Automatic Assessment of Speaking Proficiency for Language Practice Robots

Eva Verhelst [*1], Pieter Lecompte [*1], Ruben Janssens[1], Vanessa De Wilde[2], and Tony Belpaeme[1]

[1] IDLab-AIRO, Ghent University – imec
[2] Department of Translation, Interpreting and Communication, Ghent University
{firstname.lastname}@ugent.be

Abstract. Social robots have been shown to benefit learning and in particular language learning. However, existing language learning robots are limited in their dialogue and adaptation capabilities: many systems use pre-scripted lessons and adaptation options. As large language models enable open-domain dialogue, a conversation practice robot now becomes feasible. Such a robot should adapt its speech to the learner's language proficiency to make the learning more effective. For this, the robot must first accurately detect that proficiency. This study contributes to this capability by presenting a system that automatically assesses students' speaking proficiency. Based on expert knowledge and related literature, we extract relevant features from a graded student speech dataset. We train a machine learning model on this dataset, paying particular attention to its learned weights to inform future research. We then validate the model in a human-robot interaction setting, assessing how well it generalizes from human-only training data. Our findings show that a model relying on a limited set of feature types performs sufficiently well for adaptation, with minimal degradation when applied to a human-robot interaction scenario. Future work includes further automating the proposed system and integrating it into an adaptation system, enabling a fully adaptive conversational social robot for language learning.

Keywords: Robot-Assisted Language Learning (RALL), Assessment of Speech Proficiency, Machine Learning, Adaptive Social Robot, Educational Social Robot

1 Introduction

Social robots show particular promise for language learning. Prior work has shown that employing social robots in education leads to improved cognitive and affective outcomes, largely attributed to their embodiment: the social nature of interacting with these robot tutors engages students more than other educational technologies do [3]. This strength in keeping students engaged makes them particularly well-suited for language learning, especially conversational practice, as

[*] Equal contribution and joint first authors.

most of second language acquisition is based on exposure [13]. However, previous research on robots in language learning mainly focuses on vocabulary learning [5]. This makes social robots for conversational practice a promising, emerging domain for further research [20,14,16].

A crucial element for an autonomous conversational practice robot is that it should adapt to the student's language proficiency. In any learning, the content of what is taught should be in the student's zone of proximal development for the learning to be effective—meaning, the content should be within a certain range of the student's level, and should be challenging enough that the student struggles to do it alone, but can do it with some help [8]. For language learning, research has shown that matching the language level of what the student is exposed to to that of the student increases learning gains and might increase engagement [24,29]. Therefore, an effective social robot for conversational practice must be able to adapt its language complexity to the student—and for this, it first has to estimate the student's proficiency.

This research investigates how speaking proficiency of a student can be automatically assessed in an interaction with a social robot for conversation practice. We first investigate how teachers assess speaking proficiency in practice, interviewing two secondary education teachers and two language education researchers. Then, informed by these interviews and prior work in automatic speaking assessment, we select relevant features that are indicative of speaking proficiency and can be extracted from students' speech during an interaction with the robot. We train a machine learning model that uses these features to predict an expert-graded speaking proficiency score, training and evaluating this model on previously collected data from a longitudinal language development study without robots. Finally, we set up a study in a school where students do interact with a social robot and validate our model on these human-robot conversations. This research shows the validity of automatic assessment of speaking proficiency in human-robot interactions, indicating relevant features for that assessment, and is a stepping stone towards an autonomous and adaptive social robot for second language conversation practice.

2 Related work

Until recently, educational social robot tutors used mainly scripted, pre-planned lessons that taught specific concepts to students [3], with social robots in general not able to handle open-domain dialogue [4]. These limited, preplanned lessons allow for only little adaptation, often with a small number of difficulty options with a high difference in level between them.

Such adaptation is often powered by a student model, as in classical intelligent tutoring systems, tracking the cognitive and affective state of the student in relation to a domain model, which contains all relevant expert knowledge [22]. An example of a well-known student model is the Bayesian knowledge tracing model. This model keeps estimates of how well the student understands each piece of knowledge, which are updated after every student action [23]. While

this allows for modelling of student knowledge in an easy, interpretable way, it assumes that all lesson content has been predefined. Since the existence of large language models (LLMs) now allows for open-domain dialogue, social robot tutors might teach beyond a preplanned lesson and adapt in more fine-grained ways [27].

