JYX
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Al-Ghezi, Ragheb; Voskoboinik, Katja; Getman, Yaroslav; Von Zansen, Anna; Kallio, Heini; Kurimo, Mikko; Huhta, Ari; Hildén, Raili

Title: Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish

Year: 2023

Version: Published version

Copyright: © 2023 The Author(s)

Rights: CC BY 4.0

Rights url: https://creativecommons.org/licenses/by/4.0/

Please cite the original version:

Al-Ghezi, R., Voskoboinik, K., Getman, Y., Von Zansen, A., Kallio, H., Kurimo, M., Huhta, A., & Hildén, 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

Language Assessment Quarterly

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hlaq20

Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish

Ragheb Al-Ghezi, Katja Voskoboinik, Yaroslav Getman, Anna Von Zansen, Heini Kallio, Mikko Kurimo, Ari Huhta & Raili Hildén

To cite this article: Ragheb Al-Ghezi, Katja Voskoboinik, Yaroslav Getman, Anna Von Zansen, Heini Kallio, Mikko Kurimo, Ari Huhta & Raili Hildén (19 Dec 2023): Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish, Language Assessment Quarterly, DOI: 10.1080/15434303.2023.2292265

To link to this article: https://doi.org/10.1080/15434303.2023.2292265

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 19 Dec 2023.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=hlaq20

LANGUAGE ASSESSMENT QUARTERLY
https://doi.org/10.1080/15434303.2023.2292265

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish

Ragheb Al-Ghezi[a], Katja Voskoboinik[a], Yaroslav Getman[a], Anna Von Zansen[b], Heini Kallio[c], Mikko Kurimo[a], Ari Huhta[c], and Raili Hildén[b]

[a]Aalto University, Aalto, Finland; [b]University of Helsinki, Aalto, Finland; [c]University of Jyväskylä, Jyvaskyla, Finland

ABSTRACT

The development of automated systems for evaluating spontaneous speech is desirable for L2 learning, as it can be used as a facilitating tool for self-regulated learning, language proficiency assessment, and teacher training programs. However, languages with fewer learners face challenges due to the scarcity of training data. Recent advancements in machine learning have made it possible to develop systems with a limited amount of target domain data. To this end, we propose automatic speaking assessment systems for spontaneous L2 speech in Finnish and Finland Swedish, comprising six machine learning models each, and report their performance in terms of statistical evaluation criteria.

INTRODUCTION

Technology has opened the door to new possibilities in language testing and assessment. Machine learning provides a means to automatize L2 proficiency testing, but to this day applications have been more common for written than spoken language. Developing automatic systems for assessing spontaneous speech has become highly desirable in the context of L2 learning because they promote and democratize self-regulated learning, as well as serve as a facilitating tool in language proficiency assessments and teacher training programs. Such systems are typically developed for languages with a large number of learners due to the abundance of training data, yet languages with fewer learners such as Finnish and Swedish remain at a disadvantage due to the scarcity of required training data. Nevertheless, recent advancements in the field of AI manifested in self-supervised machine learning methods (Al-Ghezi et al., 2021; Devlin et al., 2019) make it possible to develop automatic speech recognition (ASR) systems with a reasonable amount of training data, which makes it feasible to develop automatic speaking assessment systems for under-resourced languages.

This article describes the development steps of the first prototype of an automatic assessment system for spontaneous L2 Finnish and Finland Swedish speech and reports the initial evaluation of the system. In addition to supporting self-regulated learning purposes, the tool could be used in a national school-leaving exam for scoring.

CONTACT Ragheb Al-Ghezi ragheb.al-ghezi@aalto.fi Signal Processing and Acoustic, Aalto University, Aalto, Finland

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

In Finland, around 30,000 students take the Matriculation Examination (ME), the school-leaving examination for general upper-secondary education each year. The ME aims to measure students' learning outcomes as well as their overall readiness for post-secondary education (Finnish Matriculation Examination Board, n.d.). The ME is the only external high-stakes examination in the Finnish school system. One of the concerns with the language tests in the ME is that they have always lacked a speaking part, although the national curriculum highlights the importance of oral language skills (Vaarala et al., 2021). The reason for this is mainly practical: Implementing an oral test in the ME would require considerable resources.

The computerization of the ME in 2017 was a significant change in itself but it might also help introduce speaking tests into the examination. Even partial automation of the assessment of speaking would improve the practicality of oral testing significantly. Therefore, the first goal of this work is to investigate the automated assessment of oral skills in the kinds of tasks that are likely to be used in the speaking tests in the ME (Xi et al., 2008). Research to date seems to support the use of a hybrid approach combining human and automated scoring in high-stakes settings (Evanini & Zechner, 2020; Xi, 2021). Therefore, we envisage automated scoring to complement human ratings rather than replace them. The second goal of this study is to support L2 Finnish and Finland Swedish learners' self-regulated learning by providing automatic feedback (Evanini & Zechner, 2020) based on the procedures developed to address the first goal above.

RESEARCH QUESTIONS

The research questions of this study are set out to support the assessment use argument of using automated speech scoring (Bachman & Palmer, 2010; Chapelle et al., 2011; Xi, 2021). We seek to investigate two important inferences in the validity argument: evaluation and explanation (Xi, 2021). RQ1 and RQ2 address the evaluation inference, (i.e., whether the automated scores are accurate indicators of the quality of the performance samples). They focus on the accuracy of the ASR, which partially impacts the accuracy of the resulting automated scores, and the human-machine score alignment in comparison to human-human score alignment. RQ3 pertains to the explanation inference, investigating the construct relevance of the top features used in the scoring models. This study is the first one to examine these questions in the context of Finnish and Finland Swedish as second languages. The present study investigated the following research questions (RQ):

RQ1: What is the quality of the ASR system measured as word and character error rate (WER/CER) in the two languages?

RQ2: How do the scoring models for the two languages perform in terms of human-machine agreement compared to human-human agreement?

RQ3: What are the most important features in the scoring models for the two languages?

BACKGROUND

Automatic assessment of non-native speech started with the focus on segmental pronunciation in read speech (Bernstein et al., 1990). Read speech has been favored due to its predictability in eliciting speech from different speakers, leading to greater accuracy in speech recognition and modeling of pronunciation features. The more unconstrained the speaking tasks are, the more unpredictable the elicited speech becomes, which causes problems for automatic scoring. The ability to produce spontaneous speech is, however, essential in human communication, and therefore researchers started to study the automatic assessment of spontaneous speech in the early 2000s, focusing 2000s, focusing on fluency features (Cucchiarini et al., 2002). Here we use the term spontaneous in comparison to read or imitated speech, although all responses to instructed tasks are still predictable to some extent.

Validation of automated speaking tests should start with identifying the role of automated scoring in the context of use (Xi, 2021). In the current study, we investigated the relationship between machine and human scoring to pave the way for their future combination in the scoring of learners' speaking performances in the ME. For pedagogical training purposes, machine scoring may serve as a useful tool to promote learner autonomy and self-efficacy. Our focus in this study is to establish a validity argument for using automated scoring in conjunction with human scoring in the future. Validation as an argumentation chain has been developed in a number of seminal works (Bachman & Palmer, 2010; Chapelle et al., 2011; Downing & Haladyna, 2006), and more recently by Xi (2021) specifically addressing validation of automated test administrations.

A few key steps are followed in this investigation including defining the construct and language use domain, designing the tasks drawing on the national core curricula, and monitoring the accuracy of the ASR and human-machine score alignment to explain and evaluate the quality of the observed score as an accurate indicator of students' speaking performance. Particular attention is paid to the relevance and accurate measurement of the construct when subjected to automated evaluation, a crucial aspect as voiced by Xi (2021).

Developing a validity argument involves attaining an acceptable balance between competing aspects. Automated tests increase the reliability in terms of the consistent scoring of test takers, test takers, which is vital in high-stakes assessments. On the other hand, the current limitations of automated task types would compromise the authenticity and construct representation of the assessment. The constrained task types that elicit read speech or short predictable phrases narrow the construct of oral competence. However, developments in the automated evaluation of more spontaneous speech samples are promising in this regard (Xi et al., 2008). Free speaking tasks, such as speaking on a given topic, picture description, and responding to verbal inputs, require learners to produce open-ended responses. Compared to constrained task types, free-speaking tasks are considered more authentic and better aligned with communication-oriented oral constructs. For such tasks, it is possible to use a task-specific language model with a vocabulary constrained to the task domain to improve the performance of the ASR.

Using tasks that generate both scripted and spontaneous speech enables a more comprehensive evaluation of learners' speaking skills: Mechanical tasks such as read-aloud and sentence repetition can be used for measuring processing speed (Van Moere, 2012) or specific pronunciation features such as stress production (Kallio et al., 2020, 2022), while

tasks eliciting unconstrained speech also enable the assessment of lexical, grammatical, and cohesion skills (Winke & Brunfaut, 2020). It is noteworthy, however, that the current automated speaking tests cannot cannot cover all aspects of L2 speaking proficiency: Current tests mainly include monologic tasks, which do not assess some of the higher-order skills such as interactional competence (Xi, 2010).

