Building the foundations for automatic assessment of verbal and nonverbal aspects of spoken interaction in Finnish as a second language

Ullakonoja, Riikka

Association for Language Testing and Assessment of Australia and New Zealand
2025-12-17

Ullakonoja, R, Lähteenmäki, I H S, Raud, N, Phan, N, Grósz, T, Suuronen, H K, Hilden, R, Kurimo, M, Kuronen, M, von Zansen, A & Kautonen, M 2025, 'Building the foundations for automatic assessment of verbal and nonverbal aspects of spoken interaction in Finnish as a second language', Studies in language assessment, vol. 14, no. 2, 2, pp. 28-57. https://doi.org/10.58379/FFNO9141

http://hdl.handle.net/10138/625212
10.58379/FFNO9141

Building the foundations for automatic assessment of verbal and nonverbal aspects of spoken interaction in Finnish as a second language

Riikka Ullakonoja[1], Ilona Lähteenmäki[2], Nora Raud[3], Nhan Phan[3], Tamás Grósz[3], Henna Suuronen[2], Raili Hilden[2], Mikko Kurimo[3], Mikko Kuronen[1], Anna von Zansen[2], Maria Kautonen[1]

[1] University of Jyväskylä, Finland
[2] University of Helsinki, Finland
[3] Aalto University, Finland

In the paper we describe the theoretical and methodological basis of a multidisciplinary project, Aasis, focusing on automatic assessment of spoken interaction in L2 Finnish. The aim is to describe the project corpora that have been built for developing ASA (Automatic Speaking Assessment), not to study detailed empirical research questions as such even though we present some preliminary results from the data. The goals of the project are novel in three ways. Firstly, we aim to develop an ASA (Automatic Speaking Assessment) system to assess oral proficiency in dialogic rather than monologic speech. Secondly, our approach includes automatic assessment of nonverbal features in interaction, extending beyond the more conventionally used ASA tools. Thirdly, the language to be assessed is Finnish, a language with scarce previous studies on ASA.

Keywords: automatic speaking assessment, nonverbal features, interaction, Finnish

Email address for correspondence: riikka.ullakonoja@jyu.fi

Introduction

The purpose of this paper is to describe the theoretical and methodological basis of a multidisciplinary research project in language assessment conducted at three Finnish universities (University of Helsinki, University of Jyväskylä, and Aalto University). The project, Aasis (Automatic assessment of spoken interaction in second language, funded by the Research Council of Finland 2023–2027), aims to improve the validity of speaking assessment in dialogue speech by including nonverbal features in the concept of speaking proficiency in addition to verbal features. Recently, the models of L2 interactional speaking skills as a component of communicative language competence of the CEFR (Council of Europe, 2020) are enriched by a broader frame of interactional competence as a collaborative endeavor, as recommended by several researchers (Galaczi & Taylor, 2018; Plough et al., 2018; Timpe-Laughlin & Youn, 2021). Notably, in Aasis, three disciplines are joining to address a broad conceptual coverage of interactional competence: a pedagogical, phonological, and technical perspective. Including nonverbal features and interactional speaking skills in automatic assessment of L2 speech is the novelty of the Aasis project, which is breaking new ground in the field of interactional competence assessment research.

The project is inspired by a previous project by the same principal investigators, that is, DigiTala (see Hilden et al., this issue), that focused on developing an ASR (automatic speech recognition) based interface for assessing monologic speech in Finnish and Swedish as an L2 (second language) for Finnish upper secondary school data. The results from the DigiTala sub-studies suggested high inter-rater reliabilities for the verbal core features of monologic speech on the one hand and sufficient consistency between human and machine ratings on the other. Differently from DigiTala, Aasis focuses on adult learners of Finnish and, instead of audio files of monologues, develops ASR for videoed interactional speech (Al-Ghezi et al., 2021). In Aasis, we aim to find the most relevant features of speech and nonverbal features in assessing interaction. First, we address the verbal and nonverbal features of interactional competence in L2 Finnish dialogues that affect human raters' assessments of L2 interactional skills, and second, we develop ways to measure these features in ASR-based assessment.

Much research on language assessment in Finland is based on two main high-stakes nationwide language tests: YKI, the National Certificates of Language Proficiency (Finnish National Agency for Education, 2025), and the matriculation examination (Matriculation Examination Board, 2025). The YKI examination measures language proficiency and is the most popular test for Finnish adult L2 learners, as the YKI certificate can be used as proof of required language skills in Finnish or Swedish for applying for Finnish citizenship. The matriculation examination takes place at the end of general upper secondary education (16–19-year-old students) and can be used to enter tertiary education. In the matriculation examination of languages, speaking proficiency is not assessed to date, but there are ongoing developments towards human assessment of speaking proficiency. In YKI, although the human assessment of speaking proficiency has been an integral part of the test for decades, steps towards automatic assessment of speaking proficiency have not been taken. Incorporating ASR-based assessment in high-stakes testing is still relatively novel and presents both practical and ethical challenges. Therefore, research on ASR-based speaking assessments is needed before large-scale implementation in high-stakes contexts.

Phonological studies of interactive speech, especially with speech material that has been reliably assessed for proficiency levels, are rare. Therefore, our results on the verbal features of interactive speech are interesting also concerning languages other than Finnish. Non-verbal features may be language- and culture-specific to some extent. However, the results concerning the nonverbal features extracted from the videoed dialogues that we will obtain in the project will be methodologically interesting also outside the Finnish context. Potentially, our findings can lay foundation on developing automatic assessment of interactional speech across different languages.

Assessment of interactional speech – Theoretical considerations and aims of Aasis

Speaking is the essence of human interaction and, hence, is the paramount teaching and learning goal at all levels of language education. The construct of spoken interaction traditionally deployed in assessment entails linguistic competences. At the same time, nonverbal cues are rarely included in rating scales used by human raters of

interactive tasks despite their crucial role in communication. Thus, we aim to improve the validity of dialogic speaking assessment by including nonverbal features in the construct of oral proficiency and in ASA. This is important because paired speaking tasks are commonplace in educational settings and should thus be increasingly applied in high-stakes contexts. The Aasis project addresses features of spoken interaction that have been out of the reach of reliable automatic assessment procedures to date. The aim is to develop an automatic speaking assessment (ASA) system for assessing interactional speech in L2 Finnish and include nonverbal features in ASA.

The Companion Volume to the CEFR (Council of Europe, 2020) depicts spoken interaction as a communicative activity comprising multiple layers of competences that our project addresses. Linguistic competences involve knowledge and use of rules regarding phonology, morphology, grammar, lexicon, and semantics. Pragmatic competences, in turn, imply the ability to build coherent spoken texts in a dialogic speaking event, ensuring that the linguistic resources are drawn on appropriately with respect to the conditions specific to each speaking situation. General competences execute the linguistic and pragmatic competences according to the premises of the speaking task.

Multiple functions on all layers of spoken interaction are mediated by verbal and nonverbal behaviors (Galaczi & Taylor, 2018). The functions mediated primarily by verbal features include turn management (interrupting, starting, maintaining, and ending); topic management (initiating, extending, shifting, and closing); and breakdown repair (joint utterance creation, recast, and correction by self or other). The same functions may also be accomplished by nonverbal means, such as facial expressions, gestures, and gaze or changes in eye contact (Galaczi & Taylor, 2018), as well as intonation, speech rhythm, and voice quality (Couper-Kuhlen & Selting, 1996).