Beyond human-robot interaction and adaptive tutoring systems, previous research has explored automatic assessment of speaking proficiency [2,28]. Demand for such automatic graders is high, as most language assessment in practice—in schools or for standardised tests such as the International English Language Testing System (IELTS) or the Test of English as a Foreign Language (TOEFL)—is scored by a trained expert, which is time-consuming and expensive. These automatic graders aim to replace the trained experts by predicting the expert-given grades as accurately as possible, based on the students' speech. Additionally, language assessment also has applications in health care, where automatic systems and even robots can play a role [25].

Many of these automatic grading systems take a classical machine learning approach: they typically extract features from the audio directly as well as from transcriptions made by an automatic speech recognition (ASR) system, and merge these as input for the grader [28]. These features are often handcrafted based on expert knowledge, focusing on fluency, pronunciation, prosody or text complexity. As part of these features are calculated on transcriptions, ASR errors can negatively impact the grading quality. Additionally, transcriptions lose crucial information about the intonation, rhythm and prosody of speech [2]. Recently, ASR systems have improved significantly, but for atypical populations like children and language learners, they tend to disappoint [15,30]. These general ASR improvements tend to hide disfluencies in the user's speech, as they are trained on fluent speech data and therefore output transcriptions of fluent, correct speech, regardless of user mistakes. Additionally, improvements in LLMs make ASR systems more useable as their context understanding lowers the impact of ASR mistakes on the conversational quality [26]. However, these improvements do not better the applicability in educational applications such as providing feedback on learner speech, as an exact transcription, errors included, is often necessary [21]. The current ASR systems are therefore generally better for conversational quality but worse for use in educational applications.

Technological advances in deep learning as well as a need to more accurately model the complexity of speech led to the emergence of end-to-end automatic grading systems. These can take audio as input directly, omitting the need for feature extraction based on domain knowledge and avoiding the errors typically introduced by ASR. An example of this is Banno and Matassoni's assessment system that is based on wav2vec 2.0 [2]. While these end-to-end systems might improve the grading accuracy, deep learning based systems typically introduce a non-negligible delay in comparison to classical machine learning approaches.

With this research, we aim to close the gap between automatic assessment research and educational social robots. We propose a system to automatically assess students' speaking proficiency in human-robot interactions, aiming to en-

able an adaptive educational social robot for second language conversational practice—which has not yet been attempted before, to the best of our knowledge.

3 Methodology

3.1 Expert interviews

We conducted interviews with experts to identify how teachers assess and adapt to students' speaking proficiency in practice during language teaching in schools. Four experts were interviewed: two secondary school teachers, both teaching English in different grades, a postdoctoral researcher in English language education and a postdoctoral researcher in French language education, both at Ghent University.

From these interviews, a number of features were identified that are used to assess the students' speaking proficiency. These features were thematically grouped into three categories: lexical diversity, lexical sophistication and pronunciation. The first group, lexical diversity, encompasses the amount of variation in the words that the student uses. The second group, lexical sophistication, focuses on the richness and rarity of the words uttered, with a specific focus on word frequencies. Finally, pronunciation was highlighted to be an important indicator for the speaking ability of a student.

The interviewed experts indicated that, while these items are usually not explicitly included in evaluation rubrics, they are often implicitly used as indicators to assess higher-level concepts such as fluency.

Besides these three categories, experts also identified grammatical correctness and cohesiveness as items that are often included in evaluation rubrics for speaking proficiency. These items are not retained in the remainder of this work, as they are more complex to objectively assess in a conversational context and while essential for evaluation with the aim of providing feedback, they are not essential when the aim is solely to adapt the language complexity to the student.

3.2 Dataset for model development

To develop the machine learning model that will predict students' speaking proficiency, a dataset is needed of language learners' speech with expert grades assessing their speaking proficiency. For this, we use a dataset collected by De Wilde and Lowie [12], comprising recordings of first-year English learners at Dutch-speaking secondary schools completing a speaking assignment. The assignment consisted of two parts: an introductory question about the student (e.g., "Can you describe your family?") and a picture narration task, where the students are shown a story depicted in multiple images and asked to describe it (see Figure 1) [11,7]. Data was collected from two schools in Flanders and one in the Netherlands, totalling 64 students, all in their first year of secondary school (n=64; mean age 11.9 years old, 5 students did not report their age; 32

girls and 32 boys). Their prior exposure to English varied: some had already received English instruction in primary school, others had just begun, and a few had none. Data collection took place weekly over 30 weeks.

