Phonetic Fluency in Finnish as a Second Language:

Acoustic Analysis of High School Students' Spontaneous Speech

Liisa Koivusalo
Master's Thesis
Phonetics
Faculty of Arts
University of Helsinki
May 2022

Abstract

Faculty: Faculty of Arts

Degree programme: Master's Programme in Linguistic Diversity and Digital Humanities

Study track: Phonetics

Author: Liisa Koivusalo

Thesis title: Phonetic Fluency in Finnish as a Second Language: Acoustic Analysis of High School Students' Spontaneous Speech

Level: Master's Thesis

Month and year: May 2022

Number of pages: 35 + 3

Keywords: fluency, proficiency, L2 speech, Finnish as L2, spontaneous speech

Supervisors: Heini Kallio, Minnaleena Toivola

Where deposited: Helsinki University Library – Helda/E-thesis

Abstract: Speaking fluently is an important goal for second language (L2) learners. In L2 research, fluency is often studied by measuring temporal features in speech. These features include speed (rate of speech), breakdown (use of silent and filled pauses), and repair (self-corrections and repetitions) phenomena. Fluent speakers generally have a higher rate of speech and fewer hesitations and interruptions than beginner language learners. In this thesis, phonetic fluency of high school students' L2 Finnish speech is studied in relation to human ratings of fluency and overall proficiency. The topic is essential for the development of automated assessment of L2 speech, as phonetic fluency measures can be used for predicting a speaker's fluency and proficiency level automatically. Although the effect of different fluency measures on perceived fluency level has been widely studied during the last decades, research on phonetic fluency in Finnish as L2 is still limited. Phonetic fluency in high school students' speech in L2 Finnish has not been studied before.

The speech samples and ratings used in this thesis are a part of a larger dataset collected in the DigiTala research project. The analyzed data contained spontaneous speech samples in L2 Finnish from 53 high school students of different language backgrounds. All samples were assessed by expert raters for fluency and overall proficiency. The speech samples were annotated by marking intervals containing silent pauses, filled pauses, corrections and repetitions, and individual words. Several phonetic fluency measures were calculated for each sample from the durations of the annotated intervals.

The contribution of phonetic fluency measures to human ratings of fluency and proficiency was studied using simple and multiple linear regression models. Speech rate was found to be the strongest predictor for both fluency and proficiency ratings in simple linear regression. Articulation rate, portion of long silent pauses, mean duration of long silent pauses, mean duration of breaks between utterances, and rate of short silent pauses per minute were also statistically significant predictors of both fluency and proficiency ratings. Multiple linear regression models improved the simple models for both fluency and proficiency: for fluency, a model with a combination of articulation rate and the portion of long silent pauses performed the best, and for proficiency, a model with a combination of speech rate and mean duration of short silent pauses.

Perceived fluency level is often affected by a combination of different phonetic fluency measures, and it seems that human raters ground their assessments on this combination, although some phonetic fluency measures might be more important on their own than others. The findings of this thesis expand previous knowledge on phonetic fluency in L2 Finnish and can benefit both language learners and teachers, as well as developers of automatic assessment of L2 speech.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiivistelmä

Tiedekunta: Humanistinen tiedekunta

Koulutusohjelma: Kielellisen diversiteetin ja digitaalisten ihmistieteiden maisteriohjelma

Opintosuunta: Fonetiikka

Tekijä: Liisa Koivusalo

Työn nimi: Foneettinen sujuvuus suomessa toisena kielenä: Lukiolaisten spontaanin puheen akustinen analyysi

Työn laji: Maisterintutkielma

Kuukausi ja vuosi: Toukokuu 2022

Sivumäärä: 35 + 3

Avainsanat: sujuvuus, taitotaso, L2-puhe, suomi toisena kielenä, spontaani puhe

Ohjaajat: Heini Kallio, Minnaleena Toivola

Säilytyspaikka: Helsingin yliopiston kirjasto – Helda/E-thesis

Tiivistelmä: Sujuvaa puhetaitoa pidetään tärkeänä tavoitteena toisen kielen (L2) oppimisessa. L2-puheen tutkimuksissa sujuvuutta tutkitaan usein puheesta mitattavilla temporaalisilla piirteillä, joita ovat esimerkiksi puheen nopeus, tauot, korjaukset ja toistot. Nopea, vähän epäröintiä ja keskeytyksiä sisältävä puhe mielletään usein sujuvaksi, ja toisen kielen oppimisen alkuvaiheessa puhe on epäsujuvampaa. Tässä tutkielmassa tutkitaan lukiolaisten L2-suomen foneettista sujuvuutta puheesta mitattavien foneettisten sujuvuuspiirteiden sekä sujuvuus- ja taitotasoarvioiden avulla. Tutkimusaihe liittyy myös puheen automaattisen arvioinnin kehittämiseen, sillä kielenoppijan sujuvuus- ja taitotasoa voidaan ennustaa automaattisesti foneettisten sujuvuuspiirteiden avulla. Vaikka sujuvuuspiirteiden ja arviointien välistä yhteyttä on tutkittu melko paljon viime vuosikymmeninä, L2-suomen foneettiseen sujuvuuteen liittyviä tutkimuksia on yhä vähän. Lukiolaisten L2-suomen foneettista sujuvuutta ei ole aiemmin tutkittu.