In language teaching and learning, interactional competence is addressed and trained by pair and group speaking tasks, but assessing the whole domain of spoken interaction in a similar setup is notoriously laborious. The higher the stakes of an assessment, the more uniform treatment is expected in terms of the validity of decisions on students' proficiency. In speaking assessment, we are often faced with the tension between authenticity and reliability. The "sameness" necessitated by reliability unavoidably results in compromising the diversity of authentic tasks since all

performances must be rated along a scale or scales addressing a set of pre-specified features comprising the most relevant task dimensions. Fluency, pronunciation, lexical range, and accuracy are examples of the most frequently rated dimensions of speaking and are addressed in numerous ready-made scales all over the world. However, these features date back to monologue types of speaking, and they risk bypassing salient interactional competences if not modified for interactional speech. Furthermore, the absence of nonverbal features is typical of almost all widely used rating scales (Galaczi & Taylor, 2018) or studies assessing interactional competence, although some features of nonverbal communication are mentioned in the mediation scales of the Companion Volume (Council of Europe, 2020). There is some evidence that low-proficiency speakers use nonverbal (or, in other words, visuo-gestural) behaviors differently than high-proficiency speakers (e.g., Gan & Davison, 2011; Plough, 2021), but there is also at least one study suggesting an absence of this relationship (Glasson, 2024). Therefore, an updated construct of spoken interaction for assessment should accommodate nonverbal features to add to the authenticity of the assessment.

Recent developments in ASR have opened new venues for automated speaking assessment. Encouraged by the progression of ASA of monologic L2 Finnish in the project DigiTala (Al-Ghezi et al., 2023), the Aasis project aims to refine the construct of oral proficiency by incorporating nonverbal features and systematically capturing them for ASA of interactional speech. In addition to traditional linguistic competences, our approach includes verbal features of pragmatic competence: turn-management, cooperating, thematic development, coherence, and interactional fluency. It also includes task-relevant nonverbal features: gaze direction, head movement, and facial gestures, as well as their functions in the jointly constructed interaction.

Thus, the project aims to answer two main research questions: 1) which verbal and nonverbal features of interactional competence in L2 Finnish dialogues affect human raters' assessments of L2 interactional skills? and 2) how can these features be measured and utilized in ASR-based assessment of interactional skills?

This paper, however, does not aim to answer these research questions. Instead, its aim is to describe and characterize the corpora we have collected for the purposes of ASA. To achieve this aim, apart from pure description, we also present some preliminary

results on raters' and speakers' views on using ASA in the language testing context and give an example of nonverbal behavior during turn-taking at the end of the paper. These preliminary results offer a glimpse into future empirical studies of the project.

The multimodal Aasis corpora of L2 Finnish dialogue speech

This section describes the Aasis corpora which consist of audio and video recordings and their annotations, questionnaires for speakers and raters, as well as human rating data. The main corpora needed in the development of ASA are the video recordings of L2 Finnish speakers completing dialogue tasks, their ratings, and annotations. In this section, we first describe the speaking tasks and their rating criteria, followed by the data collection and design, as well as the annotation and the rating procedures.

Speaking tasks

Six speaking tasks were created for the purposes of the project, using materials from Yle oppiminen (2016) and the FlowLang research group (Peltonen, 2020) as inspiration. The guiding principles in creating the tasks were (1) enabling performance on the A2–B2 levels, (2) encouraging interaction without excessively looking at the paper prompt, and (3) ensuring both speakers' equal participation (Tasks 4A and 4B). The principles derived from the project's aims: the tasks were designed to enable performances by speakers of different levels as the basis for automated speaking assessment. A negotiation task (Task 5) has been typically used in oral proficiency testing (see, e.g., Fulcher 2014) and is thus included here.

The two monologue tasks used here, a read-aloud task (Task 1, 10–20 seconds) and a free production task (Task 2, two minutes), had been successfully used in our previous project (von Zansen, 2022). The four dialogue tasks were recorded in pairs and included: a warm-up task (Task 3, three minutes), two customer service roleplay scenarios where participants alternated roles as customer and clerk (Tasks 4A and 4B, three minutes each), and a problem-solving task (Task 5, five minutes). The tasks are described in more detail online (von Zansen, 2024f).

Rating criteria

In order to assess linguistic aspects, we used CEFR-based rating criteria (Council of Europe, 2020) for a holistic criterion and analytic criteria (range, accuracy, fluency, and pronunciation, and for dialogues, also interaction) modified for the purposes of the project (von Zansen et al., 2025a). First, we chose the most appropriate scales for our purposes from the existing CEFR scales (Council of Europe, 2020, pp. 62, 72, 183–185). Since all the scales of the Companion Volume have not yet been translated into Finnish, we translated the scales using the CEFR version available in Finnish (Huttunen & Jaakkola, 2003). Next, we shortened the holistic scales (Council of Europe, 2020, p. 62, 72) by removing points that are not suitable for assessing individual speech samples and cut the Coherence column from the Qualitative features of spoken language (pp. 183–185). Different from the CEFR scales, we added short descriptions of how performances representing the A-, B-, and C-levels differ from each other.

Additionally, to complement the linguistic scales described above, we created a novel scale for assessing nonverbal features in dialogic speech. The scale was developed based on the available literature (Blanch-Hartigan et al., 2018; Council of Europe, 2020; Giri, 2009). The criteria were tested project-internally prior to the pilot rating round and the criteria to be used in the pilot rating round was a result of joint discussions of the project team (for details on the nonverbal scale development, see von Zansen, 2024b).

Audio and video data collection

The audio and video data from adult Finnish L2 learners were collected during the spring and summer of 2024 at the three participating Finnish universities. Participants were recruited using university email lists, a university personnel intranet and student website, physical ads such as posters, personal connections, and by visiting language lessons. We originally targeted Finnish learners at levels A2–B2, but we accepted anyone willing to participate, regardless of their level of proficiency in Finnish. Participants at the beginner level (A1–A2) expressed hesitation to participate due to their low Finnish proficiency (for example, some would have only been willing to participate if a read-aloud script was available). This is a common issue in data

collection processes, and it has been noted in other studies, such as the DigiTala project (Al-Ghezi et al., 2023).

Before the data collection, all participants were informed about the study in a letter (von Zansen & Kautonen, 2024a) and privacy notice (von Zansen, 2024e). They were also sent a link to the consent form (von Zansen, 2024a) to be filled in. The data collection and management were done according to the ethical regulations (GDPR, European Union, 2016; TENK, 2019) and requirements of the ethics committee of the University of Helsinki, including an ethical review statement confirming that the research is ethically acceptable.

The data collection sessions (45–65 minutes) were conducted in the following way. First, the researcher explained the structure of the session, then before recording each task, the researcher gave the instructions both on paper and by reading the task aloud in Finnish (and explaining it in English if needed). The participants were allowed to ask questions at all phases. The speakers were not allowed to make any notes prior to or during the recording, but they had some preparation time before the recording when they could read through the instructions at their own pace. During the recording, a support sheet (including, e.g., keywords and pictures of the items) was visible to the speakers. After completing the speaking tasks, participants were asked to fill in a post-test questionnaire (von Zansen, 2024d) asking about their gender, age, first language, and other languages they spoke, as well as self-ratings of their level of Finnish spoken production and spoken interaction on the CEFR scale (A1–C2). Further, the questionnaire included questions about the participants' views on the test, the tasks used, and automated speaking assessment.

Apart from the main corpus, we also collected a separate video dataset in October 2024, where speakers (n=16) wore SeeTrue eye tracking glasses to measure their eye movements and completed Task 5 and two new tasks. After the recordings, participants filled in a post-test questionnaire designed specifically for this sub-study (for further details, see Song, 2025). This extra dataset was collected to later analyze what the speaker focuses on at different stages of the conversation, which would be valuable training information for virtual agents so that they can learn to mimic human gaze behavior, thus making the conversation more natural.