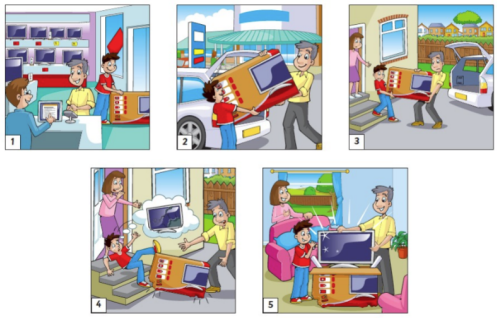

Fig. 1. Example set of pictures showing a story students were asked to describe [11,7].

The dataset consists of raw speech recordings, manual transcriptions, and an expert-graded speaking proficiency score out of twenty. For this score, the graders used a rubric with five equally-weighted categories: grammar, vocabulary, pronunciation, fluency and communication skills.

As these students are in the early stages of learning English, the recordings and their transcripts contain a mix of English and Dutch (L1) speech. English-only transcriptions were created from the originals by selecting only words that are found in the English language corpus "abc" provided by the Python library Natural Language ToolKit (NLTK) [3].

Where present, examiner speech was filtered out from the speech recordings using the speaker diarization tool provided by the Python library WhisperX [1].

3.3 Features

Inspired by prior work in automatic assessment, we decide to extract features from both the speech recordings and the transcripts. In this section, we describe how these features are extracted from the data, following the three categories identified from the expert interviews.

Lexical sophistication (LS) For calculating the lexical sophistication-related metrics, the software tool *TAALES*[4] was used [17,18]. The tool calculates frequency measures of a text by comparing against selected corpora. It also calculates other LS metrics, such as concreteness, familiarity, imageability, meaning-

[3] NLTK: https://www.nltk.org/api/nltk.html
[4] TAALES: `https://www.linguisticanalysistools.org/taales.html`

fulness and age-of-acquisition. These metrics were calculated on the transcriptions containing only the English words. In total, 251 features are extracted by TAALES.

An initial exploration revealed correlations between the expert grades and some of these automatically extracted features, showing the feasibility of this approach. Interestingly, we found that in the dataset we analyse, students with low scores used words with a higher age-of-acquisition and a lower frequency, while these metrics normally correlate with higher proficiency. This finding was also reported by De Wilde and Lowie, hypothesising that more advanced learners are able to use English in everyday contexts, while learners with lower scores are not yet sensitive to variables such as word frequency [10].

Lexical diversity (LD) For lexical diversity, 12 features were calculated by the Python library *TAALED* [5] [19] and 4 features by the library *lexical-diversity* [6]. Both libraries calculate lexical diversity-related measures such as the "*type token ratio*", which measures the variety of words. These measures were again calculated on the English-only transcriptions. Initial explorations also revealed correlations between some of these measures and the proficiency scores, warranting their inclusion in the predictive model.

Pronunciation (PR) To calculate pronunciation-related features, the Python library *myprosody* [7] was used, which is an implementation of the *Praat*-software [6]. Using the raw speech recordings, this library extracts features such as number of pauses, speaking and total duration, rate of speech and articulation rate and relevant ratios. Metrics related to the fundamental frequency ($=f0$) are calculated as well. Additionally, prosody-related comparisons to benchmarks were calculated. This resulted in 40 features in total.

Counterintuitively, metrics related to the ratio of speech time over total recording time seem to initially decrease with increasing proficiency scores. As these metrics are calculated on the audio containing both Dutch and English speech, this could be explained by low-scoring students mostly speaking Dutch, while as scores rise, students attempt to speak more English, causing hesitation and a lower speech ratio. However, the most fluent students hesitate less and less when speaking English, leading to higher scores and a higher speech rate.

