31		(0.7)	
32	S1	[så-]	
33	R	[((nods))] sister beep beep	ˈsɪstə
34	T	bra, *"Good"*	

S1's pronunciation of "sister" shows minor variation throughout the extract. However, in all his attempts, S1 lengthens the second vowel of the word, which contributes to the perception of word stress being placed on the second rather than the first syllable. The teacher's two pronunciation instructions attend to this very feature of S1's word delivery. She models the pronunciation in Lines 6 and 28 by uttering the word and accompanying it with head movements. Her nod in Line 6 (Figure 4.1 in Extract 4) shows a similar

practice as in Extract 2, whereby the downward movement of her head is temporally synchronised with the first syllable of the word (here, "sis"), and the returning movement in the upward direction with the second syllable ("ter"). In addition to the head movement, she further emphasises the word's initial stress by lengthening the first sibilant (/s/) in the word. Together, these two practices create a somewhat exaggerated emphasis on the first sound and the syllable in the teacher's model. They highlight what is problematic in S1's pronunciations and thus maximise the salience of the instruction for him. The robot eventually recognises the target word (Line 33) after S1's response (Line 30) to the teacher's second pronunciation instruction (Line 28), although S1's pronunciation still carries the stress on the second syllable.

Besides the pronunciation instructions, the teacher encourages S1 to continue with the task. Asking the learner to try "one more time" (Line 10) or "little more" (Line 23) can be seen as a way to promote perseverance with the task. As an instructional strategy, such encouragements treat the student's previous pronunciation as intelligible and position the robot as responsible for recognition failure. When the robot does not react by nodding to S1's attempt in Line 12, the teacher replays the word on the laptop (Lines 13–15). As in Extract 3, this provides S1 with yet another pronunciation model. As S1 begins to respond after the robot's model, he begins his turn (Line 16) "early" so that it slightly overlaps with the beeps in the robot's prompt. This may challenge the robot's voice recognition capabilities further, even if S1 delays his utterance by cutting off and repairing the initial attempt.

As Extract 4 illustrates, when learners need to repeat the target word multiple times, the teacher's instructional focus typically alternates between pronunciation features and encouragement to carry on with the task, even if the learner's pronunciation might still display features that could be seen as correctable (such as the stress on the second syllable of "sister"). This suggests that sometimes the teacher may withhold pronunciation instructions in order to provide positive encouragement as a way to avoid possible frustration and to promote task engagement.

As our final extract, we analyse an "unsuccessful" word repetition cycle in which the robot does not recognise any of the student's multiple utterances as the target word, which leads the participants to eventually skip the task. Extract 5 is also a deviant case in our collection in the sense that in it, the participants deal with confusion between pronunciation and meaning-related instructional foci. The extract picks up the events of Extract 3 in which S2 expressed problems beyond the pronunciation of the target word "aunt." We observed earlier that S2 pronounced the word more or less as "and" after the teacher made the robot repeat the target word. As Extract 5 begins, the teacher and S3 treat the target word as potentially unfamiliar to S2 by explaining what it means in Swedish and conveying its written form to her (Lines 26–27). Instead of clarifying the issue, this creates a new problem.