Tutkielmassa käytetty puhe- ja arviointiaineisto on osa suurempaa aineistoa, joka on kerätty DigiTala-tutkimusprojektissa. Analysoitu aineisto sisälsi 53 spontaania puhenäytettä lukiolaisilta, jotka puhuvat suomea toisena kielenä. Lisäksi jokaisen puhenäytteen sujuvuus ja yleinen taitotaso oli arvioitu. Puhenäytteisiin annotoitiin hiljaiset ja täytetyt tauot, korjaukset ja toistot sekä yksittäiset sanat. Annotoitujen intervallien kestoista laskettiin useita foneettisia sujuvuuspiirteitä jokaiselle puhenäytteelle.

Foneettisten sujuvuuspiirteiden vaikutusta ihmisarvioihin tutkittiin lineaaristen regressiomallien avulla. Puhenopeus ennusti yhden selittävän muuttujan malleissa sekä sujuvuus- että taitotasoarvioita parhaiten. Tämän lisäksi artikulaationopeus, pitkien hiljaisten taukojen osuus, pitkien hiljaisten taukojen keskimääräinen kesto, yhtenäisten puhejaksojen välisten keskeytysten keskimääräinen kesto ja lyhyiden hiljaisten taukojen suhteellinen lukumäärä olivat tilastollisesti merkitseviä ennustajia yhden selittävän muuttujan malleissa. Useamman selittävän muuttujan mallit paransivat aiempien mallien selitysvoimaa sekä sujuvuus- että taitotasoarvioissa: artikulaationopeuden ja pitkien hiljaisten taukojen osuuden yhdistelmä ennusti sujuvuusarvioita parhaiten, ja puhenopeuden ja lyhyiden hiljaisten taukojen keskimääräisen keston yhdistelmä taitotasoarvioita.

Puheen havaittuun sujuvuuteen vaikuttaa usein yhdistelmä erilaisia sujuvuuspiirteitä, vaikka yksittäisten piirteiden vaikutukset voivat olla keskenään erilaisia. Tutkielman tulokset lisäävät tietoa L2-suomen foneettisesta sujuvuudesta, ja ne ovat tarpeellisia niin kielenoppijoille, -opettajille kuin puheen automaattisten arviointityökalujen kehittäjille.

Contents

List of Tables

List of Figures

1 Introduction

Speaking fluently is an essential goal in second language (L2) learning. Although achieving fluency in a language requires practicing and knowledge of multiple areas, such as pronunciation, vocabulary, and grammar, the definition of fluency varies depending on the context. Phonetically, fluency refers to the perceived smoothness of speech, which is often quantified using temporal measurements from speech signals in L2 research. Different-leveled speakers tend to produce speech at different rates, and may pause at different frequencies, which affect the perceived fluency level of a speaker (e.g., Kallio et al., 2017). To some extent, temporal fluency measures are also connected to perceived accent (Toivola, 2011) and general language proficiency in L2 (Tavakoli et al., 2020). In this work, the term second language refers to any language that has been learned after acquiring a native language (L1) or native languages, although the concepts of second and foreign language are sometimes used separately depending on the language environment (Gass & Selinker, 2008).

This thesis is written as a part of the DigiTala research project where the aim is to develop an automatic tool for spoken language assessment and practicing (Kautonen & von Zansen, 2020). There is an increasing need for automated proficiency assessment in both language learning and testing purposes. In Finland, the Ministry of Education and Culture has proposed a spoken exam for the Matriculation Examination in the two national languages, Finnish and Swedish as L2, as well as in other languages (Ministry of Education and Culture, 2017). All exams in the Matriculation Examination have been arranged digitally since 2019, and the process of assessing spoken skills is planned to be partly automatic. Evaluations done by humans are time-consuming and might, in some cases, depend too much on the person giving the evaluation (Evanini & Zechner, 2020). By including state-of-the art methods in speech technology to the process, many practical problems are solved: the workload of teachers assessing the performances will not increase drastically, and the students will be evaluated as equally as possible (Karhila et al., 2016; Ministry of Education and Culture, 2017). Although there are already systems for automatic assessment of, for instance, spoken L2 English (Evanini & Zechner, 2020), similar tools for assessing spoken Finnish or Swedish are rare (von Zansen et al., 2022).