Percentage of English words (EP) Finally, as the LS and LD features are only calculated on the English parts of the transcripts, we also reflect the mix between English and Dutch in the feature set by dividing the number of English words by the total number of words. This is calculated by matching the words in the transcription to an English language corpus (the *abc* corpus from NLTK)

[5] Taaled: `https://pypi.org/project/taaled/`
[6] lexical-diversity: `https://pypi.org/project/lexical-diversity/`
[7] myprosody: `https://github.com/Shahabks/myprosody`

and to a Dutch corpus (*dutch-words*[8]). This feature is strongly correlated with the expert-graded proficiency score.

3.4 Proficiency prediction model

Using these extracted features, we build a machine learning model that predicts the proficiency score. We opt for a simple and traditional machine learning model, ridge regression, as we observed linear correlations between many features and the proficiency scores and due to the limited size of the training dataset. Additionally, the added computation time of deep learning models would hinder real-time adaptation.

The dataset was split between a training set containing 80% of the data and a test set of the remaining 20%. 10-fold cross-validation was used on the training set, and all splits were made ensuring all data belonging to single students remained in the same segment.

Feature selection was performed using `SelectKBest`. Normalization was applied to scale features, a critical step due to the quadratic nature of the L2 regularization in ridge regression, and to allow for direct interpretation of learned weights. A grid search was conducted to explore combinations of preprocessing pipelines and hyperparameters, resulting in a `powertransformer` followed by a `standardscaler`, with $k = 226$ and $\alpha = 0.1125$. Model performance was measured using the mean squared error (MSE), with this optimal configuration resulting in a cross-validation MSE of 7.013. Performance on the held-out test set is analysed in Section 4.

3.5 Human-robot evaluation study

To validate the model's performance in interactions with a social robot, an evaluation study was set up in six classes from two Flemish secondary schools. Most of the students ($n = 60$, mean audio duration of 75s) were 12-13 years old, having nearly completed a full year of English classes, while a small sample ($n = 4$) of older students was added to explore the model's out-of-distribution performance. This small sample consisted of students aged 14-15, nearing the end of their second year of English classes. The study was conducted according to the ethical rules presented in the General Ethics Protocol of the faculty of Engineering and Architecture of Ghent University.

The data collection set-up consisted of a Furhat robot with external microphone. To ensure consistency across experiments, the robot was teleoperated by a researcher through prescripted questions. Audio and timestamps of the student's speech was logged, to later filter out the robot's speech. The use of a robot aimed to examine its impact on model performance, which was initially trained on data without a robot. The setup is illustrated in Figure 2.

The experiment consisted of three parts. In the first, introductory part, aiming to familiarize the student with the robot, the robot asked questions such as

[8] Dutch-words: `https://pypi.org/project/dutch-words/`

Fig. 2. Experiment Set-Up

"What is your name?". This data was not used for analysis. For the second part, the participants were asked to describe their perfect weekend. This was based on advice provided by the experts in the interview, as they suggested to focus on personal questions when students were not able to prepare for the assignment. Third, the students were given the picture narration task seen in Figure 1. This was modelled after the speaking assignments used to collect the training data, to ensure transferability of the model. The audio recorded during this data collection was manually transcribed.

This data was scored either by the teacher of that class or by the English language teaching expert that also scored the training dataset. For uniform scoring, the expert first scored one first-year class ($n = 13$) and the four second-year students' data. A sample of this data and the corresponding scores, together with the filled-out rubrics used for scoring were given to the teachers. Using this as an example, they scored the remaining data. An overview of the class groups, number of students and who scored them can be found in Table 2.

4 Results

4.1 Performance on development dataset

On the held-out test data, the model reaches an MSE of 9.057, compared to a cross-validation MSE of 7.013 on the training data. For easier interpretation, we will also report the mean absolute error (MAE) in this section, as this represents the average deviation from the score on the same scale between 0 and 20. On the held-out training data, the model reached an MAE of 2.380.

To investigate the importance of the different features on the final score, we look at absolute values of the weights associated with that feature in the model. It is important to note here that the ridge regression model can distribute weights across correlated features, which can result in underestimation or dilution of the importance of any single feature. Therefore, as we did not further investigate the correlation between features, this ranking does not strictly show which features have the most influence on the final score.