Speakers

The corpus consists of 102 speakers[2] including dialogues of 47 dyads and 8 other speakers' monologue performances (who were missing a pair or whose pair did not give consent). The participants were all adults, the majority (72%) being 27–60 years old and there were more females (62%) than males (38%). As we encouraged speakers to come in pairs, 27 dyads knew each other in advance, and 20 dyads met for the first time during data collection.

The speakers came from diverse linguistic backgrounds: they had 27 different first languages in total, and spoke 33 different foreign languages, with the number of foreign languages spoken ranging from 1 to 7. Their self-reported Finnish proficiency ranged from A1 to C2 on the CEFR scale (Figure 1). Most speakers self-reported their spoken production proficiency to be at either the A2 (30%) or B1 (32%) level, while most self-reported their spoken interactional proficiency to be slightly higher, at either A2 (25%), B1 (28%) or B2 (23%).

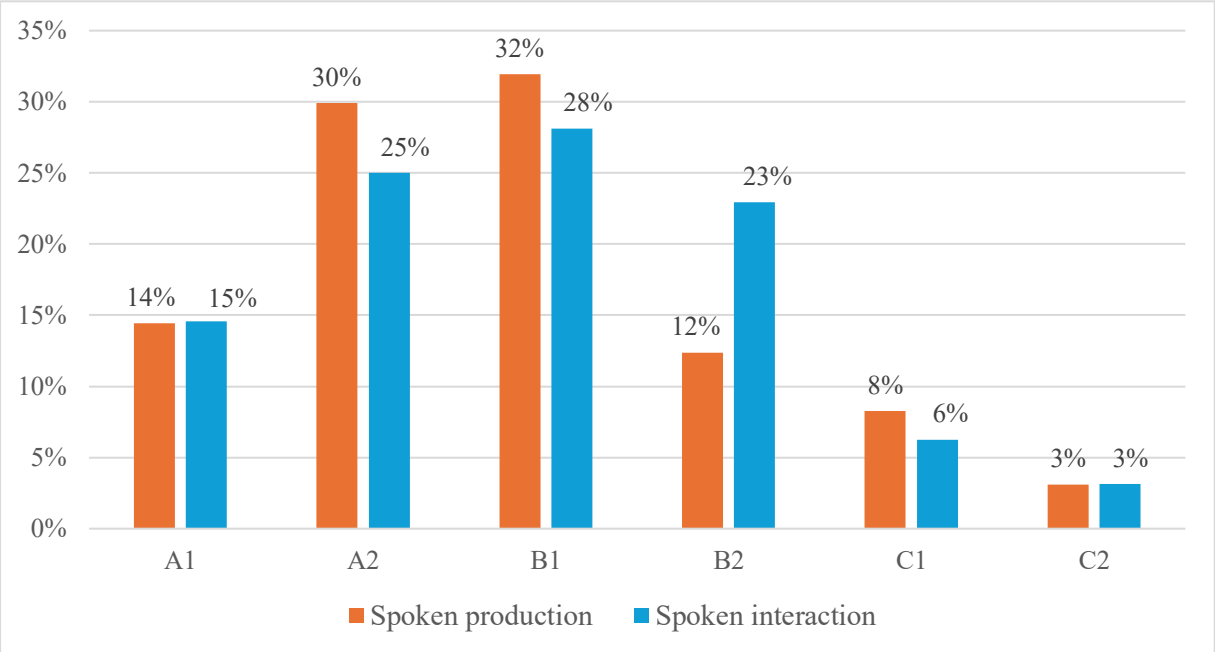

Figure 1. Finnish L2 learners' self-reported level of spoken production (n = 97) and spoken interaction (n = 96) on the CEFR scale

[2] Background information from five speakers is unavailable.

Technical setup

For building ASA for dialogic speech, we aimed to have standardized settings for video and audio recordings in all three university studios used (for the details of each recording setup, see Lähteenmäki et al., 2025) even though different equipment was used. At minimum, for each dialogue task, we collected video from three sources (two personal videos and one capturing both speakers), as well as the corresponding audio recordings. In the recordings the participants were seated facing each other with their ID numbers visible to the camera(s). At the University of Helsinki and University of Jyväskylä, the researcher was visible to the participants during the recordings, whereas in Aalto University, there was a screen between the researcher and the participants. We used three cameras: two filming each participant from the front (focusing on facial and upper body movements), and the third both speakers from the side.

Speakers had their own microphones, set to capture speech from one speaker only, but in practice the microphone sometimes also captured the other speaker's voice. In addition, both speakers were recorded with a whole-room microphone. Backup audio setups were also in place as detailed in Lähteenmäki et al. (2025).

Data annotation

All video data was transcribed and annotated using ELAN (Lausberg & Sloetjes, 2009) and a part of it (88 dialogues from 44 speakers) additionally with the computer-joystick method (Lizdek et al., 2012; Sadler et al., 2009). ELAN annotation[3] was completed by five annotators who were students in languages or teacher education, three of whom also carried out the computer-joystick coding. The annotators were trained prior to undertaking the work. In ELAN, the speech was transcribed into text on a single timestamped utterance tier, including features such as hesitations, paralinguistic units, truncations, backchanneling, incorrect utterances, overlapping speech, and pauses (our ELAN coding principles are publicly available von Zansen et al., 2024). Further, nonverbal behavior was annotated on three tiers: hand, head, and

[3] For studying their agreement in ELAN annotations, four dialogues were annotated by several annotators.

body movements. All tiers included three labels: hand movement "gesture", "pointing", and "fidgeting", head movement "nod", "shake", and "head other", and body movement "lean forward", "lean backward", and "body other". Additionally, because the project did not have the resources to annotate the participants' gaze behavior in detail, the amount of mutual eye contact in the videos was estimated by the annotators on a scale of 1–3 (little, moderate, frequent).

The computer-joystick method (Lizdek et al., 2012; Sadler et al., 2009) is a continuous time-series coding that enables investigating time-dependent processes and their changes over time and can be used to study the dyads' interpersonal behavior by exploring the dimensions of dominance and affiliation, as well as their variation, focusing on both verbal and nonverbal features (see, e.g., von Zansen et al., 2025b). These annotations enable researchers to train automatic assessment models that could simultaneously analyze users' language proficiency and interactional dynamics, opening the avenue for more informative and personalized feedback.

Data transcription and annotation with ELAN required substantial resources and time (the annotation of a 5-minute video could take around five hours) in contrast to the computer-joystick method (around 15 3–5-minute videos could be annotated in five hours). Therefore, it was important to agree on the common ELAN annotation principles. For example, the level of detail in the nonverbal annotations had to be narrowed down to only three annotation tiers and nine different codes, and gaze annotation was not performed on all the data in detail.

Finally, we wish to mention some practical issues in data collection. Firstly, planning the ethical considerations, opening the data, and addressing data security issues for personal data was time-consuming. Secondly, collecting and synchronizing multimodal data at three universities required a lot of planning and collaboration, e.g., to ensure the similarity of the data collection setups and the availability of the data for all universities. We faced some technical challenges: synchronizing all data sources (multiple cameras and microphones) caused problems, as did the sheer size of the data. Thirdly, designing the practical aspects of data collection required planning and prioritizing. For example, online data collection would have been easy and fast but lacked important features of interaction. In addition, we may have attracted more participants had the data collection been organized in language classrooms, but having

a mobile recording setup was not feasible. Thus, recruiting participants posed many challenges, as recruiting dyads and finding a suitable time for data collection were difficult, partly due to the requirement of travelling to the studio. However, we were able to tackle all these issues rather well and managed to collect extensive corpora for multiple researchers to work on, which we are aiming to open-source for limited research use as much as possible through the Language Bank of Finland (2024).