Extract 5. Distinguishing between pronunciation and meaning-related instruction

		Transcript of interaction	IPA transcript (focal word)
23	T	**det är lite tokigt,=** *"It is a bit crazy/crazily"*	
24	L4	**=mhm=**	
25	T	**=(lagt i honom) men, (0.3)** *"put in him but"*	
26		**det ska vara liksom- (.) moster eller fas[ter.]** *"It should be like a maternal or paternal aunt"*	
27	S3	**[aunt]**	aʊnt
28		**(1.2) ((S2 opens mouth and leans towards robot))**	
29	S2	**↑va ((turns to S3))** *"What"*	
30	S3	**aunt (.) (sk[rivs)]** *"Is written"*	aʊnt
31	S2	**[<au]nt>=**	aʊnt
32	S3	**=>nej<** *"No"*	
33		**(0.6)**	
34	R	**[((nods))]**	
35	T	**[ju-] ju det skrivs så. (0.3)** *"Y- yes it is written like that."*	
36		**men han- han uttalar det som (1.9) ä:n- hm (.) ää aunt** *"But he- he pronounces it like an- hm- aunt"*	ænt
37		**(0.9)**	
38	S2	**aunt**	ænt
39		**(0.7)**	
40	R	**((nods))**	
41		**(1.0)**	
42	S2	**<aunt>**	ænt
43		**(0.7)**	
44	R	**((nods))**	
45	S3	**.nh**	
46		**(1.9) ((robot adjusts posture))**	
47	S2	**AUNT**	ænt
48		**(0.5)**	
49	R	**((nods))**	
50		**(2.9) / ((T leans back and S2 shifts gaze to T))**	
51	S2	**AUNT**	ænt
52		**(0.8)**	
53	R	**((nods))**	
54		**(2.9)**	
55	S2	**AUNT**	ænt
56		**(0.8)**	
57	R	**((nods))**	
58	T	**hh hehe han bara [.hh vet att] du säger men förstår int.** *"He just knows that you say but doesn't understand"*	
59	S1	**[(°nickar bara°)]** *"Just nods"*	
60	T	**äh ((plays new word on the laptop))**	

The teacher's turn in Lines 23 and 25–26 responds to S1's prior utterance "and" (Line 18 in Extract 3). Note that unlike in Extract 1, here the teacher does not correct the word's final consonant (/d/ vs. /t/) but provides the meaning of "aunt" in the learners' first language. This treats S2's prior utterance as an indication that she may not necessarily recognise the target word. The way the teacher prefaces the instructional turn with "it is a bit crazily put in him" treats the robot's pronunciation as unusual and avoids attributing responsibility for the problem to the learner, thereby attending to face concerns. Similarly, S3 provides assistance related to word meaning by uttering the target word "as it is written" (Line 26), a practice which is designed to enable the recipient to recognise the spelling of the target word (e.g., Jakonen & Morton, 2015, pp. 87–88).

During the silence in Line 27, S2 opens her mouth to talk to the robot but does not take a turn and instead addresses yet another repair initiation to S3 (Line 29). S3 repeats the said-as-it-is-written pronunciation and clarifies that it concerns the written form, not pronunciation (Line 29). However, this comes too late, as S2 is already using that form to talk to the robot (Line 30), which suggests that she treats it as a pronunciation model. Both S3 and the teacher take steps to rectify the confusion: S3's emphatic response cry ("no" [Line 31]) and turning away without waiting for the robot's response treats S2 as having misunderstood her previous turn designed to facilitate the recognition of the word and instead having taken it as a pronunciation model. The teacher verbally orients to the distinction between the written and spoken forms of the word (Line 35), thus acknowledging partial correctness before launching a corrective action. She then first produces the word's initial vowel (/æ/) two times before providing a model for the entire word (Line 36) to S2. Reproducing the individual sounds can be seen as a way to draw S2's attention to the quite stark contrast between vowel quality in the robot's pronunciation of the word and S3's said-as-it-is-written pronunciation in Line 27.

S2's subsequent attempts to utter "aunt" (Lines 38, 42, 47, 51, and 55) demonstrate that she takes up the teacher's model pronunciation. Yet, the attempts are not ratified by the robot, receiving merely a nod of sound recognition each time. Throughout these attempts, the quality of S2's phonemes remains constant, but she further modifies the volume of her utterance, saying the word louder each time to the robot and thus orienting to the possibility of distance and volume as causes for the problem. Eventually, the teacher decides to end the "stuck" task cycle and move on to the next word. She accounts for the problem by treating the situation as the robot's failure to "understand" S2's talk (Line 58). S1's comment about the robot's nodding (Line 59) shows how the participants orient to the robot's (embodied) conduct as a relevant feature of the repetition task.

V Discussion

In this article, we have explored interactional practices in teacher-guided RALL activities involving young learners of L2 English. Our conversation analytic study has investigated interactional sequences in and through which the teacher facilitates her students' interaction with a social robot by drawing their attention to potentially troublesome word-level pronunciation features in their talk. Such situated and emergent pronunciation instructions are corrective actions that have not been a priori designed as the learning focus of the RALL activity but rather appear as a contingent by-product of

human–machine interaction as the human participants adapt their interactional conduct to technology.