The purpose of this thesis is to investigate whether phonetic fluency measures in spoken L2 Finnish are connected to perceived fluency and proficiency. A set of spontaneous speech samples of L2 Finnish is analyzed using several phonetic fluency measures, and their effect on human ratings for fluency and overall proficiency is studied statistically. The analyzed speech samples contain high school students' speech and they are a part of a larger data collection in the DigiTala research project (von Zansen et al., 2022). According to previous research on L2 fluency, particularly the rate of speech and pauses tend to have an effect on the perceived fluency level of a speaker (Kallio et al., 2017; Kallio et al., 2022; Bosker et al., 2013; Kormos & Dénes, 2004; Préfontaine et al., 2016). Although this is a well-studied area in L2 fluency research, phonetic fluency in high school students' speech in L2 Finnish has not been studied before. However, there is previous research on phonetic fluency in L2 Finnish using speech samples from adult language learners (Toivola et al., 2009; Kallio et al., 2022). The findings of this thesis expand previous knowledge on phonetic L2 fluency by explaining which aspects are most essential for fluency in high school students' spontaneous speech in L2 Finnish.

The underlying goal in studying the connection between phonetic fluency and human ratings is to advance the development of automated assessment of L2 speech. Information on how humans assess L2 speech is essential when developing automatic systems that could perform similarily for spoken L2 assessment. Phonetic fluency measures are already used in automated scoring systems for spoken language proficiency, at least for L2 English (Hsieh et al., 2020), and more research on fluency in L2 Finnish is needed when developing automated assessment specifically for Finnish language.

1.1 Fluency in L2 Speech

1.1.1 Defining Fluency

Fluency has different definitions that vary depending on the context. Fluent speech can be described as flowing, rapid, smooth, or accurate. Fluency is a common term especially in language pedagogy and language testing, where it is used for defining different levels of language proficiency. Although this thesis mainly focuses on spoken language, fluent language use is not restricted solely to speech. Reading, writing, and sign languages can all be described with fluency, and a comprehensive overview of related areas can be found in the edited book by Lintunen et al. (2019a).

Lennon (1990) introduced two descriptions for fluency in L2 speech. As a wider phenomenon, fluency is often associated with overall oral proficiency. In this sense, fluency might refer to, for instance, native-like pronunciation or advanced use of vocabulary in a second language. In a narrower sense, fluency is separate from other areas of spoken language proficiency, such as the use of vocabulary and grammar, and often refers to the temporal properties and perceived smoothness of speech. The latter definition is more commonly used in language assessment when different areas of proficiency are assessed separately (Bosker et al., 2013; Lennon, 1990). Lennon’s views are very often cited and agreed on in L2 fluency research.

The narrow definition for fluency is used in thesis as well, and the following section explains how temporal measurements and perceived fluency are connected. Temporal properties connected to phonetic fluency are considered as part of speech prosody. Prosody, or suprasegmental features, typically refer to perceived intonation, stress, and rhythm. In speech signals, these phenomena realize as changes in fundamental frequency (F_0), intensity, and duration. Prosodic features are very essential for pronunciation, although segmental properties (individual phones) have generally been more acknowledged in L2 pronunciation teaching and assessment (see e.g., Isaacs, 2018).

1.1.2 Measuring Phonetic Fluency

In L2 research, phonetic fluency is commonly studied as temporal measures (later referred to as phonetic fluency measures), which are divided into speed, breakdown and repair fluency. According to Skehan (2009), speed fluency expresses the rate of speech delivery and is often measured as speech rate and articulation rate. Breakdown fluency consists of pausing phenomena, such as the number, durations, and locations of pauses. Repairs include self-corrections, repetitions, and false starts (Skehan, 2009). The effect of different phonetic fluency measures on perceived fluency is studied using ratings from human assessors. Statistical methods are used for studying which measures are statistically significant in predicting human ratings. Previous research contains a wide range of languages, different-leveled L2 speakers, and raters from different backgrounds. Languages include, for instance, English (Lennon, 1990; Kormos & Dénes, 2004; Tavakoli et al., 2020), Swedish (Kallio et al., 2017), French (Préfontaine et al., 2016), and Dutch (Bosker et al., 2013; Cucchiarini et al., 2002). Kormos and Dénes (2004) had speakers of both advanced and low-intermediate proficiency. Studies have generally used native speakers of the target language as raters. However, non-native speakers might not necessary rate L2 speech differently from natives, as Kormos and Dénes (2004) found no significant differences between the ratings of natives and non-natives.

Speech rate and articulation rate are measures of speed fluency. Speech rate is often calculated as phones, syllables, or words per second or minute. Articulation rate is calculated similarly, only it excludes pauses of determined duration from the count. Natives tend to speak faster than language learners (Toivola et al., 2009), and more advanced L2 speakers have found to have a significantly higher speech rate than less advanced speakers (Kormos & Dénes, 2004). Higher speech rate or articulation rate usually results in higher fluency ratings (Kallio et al., 2017; Kormos & Dénes, 2004; Cucchiarini et al., 2002; Kallio et al., 2022; Préfontaine et al., 2016). However, speech type might affect the importance of these measures: as Cucchiarini et al. (2002) compared read and spontaneous speech in Dutch as L2, articulation rate was only significant in predicting the fluency ratings for read speech. Read speech is usually perceived as more fluent than sponteanous speech, which can contain a greater amount of pauses. In a speech sample where pauses are frequent but articulation rate is high, speech rate is likely to predict human ratings more effectively (Cucchiarini et al., 2002). The rate of speech can also be expressed by mean length of runs, phonation-time ratio, and number of stressed words per minute. In the studies by Kormos and Dénes (2004) and Préfontaine et al. (2016), mean length of runs was measured by counting the average number of syllables in a continuous speech run between pauses above 250 ms. Phonation-time ratio, in turn, reflects the portion of a sample that was spent speaking. Mean length of runs, phonation-time ratio, and number of stressed words per minute have all proved to be significant predictors of fluency level (Kormos & Dénes, 2004; Préfontaine et al., 2016). Cucchiarini et al. (2002) found mean length of runs to be exceptionally good at predicting fluency ratings in spontaneous speech, as it also considers the frequency of pauses.