The five features with the largest learned weights in absolute value are shown in Table 1. The first three are the percentage of English used by the student and

the lexical diversity-related metrics *number of types* (which is a measure for the number of unique words used) and *word count* (which measures the total amount of words). These latter two are correlated, as more words generally means more unique words. Number four, articulation rate, is the only pronunciation metric in the top ten. It measures the number of syllables per unit of time. The fifth feature listed here is meaningfulness. It is a dimension of lexical sophistication, which measures how related a word is to other words and how many associations it evokes. Therefore, words relating to physical objects will have a high meaningfulness, while abstract concepts will have a lower meaningfulness. Therefore, the negative weight can be explained, as speakers with a lower proficiency will use more literal, meaningful words, while the meaningfulness will decrease slightly when learning [9]. The sixth to tenth largest weights, not listed here, correspond to frequency-related lexical sophistication metrics based on different corpora.

Table 1. Five Largest Feature Weights in Trained Model Ranked by Absolute Weight

Rank	Feature	Weight	Group
1	English percentage	47.227	EP
2	Number of types	18.589	LD
3	Word Count	15.181	LD
4	Articulation rate	-8.772	PR
5	Meaningfulness	-7.712	LS

4.2 Performance in human-robot evaluation study

Table 2 shows the MAE for each class group as well as who scored them. Class 3A is the small sample of older students. The MAE of the full first-year group as well as the small sample of older students is not far from the 2.380 that was found for the held-out training data. The scatter plot provided in Figure 3 shows each student's teacher- or expert-graded score against the model's predicted score.

Deviations in MAE between classes can be explained by differences in scoring strategy. The lowest MAE was found for class 1A, which was scored by the expert, meaning this data is scored most similarly to the development dataset. The MAE of classes 1C and 1D is slightly higher, while the error for class 1B is much higher. Further investigation showed that, when scoring this specific class, the scorer changed their scoring strategy compared to the other classes they scored. The harsher scores for this class are clearly visible in Figure 3.

5 Conclusion

This research presented a system that automatically assesses a language learner's speaking proficiency in an interaction with a social robot. This system aims to enable a second language conversation practice robot that adapts its speech to the

Table 2. Model Performance on Human-Robot Evaluation Study Per Class

Class	MAE	Students	Evaluator
1A	1.321	13	Expert
1B	5.615	17	Teacher1
1C	2.378	15	Teacher2
1D	2.677	15	Teacher1
All first-years	3.140	60	-
3A	2.854	4	Expert

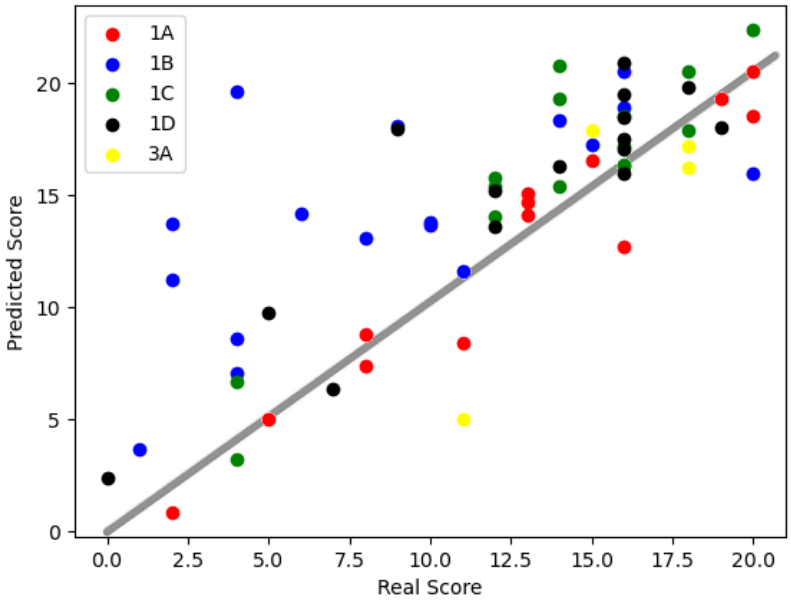

Fig. 3. Scatter plot showing individual students' expert- or teacher-graded proficiency score against the model's predicted score in the human-robot evaluation study. First diagonal shown in grey, representing a perfect model. Note that some model predictions exceed the maximum score of 20, as the model outputs continuous values without an upper limit constraint.