Nevertheless, our analysis has shown how encountering word-recognition-related problems can lead participants to identify and work on pronunciation-focused "teachables" and "learnables" (Eskildsen & Majlesi, 2018). Our data highlight the key role of the teacher in identifying these opportunities. The teacher can help learners, for example, by teasing apart pronunciation-related troubles from other kinds of interactional troubles, pinpointing problematic sounds verbally and using embodied means (see also e.g., Baills & Prieto, 2023; Nguyen, 2016; Smotrova, 2017), as well as modelling the focal utterances. Besides these direct instructional strategies that aim to draw learners' attention to phonetic phenomena to eventually modify their pronunciation, the teacher in our data also recurrently facilitates learners' motivation with the task. This is evident in situations where a learner ends up uttering the target word several times without success. Instructing pronunciation in such instances of "getting stuck" is interactionally sensitive because, for the learner, having to repeat the same word multiple times may feel awkward and imply a pronunciation failure.

The teacher's instructional practices analysed in this article constitute scaffolding that is in many ways carefully designed for young learners. Firstly, she seems to mitigate any impression of failure, for example, by avoiding correcting all deviations from the robot's target pronunciation, encouraging learners to try again (Extracts 2, 4), and attributing communicational failures to the way the robot has been programmed to recognise different accents (Extract 5). Such interactional work contributes to creating a safe space for young learners to practice and learn pronunciation in the face of potential frustrations with a new and somewhat opaque technology. Another striking feature in the teacher's instructions is the lack of technical pronunciation-related terminology. Instead, the teacher often offers direct pronunciation models by saying the focal word or sound aloud or by repeating it on the laptop. These strategies can be seen as ways of avoiding instructional ambiguity and ensuring that the technical level of pronunciation instruction is age-appropriate for young learners (see also Tergujeff, 2022).

The analysis sheds light on the interactional consequences of typical recognition problems of ASR technologies in language teaching, particularly when used in contexts involving young children and L2 speakers (e.g., Alharbi et al., 2021; Cumbal, 2024). In our data, the range of pronunciations that the robot recognises and accepts as "correct" is much narrower than what the teacher treats as intelligible within the context. The robot did not systematically recognise a specific pronunciation as the targeted word during an ongoing activity (e.g., Extracts 1, 2, 4). For example, in Extracts 3 and 5, the learner modifies her pronunciation into one that is arguably more target-like (and nearly identical to the way the robot pronounces the target word), but the learner's utterance is never recognised by the robot before the teacher ends the dragging task cycle. Conversely, in other situations in our dataset, audible deviations from the robot's pronunciation and those typical in the target language (e.g., concerning the second vowel in the word "cousin") do not necessarily prevent robot recognition (e.g., Extract 4). Moreover, the robot seems to more readily recognise talk in an American English accent over other varieties. This can lead to false-negative error detection with words such as "aunt," which can frustrate learners (e.g., Extracts 1, 3, 5). To be sure, an ASR-based robot's

recognition of human talk can be influenced by many factors besides pronunciation, including distance (Extract 1) or volume of talk (Extracts 4, 5), which the participants in our data observably orient to as potential trouble sources that may prevent progress of the task (see also Jakonen et al., 2024). Nevertheless, in order to maximise the recognition likelihood of ASR-based technologies in educational settings, the recognisability of children's talk and learner accents should be addressed in technical development work.

As we have argued in this study, in many ways, it is the teacher who converts problems into meaningful learning opportunities in our dataset. From an interactional perspective, perhaps an even more serious limitation than ASR in the combination of the NAO robot and the Elias learning app is that it is simply not able to perform the kind of situated, adaptive, and sensitive instructional work that the teacher in our data routinely does. In contrast, the robot is not capable of initiating repair to secure mutual understanding or adapting instructional strategies to the needs and performance of individual children (for an exploration of AI personhood and alignment, see also Ward, 2025) but instead operates through prepackaged, overly explicit, and impersonal talk much like other voice chatbots (e.g., Voss & Waring, 2025). Had this task been conducted in dyadic interaction between one learner and the robot, the learner would have had fewer resources at their disposal to deal with a robotic application that simply indicates there may be a (pronunciation) problem but does not teach how to overcome it. In terms of pedagogical implications, our findings therefore suggest a need to attend closely to the "pedagogy–technology interface" (Neri et al., 2002), including a broad range of task design elements, when introducing social robots or other kinds of conversational agents in language classrooms. In light of our observations, rather than a dyadic child–robot interaction, a far more feasible option is to use robots in the role of a tutor or a peer, whose interaction with learners is steered, facilitated, and made understandable by a human being. This would allow participants to use the availability and sense-making of other humans as a resource for language learning.