Human ratings

The audio and video data were rated by human raters, who were recruited on a volunteer basis from the YKI (Finnish National Agency for Education, 2025) rater pool and paid compensation for their work.

Rating process

The rating criteria and process were piloted in September 2024 prior to the rating proper by eight volunteer YKI raters and three project team members. In the pilot all raters rated the same two monologic and six dialogic samples, which were chosen to cover a range of different speakers and assumed proficiency levels as well as all dialogue task types. No benchmark samples were used in the pilot as one of the purposes of the pilot was to create benchmarks for the rating proper. Rater training was organized via Zoom (where it was recorded, and later made available to the raters in Microsoft Teams). Raters received the instructions, consent forms, and criteria through email. The fact that the Moodle platform we planned to use for the rating proper was not ready by the time of piloting created some extra steps for the raters: they gave their ratings in Webropol (von Zansen & Kautonen, 2024b) but accessed the videos through a separate server requiring a separate password and login.

For the rating proper, we recruited 20 volunteers from the YKI Finnish rater pool, five of whom had already participated in the pilot. All were native Finnish speakers, with Swedish as their second or other first language. Their prior experience varied: on average, they had 19 years of experience (SD = 7.7, range: 0–34) in teaching L2 Finnish, 14 years of experience (SD = 9.0, range: 1–33) in assessing language examinations and 18 years of experience (SD = 8.1, range: 8–33) in assessing spoken proficiency.

The actual rating round took place in January and February 2025, and similarly to the pilot round, it started with a rater training that was recorded for later use. The Moodle platform was used for accessing speech performances (von Zansen et al., 2025c), for sharing the criteria and instructions, and for giving the ratings. All raters rated the same anchor samples (three monologues and three dialogues), and each rater had their own bundle to be rated, ranging from 28 to 34 samples (which took them approximately 16 hours according to our estimation).[4] After the rating round, a project member (a trained YKI Finnish rater) rerated some of the samples the raters had skipped, and the rating was completed in March 2025.

The rating criteria were almost the same in both rating rounds for the verbal features (von Zansen et al., 2025a), as we found the original scales to function rather well in the pilot. We only made minor changes to wordings; for instance, the differences between the CEFR grades for holistic assessment were made clearer with short descriptions of the main differences between A2 and B1, and B2 and C1, respectively. The scales for the nonverbal features were simplified after the pilot (there was no numeric rating in the rating proper), as the scales used in the pilot were considered too complicated by the raters (for further information on the simplification, see von Zansen, 2024b).

Description of the rating data

After the rating, we conducted an initial analysis of the ratings. First, we noted that in all five rating categories (holistic, range, accuracy, fluency, and pronunciation), the CEFR levels between A2 and B2 were over-represented, amounting to approximately 80% of the data, while samples rated as A1 or C1–C2 were relatively rare. This phenomenon of intermediate speakers being over-represented in the collected data is quite well known and often causes issues during the development of automatic systems (Voskoboinik et al., 2025).

[4] Task 1 (a monologue task consisting of reading aloud a text) was not rated at this point for practical reasons.

The anchor samples (six speakers in three dialogues), rated by all raters, allowed us to estimate the inter-rater agreement to see how reliable the assigned scores were. For the holistic score and range, we saw relatively low disagreement (see Table 1). Ratings in these categories often varied by just one level (e.g., between A2 and B1 for anchor speaker 114). In contrast, accuracy, fluency and pronunciation proved harder to assess, thus leading to wider variation in the ratings.

Table 1. Rater agreement for the six anchor speakers[5] (the conversion of the numerical scale to CEFR: A1=1, A2=2, B1=3, B2=4, C1=5, C2=6)

Speaker ID	Holistic		Range		Accuracy		Fluency		Pronunciation	
	mean	std.	mean	std.	mean	std.	mean	std.	mean	std.
23	5.76	0.44	5.72	0.45	5.62	**0.56**	5.76	0.44	5.69	0.54
24	3.9	0.72	3.72	0.88	3.83	0.71	4.0	0.76	4.38	**0.94**
39	3.54	0.5	3.46	0.55	3.59	0.54	3.63	0.53	3.8	**0.72**
40	3.52	0.62	3.43	0.69	3.41	0.69	3.44	0.66	3.85	**0.73**
114	2.8	0.45	2.91	0.46	2.54	0.5	2.89	0.38	2.76	**0.67**
115	2.86	0.45	2.79	0.5	2.75	0.52	2.71	0.53	3.5	**0.64**

Next, we investigated the nonverbal categories. It should be noted that not all raters were able to assess these categories (the reasons for this are explained in the last section of the paper), which resulted in 8–22% of the ratings being categorized as missing or "cannot decide". Overall, rating the eye contact and hand movement seemed the easiest, based on the low number of missing ratings, whereas assessing facial expressions and head movement seemed the most challenging. The overall data exhibit a heavy bias towards rating nonverbal features as positively contributing towards the conversation, with 44–76% of the ratings belonging to this category, depending on the rated aspect. The second most common label is the neutral one. Only a small percentage of the speakers were rated with the negative label for disrupting the conversation with their nonverbal communication, which was mostly due to vocalizing (3.1%) or bad eye contact (2.3%).

[5] Bolded values highlight the category with the highest variation per speaker.

The analysis of the anchor samples revealed a 35–90% agreement between the 21 raters, almost always leading to a majority decision. As Figure 2 illustrates, the largest disagreement was observed in the hand and head movement ratings of the three anchor speakers, which highlights the difficulty of the task. Overall, the high agreement on the 3-tier scale signals that the ratings are usable for AI training, but the unbalanced distributions call for appropriate solutions that could still learn to recognize signs of negative nonverbal cues.

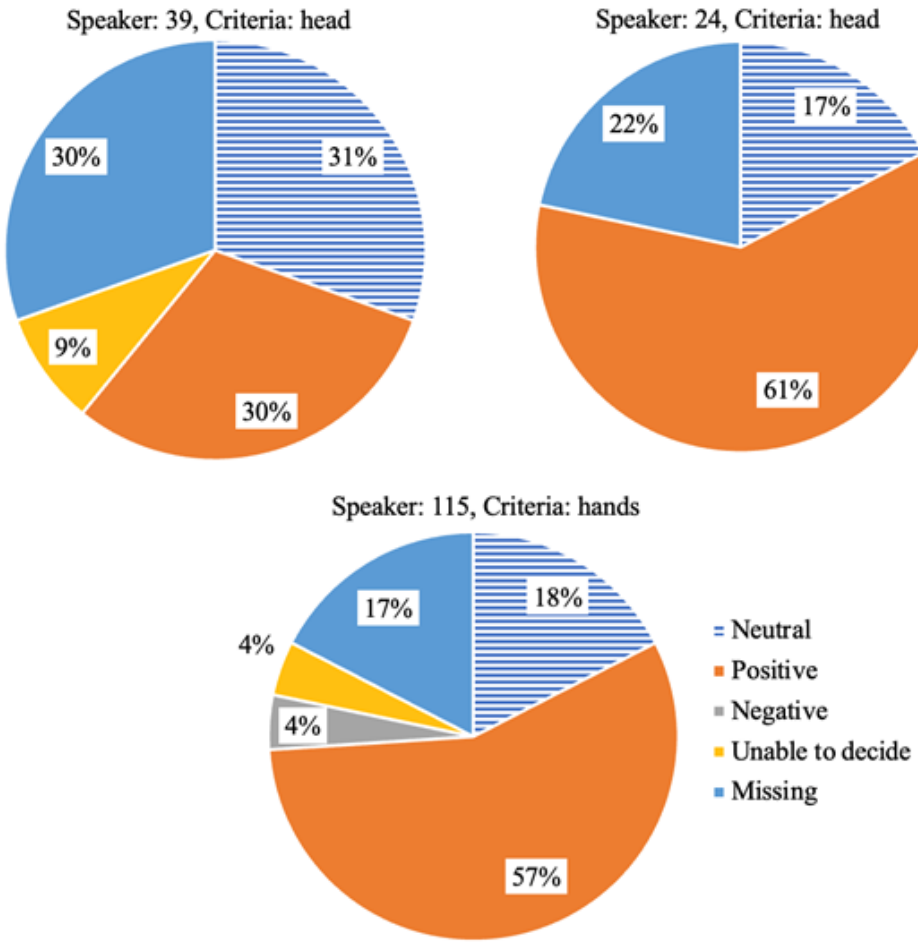

Figure 2. Rater disagreement on the hand and head movement aspect of three anchor speakers

Automatic spoken interaction assessment

The data described in the previous sections is being used to develop an automatic spoken interaction assessment system. The goal of this system is to use machine learning techniques to predict human raters' evaluations of L2 interaction, particularly focusing on how both verbal and nonverbal cues contribute to proficiency scores. This

section outlines our plans for developing the AI models that will learn to mimic human rating behavior.