student's proficiency in open-domain dialogue, whereas previous educational social robots were limited in their adaptivity to course-grained adaptation in fixed lesson plans. The system uses an architecture based on prior work in automatic assessment for standardised language tests, predicting a proficiency score based on features extracted from the user's speech and from a transcript thereof. These features were informed by domain knowledge, through interviews conducted with teaching experts. We trained a machine learning model, ridge regression, using a previously collected dataset of language learners, and saw that the model's predicted proficiency scores are close to expert-graded scores, with an MAE of 2.38 on a scale from 0 to 20. Finally, we validated the model's performance in human-robot interactions by running a study where language learners in a school interacted with a social robot, finding that the model's performance transfers well from the development dataset to the real-world interactions, achieving an MAE of 3.14, confidently demonstrating the usability of this system. Grading style was found to have a non-negligible impact on this metric. Additionally, we investigated which features have the highest impact on the predicted scores.

A limitation of this research is that not all processing steps were automated yet. Most importantly, manual transcriptions were used instead of automatically generated ones. This choice was motivated by poorer ASR accuracy for low-proficiency speakers and by the learners in the development dataset and evaluation study speaking a mix of English and Dutch in the recordings. As ASR systems typically aim to recognize one language, this strongly reduced transcription accuracy. A dedicated system to recognise language switching during the transcription processes could mitigate this issue. Besides transcription, automatic processing was hindered by the unavailability of a programmatic interface for the TAALES tool.

Future work should integrate this model into a full adaptation pipeline, investigating whether the model's performance on predicting expert grades transfers well to adaptation. Furthermore, future work can evaluate how well this model transfers to other language learning interactions, investigating how features should be differently weighted when the student completes a different learning assignment than the one used in this training dataset and evaluation study.

In conclusion, this paper contributes a speech proficiency assessment model that is based on domain knowledge and prior automatic assessment work, and shows that its performance transfers well to a human-robot interaction context. This model is a stepping stone to a second language conversation practice robot that dynamically adapts its speech to the learner's speaking proficiency, a highly promising avenue of future work in educational social robotics.

Acknowledgments. This research is funded by imec Smart Education, the Research Foundation Flanders (FWO Vlaanderen, 1S50425N) and the Flanders AI Research 2 initiative. We are indebted to the authors of [12] for making the recordings and transcriptions available to us. This data was collected during research funded by the Research Foundation Flanders (FWO Vlaanderen, 1203923N).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bain, M., Huh, J., Han, T., Zisserman, A.: Whisperx: Time-accurate speech transcription of long-form audio (2023)
2. Bannò, S., Matassoni, M.: Proficiency assessment of l2 spoken english using wav2vec 2.0. In: 2022 IEEE Spoken Language Technology Workshop (SLT). pp. 1088–1095. IEEE (2023)
3. Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., Tanaka, F.: Social robots for education: A review. Sci Robot **3**(21), eaat5954 (8 2018). https://doi.org/10.1126/scirobotics.aat5954
4. Belpaeme, T., Tanaka, F.: Social robots as educators. In: OECD digital education outlook 2021 pushing the Frontiers with artificial intelligence, blockchain and robots: pushing the Frontiers with artificial intelligence, blockchain and robots, p. 143. OECD Publishing Paris (2021)
5. van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., Leseman, P.: Social robots for language learning: A review. Review of Educational Research **89**(2), 259–295 (2019). https://doi.org/10.3102/0034654318821286
6. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program] (2024), version 6.4.12, retrieved 2024 from https://www.praat.org
7. Cambridge English Language Assessment: Cambridge english: Young learners: Flyers. (2014), http://www.cambridgeenglish.org/exams/young-learnersenglish/, retrieved September 5, 2017
8. Chaiklin, S., et al.: The zone of proximal development in vygotsky's analysis of learning and instruction. Vygotsky's educational theory in cultural context **1**(2), 39–64 (2003)
9. Crossley, S.A., Skalicky, S.: Examining lexical development in second language learners: An approximate replication of salsbury, crossley & mcnamara (2011). Language teaching **52**(3), 385–405 (2019)
10. De Wilde, V.: Lexical characteristics of young l2 english learners' narrative writing at the start of formal instruction. Journal of Second Language Writing **59** (12 2022). https://doi.org/10.1016/j.jslw.2022.100960
11. De Wilde, V., Lowie, W.: Longitudinal L2 Speaking development - Exploring groups (2022), https://osf.io/qytmd/
12. De Wilde, V., Lowie, W.: The forest and the trees: Investigating groups and individuals in longitudinal second language english speaking development. Language Learning (2024)
13. Ellis, N.C., Wulff, S.: Usage-based approaches to l2 acquisition. In: Theories in second language acquisition, pp. 63–82. Routledge (2020)
14. Engwall, O., Lopes, J., Åhlund, A.: Robot interaction styles for conversation practice in second language learning. International Journal of Social Robotics **13**(2), 251–276 (2021)
15. Janssens, R., Verhelst, E., Abbo, G.A., Ren, Q., Bernal, M.J.P., Belpaeme, T.: Child speech recognition in human-robot interaction: Problem solved? In: International Conference on Social Robotics. pp. 476–486 (2024)
16. Kamelabad, A.M., Inoue, E., Skantze, G.: Comparing monolingual and bilingual social robots as conversational practice companions in language learning. In: 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 829–838. IEEE (2025)
17. Kyle, K., Crossley, S.A.: Automatically assessing lexical sophistication: Indices, tools, findings, and application. TESOL Quarterly **49**(4), 757–786 (2015). https://doi.org/10.1002/tesq.194