VI Conclusion

This study set out to expand the predominantly experimental and quantitative research on technology-mediated pronunciation instruction with a qualitative exploration of social interaction during RALL activities in small groups consisting of a teacher and young learners of L2 English. We used multimodal conversation analysis to investigate micro-level interactional practices for resolving interactional problems generated by the robot's ASR capabilities through pronunciation-focused repair sequences. Analysing in sequential detail how human participants conducted such situated and emergent pronunciation work, we identified distinct multimodal practices combining talk, gesture, and technological-material resources, which were used to highlight and model word-level pronunciation detail to learners. These practices were designed to sustain and facilitate learner interaction with the robot and to offer them opportunities for correcting their pronunciation. In all our analysed data extracts, the learners also showed signs of local immediate uptake of these instructional practices in the form of pronunciation modification, even if our naturalistic and video-based research orientation has the methodological limitation of not being able to assess potential long-term learning gains resulting from

these practices. Another limitation of the study is that we have only analysed interactions involving the specific combination of a learning application (Elias) with the social robot NAO6. Further research in this area is needed to explore a broader range of interactional practices in RALL, including phenomena such as routinisation and ways of adjusting one's talk to a machine over longer timescales, and the development of other aspects of L2 oral skills beyond pronunciation.

The main contributions of our exploratory study to research on language education thus include (a) shedding light on how RALL activities are interactionally organised and (b) describing the real-life consequences of limitations in speech recognition technologies for L2 teaching and learning interactions with conversational agents. Methodologically, our findings underscore the need to adopt multimodal approaches for a richer understanding of technology-mediated teaching and learning interactions and teachers' professional embodiment. More generally, the findings indicate that there is a need for human and teacher agency to lead and support L2 instructional activities in which technology is a participant, as such activities involve distinct social and interactional affordances and constraints compared to human–human instructional interaction, which alter the dynamics of language learning. Human effort is needed to provide pedagogically meaningful support to learners as they face the task of adapting themselves to patterns and expectations of human–machine interaction.

Acknowledgements

We are grateful to the anonymous reviewers for their constructive suggestions for improving earlier versions of this article. We have also benefitted greatly from comments received at various data sessions and conferences where we have presented this work. Any remaining argumentative errors and shortcomings are our own.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Research Council of Finland (Grant Nos. 343480 and 363643).

ORCID iDs

Teppo Jakonen https://orcid.org/0000-0002-7250-0042
Derya Duran https://orcid.org/0000-0002-5416-0339
Pauliina Peltonen https://orcid.org/0000-0002-8743-0032

Note

1. Both the teacher and L2 consistently pronounce the second vowel and the sibilant in "cousin" somewhat differently from the robot, the latter as an voiceless /s/ as opposed to the target voiced /z/. We suspect that the difference in the voicing may be a sign of transfer from Swedish, the participants' first language. Nevertheless, voicing is here neither the target of the teacher's instruction nor something that would prevent the robot from recognising the word.

References

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., & Alharbi, R. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access*, *9*, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Amioka, S., Janssens, R., Wolfert, P., Ren, Q., Bernal, M.J.P., & Belpaeme, T. (2023). Limitations of audiovisual speech on robots for second language pronunciation learning. In Castellano, G, & Riek, L. (General Chairs), *Proceedings of the International Conference on Human–Robot Interaction* (pp. 359–367). Association for Computing Machinery.