To date, several automated systems for semi-direct spoken L2 assessment have been introduced. The dominant automatic assessment systems have been developed for English as a second or foreign language (Educational Testing Service, n.d.; Pearson, 2017; Xi et al., 2008; Xu et al., 2021), which often use massive data sets, enabling advanced and elaborate machine learning. Versant has also developed tests for Spanish, Dutch, French, and Arabic, the last of which could be classified as a low-resourced language (Pearson, 2023). The target languages in the present study, Finnish and Finland Swedish as L2, are also considered low-resourced languages. Due to data scarcity, the possibilities to develop automated L2 speech assessments for these languages have been limited compared to the well-known systems for English. Research on features related to oral proficiency in Finnish and Finland Swedish has also been quite rare (Kallio et al., 2017, 2020, 2021, 2022, 2023).

Gu and Davis (2020) described a more comprehensive, automated diagnostic feedback system developed for spontaneous speech. In addition to reporting proficiency level, the system aims at providing both analytic and holistic feedback to the learners (von Zansen & Heijala, 2023; von Zansen & Huhta, 2022). Similarly, the Versant tests for Arabic and English (Pearson, 2018, 2020) report an overall score and four diagnostic scores while the Linguaskill Speaking Test (Xu et al., 2020) provides a proficiency level as feedback to the learner. Automated feedback is beyond the scope of this paper as we will focus on automated assessment of vocabulary, grammar, pronunciation, and fluency in developing a comprehensive spoken assessment system.

A combination of human and machine scoring seems to be the most appropriate alternative to human scoring to date (Evanini & Zechner, 2020), since some speech features, such as pronunciation and fluency (particularly in short phrases) are measured more accurately by a machine, whereas more extended spontaneous speech, let alone complex verbal interactions, are most appropriately left for humans to judge. Furthermore, strong correlations between machine-scored and human-scored tests have been found (Bernstein et al., 2010), which is promising for using machine scoring in large-scale assessments. Bernstein et al. (2010), however, compared scorings from two distinct constructs: Automated scores Automated scores were derived from test-takers constrained speech, while human scores were from interactive speech tasks.

L2 automatic speech recognition

In the context of automated spoken language assessment, the development of high-performance ASR systems is crucial, and for this purpose, large amounts of transcribed speech data should be used for training. Unlike languages with more learners such as English and Spanish, adequate training data for languages with fewer learners such as Swedish and Finnish may not always be feasible.

Due to data scarcity, low-resourced L2-ASR systems are often developed using a pipeline ASR paradigm, where custom-engineered solutions are applied at each stage of the pipeline to improve performance. For instance, in acoustic modeling, each word has multiple

pronunciations to accommodate the mispronunciation of words by L2 speakers. In some cases, custom solutions for acoustic and language modeling require specialized language expertise and are cost ineffective. Therefore, developing end-to-end L2-ASR systems that do not require separate pronunciation or external language modeling is highly desirable in the context of L2 ASR for low-resourced languages. However, developing end-to-end ASR systems requires a large amount of labeled (transcribed) data, which is not always attainable.

Self-supervised learning (SSL) has emerged as an effective technique to bridge this gap. The key idea is to learn general representations in settings where large amounts of unlabeled (untranscribed) source data are available, thereby leveraging them to improve the performance of downstream target tasks with limited amounts of labeled data. This is especially interesting for tasks that require considerable effort to obtain labeled data, such as speech recognition. In this work, an SSL acoustic model called *Wav2Vec2* (Baevski et al., 2020) was incorporated in an end-to-end ASR pipeline for both Finnish and Finland-Swedish.

Scoring features in automated speaking assessment systems

In this section, we provide some background In this section, we provide some background for designing an automatic speaking assessment system for spontaneous L2 speech of Finnish and Finland Swedish is provided. In many L2 tests and studies, the main aspects of speaking proficiency relate to speech fluency, pronunciation, vocabulary, and grammar (Baker-Smemoe et al., 2014; Educational Testing Service, n.d.; Gretter et al., 2019; Kallio et al., 2022, 2023; Kang & Johnson, 2018; Pearson, 2017, 2018, 2020). Machine-derived measures also cover these dimensions of speaking proficiency, and although L2 proficiency is a broad concept, these measures have proved to be good predictors of overall oral proficiency (Baker-Smemoe et al., 2014; Cox & Davies, 2012; Kang & Johnson, 2018). Further explanations are provided below on the dimensions of speaking proficiency regarding measures that can be integrated into automated systems assessing spontaneous L2 speech.

First, vocabulary and grammar are two crucial dimensions in assessing L2 speaking. Lexical diversity or range refers to the learners' sophistication in vocabulary use (Lu, 2012). Range has been commonly measured using features like "number of different words/tokens" and "type-token ratio" (Zechner & Evanini, 2019). Other variant features include the OVIX lexical diversity measure (Hultman, 1994) for automated assessment of L2 Swedish (Östling et al., 2013). Grammatical accuracy has been quantitatively measured using a set of syntactic features derived from automated text annotation tools such as part-of-speech tagging and constituency and dependency parse trees (Östling et al., 2013; Zechner & Evanini, 2019).

Research on automatic assessment of spontaneous L2 speech has initially focused on fluency features measured by temporal features, some of which are associated with strong fluency and others weak fluency (Cucchiarini et al., 2002). A ternary division by Tavakoli and Skehan (2005) introduced three components related to speech fluency: (1) speed fluency, generally measured as speech rate, articulation rate, or mean length of syllables; (2) breakdown fluency, generally measured as the frequency, length, and/or relative amount of silent and filled pauses in an utterance; utterance; and (3) repair fluency, referring to the partial or full repetition of words, syllables, or entire phrases, false starts, or reformulations. entire phrases, false starts, or reformulations. These three dimensions have guided research on L2 fluency,

and measures of speed and breakdown fluency, in particular, have been found to correlate with assessments of fluency (Bosker et al., 2013; Cucchiarini et al., 2002; Derwing et al., 2004; Kormos & Dénes, 2004; Lennon, 1990; Préfontaine et al., 2016) and oral proficiency (Iwashita et al., 2008; Kallio et al., 2017, 2022; Kang & Johnson, 2018). Recently, researchers have also called for accounting for the locations of pauses with respect to syntactic constituents in modeling fluency (De Jong, 2018; Hsieh et al., 2020; Kallio et al., 2022).

As for pronunciation, automated assessment is generally based on acoustic model scores or phone likelihood measures (Hsieh et al., 2020; Loukina et al., 2017). A phone from the output of a speech recognizer is compared to the corresponding phone in a pronunciation model trained on a large corpus of (generally) native speech. The more the L2 pronunciation differs from the model, the lower the assumed proficiency. Automatic measures of prosodic features aim to capture the relevant rhythmic and tonal properties of speech, such as realizations of prominence and intonation (Kang & Johnson, 2018).

DATA COLLECTION AND PROCESSING

L2 speech corpora for training the automatic speaking assessment systems were human labeled. First, diverse sets of Swedish and Finnish-speaking tasks were designed. Second, grading rubrics were developed to evaluate L2 learners' proficiency on multiple dimensions. Next, learners' speech samples were collected, manually transcribed, and scored by human raters using the rubrics. Finally, data were selected for our machine-learning experiments.

Test format and task design

Four speaking tests were used: one for L2 Finland Swedish, developed in an earlier project (Karhila et al., 2016), and three for L2 Finnish, developed in the current project (von Zansen, 2022a, 2022b, 2022c). All tests were computer delivered computer delivered and included both read-aloud and freeform tasks. This study focused on the freeform tasks including semi-structured and open-ended ones (Luoma, 2004). The original descriptions of the freeform tasks are listed in Tables A1 and A2 of Appendix A for Finland Swedish and Finnish, respectively.

The Swedish-speaking test included a total of six tasks with subtasks designed to cover various dimensions of speaking proficiency, differing in formality and complexity. In addition to the read-aloud tasks, the test included three other task types: situational situational reacting tasks that involve reacting to different situations based on written prompts (10 seconds time limit/response) or a picture with text clue (30 seconds time limit/response); a simulated video phone call with pre-recorded questions and replies from a native speaker of the target language (10 seconds time limit/answer and 20 seconds time limit/question); and a live dialogue task with a peer.

The three Finnish-speaking tests were delivered using Moodle's Quiz module. Due to the COVID-19 pandemic, the data were collected remotely by giving instructions via Zoom and Teams (see von Zansen et al. (2022a), and von Zansen and Hilden (2022)).

First, two test versions targeting B1 (von Zansen, 2022c) and B2 (von Zansen, 2022d) levels were designed following the goals, contents, and target-level descriptions of the National Core Curriculum (NCC) for upper secondary education (Finnish National Agency for Education, 2015) and (Finnish National Agency for Education, 2019). Both

tests included read-aloud and production tasks (semi-structured and open-ended) with a 40-minute time limit. In the semi-structured tasks, the learner was, for example, asked to briefly reply to a comment in a webinar or answer a question during a simulated phone call (time limit 15 seconds/response). The open-ended tasks included talking about a given topic for 1 minute and describing and comparing pictures. The open-ended topics ranged from the B1 version's everyday life situations to genetically manipulated foods in B2 (see also von Zansen et al. (2022b)).

Second, the A-level speaking tasks (von Zansen, 2022a) were developed in cooperation with the university teachers teaching beginner level Finnish courses to match the course learning objectives. However, as an important starting point for the A-level task design, B1 and B2 speaking tasks were used. This approach was justified since the usefulness of the tasks targeting B-level speakers were investigated from the perspectives of human rating (von Zansen, Kallio et al., 2022), language learners' perceptions (von Zansen et al., 2022b) and functioning of the rating scales (von Zansen & Huhta, 2022). As a result, the A-level tasks included read-aloud and semi-structured and open-ended tasks (see also von Zansen and Hilden (2022).