Breakdown fluency concerns pause-related phenomena. Pauses are divided into silent (or unfilled) and filled ones and their durations are measured in milliseconds (ms). Filled pauses are usually perceived as non-lexical hesitations (Lennon, 1990; Kallio et al., 2017). Pauses are an essential part of all human speech, but there are differences in the ways native speakers and language learners pause. For language learners, it is typical to produce longer pauses than for natives (de Jong & Bosker, 2013). Placing pauses in the middle of a clause is also more common for L2 speakers than for natives (Skehan, 2009). Speech type affects pausing as well, as filled pauses tend to be more common in spontaneous than in read speech (Kallio et al., 2017).

The lower threshold for silent pause durations has varied in L2 fluency research, and it is not always clear what is the minimum duration for silences that affect the perceived fluency level of a speaker. For instance, Kormos & Dénes (2004) and Lennon (1990) used a threshold of 200 ms, and Derwing et al. (2004) used a threshold of 400 ms. De Jong and Bosker (2013) searched for an optimal silent pause threshold for L2 fluency research by studying the relation of the number of silent pauses and an approximate of overall proficiency in spontaneous speech in L2 Dutch. For proficiency approximate, they used a measure of vocabulary knowledge, which had previously proven to be fairly accurate (Zareva et al., 2005). They found out that a threshold of

250–300 ms for silent pauses resulted in highest correlations between vocabulary size and the number of silent pauses. The researchers added that although approximately one fourth of silent pauses in their speech data were less than 250 ms, they were not related to L2 proficiency in this setting. The silent pause threshold of 250 ms has been used by many researchers since (Préfontaine et al., 2016; Kallio et al., 2022). However, even shorter silences might affect perceived fluency as well, as Kallio et al. (2017) found silent pauses between 50 ms and 200 ms to be significant predictors of fluency ratings in read speech in Swedish as L2. Silences this short were not significant predictors of fluency ratings in spontaneous speech, which suggests that brief silences are more accepted by listeners in more conversational speech types.

Pause durations can also be language-dependent. Campione & Véronis (2002) compared silent pause durations in L1 English, French, German, Italian, and Spanish. They used a corpus of read speech including all five languages, together with a smaller corpus of spontaneous speech in French. The average duration of silent pauses was lower in Italian and higher in Spanish than in the other languages. Based on their results, silent pauses could be divided into two categories in read speech: brief pauses (below 200 ms) and medium pauses (200–1,000 ms). In spontaneous speech, there could also be a third category for long pauses (above 1,000 ms). Their data also contained very short pauses of 60 ms, but they claimed that pauses this short have mainly a physiological or respiratory function in speech. In Toivola et al.'s (2009) research, native Finnish speakers produced longer pauses in read speech than non-natives (Russian, Thai, Turkish, and Vietnamese as L1); in this case, mean duration of pause might not affect perceived fluency as much as, for instance, pause frequency. They also found variation in pausing among the natives: one native Finnish speaker produced much shorter pauses than the rest, whereas another native speaker spoke very calmly, producing pauses with more varying durations. Pause durations might, in some cases, have more to do with individual differences between speaking styles and the language spoken rather than fluency (Toivola et al., 2009).

Kallio et al. (2017) discovered that for L2 Swedish, the portion of filled pauses (> 50 ms) as well as long silent pauses (> 1,000 ms) in a sample affected the perceived fluency level significantly in both read and spontaneous speech: a greater amount of filled and long silent pauses resulted in lower fluency ratings. However, the portion of silences between 200 ms and 1,000 ms remained insignificant in predicting fluency level in both speech types. In the study by Kallio et al. (2022) concerning spontaneous L2 Finnish speech of adult language learners, the mean duration of filled pause and the rate of long silent pauses (250–5,000 ms) per minute had negative effects on fluency ratings. In Préfontaine et al.'s research concerning L2 French

(2016), mean duration of silent pause (> 250 ms) was a significant predictor of fluency ratings, but the effect of pause frequency remained weak. The mean duration of silent pauses (> 250 ms) was also a significant predictor of fluency in L2 Dutch in Bosker et al.'s paper (2013). The number of silent pauses affected fluency ratings more than the number of filled pauses. Cucchiarini et al. (2002) found number of silent pauses (> 200 ms) per minute to predict fluency ratings well in both spontaneous and read speech.