Baills, F., & Prieto, P. (2023). Embodying rhythmic properties of a foreign language through hand-clapping helps children to better pronounce words. *Language Teaching Research*, *27*(6), 1576–1606. https://doi.org/10.1177/1362168820986716

Baker, A., & Burri, M. (2016). Feedback on second language pronunciation: A case study of EAP teachers' beliefs and practices. *Australian Journal of Teacher Education*, *41*(6), 1–19. https://doi.org/10.14221/ajte.2016v41n6.1

Chun, D.M., & Jiang, Y. (2022). Using technology to explore L2 pronunciation. In J.M. Levis, T.M. Derwing, & S. Sonsaat-Hegelheimer (Eds.), *Second language pronunciation: Bridging the gap between research and teaching* (pp. 129–150). John Wiley & Sons.

Cumbal, R. (2024). *Robots beyond borders: The role of social robots in spoken second language practice* [Unpublished doctoral dissertation]. KTH Royal Institute of Technology.

Derwing, T. (2017). The efficacy of pronunciation instruction. In O. Kang, R.I. Thomson, & J.M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 320–334). Routledge.

Dizon, G., Tang, D., & Yamamoto, Y. (2022). A case study of using Alexa for out-of-class, self-directed Japanese language learning. *Computers and Education: Artificial Intelligence*, *3*, Article 100088. https://doi.org/10.1016/j.caeai.2022.100088

Duran, D., & Kääntä, L. (2023). Building word knowledge through integrated vocabulary explanations in ESL tutorials. *Linguistics and Education*, *75*, Article 101182. https://doi.org/10.1016/j.linged.2023.101182

Ellis, R. (2009). Corrective feedback and teacher development. *L2 Journal*, *1*(1), 3–18. https://doi.org/10.5070/l2.v1i1.9054

Eskildsen, S.W., & Majlesi, A.R. (2018). Learnables and teachables in second language talk: Advancing a social reconceptualization of central SLA tenets. Introduction to the special issue. *The Modern Language Journal*, *102*(S1), 3–10. https://doi.org/10.1111/modl.12462

Galante, A., & Piccardo, E. (2022). Teaching pronunciation: Toward intelligibility and comprehensibility. *ELT Journal*, *76*(3), 375–386. https://doi.org/10.1093/elt/ccab060

García, C., Nickolai, D., & Jones, L. (2020). Traditional versus ASR-based pronunciation instruction: An empirical study. *CALICO Journal*, *37*(3), 213–232. http://doi.org/10.1558/cj.40379

Gómez Lacabex, E., & Gallardo-del-Puerto, F. (2014). Two phonetic-training procedures for young learners: Investigating instructional effects on perceptual awareness. *Canadian Modern Language Review*, *70*(4), 500–531. https://doi.org/10.3138/cmlr.2324

Gómez Lacabex, E., Gallardo-del-Puerto, F., & Gong, J. (2022). Perception and production training effects on production of English lexical schwa by young Spanish learners. *Journal of Second Language Pronunciation*, *8*(2), 196–217. https://doi.org/10.1075/jslp.20043.gom

Gordon, J. (2022). Making the teaching of segmentals purposeful. In V.G. Sardegna & A. Jarosz (Eds.), *English pronunciation teaching: Theory, practice and research findings* (pp. 61–84). Multilingual Matters.

Han, Z. (2024). ChatGPT in and for second language acquisition: A call for systematic research. *Studies in Second Language Acquisition*, *46*(2), 301–306. https://doi.org/10.1017/S0272263124000111

Hsu, H.L., Chen, H.H.J., & Todd, A.G. (2023). Investigating the impact of the Amazon Alexa on the development of L2 listening and speaking skills. *Interactive Learning Environments*, *31*(9), 5732–5745. https://doi.org/10.1080/10494820.2021.2016864

Iio, T., Maeda, R., Ogawa, K., Yoshikawa, Y., Ishiguro, H., Suzuki, K., Aoki, T., Maesaki, M., & Hama, M. (2019). Improvement of Japanese adults' English speaking skills via experiences speaking to a robot. *Journal of Computer Assisted Learning*, *35*(2), 228–245. https://doi.org/10.1111/jcal.12325

Immonen, K., Alku, P., & Peltola, M.S. (2022). Phonetic listen-and-repeat training alters 6–7-year-old children's non-native vowel contrast production after one training session. *Journal of Second Language Pronunciation*, *8*(1), 95–115. https://doi.org/10.1075/jslp.21005.imm

In, J.Y., & Han, J.H. (2015). The acoustic-phonetics change of English learners in robot assisted learning. In Adams, J.A., & Smart, W. (General Chairs), *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human–Robot Interaction* (pp. 39–40). ACM.