Rating scales

Since the current NCCs (Finnish National Agency for Education, 2015, 2019) highlight communication and interaction skills, which cannot be properly measured automatically, the level descriptors from the previous NCC scale (Finnish National Agency for Education, 2003) were applied, which describe speaking skills in more detail, for scale validation see (Hildén & Takala, 2007). The Finnish NCC scales are local applications of the CEFR (Council of Europe, 2001) and divide the CEFR levels into two or even three sublevels. For example, level A2 is split into A2.1 and A2.2. However, when piloting the scales, the need to simplify the rating of learners' overall speaking proficiency was noted noted. As a result, only the main levels A1-C1 were used with no further division into sub-classes. In addition, C2 was added to the scale using the CEFR descriptors, since the NCC only goes up to C1. In addition, an extra level below A1 was included, resulting in a 7-level rating scale. The rating scales can be found in von Zansen (2022a).

For the rating of specific dimensions of speaking, five analytic scales were designed by selecting key descriptors for each dimension from the overall NCC scale (Finnish National Agency for Education, 2003). To simplify the raters' task, the analytic scales were shorter (three levels/points) and simpler (few descriptors per level; see Table B1 in Appendix B). As an initial step in validating the scales, feedback was gathered from the raters in an earlier phase of the project, which led to the addition of a fourth point to some of the scales to allow finer distinctions to be made. The final analytic dimensions were task completion (3 points), fluency (4 points), pronunciation (4 points), range (3 points), and accuracy (4 points).

Speech data and human ratings

In this study, recordings from non-native Finland Swedish ($n = 181$) and Finnish ($n = 325$) speakers were used. The Swedish data were collected from upper secondary school students in 2015 (Karhila et al., 2016), while the Finnish data were collected in 2021 from upper

secondary school (von Zansen et al., 2022b) and university students (von Zansen & Hilden, 2022). Table 1 presents the detailed characteristics of the two datasets.

The participants took the test either in classroom environments using school headsets or at home using their own microphones. As a result, the recordings varied varied in terms of audio quality and the amount of background noise. Some of them were rejected by human transcribers before starting the rating process.

Human raters were recruited and trained to assess the collected speech samples by using one holistic and five analytic rating scales. The analytic scores were given independently after giving the holistic score. The Swedish recordings were assessed by 18 raters in 2020 and the Finnish recordings in 2021–2022 by 26 raters (for details concerning human ratings, see von Zansen, Kallio et al. (2022)). The raters participated in thorough online training, after which they rated the samples between December 2020 (Swedish) and April 2022 (Finnish). The raters proceeded by scoring one sample at a time: that is, at a time: that is, rating all the dimensions for one sample before moving to the next sample.

An overlapping rating design was used, with most performances rated by at least two raters, which allowed the ratings to be analyzed by Facets (Linacre, 2020) to check for the quality of the ratings and the rating scales (von Zansen & Huhta, 2022). Fair average values for each sample were used to represent human ratings instead of raw means as they were adjusted for rater severity.

The analytic scales were further validated in our other study, which showed that the raters and the rating scales functioned as expected (von Zansen & Huhta, 2022). Furthermore, the Facets analyses of the ratings reported in the study demonstrated the scales functioned adequately when there were enough samples per scale point.

Data preparation

In this work, only the samples with a human score for every rating criterion (no criterion was marked as "non-gradable" by any of the raters) were used to eliminate potentially problematic samples. In addition, only tasks that generated freeform speech were analyzed. After filtering, the size of the Finland Swedish and the Finnish datasets was reduced to 1,542 1,542 recordings (5.5 h) and 2,112 2,112 recordings (14.1 h), respectively.

Most samples in the remaining recordings were rated by at least two human raters. In the Finland Swedish subset, 1,360 1,360 out of 1,560 1,560 speech samples were rated by two raters, 42 by three raters, 39 by five raters, and the final 101 recordings by one rater. In the Finnish subset, 1,785 1,785 out of 2,112 2,112 recordings were rated by two raters and 288 by 1 rater. In

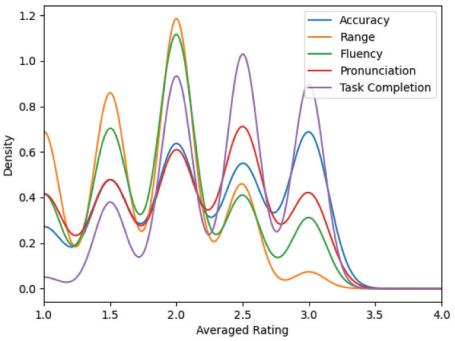
Table 1. Characteristics of the rated Swedish and Finnish data.

	Swedish L2 Data	Finnish L2 Data
Total Duration, h	7.12	18.58
Average Duration, s	12.67	15.18
# of recordings	2,025	4,405
# of ratings	4,134	9,360
# of students	181	325
# of raters	18	26
# of tasks	22	43

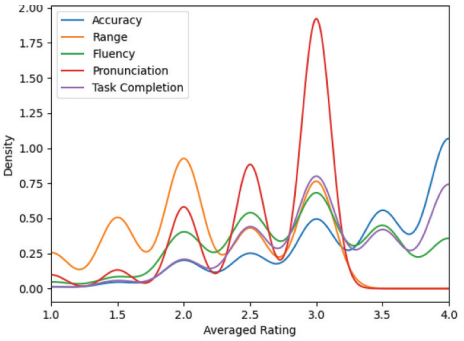
addition, there was a control set of L2 Finnish recordings which was distributed to all raters. As a result, three recordings were rated by 24 raters, and 17 samples were rated by 25 raters.

In this research project, a partial overlapping rating design was used to save resources and ensure the quality of ratings (more details are provided below in *Speech data and human ratings*). To combine the ratings from several raters into a unique score per rating criteria, the scores were averaged between raters and then rounded.

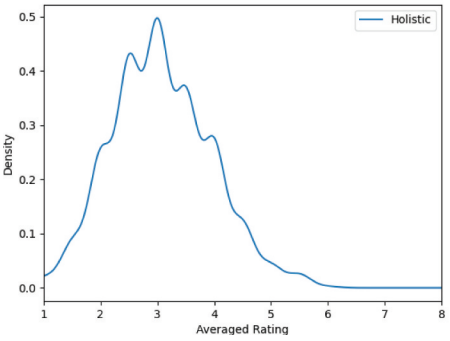
For evaluation, several factors should be considered. First, there was a limited amount of data. In addition, our data were heavily imbalanced in terms of human ratings (see Figure 1) (see Figure 1). For example, the ratings of most of the samples centered around level 2 for both languages. Moreover, the diversity of tasks and corresponding experiments should be taken into account: For each language, systems for ASR, as well as for speech rating on the holistic scale and classification on the analytic dimensions are needed. As a result, it was not possible to design a universal test set that would showcase real model performance in all our experiments. Therefore, cross-validation (CV) was used with the data split by speaker into four folds with no overlap between folds. One fold was used for testing in each training iteration of the 4-fold CV and the predictions on the test folds of the four models were aggregated when running the evaluation on the entire dataset.



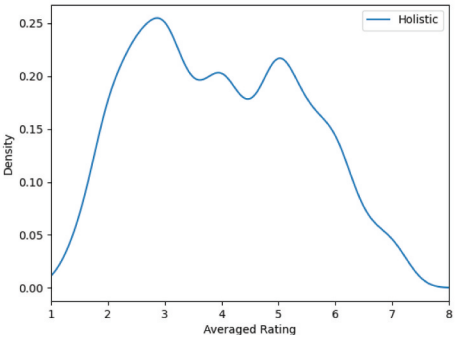
(a) *Analytic ratings of Swedish samples*



(b) *Analytic ratings of Finnish samples*



(c) *Holistic ratings of Swedish samples*



(d) *Holistic ratings of Finnish samples*

Figure 1. Distribution of speech samples between ratings. the horizontal axis represents averaged scores and the vertical axis represents normalized number of samples.

SYSTEM DESIGN

Our main goal was to develop a system to help teachers assess spontaneous L2 Swedish and Finnish short utterances in an automatic or semi-automatic fashion. The automatic assessment system, shown in Figure 2, includes a Wav2vec2-based ASR model (Al-Ghezi et al., 2021) and five main scoring models that worked concurrently to produce a score for each analytic dimension (task completion, lexico-grammatical competence, pronunciation, and fluency) and predict the overall spoken language proficiency level. Each scoring model predicted individual scores using a set of textual and acoustic features (more details in Table 5). In addition to human-designed measures, deep acoustic embeddings from the ASR (*Hidden Representations* in Figure 2) were extracted. The task completion scoring model served two purposes: to filter out responses that do not pertain to a task, and to evaluate the content of a response.