Traditional automatic speaking assessment treats speech as a monologue: one learner, one microphone, no overlap. Dialogic tasks in L2 speech are very different: they are co-constructed, often simultaneous, and much of the communicative meaning is carried by nods, gaze, and timing as well as paralinguistic features such as laughter, rather than words alone. To approximate human raters in this richer setting, we capture who spoke, when, what they said, and how they behaved nonverbally. Then we let AI models discover how these cues predict proficiency.

Figure 3 provides an overview of the proposed automatic spoken interaction assessment pipeline, which combines speech and visual modalities to predict proficiency scores. Since the dataset contains overlapping dialogic speech, we use speaker diarization, i.e., separating each recording segment according to speaker identity, which is needed to identify speech samples for each individual (McKnight et al., 2023). Then, we utilize automatic speech recognition (ASR) models, such as wav2vec 2.0 (Baevski et al., 2020) or Whisper (Radford et al., 2022), to extract both linguistic features (e.g., fluency, lexical range, and grammatical accuracy) and prosodic features (e.g., pitch, intensity, and tonal properties). Linguistic features can be captured indirectly by the ASR models through text transcripts or directly from the audio recording. These features will then be fed into machine learning models – either text-based architectures like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) or multitask ASR setup that perform transcription and proficiency estimation simultaneously – to estimate scores on holistic and analytic proficiency scales (range, fluency, accuracy, pronunciation, and interaction).

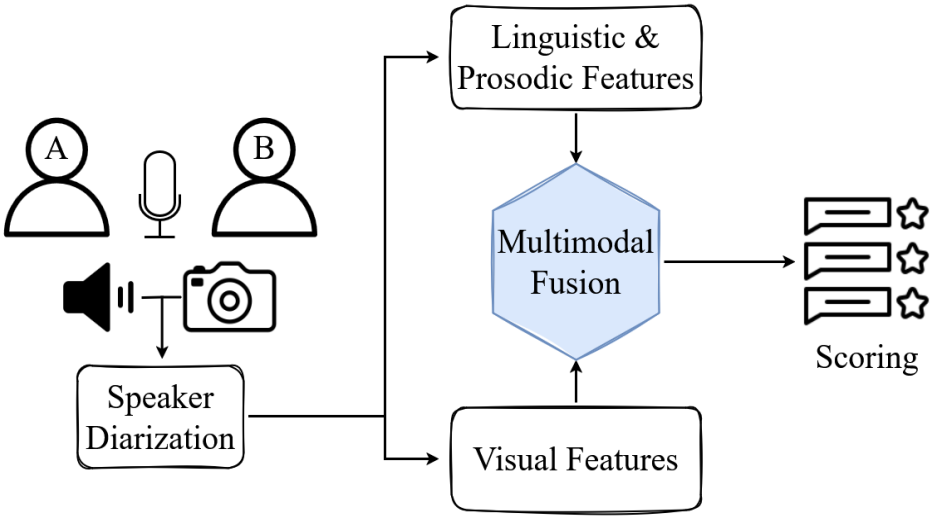

Figure 3. Automatic Spoken Interaction Assessment pipeline.

For nonverbal behavior, we plan to leverage computer vision techniques for automatically detecting gestures, head movements, and facial expressions from video recordings. The model will be trained to identify salient cues that align with communicative effectiveness, e.g., frequency and appropriateness of nods, hand gestures, or gaze patterns. By extracting and training using these visual signals, we aim to capture key aspects of speakers' interactional competence that are typically overlooked in purely audio-based systems.

Finally, we will combine the verbal and nonverbal features into a multimodal framework, covering both speech and visual cues. This system will be trained and validated using the human rating data to learn the mappings from multimodal inputs to proficiency scores. By developing this comprehensive automatic system, our goal is to provide a more comprehensive picture of learner interactional competence in L2 Finnish, especially for dialogic tasks where nonverbal communication plays a central role.

Preliminary results on nonverbal behavior during turn-taking and rater and speaker experience

We next present preliminary results, focusing on two topics: (1) nonverbal behavior during turn-taking based on ELAN annotations, and (2) raters' and speakers' experiences of automatic speaking assessment based on the raters' post-rating

questionnaire, five rater interviews, and the speakers' post-test questionnaire. Both topics are important in the development of ASA systems.

Description of nonverbal behavior during turn-taking in a dialogue task

First, we will give an example of the analysis (Raud, 2025) of the nonverbal behavior in the annotated video data in one dialogue task (Task 5). Such analysis can support the development of ASA systems by providing training data for speaker change detection. For the purposes of the analysis of this sub-study, ELAN annotations of a total of 47 dialogues were used. As four of the dialogues were annotated twice, the total number of speaker ELAN annotations was 102 (247 minutes of annotations, with 17 minutes of that overlapping[6]).

Figures 4 and 5 show the nonverbal behavior based on annotations during turn-taking, illustrating the 20-second context around the speaker change on the x-axis and the mean frequency of occurrence of nonverbal features on the y-axis. The zero crossing on the x-axis is the moment of turn-taking. An outgoing speaker is the speaker who holds the floor first, whereas the incoming speaker means the speaker taking the floor. These figures show evidence of specific nonverbal behaviors undergoing change during turn-taking. Of interest to us are the segments of recordings where behavior is divergent among the speakers. This can be found in hand movements (fidgeting and gesturing) and other nonverbal behavior (nodding, laughter, and backchannelling). In Figure 4, the divergence between the incoming and outgoing speaker increases during and immediately following a turn change in the way that the outgoing speaker uses more fidgeting before taking the turn than the incoming speaker. The incoming speaker uses more fidgeting shortly after finishing their turn. For hand gestures, the situation is different: the outgoing speaker uses slightly fewer gestures before and after the turn-taking than the incoming speaker.

[6] For the purposes of this sub-study, overlapping annotations are considered as independent samples.