Inceoglu, S., Lim, H., & Chen, W.H. (2020). ASR for EFL pronunciation practice: Segmental development and learners' beliefs. *Journal of Asia TEFL*, *17*(3), 824–840. https://doi.org/10.18823/asiatefl.2020.17.3.5.824

Jakonen, T., & Morton, T. (2015). Epistemic search sequences in peer interaction in a content-based language classroom. *Applied Linguistics*, *36*(1), 73–94. https://doi.org/10.1093/applin/amt031

Jakonen, T., Veivo, O., Mutta, M., Maijala, M., Honkalammi, H.-M., & Johansson, M. (2024). "Am I saying it wrong?" Progressivity-related troubles and instructional opportunities in child–robot L2 interaction. *Prologi*, *20*(1), 14–34. https://doi.org/10.33352/prlg.120961

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G.H. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp. 13–31). John Benjamins.

Kääntä, L. (2017). In search of proper pronunciation: Students' practices of soliciting help during read-aloud. *AFinLA-teema*, *10*, 61–81. https://doi.org/10.30660/afinla.73125

Kern, R. (2024). Twenty-first century technologies and language education: Charting a path forward. *Modern Language Journal*, *108*(2), 515–533. https://doi.org/10.1111/modl.12924

Kirkova-Naskova, A. (2019). Second language pronunciation: A summary of teaching techniques. *Journal for Foreign Languages*, *11*(1), 119–136. http://doi.org/10.4312/vestnik.11.119-136

Krisdityawan, E., Yokota, S., Matsumoto, A., Chugo, D., Muramatsu, S., & Hashimoto, H. (2022). Effect of embodiment and improving Japanese students' English pronunciation and prosody with humanoid robot. In Daswin, D.S., Jacek, R, & Milos, M. (General Chairs), *Proceedings of the 15th International Conference on Human System Interaction* (pp. 1–6). IEEE.

Lee, H., & Lee, J.H. (2022). The effects of robot-assisted language learning: A meta-analysis. *Educational Research Review*, *35*, Article 100425. https://doi.org/10.1016/j.edurev.2021.100425

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Lee, S., Noh, H., Lee, J., Lee, K., Lee, G.G., Sagong, S., & Kim, M. (2011). On the effectiveness of robot-assisted language learning. *ReCALL*, *23*(1), 25–58. https://doi.org/10.1017/S0958344010000273

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, *32*(1), 1–25. http://doi.org/10.1558/cj.v32i1.25962

Lintunen, P., Mäkilähde, A., & Peltonen, P. (2023). L2 pronunciation feedback: Pre-service teachers' beliefs and practices. In V.G. Sardegna & A. Jarosz (Eds.), *English pronunciation teaching: Theory, practice and research findings* (pp. 185–201). Multilingual Matters.

Loewen, S., Crowther, D., Isbell, D.R., Kim, K.M., Maloney, J., Miller, Z.F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293–311. http://doi.org/10.1017/S0958344019000065

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation*, *5*(1), 98–118. https://doi.org/10.1075/jslp.16034.mcc

Mlynár, J., de Rijk, L., Liesenfeld, A., Stommel, W., & Albert, S. (2025). AI in situated action: A scoping review of ethnomethodological and conversation analytic studies. *AI & Society*, *40*, 1497–1527. https://doi.org/10.1007/s00146-024-01919-x

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy–technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, *15*(5), 441–467. http://doi.org/10.1076/call.15.5.441.13473

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, *21*(5), 393–408. https://doi.org/10.1080/09588220802447651

Nguyen, M. (2016). A micro-analysis of embodiments and speech in the pronunciation instruction of one ESL teacher. *Issues in Applied Linguistics*, *20*(1), 111–134. https://doi.org/10.5070/L4200024274