RESULTS

Research question 1: ASR

For ASR experiments, publicly available[1][2] pre-trained *Wav2Vec2-Large* (317 M parameters) models were used and fine-tuned fine-tuned using the L2 data. For Finland Swedish, the pre-trained monlingual *Wav2Vec2* model fine-tuned fine-tuned by the data lab at the National Library of Sweden was used.[3] The 11.5K hours pre-training data include unlabeled speech from Swedish local radio broadcasts and audiobooks, while the labeled finetuning

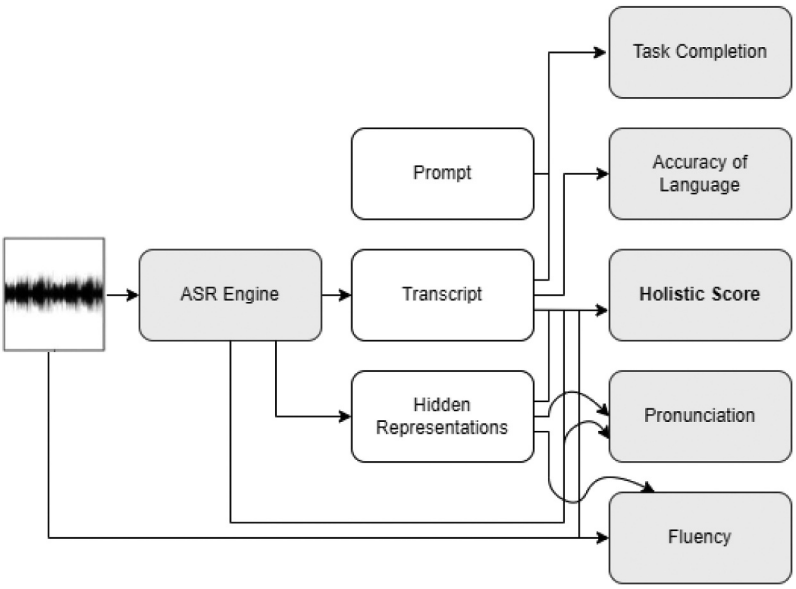


Figure 2. Schematic diagram illustrating the components of the rating system.

[1]https://github.com/facebookresearch/voxpopuli#pre-trained-models
[2]https://huggingface.co/KBLab/wav2vec2-large-voxrex-swedish
[3]https://www.kb.se/in-english/research-collaboration/kblab.html

data is composed of Swedish Common Voice (Ardila et al., 2020), Nordisk Sprakteknologi (NST) and Swedish local radio recordings.

For Finnish, a multilingual *Wav2Vec2* model pre-trained on the Uralic (Finnish, Estonian and Hungarian) part of the *VoxPopuli* corpus (Wang et al., 2021) was used which consists of 400K hours of unlabeled speech from European Parliament plenary session recordings. The Uralic subset includes in total 42.5K hours of recordings, out of which 14.2K hours are Finnish speech. Before adapting the model to our target data, it was finetuned on a transcribed 100-hour subset of the Lahjoita puhetta (Donate Speech) corpus (Moisio et al., 2023) which consists of colloquial Finnish speech.

The models were finetuned by following the 4-fold CV setup described in Section *Data preparation*, resulting in 4 sub-models for each language. Each of them was trained for 20 epochs with a learning rate of 1e-4 and an effective batch size of 4. Table 2 summarizes the results of the ASR experiments. After aggregating the test results of the sub-models, 17.71%/9.08% WER/CER and 21.89%/7.06% WER/CER were obtained on the entire data for L2 Finland Swedish and L2 Finnish, respectively. Model predictions for ASR error analysis were made, and the findings are reported in the section *Error analysis of ASR*. In this study, character error rate (CER) was used in addition to WER, because in long words the CER reflects better the number of completely wrong words compared to small errors or mispronunciations. This is particularly relevant for agglutinative languages such as Finnish where the words are often quite long.

Research questions 2 and 3: results of the scoring models

As discussed in the *Data preparation section* and shown in Figure 1, some rating dimensions included empty or heavily underrepresented categories, and modeling them for an automatic scoring system is impossible and impractical. Thus, the training and the evaluation were limited to sufficiently represented ones by cutting thin tails from the distributions. For example, for the Swedish holistic classifier, ratings from 2 to 5 were retained and the rest removed, since only 36 out of 1,542 recordings were in these removed categories. For the same reason, accuracy and range were aggregated into one analytic dimension, lexicogrammatical competence. The updated ranges for each dimension are reported in Tables 3 and 4. It should be noted that the values of the evaluation metrics were not comparable between various dimensions and between different datasets that may have different scales.

Each scoring model (see Figure 2), except for task completion, was a 6-hidden-layer neural classifier of 300 hidden units optimized by Adam optimizer with a learning rate of 1e-3. The models were trained for 600 epochs with a batch size of 100.

The task completion model served two purposes: filtering filtering and content evaluation. In the first step, the model checked if the transcript belonged to the predefined task. It

Table 2. ASR experiments on L2 Finland Swedish and L2 Finnish.

Model	# of recordings	Duration, h	WER, %	CER, %
L2 Swedish Wav2Vec2	1,542	5.6	17.71	9.08
L2 Finnish Wav2Vec2	2,112	14.1	21.89	7.06

Columns represent the developed models, the amount of finetuning data used, as well as corresponding WER and CER.

Table 3. Comparison of human-human and machine-human evaluation metrics for Swedish scoring models.

Criterion and Range of Classes	Human-to-Human			Machine-to-Human					
	Kappa	Correlation	MAE	Kappa	Correlation	MAE	P, %	R, %	F1, %
Holistic (**2–5**)	0.496	0.490	0.613	0.524	0.524	0.461	56.13	47.33	49.76
Fluency (**1–3**)	0.498	0.490	0.425	0.560	0.574	0.305	63.06	59.41	60.53
Pronunciation (**2–3**)	0.162	0.162	0.419	0.276	0.290	0.343	66.97	67.53	66.85
Lex.-Gram. (**1–3**)	0.427	0.435	0.516	0.246	0.259	0.460	47.10	42.84	43.33
Task Achiev. (**1–3**)	0.376	0.371	0.621	0.582	0.636	0.366	59.31	58.90	58.92

Table 4. Comparison of human-human and machine-human evaluation metrics for Finnish scoring models.

Criterion and Range of Classes	Human-to-Human			Machine-to-Human					
	Kappa	Correlation	MAE	Kappa	Correlation	MAE	P, %	R, %	F1, %
Holistic (**2–7**)	0.732	0.751	0.782	0.807	0.803	0.612	46.85	39.95	39.05
Fluency (**2–4**)	0.393	0.392	0.575	0.507	0.522	0.359	63.95	55.67	57.62
Pronunciation (**2–4**)	0.513	0.531	0.445	0.583	0.612	0.269	66.51	54.62	55.18
Lex.-Gram. (**1–3**)	0.576	0.580	0.404	0.529	0.546	0.265	47.56	49.14	48.18
Task Achiev. (**1–3**)	0.340	0.298	0.410	0.323	0.289	0.359	49.61	45.12	46.53

used the cosine similarity metric to compare the embedding of the transcript to the centroids of each task and returned a binary output indicating whether the transcript belonged to the closest centroid. In the second step, the model compared a transcript to other responses of the same task and assigned it a score of its closest neighbor.

Many tasks in our dataset had very imbalanced score distributions. For example, if the task scores were put into three bins, one Swedish task would contain 92 responses in the highest score bin and only two responses in the lowest score bin. Choosing more than one neighbor would leave our system no chance of giving out the underrepresented score interval. For our procedure to be successful, the vector spaces needed to have the following properties. The task classification space should keep responses to the same tasks close to each other and far away from other responses. The space for content scoring of tasks should keep vectors of responses with similar score ranges close to each other and far away from responses in other score ranges. To get such vector spaces, monolingual BERT (Devlin et al., 2019) models were first fine-tuned in a Siamese manner (Reimers & Gurevych, 2019) to cluster responses from the same task together, provided with positive and negative examples of responses to the same task. For each response one positive pair and five negative pairs were formed. The negative pairs were chosen from responses to tasks other than the task of the response that was closest to the current response in the vector space.

The models were trained for five epochs using Contrastive Loss (Chopra et al., 2005) with a margin of 0.5 and cosine similarity as the similarity metric, and representations were obtained with mean-pooling. Then, the resulting models were fine-tuned to place responses with similar scores together. To achieve this, for a response in our dataset, one positive and five negative examples were sampled from the same task as this response. Positive examples were responses that received the scores from the same score bin. Negative examples were responses from other score ranges. The Swedish model was trained for five epochs, and the Finnish model was trained for two epochs using Contrastive Loss, a margin of 0.5, and cosine similarity as the similarity metric.

Tables 3 and 4 compare the results of human-human and machine-human evaluations of the five main scoring models for L2 Finland Swedish and L2 Finnish, respectively, using Weighted quadratic Kappa, Spearman's correlation, and mean absolute error (MAE) in addition to Precision (P), Recall (R), and F1 scores. It should be noted that the machine-human and human-human measurements were not exactly comparable, because human-human comparison was only possible on a smaller random subset that had more than one human score.

Table 5 shows the top-performing expert-designed features. While some measures were important for one language only, others proved to be beneficial for both languages. It should be noted that expert-designed features were combined with deep neural acoustic representations for predicting the holistic score. These embeddings were not intuitive from a pedagogical standpoint, but were useful in practice (Al-Ghezi et al., 2023; Bannò & Matassoni, 2023).

DISCUSSION

ASR performance

The ASR model for L2 Finland Swedish achieved 17.71%/9.08% WER/CER and for Finnish 21.89%/7.06%. For the ASR error analysis, the utterances from the test sets with the highest word and character error rates were analyzed. Table 6 shows some examples of ASR outputs with the corresponding reference transcriptions, as well as WERs and CERs.