Repair fluency phenomena include self-corrections, repetitions, and restarts. Speech that frequently contains repairs is often considered disfluent (Lintunen et al., 2019b: 6). Among speed, breakdown, and repair fluency, repair measures have usually had the weakest connection with perceived fluency (Bosker et al., 2013; Cucchiarini et al., 2002). Kallio et al. (2017) found that the portion of self-corrections and self-repetitions in a speech sample was significant for read speech but not for spontaneous speech in L2 Swedish. This could suggest that disfluencies of this sort are better tolerated in spontaneous speech than read speech that does not require content planning. However, a combination of speed, breakdown, and repair fluency measures have predicted the fluency ratings successfully; raters tend to ground their assessments on the sum of all disfluency phenomena present in speech (Bosker et al., 2013; Kormos & Dénes, 2004). Kallio et al. (2022) studied the contribution of different utterance fluency measures in L2 Finnish to both perceived fluency and proficiency. They included a new temporal measure, utterance break, which combined breakdown and repair fluency. Utterance breaks were defined as interruptions of over 250 ms in duration between continuous speech runs, consisting of both silent and filled pauses, hesitations, corrections, and repetitions. Of the used fluency measures, the mean duration of utterance break was the most significant predictor of fluency ratings, and had a smaller yet significant effect on predicting the proficiency ratings. Longer breaks between continuous speech runs in a speech sample would thus result in lower fluency and proficiency ratings.

1.2 Automated Assessment of Fluency

Language proficiency tests are used, for instance, for assessing a learner's progress, or as entrance tests for professions and educational institutions (Brown, 2013). In Finland, the Ministry of Education and Culture has proposed a spoken exam for the Matriculation Examination in the two national languages, Finnish and Swedish as L2, as well as in other second languages (Ministry of Education and Culture, 2017). The main purpose of the Matriculation Examination is to test how well high school students in Finland have reached different learning goals set in the high school curriculum. The assessment procedure for the spoken exam is planned to be partly automatic to overcome many practical problems and to achieve more standardized assessment (Ministry of Education and Culture, 2017; Kautonen & von Zansen, 2020). A system for automated spoken language assessment is currently under development in the DigiTala research project, with an emphasis on L2 Finnish and Swedish (Kautonen & von Zansen, 2020). However, there is still far more research on automated assessment of spoken L2 English than of Finnish or Swedish.

According to Evanini and Zechner (2020), the main application areas for automated speech scoring include assessment of non-native speaking proficiency, feedback for language learning, and reading tutoring and assessment. Automated speech scoring system pipeline usually consists of an automatic speech recognition (ASR) system, speech feature extraction, a filtering model, and a scoring model (Evanini & Zechner, 2020: 9–10). SpeechRater is an automated system for scoring non-native English speaking proficiency developed in the Educational Testing Service. It applies a scoring rubric from the TOEFL iBT speaking test, where proficiency score is based on speech delivery, language use, and topic development (Evanini & Zechner, 2020: 8–9). Fluency is assessed as a part of the speech delivery dimension using several measures related to pauses, speaking rate, repair, and repetitions (Hsieh et al., 2020: 101). This follows the common division of fluency into three aspects; breakdown, speed, and repair (see Skehan, 2009). When compared with other automatically measured dimensions, such as grammatical accuracy, fluency features can be computed reliably since they don’t rely on the content as much and are less sensitive to ASR errors. They have also proved to have some of the highest correlations with human scores regarding speaking proficiency. (Hsieh et al., 2020)

De Wet et al. (2009) described the development of automated spoken language proficiency assessment for L2 English in reading and repeating tasks. They examined correlations between human ratings and ASR-based measures for speech samples from speakers of intermediate to advanced proficiency levels. In addition to overall proficiency scores, humans assessed the read speech samples on the degree of hesitation, pronunciation, intonation, and the repeated speech samples on the degree of success and accuracy. An ASR system was used for measuring the rate of speech, goodness of pronunciation, and repeat accuracy. They found that the rate of speech and accuracy had relatively high correlations with human ratings for proficiency, whereas goodness of pronunciation was poorly correlated (de Wet et al., 2009).

Since ASR-based spoken language assessment requires lots of speech data, especially from language learners, alternative methods for automated assessment have been suggested. Fontan et al. (2018) used two different algorithms to predict the fluency levels of language learners: one for segmenting speech into subphonemic units (a forward-backward divergence segmentation algorithm) and another to track the first formant. They had read speech samples spoken by Japanese learners of French from beginner, intermediate, and advanced proficiency levels. Native French speakers assessed the samples for global fluency, speech rate, regularity of speech rate, and speech fluidity. The two algorithms were used for measuring speech rate, regularity of speech rate, speech fluidity, the length and regularity of silent pauses, and percentage of speech. Their results showed that correlations were high between human assessments and automatic measures, especially for speech rate and regularity of speech rate.