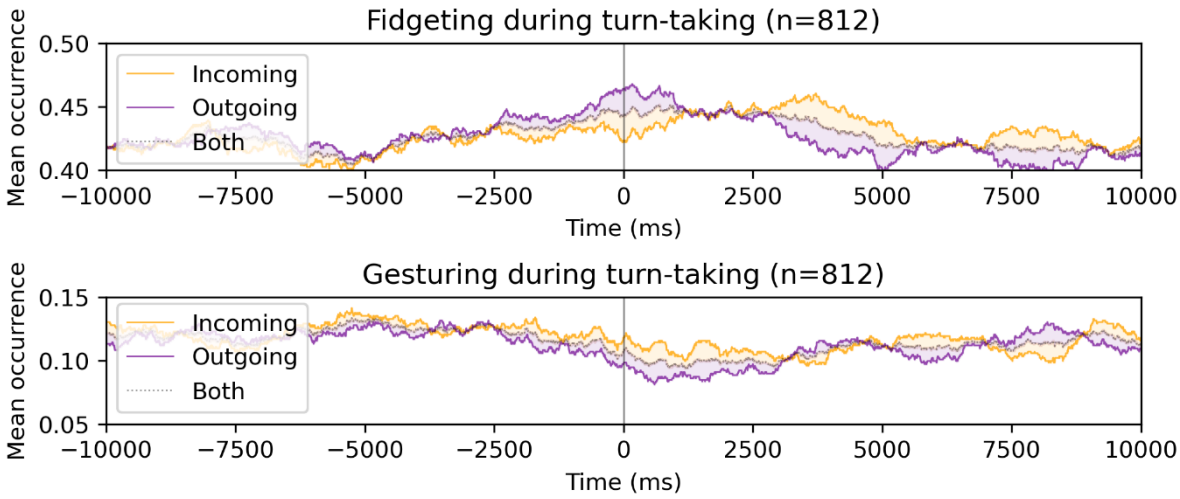

Figure 4. Mean occurrence of hand movements ("fidgeting" and "gesturing") during turn-taking

Similarly to Figure 4, Figure 5 illustrates nonverbal behavior during turn-taking. Laughing and backchannelling can be seen as signaling an outgoing speaker, while nodding seems to be a feature related to becoming a speaker.

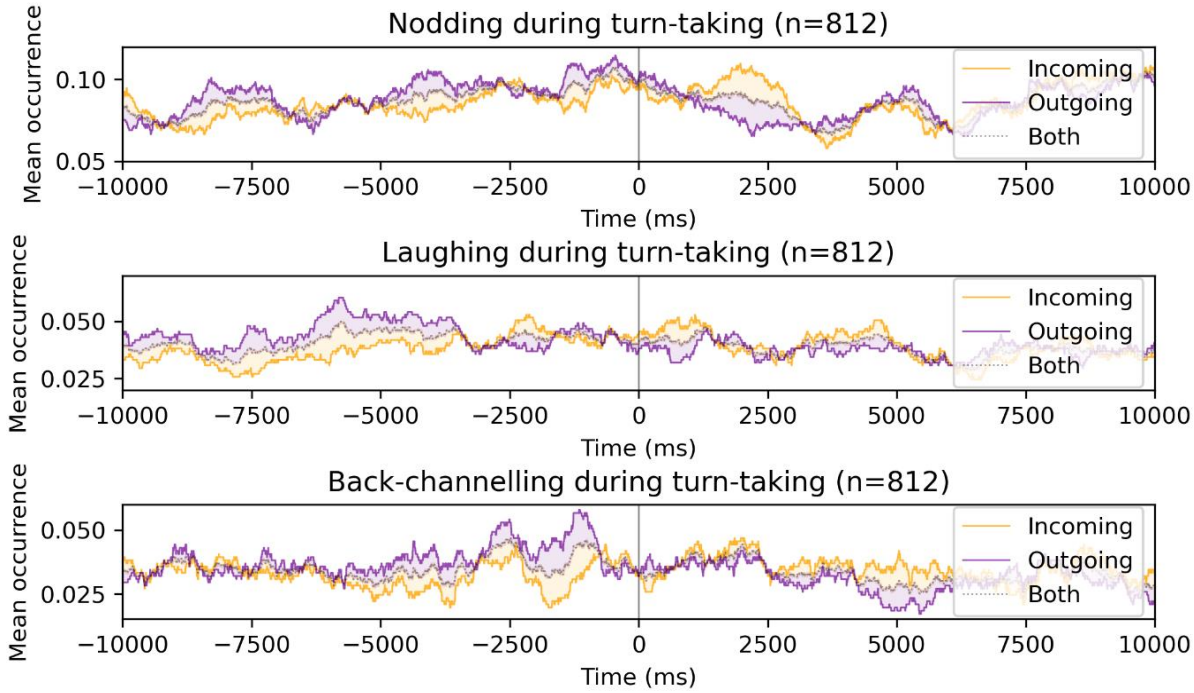

Figure 5. Mean occurrence of "nodding", "laughing", and "back-channeling" during turn-taking

Naturally, this dataset, with its rich annotations, is valuable not just for the development of more accurate assessment tools but for creating conversational agents that can control virtual avatars. These systems require multimodal (video, as well as audio or text) input, and they learn to understand and mimic human behavior from it,

as reported in Althubyani et al. (2025). Recently, it has also been shown that creating such a system is beneficial in educational scenarios (Zhu et al., 2024). Based on this, we hypothesize that our collected data will pave the way for the development of a virtual language learning tutor, such as the InteLLA-tool (Saeki et al., 2024), that can simultaneously assess and converse with users. The challenges for designing such systems are numerous, including the important decision of when to take turns in the conversation, what kind of nonverbal cues to use and observe, and how to resolve conflicts (i.e., interruptions, overlapping speech, etc.). Fortunately, our collected corpus is rich in this type of data, as demonstrated by our analysis, and could serve as training data for such systems.

Rater and speaker experience

For developing ASA systems, it is also important to take into account the potential end users. Thus, we next present some preliminary results on rater and speaker experience, which will be reported in upcoming publications in more detail. We report on the raters' views based on the questionnaire and interviews, followed by presenting an interesting result from the speakers' questionnaire.

The responses to the post-rating questionnaire (von Zansen, 2024c) showed that, overall, the raters viewed automated assessment of spoken interaction positively. When asked about their views on the topic, 65% of raters described them as "somewhat positive" or "very positive," 30% were neutral, and 5% reported "somewhat negative" views. Notably, none of the raters reported their views to be "very negative".

Raters' views on automatic assessment were studied more in depth by analyzing interviews conducted after the pilot rating round using qualitative content analysis (Suuronen, 2025). The interviewed five raters were rather unanimous about their views. For the most part, they raised similar issues about ASA: its objectivity and challenges with its inflexibility. Additionally, the overall attitude towards automatic assessment varied, with some raters expressing a more negative stance than others. The key advantages of automated assessment identified by the raters were cost-efficiency, the ability to provide more frequent assessments to learners because of the untiring nature of the machine, and objectivity, as illustrated in the excerpt below:

> et se pystyttäis ehkä sit sil poistamaan, että ei, et me ei arvioitais esimerkiks virolaistaustasia puhujia paremmaks puhujaks sen takia, et he jotenki luontevasti nappaa sen kielen tai, tai vaikka vietnamilaistaustasii puhujia helposti niin ku huonommiks puhujiks, koska heille se ääntäminen on todella vaikeaa tosi kauan.
>
> that it could maybe be avoided, that we wouldn't rate for example speakers with an Estonian background to be better speakers because they learn the language somehow naturally, or, or for example Vietnamese speakers to be weaker because for them pronunciation is very difficult for a long time. (Rater 4) [translation by the authors]

These qualities made automated assessment seem fairer and more suitable for high-stakes tests.