As can be seen from L2 Swedish examples, some hesitations or complete words were missing from reference transcriptions, which resulted in relatively high error rates. For instance, the human transcription for example #1 missed the word "ledigheten" and its repetitions. In addition, sometimes the ASR model merged separate words (see example #4) possibly due to the lack of language modeling (LM) component. Another possible reason for that error might be background noise in the recordings or the speech rate of a speaker being too rapid.

Like L2 Swedish ASR, L2 Finnish ASR models often had high word and character error rates in sentences where words such as proper names were missing from the reference transcriptions. Also, single-character errors in words were quite common in L2 Finnish

Table 5. Most important (L)exical, (G)grammatical, (P)ronunciation, and (F)luency features for each language.

Feature	Type	Description	Swe	Fin
n_tokens	L	number of tokens in the ASR transcript	0.2513	0.1205
dep_dist_root	G	average of word distances to the root (Liu, 2008)	0.2533	0.1586
types	L	count of unique words in the response	0.2872	0.1842
rootTTR	L	lexical diversity root of token-type ratio	0.3088	0.2367
ovix	L	lexical diversity measure (Hultman, 1994)	0.2961	0.2508
AMscore	P	Acoustic Model score; sum of log-probabilities of the ASR output (Zechner & Evanini, 2019)	0.2381	0.3175
rPVI_vowels	P	difference in duration between immediately consecutive vowels (de Jong et al., 2021)	0.2631	0.3112
speech_rate	F	words per second in total response duration (Zechner & Evanini, 2019)	0.2202	0.3735
rPVI_cons	P	difference in duration between immediately consecutive consonants (de Jong et al., 2021)	0.0605	0.2435
articulation_rate	F	words per second in articulation time (Zechner & Evanini, 2019)	0.1687	0.2412
ASD	F	Average Syllable Duration; ratio of speaking rate to the number of syllables	0.1251	0.2962
voiced_fraction	P	ratio of voiced frames	0.1336	0.2161

Table 6. Example ASR outputs with corresponding human transcriptions, WERs and CERs.

Reference	ASR Output	WER, %	CER, %
L2 Swedish ASR			
när den här fri	när den här ledig le ledigheten fri	75.00	141.67
okej först gå gå till norragatan	okej först gå öö vad det mm gå till norragatan	66.67	37.04
aa kan kan jag har mot list	kan kan jag ha de mot list	42.86	19.05
ööm ha en bra födelsedag	öömha en bra födelsedag ningen	60.00	30.00
L2 Finnish ASR			
moi maria hauska tutustua minun nimeni on minulle kuuluu hyvää	moi marja hauska tutustuaa minun nimeni on belttama koko minulle kuuluu hyvää	35.29	20.93
öö vähä väsyttää vielä mutta entäs itse	öövähä väsyyttää tiällä mutta entäs itse		
se kuuluu tähän vuodenaikaan oletko levännyt tarpeeksi ja yrittänyt ottaa lääkkeitä	ja kuuluu tähän vuoden aikaan oletko hävennyt tarpeeksi ja yrittänyt ottaa lääkkeitä	36.36	6.85
öö tänään meillä on o olet ollut koulussa meillä oli ensin psykologiaa ja me aloitettiin yhden projektin öö sitten meillä oli ruokailu ja nyt meil on tämä suomen juttu	öö tänään millä on o olet ollut koulussa me olin sin sykologiaa ja me aloitettiin projekkin öa sitte meil oli ruokailla ja nyt meil on me suomen jutt	44.83	16.55

ASR outputs. For example, words "tutustua" and "väsyttää" were recognized as "tutustuaa" and "väsyyttää" in sentence #5, and words "psykologiaa" and "projektin" were recognized as "sykologiaa" and "projekkin" in utterance #7. In addition, the Finnish L2 ASR system recognized some words as completely different words. Even though no external language model was used in this work, these words were grammatically correct but had different meanings. For instance, the word "levännyt" ("have had a rest") is recognized as "hävennyt" ("have been ashamed") in sentence #6.

Scoring models

Tables 3 and 4 show that the machine-human agreement was higher than the human-human in nearly all analytical dimensions for both languages, except for lexico-grammatical (for both languages). One possible reason for the low performance of lexico-grammatical features was the absence of an external language model in the ASR, which not only led to minor, character-level spelling mistakes, but also led to inaccuracies in lexico-grammatical feature calculations. Lexical features such as TTR or types rely on existing words in external corpora, and any slight mismatch would affect the calculations. Similarly, features like depth root distance rely on dependency parsers which are trained on well-edited text. In addition, it should be noted that, unlike other scoring models, the lexico-grammatical ones did not use any deep embeddings, which suggests possibilities in conducting further experiments to improve the model performance by incorporating neural textual representations solely or combined with the acoustic ones.

Higher machine-human agreement than the human-human agreement was expected since the scoring models were trained with the human average scores using a range of fluency, pronunciation, lexico-grammatical, and task achievement features. Using average scores may reduce human-related assessment bias, and modeling the human rater behavior, such as analyzing rater severity/leniency, can further improve assessment reliability. Re-training our scoring models with fair average scores that are based on many-facet Rasch models would be a useful extension of the present study linacre (Linacre 1989).

Limitations of the system

The automatic speaking system has several limitations that could be further addressed to improve its accuracy, reliability, and generalizability. As discussed previously, the ASR did not use an external language model, which may have resulted in some errors in the ASR outputs and negatively impacted the calculations of lexicogrammatical features. Another limitation was the system's lack of robustness against very noisy or other low-quality recordings, which is a common issue for all automatic systems. Additionally, the scoring models were more accurate at distinguishing between intermediate levels, rather than extreme ones, which were not adequately represented in the collected data. Furthermore, the deep acoustic representations, while having the potential to complement interpretable human-designed features, may seem unusable from a pedagogical standpoint since they were not interpretable. Finally, due to the incorporation of multiple deep neural models, the system was computationally complex, which could affect real-time interaction with users. Therefore, future engineering endeavors are required to compress them or use techniques such as knowledge distillation to reduce the latency and improve computational efficiency.

CONCLUSIONS

This work focused on developing automatic speaking assessment systems for spontaneous L2 Finnish and L2 Finland Swedish. The steps involved in designing the assessment tasks, collecting and processing the training data, training and evaluating the ASR systems, as well as the scoring models for pronunciation, fluency, lexico-grammatical competence, and task completion, were discussed.

Self-supervised deep acoustic models, including Wav2vec2, were utilized to develop ASR systems with a relatively small amount of training data, which is particularly useful in the context of low-resourced languages. The scoring models for analytic and holistic dimensions exhibited a high degree of human-machine agreement for the targeted skill levels, indicating their potential for automated speaking assessment. In addition, the high performing expert-designed features were identified, and an additional type of feature, namely deep acoustic embeddings was integrated.

Our work contributes to the development of automated speaking assessment systems, especially for low-resourced languages, which could provide benefits to language learners and educators. Future research could explore providing diagnostic feedback from automated speaking assessment systems to learners for the purposes of formative assessment.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Academy of Finland [322619, 322625, 322965].

ORCID

Ragheb Al-Ghezi http://orcid.org/0000-0003-0141-5753
Yaroslav Getman http://orcid.org/0000-0003-4680-8294
Anna Von Zansen http://orcid.org/0000-0002-6444-7667
Heini Kallio http://orcid.org/0000-0002-1263-9624
Mikko Kurimo http://orcid.org/0000-0001-5278-7974
Ari Huhta http://orcid.org/0000-0003-3124-449X
Raili Hildén http://orcid.org/0000-0002-5114-5600

References

Al-Ghezi, R., Getman, Y., Rouhe, A., Hildén, R., & Kurimo, M. (2021). Self-supervised end-to-end ASR for low resource L2 Swedish. *Proceedings Interspeech*, *2021*, 1429–1433. https://doi.org/10.21437/Interspeech.2021-1710

Al-Ghezi, R., Getman, Y., Voskoboinik, E., Singh, M., & Kurimo, M. (2023). Automatic rating of spontaneous speech for low-resource languages. *2022 IEEE Spoken Language Technology Workshop (SLT)*, 339–345. https://doi.org/10.1109/SLT54892.2023.10022381

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common voice: A massively-multilingual speech corpus. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4218–4222. /https://aclanthology.org/2020.lrec-1.520 .

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 12449–12460). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, *47*(4), 707–728. https://doi.org/10.1111/flan.12110

Bannò, S., & Matassoni, M. (2023). Proficiency assessment of L2 spoken English using wav2vec 2.0. *2022 IEEE Spoken Language Technology Workshop (SLT)*, 1088–1095. https://doi.org/10.1109/SLT54892.2023.10023019

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. *First International Conference on Spoken Language Processing (ICSLP 1990)*, 1185–1188. https://doi.org/10.21437/ICSLP.1990-313

Bernstein, J., Moere, A. V., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–377. https://doi.org/10.1177/0265532210364404

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2011). *Building a validity argument for the test of English as a foreign language*. Routledge. https://doi.org/10.4324/9780203937891

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *1*, 539–546 vol. 1. https://doi.org/10.1109/CVPR.2005.202

Council of Europe. (2001) . *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Cox, T., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal*, *29*(4), 601–618. https://doi.org/10.11139/cj.29.4.601-618

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862–2873. https://doi.org/10.1121/1.1471894

De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237–254. https://doi.org/10.1080/15434303.2018.1477780

de Jong, N. H., Pacilly, J., & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education Principles, Policy & Practice*, *28*(4), 456–476. https://doi.org/10.1080/0969594X.2021.1951162

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, Minneapolis, Minnesota (pp. 4171–4186).