Automatically calculated phonetic fluency measures, particularly those related to speed and pauses, have previously had relatively high correlations with human ratings for fluency and proficiency (Hsieh et al., 2020; de Wet et al., 2009; Fontan et al., 2018). However, there might be differences between languages in, for instance, pause durations (e.g., Campione & Véronis, 2002; Toivola et al., 2009), which means that the same methods for automated assessment of L2 speech will not necessarily work for different languages. In order to develop automated assessment for fluency in L2 Finnish, it is important to study fluency and language assessment in the context of Finnish language.

1.3 Research Questions

The main objective of this thesis is to study whether phonetic fluency measures calculated for L2 Finnish speech samples can predict human ratings for fluency and proficiency, and if so, which measures are the most effective in this. A set of high school students' spontaneous speech samples in L2 Finnish is annotated, and different phonetic fluency measures are calculated for each sample. All speech samples are rated for fluency and proficiency by expert assessors. Research questions are as follows:

1. Can the selected phonetic fluency measures predict human ratings for L2 Finnish fluency and proficiency?
2. How do the phonetic fluency measures differ in predicting human ratings for fluency and proficiency?

The connection between phonetic fluency measures and human ratings for fluency and overall proficiency is studied using linear regression models. Linear regression was used in similar L2 fluency experiments by Kallio et al. (2022) and Bosker et al. (2013). Different linear regression models are compared to each other to find out which phonetic fluency measures are the best predictors of the ratings. As assessments for overall proficiency are affected by multiple dimensions of language, and not only phonetic fluency, it is very likely that the fluency measures have a greater impact on fluency ratings than for proficiency.

The used phonetic fluency measures are based on previous research in L2 fluency, and they include measures related to speed, breakdown, and repair fluency. Articulation rate and speech rate are expected to perform well in predicting the fluency and proficiency ratings, since faster speech has previously resulted in higher ratings (Kallio et al., 2017; Kormos & Dénes, 2004; Préfontaine et al., 2016). Based on previous research, silent pauses are expected to affect the perceived fluency level as well (Kallio et al., 2017; Kallio et al., 2022; Préfontaine et al., 2016; Bosker et al., 2013). The speech samples will likely contain longer silences as they are typical for spontaneous L2 speech. Shorter silent pauses are separated from longer ones since they have also predicted fluency ratings (Kallio et al., 2017). Multiple pause-related measures are included to find out how they affect perceived fluency and proficiency in this speech data. Cucchiarini et al. (2002) categorized filled pauses and repair fluency measures as “secondary variables” which do not appear in speech as often as silences. They might not predict fluency and proficiency ratings as effectively as measures related to speed and breakdown. Pause and repair phenomena can be examined separately, but also combined into one phonetic fluency measure: Kallio et al. (2022) found the utterance break to be a significant predictor for both fluency and proficiency ratings in L2 Finnish, and it is expected to perform similarly in this research. A combination of different phonetic fluency measures is expected to improve the strength of prediction, as human assessors tend to rate speakers’ performances by considering multiple aspects (Bosker et al., 2013; Kormos & Dénes, 2004).

2 Methods

2.1 Data Processing

The aim of this research is to investigate whether phonetic fluency measures calculated for L2 Finnish speech samples can predict human assessments of fluency and overall proficiency. The analyzed data consist of two components: speech samples and their ratings. This section covers all phases in preparing the data for statistical analyses.

2.1.1 Description of Data

The speech recordings and ratings used for this research were collected in the DigiTala research project in 2021 (see, for instance, von Zansen et al., 2022; Kautonen & von Zansen, 2020). The speakers were high school students of ages between 15 and 21. Each student completed an oral examination where they produced both read and spontaneous speech in Finnish as L2 in eight different tasks. The examination took place in an online Moodle environment; most students participated remotely at their homes, and the recording equipment were not controlled for. The audio quality of the recordings was mainly sufficient, although some samples remained too noisy for acoustic analysis.

The speakers had different language backgrounds: there were 21 native speakers of Swedish, nine bilinguals in Finnish and Swedish, and the others spoke Russian, Finnish, Estonian, Kurdi, Chinese, Vietnamese, Punjabi, Nepalese, Spanish, German, Turkish, Lingala, English, or Arabic as their native language. The speakers had learnt Finnish on different levels in school and there was variety in terms of fluency and proficiency. Majority of them had studied two to six courses of Finnish in high school before this examination.

For this research, task 2 from the exam's eight tasks was chosen since it had fairly long responses and unique answers from each speaker, resulting in spontaneous speech. In the task, the speakers described an important place by answering questions, such as "why is this place important to you?", "what is best about this place?", and "what do you do in this place?" The speakers were instructed to continue speaking for one minute. In practice, the durations of the speech samples varied between 15.8 and 59.5 seconds. The original Finnish task instructions from the examination are in Appendix A. The total number of speech recordings for this task was 61. However, 8 samples were discarded, as they were perceived as read speech – these samples had very few hesitations, contained specific and well-planned vocabulary, and the intonation was repetitive for each sentence. The remaining 53 speech samples were annotated and included in the analyses.