18. Kyle, K., Crossley, S.A., Berger, C.: The tool for the analysis of lexical sophistication (taales): Version 2.0. Behavior Research Methods **50**(3), 1030–1046 (2018). https://doi.org/10.3758/s13428-017-0924-4
19. Kyle, K., Crossley, S.A., Jarvis, S.: Assessing the validity of lexical diversity using direct judgements. Language Assessment Quarterly **18**(2), 154–170 (2021). https://doi.org/10.1080/15434303.2020.1844205
20. Lin, V., Yeh, H.C., Chen, N.S.: A systematic review on oral interactions in robot-assisted language learning. Electronics **11**(2), 290 (2022)
21. Lu, Y., Gales, M.J., Knill, K.M., Manakul, P., Wang, L., Wang, Y.: Impact of asr performance on spoken grammatical error detection. ISCA (2019)
22. Pavlik, P., Brawner, K., Olney, A., Mitrovic, A.: A review of student models used in intelligent tutoring systems. Design recommendations for intelligent tutoring systems **1**, 39–68 (2013)
23. Pelánek, R.: Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. User modeling and user-adapted interaction **27**, 313–350 (2017)
24. Randall, N.: A survey of robot-assisted language learning (rall). ACM Transactions on Human-Robot Interaction **9**(1), Article 7 (12 2019). https://doi.org/10.1145/3345506
25. Seok, S., Choi, S., Kim, K., Choi, J., Sung, J.E., Lim, Y.: Robot-assisted language assessment: development and evaluation of feasibility and usability. Intelligent Service Robotics **17**(2), 303–313 (2024)
26. Verhelst, E., Belpaeme, T.: Large language models cover for speech recognition mistakes: Evaluating conversational ai for second language learners. In: Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction. pp. 1705–1709 (2025)
27. Verhelst, E., Janssens, R., Belpaeme, T.: Enabling autonomous and adaptive social robots in education: A vision for the application of generative ai. In: Social Robots in Education: How to Effectively Introduce Social Robots into Classrooms, pp. 17–42. Springer (2025)
28. Wang, Y., Gales, M.J., Knill, K.M., Kyriakopoulos, K., Malinin, A., van Dalen, R.C., Rashid, M.: Towards automatic assessment of spontaneous spoken english. Speech Communication **104**, 47–56 (2018)
29. Westlund, J.K., Breazeal, C.: The interplay of robot language level with children's language learning during storytelling. In: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts. pp. 65–66 (2015)
30. Wills, S., Bai, Y., Tejedor-García, C., Cucchiarini, C., Strik, H.: Automatic speech recognition of non-native child speech for language learning applications. In: The 33rd Meeting of Computational Linguistics in The Netherlands (CLIN 33) (2023)