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Lawrence Erlbaum.

Educational Testing Service. (n.d.) *TOEFL iBT speaking Section Scoring Guide*. https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf

Evanini, K., & Zechner, K. (2020). Overview of automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 3–20). Routledge. https://doi.org/10.4324/9781315165103-1

Finnish Matriculation Examination Board. (n.d.). *The Finnish Matriculation Examination*. https://www.ylioppilastutkinto.fi/en/

Finnish National Agency for Education. (2003). *Lukion Opetussuunnitelman Perusteet 2003 [National Core Curriculum for General Upper Secondary Schools 2003]*. https://www.oph.fi/sites/default/files/documents/47345lukionopetussuunnitelmanperusteet2003.pdf

Finnish National Agency for Education. (2015). *Lukion Opetussuunnitelman Perusteet 2015 [National Core Curriculum for General Upper Secondary Schools 2015]*. https://www.oph.fi/sites/default/files/documents/172124lukionopetussuunnitelmanperusteet2015.pdf

Finnish National Agency for Education. (2019). *Lukion Opetussuunnitelman Perusteet 2019 [National Core Curriculum for General Upper Secondary Schools 2019]*. https://www.oph.fi/sites/default/files/documents/lukionopetussuunnitelmanperusteet2019.pdf

Gretter, R., Matassoni, M., Allgaier, K., Tchistiakova, S., & Falavigna, D. (2019). Automatic assessment of spoken language proficiency of non-native children. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7435–7439. https://doi.org/10.1109/ICASSP.2019.8683268

Gu, L., & Davis, L. (2020). Providing SpeechRater feature performance as feedback on spoken responses. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 159–175). https://doi.org/10.4324/9781315165103-10

Hildén, R., & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. In A. Koskensalo, J. Smeds, P. Kaikkonen, & V. Kohonen (Eds.), *Foreign languages and multicultural perspectives in the European context: Fremdsprachen und multikulturelle perspektiven im europäischen kontext* (pp. 291–300). LIT Verlag.

Hsieh, C.-N., Zechner, K., & Xi, X. (2020). Features measuring fluency and pronunciation. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment* (pp. 101–122). Routledge.

Hultman, T. G. (1994). Hur gick det med OVIX? I: Språkbruk, grammatik och språkförändring. *En festskrift till Ulf Teleman*, *13*(1), 55–64.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Kallio, H., Kautonen, M., & Kuronen, M. (2023). Prosody and fluency of Finland Swedish as a second language: Investigating global parameters for automated speaking assessment. *Speech Communication*, *148*, 66–80. https://doi.org/10.1016/j.specom.2023.02.003

Kallio, H., Koivusalo, L., & Kuronen, M. (2022). The role of pause location in perceived fluency and proficiency in L2 Finnish. *ISAPh 2022: Proceedings of the 11th International Conference on Speech Prosody*, 777–781. https://doi.org/10.21437/SpeechProsody.2022-158

Kallio, H., Kuronen, M., & Kautonen, M. (2021). Differences in acoustically determined sentence stress between native and L2 speakers of Finland Swedish. *Working Papers-Lund University, Department of Linguistics, General Linguistics, Phonetics*, *56*, 42–47. https://journals.lub.lu.se/LWPL/issue/view/3250/794

Kallio, H., Šimko, J., Huhta, A., Karhila, R., Vainio, M., Lindroos, E., Hildén, R., & Kurimo, M. (2017). Towards the phonetic basis of spoken second language assessment: Temporal features as indicators of perceived proficiency level. *AFinLA-e Soveltavan Kielitieteen Tutkimuksia*, *10*(10), 193–213. https://doi.org/10.30660/afinla.73137

Kallio, H., Suni, A., & Šimko, J. (2022). Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of English with different language backgrounds. *Language and Speech*, *65*(3), 571–597. https://doi.org/10.1177/00238309211040175

Kallio, H., Suni, A., Šimko, J., & Vainio, M. (2020). Analyzing second language proficiency using wavelet-based prominence estimates. *Journal of Phonetics*, *80*, 1–12. https://doi.org/10.1016/j.wocn.2020.100966

Kang, O., & Johnson, D. (2018). The roles of suprasegmental features in predicting English oral proficiency with an automated system. *Language Assessment Quarterly*, *15*(2), 150–168. https://doi.org/10.1080/15434303.2018.1451531

Karhila, R., Rouhe, A., Smit, P., Mansikkaniemi, A., Kallio, H., Lindroos, E., Hildén, R., Vainio, M., & Kurimo, M. (2016). Digitala: An augmented test and review process prototype for high-stakes spoken foreign language examination. *Interspeech*, 784–785. https://doi.org/10.21437/Interspeech.2016

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2020). *Facets Computer Program for Many-Facet Rasch Measurement*. (Version 3.83.2). https://www.winsteps.com/facets.htm

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191. https://doi.org/10.17791/jcs.2008.9.2.159

Loukina, A., Davis, L., & Xi, X. (2017). Automated assessment of pronunciation in spontaneous speech. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 153–171). Routledge.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Moisio, A., Porjazovski, D., Rouhe, A., Getman, Y., Virkkunen, A., AlGhezi, R., Lennes, M., Grósz, T., Lindén, K., & Kurimo, M. (2023). Lahjoita puhetta: A large-scale corpus of spoken Finnish with some benchmarks. *Language Resources and Evaluation*, *57*(3), 1295–1327. https://doi.org/10.1007/s10579-022-09606-3

Östling, R., Smolentzov, A., Hinnerich, B. T., & Höglin, E. (2013). Automated essay scoring for Swedish. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia (pp. 42–47).

Pearson. (2017). *PTE Academic Score Guide*. https://pearson.com.cn/file/PTEA_Score_Guide.pdf

Pearson. (2018). *Versant™ Arabic Test. Test Description and Validation Summary*. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Versant/ValidationSummary/Versant-Arabic-Test -Description-Validation-Summary.pdf

Pearson. (2020). *Versant™ English test. Test description and validation summary*. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Versant/ValidationSummary/Versant-English-Test -Description-Validation-Report.pdf

Pearson. (2023). *Versant Automated Oral Proficiency Tests*. https://www.pearsonhighered.com/versant/

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, *33*(1), 53–73. https://doi.org/10.1177/0265532215579530

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Tavakoli, P., & Skehan, P. (2005). *Strategic planning, task structure and performance testing* (R. Ellis, Ed.). Benjamins. https://doi.org/10.1075/lllt.11

Vaarala, H., Riuttanen, S., Kyckling, E., & Karppinen, S. (2021). Language reserve. Now! Follow-up on Pyykkö's report Multilingualism into a strength (2017): Summary in English. University of Jyväskylä. https://www.kieliverkosto.fi/fi/toiminta/kieliverkostossa-tapahtuu/language-reserve-now-english-summary-of-report-now-available

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29* (3), 325–344. https://doi.org/10.1177/0265532211424478

von Zansen, A. (2022a, April). *DigiTala's rating criteria: Holistic and analytic scales for assessing L2 speaking*. Zenodo. https://doi.org/10.5281/zenodo.6477089

von Zansen, A. (2022b, June). *DigiTala's speaking tasks and questionnaire for L2 Finnish learners (proficiency level A)*. Zenodo. https://doi.org/10.5281/zenodo.6976044

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

von Zansen, A. (2022d, May). *DigiTala's speaking tasks for L2 Finnish learners (proficiency level B2)*. Zenodo. https://doi.org/10.5281/zenodo.6562865

von Zansen, A., & Heijala, M. (2023). Miten suomen ja ruotsin opettajat käyttäisivät puheen automaattiseen arviointiin kehitettyä työkalua? [How would Finnish and Swedish teachers use a tool developed for automated L2 speaking assessment? *AFinLA-teema*, (15). https://doi.org/10.30660/afinla.124822

von Zansen, A., & Hilden, R. (2022). It was cool and comfortable!" Akateemisten alkeistason S2-opiskelijoiden kokemuksia tietokoneella suoritettavasta puhumisen kokeesta [Experiences of academic beginning learners of a computerised speaking test]. In S. Routarinne, P. Heinonen, T. Kärki, A. Roiha, M.-L. Rönkkö, & A. Korkeaniemi (Eds.), *Ainedidaktiikka ajassa: Laajenevat oppimisympäristöt ja eri-ikäiset oppijat [Subject didactics today: Expanding learning environments and learners of different age]* (pp. 72–90). University of Turku. http://hdl.handle.net/10138/353562

von Zansen, A., & Huhta, A. (2022). Developing automated feedback on spoken performance: Exploring the functioning of five analytic rating scales using many-facet rasch measurement. *Digital Research Data and Human Sciences*, 211–229. http://urn.fi/URN:ISBN:978-951-39-9450-1

von Zansen, A., Kallio, H., Sneck, M., Kuronen, M., Huhta, A., & Hildén, R. (2022a). Ihmisarvioijien näkemyksiä suullisen kielitaidon automaattisesta arvioinnista, digitaalisesta arviointiprosessista sekä puhesuorituksista arvioitavista ulottuvuuksista [Human raters' perceptions of the automated assessment of oral language skills, the digital assessment process and the dimensions to be assessed from speaking performances]. *AFinLa Year Book*, 370–394. https://doi.org/10.30661/afinlavk.114821

von Zansen, A., Sneck, M. M., & Hilden, R. (2022b). Lukiolaisten käsitykset ja heidän antamansa palaute suullisen kielitaidon automaattisesta arvioinnista: Ainedidaktinen symposium 2021 [Language learners' perceptions and their feedback on automated assessment of oral language skills]. In R. Kantelinen, M. Kautonen, & Z. Elgundi (Eds.), *Linguapeda 2021* (Vol. 21, pp. 176–205). University of Eastern Finland. http://hdl.handle.net/10138/352128