However, the raters also acknowledged the potential challenges of automation, including its mechanical and inflexible nature, lack of cultural understanding, and the risk of overwhelming learners with excessive feedback. This is clearly visible in the comment by Rater 2:

> vaikka joku sellanen, että, et jos joku haluaa hetken pohdiskella, et mitä hän vastaa johonkin spontaaniin kysymykseen - - - saattaa olla et kone kattoo, että mmmm ei kuulu mitään. Ja sit taas me suomalaiset tiedetään et no se nyt, nythän se miettii. Tai siis kuhan se puhe alkaa sit sielt tulla.
>
> *something like that, that if someone wants to take time to think what they want to answer to a spontaneous question, that it - - - it may be that the machine misinterprets that there is nothing being said. And then we as Finnish people know that, oh now the speaker is thinking. I mean if the speech begins after the pause.* [translation by the authors]

The raters suggested that the optimal assessment would combine automated and human assessment, balancing the efficiency and objectivity of automation with the judgement of human assessors. In addition, Rater 5 underlined the need for human verification of the machine assessments:

> - - - se ois ehkä ideaali tilanne niinkö että, että se niin kun riittäis, että vaan kattois läpi, että onko tämä hyvä arvio ja sitte tsekkais sen videon ja, että onko nämä niinku yks yhteen. Että miksi ei sitte niin ku ihan parhaassa tilanteessa niin se vois olla riittävä pelkästäänki se konearvio, mut aina se pitäis niin ku tarkistaa kuitenki.
>
> *- - - it would be an ideal situation that, that it would be enough that we could go through it, that is it a good rating and then watch the video and check that the rating and the video are compatible. Maybe in the best scenario the automatic assessment would be enough on its own, but it should always be checked by a human.* (Rater 5) [translation by the authors]

Finally, we get to the promised interesting result about the speakers' views on ASA reported in the post-test questionnaire (von Zansen, 2024d). The close-ended answers suggest that speakers display very little concern regarding automated assessment of spoken interaction. In fact, to the question "Would you be worried if you knew that your test performance was being assessed by a computer instead of a human?", a clear majority (81% of the participants) answered "No". This indicates that participants display a generally very positive outlook on ASA. This result seems surprisingly positive and is very promising regarding our goal of developing ASA in high-stakes contexts.

Concluding remarks

As described above, the Aasis project focuses on ASA of interactional speech, building on the previous DigiTala project (see Hilden et al., this issue). Its novelty lies in incorporating nonverbal features into the concept of L2 interactional proficiency by assessing nonverbal features both automatically and manually, thus adding to our understanding of the nonverbal behavior that is relevant to interactional competence.

At the time of writing, we are at the stage of having collected corpora that are very suitable for our goals and are about to start developing the ASA system. Furthermore, we are already conducting studies on multiple aspects of the data. Apart from the studies mentioned in this paper, we are studying for example, integrating Large Language Models (LLMs) for Finnish (Phan et al., 2024; Voskoboinik et al., 2025), investigating how facial expressions can be linked to various ratings, examining L2 Finnish learners' interpersonal behavior (von Zansen et al., 2025b) and the relationship between visual and verbal aspects of L2 Finnish fluency (Ullakonoja et al., submitted). Finally, it looks very promising that we will reach our broader aim, that is, to develop the first ASA system for L2 interactional Finnish, which will serve to further develop assessment of interactional speech in the future.

Acknowledgements

We wish to acknowledge the invaluable work of our project team members Yaroslav Getman and Ekaterina Voskoboinik for designing and managing the Moodle platform for data collection.

Author disclosures

The authors reported no conflicts of interest in undertaking this research.

The project Aasis, “Automatic assessment of spoken interaction in second language”, is supported by the Research Council of Finland (grant numbers 355586, 355587, 355588). We also received grants from the Helsinki Institute for Social Sciences and Humanities (Catalyst Grant 2024) and The Language Bank of Finland.

The authors had the following roles respectively in conducting the research and writing the article:

Riikka Ullakonoja: writing – original draft, writing – review & editing, methodology, investigation, data curation, resources, project administration, supervision

Ilona Lähteenmäki: writing – original draft, writing – review & editing, investigation, visualization, methodology, data curation

Nora Raud: writing – original draft, investigation, data curation, software, visualization

Nhan Phan: data curation, investigation, methodology, resources, software, visualization, writing – original draft, writing – review & editing,

Tamás Grósz: writing – original draft, writing – review & editing, methodology, investigation, data curation, software, visualization, supervision, conceptualization

Henna Suuronen: writing – original draft, writing – review & editing, investigation

Raili Hilden: methodology, funding acquisition, conceptualization, project administration, writing – original draft

Mikko Kurimo: methodology, funding acquisition, conceptualization, project administration, writing – original draft, data curation, resources, supervision

Mikko Kuronen: writing – review & editing, methodology, funding acquisition, conceptualization, project administration

Anna von Zansen: conceptualization, funding acquisition, data curation, investigation, methodology, project administration, resources, supervision, validation, writing – original draft, writing – review & editing

Maria Kautonen: investigation, methodology, resources, writing – original draft

ORCID iDs

Riikka Ullakonoja https://orcid.org/0000-0002-9421-3706

Ilona Lähteenmäki https://orcid.org/0009-0004-6014-9584

Nora Raud https://orcid.org/0009-0006-0987-4574

Nhan Phan https://orcid.org/0000-0003-2040-9834

Tamás Grósz https://orcid.org/0000-0001-7918-9579

Henna Suuronen https://orcid.org/0009-0002-2675-5148

Raili Hilden https://orcid.org/0000-0002-5114-5600

Mikko Kurimo https://orcid.org/0000-0001-5278-7974

Mikko Kuronen https://orcid.org/0000-0001-5971-7063

Anna von Zansen https://orcid.org/0000-0002-6444-7667

Maria Kautonen https://orcid.org/0000-0002-0688-7445

References

Al-Ghezi, R., Getman, Y., Rouhe, A., Hilden, R., & Kurimo, M. (2021). Self-supervised End-to-End ASR for Low Resource L2 Swedish. *Proceedings of the Interspeech 2021*, 1429–1433. ISCA International Speech Communication Association. https://doi.org/10.21437/Interspeech.2021-1710

Al-Ghezi, R., Voskoboinik, K., Getman, Y., Von Zansen, A., Kallio, H., Kurimo, M., Huhta, A., & Hilden, R. (2023). Automatic speaking assessment of spontaneous L2 Finnish and Swedish. *Language Assessment Quarterly, 20*(4–5), 421–444. https://doi.org/10.1080/15434303.2023.2292265

Althubyani, M., Meng, Z., Xie, S., Seung, C., Razzak, I., Sandoval, E.B., Kocaballi, B., & Cruz, F. (2025). MERCI: Multimodal emotional and personal conversational interactions dataset. *Human-Computer Interaction*. https://doi.org/10.48550/arXiv.2412.04908

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460. https://doi.org/10.48550/arXiv.2006.11477

Blanch-Hartigan, D., Ruben, M. A., Hall, J. A., & Mast, M. S. (2018). Measuring nonverbal behavior in clinical interactions: A pragmatic guide. *Patient Education and Counseling, 101*(12), 2209–2218. https://doi.org/10.1016/j.pec.2018.08.013

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. www.coe.int/lang-cefr

Couper-Kuhlen, E., & Selting, M. (1996). Towards an interactional perspective on prosody and a prosodic perspective on interaction. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 11–56). Cambridge University Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L 119, 1–88. http://data.europa.eu/eli/reg/2016/679/oj

Finnish National Agency for Education (2025). *National Certificates of Language Proficiency (YKI)*. Retrieved March3 1, 2025, from https://www.oph.fi/en/national-certificates-language-proficiency-yki

Fulcher, G. (2014). *Testing second language speaking*. Routledge. https://doi.org/10.4324/9781315837376

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Gan, Z., & Davison, C. (2011). Gestural behavior in group oral assessment: A case study of higher- and lower-scoring students. *International Journal of Applied Linguistics*, *21*(1), 95–120. https://doi.org/10.1111/j.1473-4192.2010.00264.x

Glasson, N. (2024). *Left to their own devices: Exploring interactional practices in an online group speaking task* [Doctoral dissertation, University of Bedfordshire]. University of Bedfordshire Repository. http://hdl.handle.net/10547/626396

Giri, V. N. (2009). Nonverbal communication theories. In S. W. Littlejohn & K. A. Foss (Eds.), *Encyclopedia of communication theory* (pp. 690–694). SAGE.