O'Brien, M.G., Derwing, T.M., Cucchiarini, C., Hardison, D.M., Mixdorff, H., Thomson, R.I., Strik, H., Levis, J.M., Munro, M.J., Foote, J.A., & Levis, G.M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, *4*(2), 182–207. http://doi.org/10.1075/jslp.17001.obr

Randall, N. (2019). A survey of robot-assisted language learning (RALL). *ACM Transactions on Human–Robot Interaction*, *9*(1), 1–36. https://doi.org/10.1145/3345506

Sacks, H., Schegloff, E.A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735. http://doi.org/10.2307/412243

Saito, K., & Akiyama, Y. (2017). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning*, *67*(1), 43–74. http://doi.org/10.1111/lang.12184

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development: The case of English /ʔ/ by Japanese learners of English. *Language Learning*, *62*(2), 595–633. https://doi.org/10.1111/j.1467-9922.2011.00639.x

Sardegna, V.G., & McGregor, A. (2022). Classroom research for pronunciation. In J.M. Levis, T.M. Derwing, & S. Sonsaat-Hegelheimer (Eds.), *Second language pronunciation: Bridging the gap between research and teaching* (pp. 107–128). John Wiley & Sons.

Shadiev, R., Sun, A., & Huang, Y.-M. (2019). A study of the facilitation of cross-cultural understanding and intercultural sensitivity using speech-enabled language translation technology. *British Journal of Educational Technology*, *50*(3), 1415–1433. https://doi.org/10.1111/bjet.12648

Smotrova, T. (2017). Making pronunciation visible: Gesture in teaching pronunciation. *TESOL Quarterly*, *51*(1), 59–89. https://doi.org/10.1002/tesq.276

Tejedor-García, C., Cardeñoso-Payo, V., & Escudero-Mancebo, D. (2021). Automatic speech recognition (ASR) systems applied to pronunciation assessment of L2 Spanish for Japanese speakers. *Applied Sciences*, *11*(15), Article 6695. https://doi.org/10.3390/app11156695

Tergujeff, E. (2022). Pronunciation teaching in EFL K-12 settings. In J.M. Levis, T.M. Derwing, & S. Sonsaat-Hegelheimer (Eds.), *Second language pronunciation: Bridging the gap between research and teaching* (pp. 235–253). John Wiley & Sons.

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., & Leseman, P. (2019). Social robots for language learning: A review. *Review of Educational Research*, *89*(2), 259–295. https://doi.org/10.3102/0034654318821286

Veivo, O., & Mutta, M. (2025). Dialogue breakdowns in robot-assisted L2 learning. *Computer Assisted Language Learning*, *38*(1–2), 30–51. https://doi.org/10.1080/09588221.2022.2158203

Voss, E., & Waring, H.Z. (2025). When ChatGPT can't chat: The quest for naturalness. *TESOL Quarterly*, *59*(2), 1064–1075. https://doi.org/10.1002/tesq.3374

Ward, F.R. (2025). *Towards a theory of AI personhood*. ArXiv. https://arxiv.org/abs/2501.13533

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages*, *8*(3), Article 212. https://doi.org/10.3390/languages8030212

Appendix. Conversation analysis transcription conventions.

Symbol	Explanation
(0.8)	Numbers in parentheses = length of silence in 10ths of a second
[	Start of overlapping talk
]	End of overlapping talk
.	Falling intonation
,	Rising intonation, suggesting continuation
?	Rising intonation. Questioning inflection, but not necessarily a question
word	Underlining = stress/emphasis
°°‾	Degree signs = talk between these is markedly quieter than the surrounding talk
↑	Up arrow = sharp intonation rise
↓	Down arrow = sharp intonation fall
.hh	Audible in-breath
(( ))	Double parentheses enclose description of environment or nonverbal behaviour
( )	Empty parentheses enclose unintelligible talk
(word)	Words in parentheses indicate transcriber's "best guess" utterance
> <	Talk between symbols is rushed
:	Prolongation/stretching out of sound
=	Contiguous utterances with no interval between talk
$	Smiley voice
->	Highlights point of analysis