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, *1*(1), 993–1003. https://doi.org/10.18653/v1/2021.acl-long.80

Winke, P., & Brunfaut, T. (2020). *The Routledge handbook of second language acquisition and language testing*. Routledge.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291–300. https://doi.org/10.1177/0265532210364643

Xi, X. (2021). Validity and the automated scoring of performance tests. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 513–529). Routledge. https://doi.org/10.4324/9781003220756-40

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). Automated scoring of spontaneous speech using SpeechRaterSM v1. 0. *ETS Research Report Series*, *2008*(2), i–102. https://doi.org/10.1002/j.2333-8504.2008.tb02148.x

Xu, J., Brenchley, M., Jones, E., Pinnington, A., Benjamin, T., Knill, K., Seal-Coon, G., Robinson, M., & Geranpayeh, A. (2020). *Linguaskill: Building a validity argument for the speaking test*. Cambridge Assessment English. https://www.cambridgeenglish.org/Images/589637-linguaskill-building-a-validity-argument-for-the-speaking-test.pdf

Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in Education Principles, Policy & Practice*, *28*(4), 411–436. https://doi.org/10.1080/0969594X.2021.1979467

Zechner, K., & Evanini, K. (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

Appendix A Task descriptions

Table A1. L2 Finland Swedish freeform tasks.

Task 4: Mitä sanot ruotsiksi seuraavissa tilanteissa? Reagoi, kun näet tilanteen kuvauksen. Sinulla on 10 sekuntia aikaa sanoa vastauksesi.	
4.1	kysyt milloin asunto vapautuu
4.2	kehotat kaveriasi tarkistamaan viimeisin instagram-päivityksesi
4.3	sanot odottavasi innokkaasti seuraavaa tapaamista
4.4	kiität opettajaa hyvästä tunnista
4.5	kiität kivoista juhlista viime kerralla
4.6	tiedustelet näytöksen alkamisaikaa
4.7	pyydät ruotsalaista tuttavaasi toistamaan sanomansa
4.8	toivotat onnea syntymäpäiväsankarille
4.9	pahoittelet ettet voi tulla, koska olet sairastunut
4.10	pyydät ruokalistaa ravintolassa
Task 5: Nyt näet kuvan kustakin tilanteesta. Sen jälkeen sinulla on 30 sekuntia aikaa puhua.	
5.1	kuvailet ja kommentoit retkipäivän säätä
5.2	suostuttelet ystävääsi mukaan konserttiin
5.3	neuvot tien apteekkiin
5.4	soitat poliisille ja kerrot havaitsemastasi vaarasta
5.5	kuvailet vartijalle näkemäsi henkilön
5.6	kuvailet tilaamaasi ruoka-annosta
Task 6: Keskustelet Skypen välityksellä ruotsinkielisen ystävyyskoulun opiskelijan kanssa. Vastaa hänen kysymyksiinsä. Sinulla on 10 sekuntia vastausaikaa kunkin kysymyksen jälkeen.	
6.1	Vad kul att träffa dig! Hur har du det i dag?
6.2	Vad har du gjort i dag?
6.3	Jag körde just hem från skolan. Har du körkort?
6.4	(Mjau!) Oj, min katt vill också prata med dig. Har du ett keldjur?
Task 7: Haluat vielä tietää muutaman asian. Reagoi, kun näet aiheen. Sinulla on 20 sekuntia aikaa esittää kysymys. (Aiheet näkyvät kuvaruudussa yksi kerrallaan 25 sekunnin ajan)	
7.1	Suunnitelmat lukion jälkeen
7.2	Toisen kotipaikkakunta ja siellä viihtyminen
7.3	Mielipide Ruotsin kuninkaallisista
Tämä on parikeskustelu. Asettukaa webkameran eteen niin, että molemmat näkyvät. Sinä ja kaverisi haluatte lukion päätyttyä hakea samaan kesätyöpaikkaan. Olette löytäneet kaksi kiinnostavaa ilmoitusta. Keskustelkaa vaihtoehdoista ja päättäkää, haetteko kalanperkaajiksi Norjaan vai puistotyöntekijöiksi Ruotsiin? Aikaa on 7 minuuttia.	
9.1	Paritehtävä: kesätyö

Table A2. L2 Finnish freeform tasks.

Task 2: Tärkeä paikka. Teette puheharjoituksia suomen kurssilla. Tänään aiheena ovat tärkeät paikat. Kerro sinulle tärkeästä paikasta alla olevien kysymysten avulla. Sinun ei tarvitse vastata kaikkiin kysymyksiin. Valmistaudu lukemalla kysymykset, mieti mistä paikasta haluat puhua, ja paina sitten vasta Start recording-painiketta. Yritä pitää yllä puhetta noin 1 minuutin ajan. HUOM! Älä kerro nimiä, tarkkaa osoitetta tai henkilökohtaisia asioita. Miksi paikka on sinulle tärkeä? Mikä on tässä paikassa parasta? Miksi? Mitä teet tässä paikassa? Millainen tämä paikka on? Kuinka kauan tämä paikka on ollut sinulle tärkeä?

Task 3: Webinaari. Järjestät webinaarin ystävyyskoulun opiskelijoille ja opettajille. Käytte alkuun läpi osallistujien kuulumisia, joita he kirjoittavat chattiin. 3a-b,e-f) Reagoi kunkin osallistujan kommenttiin suullisesti (á 15 sek). 3c-d) Esitä osallistujille kysymys annetusta aiheesta. (á 15 sek)

3a Äänitä alla olevaan kommenttiin sopiva vastaus (15 sek). ANTON: Ihan hyvää, paitsi heräsin juuri – nukuin melkein pommiin!

3b Äänitä alla olevaan kommenttiin sopiva vastaus (15 sek). MILLA: Olen ollut flunssassa koko viikon. . .

3c Esitä webinaarin osallistujille kysymys (15 sek) seuraavasta aiheesta: suunnitelmat lukion jälkeen

3d Esitä webinaarin osallistujille kysymys (15 sek) seuraavasta aiheesta: kotipaikkakunta ja siellä viihtyminen

3e Äänitä sopiva vastaus (15 sek) alla olevaan kommenttiin. ELINA sanoo: Mulla on tänään syntymäpäivä! Miten vastaat?

3f Äänitä sopiva vastaus (15 sek) alla olevaan kommenttiin. JOONA sanoo: Millon alotetaan, ehdinkö hakea kahvia? Miten vastaat?

Task 4: Jokaisella oppilaalla on kummioppilas ystävyyskoulussa. Teillä on puhelu ja vastailet hänen kysymyksiinsä (15 sek), jotka kuulet. HUOM! Älä kerro nimiä tai henkilökohtaisia asioita.

4a Kuuntele äänite ja vastaa kuulemaasi kysymykseen (15sek). HUOM! Älä kerro nimiä tai henkilökohtaisia asioita. Kysymys: [Moikka, mä olen Maria Erikäinen. Kiva tavata! Mitä sulle kuuluu tänään?]

4b Kuuntele äänite ja vastaa kuulemaasi kysymykseen (15sek). HUOM! Älä kerro nimiä tai henkilökohtaisia asioita. Kysymys: [Millainen sää siellä on?]

4c Kuuntele äänite ja vastaa kuulemaasi kysymykseen (15sek). HUOM! Älä kerro nimiä tai henkilökohtaisia asioita. Kysymys: [Mitä olet tehnyt tänään?]

4d Kuuntele äänite ja vastaa kuulemaasi kysymykseen (15sek). HUOM! Älä kerro nimiä tai henkilökohtaisia asioita. Kysymys: [Millaisia suunnitelmia sinulla on viikonlopulle?]

4e Kuuntele äänite ja vastaa kuulemaasi kysymykseen (15sek). HUOM! Älä kerro nimiä tai henkilökohtaisia asioita. Kysymys: [Moikka, mä oon Maria Erikäinen. Mitä sulle kuuluu tänään?]

Task 5: Jatko-opiskelupaikka. Näet seuraavaksi kuvan korkeakoulusta, johon olet aikeissa hakea lukion jälkeen. Kerro, mitä näet kuvassa (max 1 min). Voit kuvailla rakennusta/tilaa/huonekaluja/ihmisiä tiloissa. Voit kertoa väreistä ja muodoista, kuvakulmasta, valaistuksesta. Voit suunnitella vastaustasi hetken ennen äänittämistä, jotta sinulla riittää sanottavaa. Yritä pitää yllä puhetta 1 minuutin ajan.