The speech samples were rated by 14 trained assessors. Four of them were researchers in the DigiTala project. 13 assessors were native speakers of Finnish, and one was a non-native speaker with C2-level proficiency in Finnish. All assessors were experienced in evaluating spoken language skills. Ratings were provided concerning task completion, fluency, pronunciation, scope of expression, accuracy of vocabulary and grammar, and overall proficiency (Kautonen & von Zansen, 2020; von Zansen et al., 2022). As the focus of this research is in the phonetic fluency measures, ratings for fluency and overall proficiency are of interest. Fluency (speech fluency and effortlessness) was rated on a scale from zero to four, where zero referred to a sample that could not be graded for fluency (for instance, if the speaker had not produced enough speech), and four referred to a very fluent and effortless speech sample. Table 1 presents fluency levels' descriptions, which are translated from the original Finnish version. Overall proficiency was assessed using an application of the CEFR scale, ranging from below A1 to C2. The original Finnish descriptions for all assessed dimensions are in Appendices B and C.

Table 1. Rating criteria for fluency assessments (speech fluency and effortlessness).

Fluency level	Description
0	Cannot be graded. / I am unable to tell.
1	Disfluent; several disturbing breaks, repetitions, interruptions, and hesitations.
2	Moderately fluent; a few disturbing breaks, repetitions, interruptions, and hesitations.
3	Fluent and effortless; no disturbing breaks, repetitions, interruptions, or hesitations.
4	Very fluent and effortless; no disturbing breaks, repetitions, interruptions, or hesitations.

Every speech sample in the data was assessed by at least one of the 14 raters. For the 53 speech samples, the total number of ratings was 277, of which 224 formed a subset that was also used for studying the agreement between the raters. In the subset, all 14 raters assessed the same 16 samples. The speech samples and ratings used in this thesis are a part of a larger L2 data collection. The assessments were collected on an online Moodle environment after providing the raters with instructions and the assessment criteria. The raters could proceed with the evaluation at their own pace and discuss the criteria with each other on the Moodle environment. (von Zansen et al., 2022)

2.1.2 Annotations

Each speech sample was annotated in Praat (Boersma & Weenink, 2021) to obtain durations of different types of disfluencies, pauses, and individual words. There were two tiers in the annotations: utterance level and word level.

Individual words, pauses, and disfluencies were marked in the word level. Silent pauses were divided into two categories based on their durations; silent pauses above 250 ms were labeled as LSP (long silent pause) and silent pauses between 50 ms and 250 ms were labeled as SP (short silent pause). Short silent pauses were labeled due to previous research that have found a connection between shorter silences and perceived fluency (Kallio et al., 2017). The long silent pause threshold was set to 250 ms based on previous research on silent pause durations in L2 speech (de Jong & Bosker, 2013; Préfontaine et al., 2016; Kallio et al., 2022).

In addition to silent pauses, intervals that contained filled pauses, hesitations, corrections, or repetitions were labeled in the word level. Filled pauses and hesitations were labeled as FP, and corrections and repetitions as CR. Laughter and coughing were labeled as EL (extralinguistic). For labels SP, FP, and CR, the threshold level was set to 50 ms, following the methodology by Kallio et al. (2017) – disfluencies below this were not marked.

Regarding the different disfluency types, a few specific rules were determined to maintain consistency throughout the annotations. For corrections and repetitions, only intervals containing parts of words were labeled: A partial word or a sequence of those was replaced with CR in the annotations whenever a speaker began to pronounce a word but changed it to something else in the middle, or began to pronounce the same word from the beginning. If a speaker uttered the same, complete word more than once, this was not marked as a disfluency. However, in Finnish, this is not always straightforward, since spoken language commonly contains words that are shortened versions of the complete word. In the annotations, the surrounding words were observed when making these decisions.

The assessors were instructed to ground their assessments on the parts that the speakers produced themselves. Some students read aloud questions from the task instructions, and these parts were thus excluded from the annotations and the analysed speech data. Silent pauses inside these read intervals were not annotated, but the break intervals preceding and following the read intervals were annotated. Silences in the beginning and end of a sample were excluded from the annotations as well; as the students recorded their responses themselves, it is possible they started to record accidentally before being ready to speak. The first and last annotated intervals always contained either speech or filled pauses. If the recording did not capture the first or last uttered word completely, it was still annotated as a complete word.

For most parts, the boundaries were placed by hand in the annotations. A Praat script was used in the beginning to detect silences above 250 ms, but the outcome was manually checked – consonant lengthening is very common in Finnish, and the script sometimes captured pauses related to articulation in the silences. Individual words were added to the annotations using previously produced transcripts and a Praat script. The result was again checked manually, since the transcripts sometimes had disfluencies written out and they had to be changed into their respective disfluency labels.