Huttunen, I., & Jaakkola, H. (2003). *Eurooppalainen viitekehys: Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys*. WSOY.

Lähteenmäki, I., von Zansen, A., Kautonen, M., & Phan, N. (2025). Aasis data collection recording setup. Zenodo. https://doi.org/10.5281/zenodo.15350034

The Language Bank of Finland. (2024). *Language Bank*. Kielipankki. Retrieved March, 31, 2025, from https://www.kielipankki.fi/language-bank/

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, 41*, 841–849. https://doi.org/10.3758/BRM.41.3.841

Lizdek, I., Sadler, P., Woody, E., Ethier, N., & Malet, G. (2012). Capturing the stream of behavior: A computer-joystick method for coding interpersonal behavior continuously over time. *Social Science Computer Review, 30*(4), 513–521. https://doi.org/10.1177/0894439312436487

Matriculation Examination Board (2025). *Welcome to the Matriculation Examination website!* Ylioppilastutkinto. Retrieved March, 31, 2025, from https://www.ylioppilastutkinto.fi/en

McKnight, S. W., Civelekoglu, A., Gales, M., Bannò, S., Liusie, A., & Knill, K. M. (2023). Automatic Assessment of Conversational Speaking Tests. *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 99–103. https://doi.org/10.21437/SLaTE.2023-19

Peltonen, P. (2020). *Individual and interactional speech fluency in L2 English from a problem-solving perspective: A mixed-methods approach*. [Doctoral dissertation, University of Turku]. UTUPub. https://urn.fi/URN:ISBN:978-951-29-8137-3

Phan, N., von Zansen, A., Kautonen, M., Voskoboinik, E., Grósz, T., Hilden, R., & Kurimo, M. (2024). Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task. *Proceedings of Interspeech 2024*, 317–321. ISCA International Speech Communication Association. https://doi.org/10.21437/Interspeech.2024-1166

Plough, I. (2021). A case for nonverbal behavior: Implications for construct, performance, and assessment. In M. R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context—Expanding the construct and its applications* (pp. 50–72). Multilingual Matters. https://doi.org/10.21832/9781788923828-004

Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing, 35*(3), 427–445. https://doi.org/10.1177/0265532218772325

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2212.04356

Raud, N. (2025). *Automatic assessment of L2 interactional competency*. [Master's thesis, Aalto university]. Aaltodoc. https://urn.fi/URN:NBN:fi:aalto-202508196328

Sadler, P., Ethier, N., Gunn, G., Duong, D., & Woody, E. (2009). Are we on the same wavelength? Interpersonal complementarity as shared cyclical patterns during interactions. *Journal of Personality and Social Psychology, 97*(6), 1005–1020. https://doi.org/10.1037/a0016232

Saeki, M., Takatsu, H., Kurata, F., Suzuki, S., Eguchi, M., Matsuura, R., Takizawa, K., Yoshikawa, S., & Matsuyama, Y. (2024). InteLLA: Intelligent Language Learning Assistant for Assessing Language Proficiency through Interviews and Roleplays. *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 385–399. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.sigdial-1.34

Song, L. (2025). *The Role of Visual Input and Joint Attention in Second Language Dyadic Speaking Tasks: An Eye-Tracking Study on the Impact of Manipulable Real Objects and Static Images*. [Master's thesis, University of Helsinki]. Helda. http://urn.fi/URN:NBN:fi:hulib-202509103730

Suuronen, H. (2025). *Ihmisarvioijien näkemyksiä suullisen vuorovaikutuksen automaattisesta arvioinnista*. [Master's thesis, University of Helsinki]. Helda. http://urn.fi/URN:NBN:fi:hulib-202507023400

TENK, Finnish National Board on Research Integrity. (2019). *The ethical principles of research with human participants and ethical review in the human sciences in Finland* [PDF]. Finnish National Board of Research Integrity TENK publications 3/2019. https://tenk.fi/sites/default/files/2021-01/Ethical_review_in_human_sciences_2020.pdf

Timpe-Laughlin, V., & Youn, S. J. (2021). Measuring L2 pragmatics. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 254–264). Routledge. https://doi.org/10.4324/9781351034784-28

Ullakonoja, R., Peltonen, P. & Kuronen, M. (submitted). Visuo-gestural behavior during pausing in L2 Finnish at different fluency levels.

Yle Oppiminen. (2016). *Asiointisuomea*. Yle. Retrieved March 31, 2025, from https://yle.fi/a/20-146213

von Zansen, A. (2022). DigiTala's speaking tasks for L2 Finnish learners (proficiency level B1). Zenodo. https://doi.org/10.5281/zenodo.6562855

von Zansen, A. (2024a). Aasis consent form for L2 Finnish learners (2024). Zenodo. https://doi.org./10.5281/zenodo.11066027

von Zansen, A. (2024b). Aasis nonverbal rating scale (2024). Zenodo. https://doi.org/10.5281/zenodo.14214966

von Zansen, A. (2024c). Aasis post-rating questionnaire for human raters (2024). Zenodo. https://doi.org/10.5281/zenodo.14754353

von Zansen, A. (2024d). Aasis post-test questionnaire for L2 Finnish learners (2024). Zenodo. https://doi.org./10.5281/zenodo.11066137

von Zansen, A. (2024e). Aasis privacy notice (2024). Zenodo. https://doi.org/10.5281/zenodo.11046434

von Zansen, A. (2024f). Aasis speaking tasks for Finnish L2 learners (2024). Zenodo. https://doi.org./10.5281/zenodo.11046534

von Zansen, A., & Kautonen, M. (2024a). Aasis information notice for L2 Finnish learners (2024). Zenodo. https://doi.org/10.5281/zenodo.11045461

von Zansen, A., & Kautonen, M. (2024b). Webropol rating form (pilot rating 2024). Zenodo. https://doi.org/10.5281/zenodo.14438909

von Zansen, A., Kautonen, M., Lähteenmäki, I., Kuusela, J., & Kurimo, M. (2024). Aasis ELAN Transcription and Annotation Instructions. Zenodo. https://doi.org/10.5281/zenodo.13944045

von Zansen, A., Kautonen, M., Lähteenmäki, I., Huhta, A., Hilden, R., Kuronen, M., & Ullakonoja, R. (2025a). Rating criteria for monologue and dialogue speech. Zenodo. https://doi.org/10.5281/zenodo.15074770

von Zansen, A., Lähteenmäki, I., Juselius, J., & Henttonen, P. (2025b). Beyond monologues – Examining L2 Finnish learners' interpersonal behavior during dialogue speaking tasks using the computer-joystick method. *System*, 134. https://doi.org/10.1016/j.system.2025.103803

von Zansen, A., Voskoboinik, E., Ullakonoja, R., Getman, Y., & Phan, N. (2025c). Aasis Moodle rating platform (2025). Zenodo. https://doi.org/10.5281/zenodo.15077384

Voskoboinik, E., von Zansen, A., Phan, N., Getman, Y., Grósz, T., & Kurimo, M. (2025). Enhancing Second Language Speech Assessment: Integrating Large Language Models for Finnish and Finland Swedish Proficiency Scoring. *Language Testing, 42*(4), 508–538. https://doi.org/10.1177/02655322251351648

Zhu, Y., Guo, L., Sun, J., & Hei, X. (2024). Designing a LLM-driven avatar system to enhance social skills for autistic children in DTT learning. *Proceedings of the 2024 International Conference on Intelligent Education and Intelligent Research (IEIR),* 1–8. IEEE. https://doi.org/10.1109/IEIR62538.2024.10959902