Task 6: Instagram. Teillä on ollut koulujen välinen kuvakilpailu Instagramissa. Voit suunnitella vastaustasi hetken ennen äänittämistä, jotta sinulla riittää sanottavaa. Yritä pitää yllä puhetta ainakin 1 minuutin ajan (max 1,5 min): Mitä pidät kuvista 1–3? Millaisia ajatuksia kuvista 1–3 herää? Mikä on sinun suosikkisi? Miksi? HUOM! Älä kerro nimiä tai henkilökohtaisia asioita.

6a [kuva 1:] Huomenta! #aamuaurinko #kaupunkipyöräily (Kuva: Lauri Manninen) [kuva 2:] Kesä on täällä! #kukkaniitty #aurinko #kesä (Kuva: 123rf) [kuva 3:] Päivän tärkein ateria #aamiainen (Kuva: 123rf)

6b [kuva 1:] Jäiden lähtö #kevät #luonto (Kuva: Mostphotos) [kuva 2:] Uusi perheenjäsen #koiranpentu (Kuva:123rf) [kuva 3:] Sataa sataa ropisee #syksy (Kuva 123rf)

Task 7: Uutinen. Katsele lyhyt uutinen ja ota sen pohjalta kantaa. Yritä pitää puhetta yllä ainakin minuutin ajan (max 1,5 min). Voit suunnitella vastaustasi hetken, jotta sinulla riittää sanottavaa. Apukysymyksiä – kaikkiin ei tarvitse vastata: Onko geenimuunneltu ruoka sinulle aiheena tuttu? Söisitkö itse geenimuunneltua ruokaa? Miksi/miksi et? Mitä hyviä puolia geenimuunnellussa ruoassa on? Mitä huonoja puolia geenimuunnellussa ruassa on? Mitä haluaisit tietää geennimuunnellusta ruoasta? Huom! Älä kerro oikeita nimiä tai henkilökohtaisia asioita Geenimuunneltua lohta, olkaa hyvä! Ylen kuvausryhmä kävi lohen kehittäneen bioteknologiayhtiö.

Task 8: Puhelimessa. Alla on neljä tilannetta, jotka liittyvät ystävyyskoulun vierailuun. Tutustu tilanteeseen, kuuntele äänite ja esitä asiasi kohteliaasti (max 30sek).

8a Tilanne: Olette lähdössä ystävyyskoulun kanssa retkelle. Bussi on varattu väärälle päivälle! Oikea ajankohta on 31.5., lähtö klo 9 ja paluu klo 16. Soitat koulusihteerille ja kuulet vastaajaviestin: [Koulusihteeri Kaisa tässä hei! En juuri nyt pääse vastaamaan, mutta jätäthän viestin ja yhteystietosi niin otan sinuun yhteyttä. Kiitos, kuulemiin!] Jätä kohtelias viesti (max 30sek) vastaajaan: Esittele itsesi. Kerro ongelma. Pyydä häntä korjaamaan asia. HUOM! Älä kerro oikeaa nimeäsi tai henkilökohtaisia asioita. Oman nimesi sijaan voit käyttää nimeä Maija/Matti Meikäläinen

8b Tilanne: Vastaat puhelimeen. Soittaja sanoo: [No hei, se on paikallislehden toimittaja täällä. Teen juttua lukiolaisten ajankäytöstä. Voisinko haastatella sinua 5–10 minuutin ajan?] Vastaa soittajalle (max 30sek): Et juuri nyt voi puhua. Kieltäydy kohteliaasti. Ehdota toista ajankohtaa.

(*Continued*)

Table A2. (Continued).

8c Tilanne: Suunnittelitte eilen illalla ystävyyskoulun vierailua kahvilassa. Kotona huomaat hupparisi kadonneen. Soitat kahvilaan (max 30sek): Esittele itsesi. Kerro asiasi kohteliaasti. Kuvaile hupparia, se näyttää tältä: [kuva henkilöstä ko. huppari päällä] HUOM! Älä kerro oikeaa nimeäsi tai henkilökohtaisia asioita. Oman nimesi sijaan voit käyttää nimeä Maija/Matti Meikäläinen

8d Tilanne: Kaverisi onkin ottanut hupparisi talteen ja tuonut sen kotiisi. Jätä ääniviesti (max 30sek) kaverillesi: Kerro, miltä sinusta tuntuu. Kiitä kaveriasi. Kysy kaveriasi ulos huomenna.

Task 10: Minun päiväni. Kerro päivästäsi. Älä kerro nimiä tai osoitteita. Paina Start recording -painiketta ja yritä puhua minuutin ajan. Apukysymyksiä: (Tell about your day. Do not use real names or addresses. Click on "Start recording" and try to speak for 1 minute. Supporting questions:) Mihin aikaan sä heräät? Mitä sä syöt ja juot aamulla? Mitä sä teet päivällä? Kenen kanssa sä olet? Missä sä olet illalla? Mitä sä teet illalla? Mihin aikaan sä meet nukkumaan?

Task 11: Mitä sanot? Tapaat uuden kurssikaverin kahvilassa. Mitä voit sanoa eri tilanteissa? Vastaa kokonaisella lauseella. (You meet your new classmate in a cafe. What can you say in the following situations? Reply using full sentences.)

11a Paina Start recording -painiketta ja tilaa kahvi (Click on "Start recording" and order a coffee)

11b Paina Start recording -painiketta ja kysy kahvin hintaa (Click on "Start recording" and ask how much the coffee costs)

11c Paina Start recording -painiketta ja vastaa, kun kaveri soittaa ja kysyy, missä olet (Click on "Start recording" and answer to your friend who calls and asks where you are.)

11d Paina Start recording -painiketta ja kysy vähintään kaksi kysymystä kaverilta. Haluat tutustua kaveriin. (Click on "Start recording" and ask your friend at least two questions. You want to get to know him/her.)

Task 12: Kurssikaveriin tutustuminen. Saatte läksyksi tutustua kurssikaveriin. Puhutte puhelimessa ja tutustutte enemmän. Kuuntele ja vastaa kaverin kysymyksiin. Älä kerro oikeita osoitteita tai nimiä. (Your homework is to interview a classmate. You talk on the phone and get to know each other. Listen to your classmate and answer the questions. Do not give real addresses or names.)

12a Kuuntele ja nauhoita vastaus. (Listen to the question and record your answer.) "Moi, Anna tässä. Mitä sulle kuuluu?"

12b Kuuntele ja nauhoita vastaus. (Listen to the question and record your answer.) "Nii missä sä asut?"

12c Kuuntele ja nauhoita vastaus. (Listen to the question and record your answer.) "Millainen sun asunto on?"

12d Kuuntele ja nauhoita vastaus. (Listen to the question and record your answer.) "Onks sulla sauna? Tykkäätkö saunoa?"

Task 13: Kerro kuvasta. Kerro, mitä kuvassa on. Yritä puhua minuutin ajan. Suunnittele hetki, mitä aiot sanoa. Paina sitten Start recording -painiketta ja nauhoita vastaus. (Tell about the picture. What do you see? Try to speak for 1 minute. Plan for a moment what you are going to say. Then click on "Start recording".) Apukysymyksiä (Supporting questions): Millainen perhe on? (esimerkiksi kuka on kuka, nimi, ikä) Mitä he tekevät? Mitä he syövät? Mitä he juovat? Missä he ovat? Missä he asuvat?

Tasks 2–8 are part of B1 or B2 test for high school students. Tasks 10–13 are part of the A-level test for university students. The A-level tasks have been translated to English for students.

Appendix B Rating rubrics

Table B1. Rating rubrics used by human raters.

Holistic
0. Pre-A1, 1.A1, 2.A2, 3.B1, 4.B2, 5.C1, 6.C2, 7.cannot judge
Completion of the task (does the speaker answer the exam question)
(0) Cannot be evaluated./I can not say.
(1) Only partially responds to the assignment, there are many significant shortcomings in the response.
(2) Responds well to the assignment, but there are some significant shortcomings in the response.
(3) Excellent response to the assignment, there are no significant deficiencies in the response.
Fluency (fluency and ease of speech)
(0) Cannot be evaluated./I can not say.
(1) Irregular; lots of disturbing breaks, repetitions, breaks and hesitations.
(2) Moderately smooth; some disturbing breaks, repetitions, breaks, and hesitations.
(3) Smooth and effortless; no disturbing breaks, repetitions, breaks or hesitation.
(4) Really smooth and effortless; no disturbing breaks, repetitions, breaks or hesitation.
Pronunciation (control of sounds and prosodic features and comprehensibility of pronunciation)
(0) Cannot be evaluated./I can not say.
(1) Weak, difficult to understand, a lot of pronunciation problems.
(2) Moderate, fairly easy to understand, but with some pronunciation problems.
(3) Good, understandable, no major pronunciation problems.
(4) Really good, clear and natural pronunciation.
Extent of expression (how extensive vocabulary, structures and expressions the speaker uses)
(0) Cannot be evaluated./I can not say.
(1) Concise (eg single words, schematic expressions)
(2) Adequate (basic vocabulary, eg sentences)
(3) Extensive (diverse vocabulary and expression)
Vocabulary and grammar accuracy (effect of vocabulary and grammar errors on comprehensibility)
(0) Cannot be evaluated./I can not say.
(1) Much vocabulary and grammatical errors that impair comprehensibility.
(2) Some vocabulary and grammatical errors that impair comprehensibility.
(3) There are few vocabulary and grammatical errors that impair comprehensibility.
(4) No disturbing vocabulary or grammar errors, or the speaker corrects the errors himself.