The utterance level consisted of two types of labeled intervals, U (utterance) and B (break). The U intervals contained all orthographic words and brief (< 250 ms) silences or disfluencies inside utterances. The B intervals represented interruptions between utterances, including both silences and disfluencies that were at least 250 ms in duration combined. A similar compound measure was previously found significant in predicting both fluency and proficiency ratings in L2 Finnish (Kallio et al., 2022). Figure 1 shows an example from an annotated sample, where a B interval contains a filled pause between two long silent pauses in the word level.

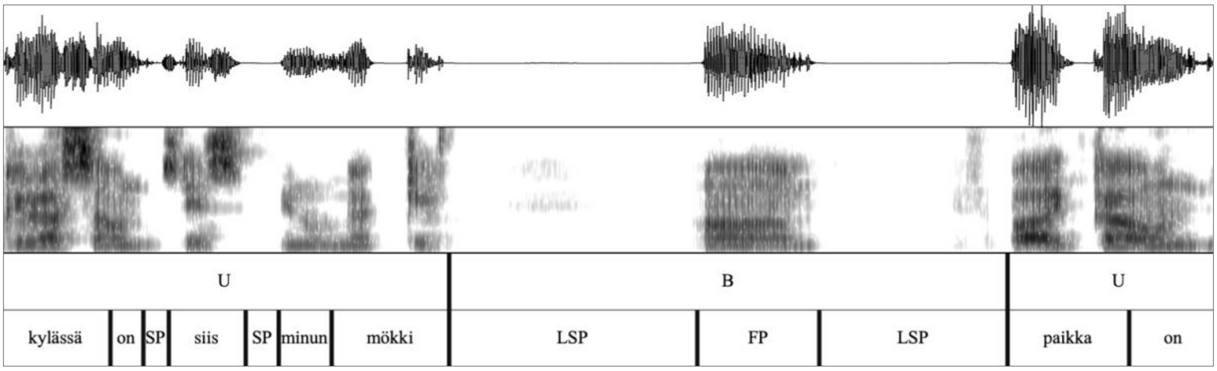


Figure 1. An example piece of an annotated speech sample in Praat. All pronounced words are included in the utterance intervals (U). In this example, the break interval (B) contains two long silent pauses (LSP) and a filled pause (FP) in the middle. The U interval on the left contains two individual short silent pauses (SP), as their durations are less than 250 ms.

2.1.3 Calculations for Phonetic Fluency Measures

Labeled intervals' durations were obtained from the annotations using a Praat script. As there were two annotation tiers, two text files were created containing filenames (speaker IDs) and each annotated interval's label and duration. These text files were imported to RStudio as data frames, and R programming language was used for calculating several phonetic fluency measures for each speech sample from the interval durations (RStudio Team, 2022; R Core Team, 2021).

The speech samples were only annotated to word-level and therefore it was decided to calculate articulation rate and speech rate using the number of characters in the annotations, which is an approximate of the number of phones produced. This is a simplification, but it helps to express the rate speakers produced speech. In previous fluency research, the number of phones was used, for instance, by Cucchiarini et al. (2002) to count articulation rate and speech rate. Only orthographic words were included in the character count, and not the labeled disfluencies (LSP, SP, FP, CR, and EL). Dividing the character count by the total duration of a sample produced the variable *speech rate*. Dividing the character count by the total duration of the annotated intervals containing only spoken words (excluding pauses and disfluencies) produced the variable *articulation rate*.

The relative numbers of pauses, disfluencies, utterances, and utterance breaks were expressed as their average rates of occurrence per minute. Using absolute frequencies would have made comparisons between samples unreliable, as some speech samples were less than 20 seconds and some close to 60 seconds in duration. The relative counts were calculated by dividing the frequency of an interval type by the total duration of a sample and multiplying this with 60. The numbers were stored under variables *SPrate, LSPrate, FPrate, CRrate, ELrate, Urate,* and *Brate*.

Relative proportions for each disfluency type per sample were calculated by dividing the disfluency type's total duration in a sample by the total duration of a sample. These were stored under variables *SPratio, LSPratio, FPratio, CRratio,* and *ELratio.*

Finally, mean durations in seconds were calculated for each inverval-based measure, and they were stored under variables *meanSP, meanLSP, meanFP, meanCR, meanEL, meanU,* and *meanB.*

R packages plyr (Wickham, 2011), tidyverse (Wickham et al., 2019), data.table (Dowle & Srinivasan, 2021), and reshape2 (Wickham, 2007) were used in data processing. The calculated phonetic fluency measures and their descriptions are listed in Table 2.

Table 2. Phonetic fluency measures calculated for each speech sample.

Phonetic fluency measure	Description
speech rate	rate of produced characters (~phones) per sample
articulation rate	rate of produced characters (~phones) per sample excluding pauses and disfluencies
SPrate	rate of short silent pauses per minute
LSPrate	rate of long silent pauses per minute
FPrate	rate of filled pauses per minute
CRrate	rate of corrections and repetitions per minute
ELrate	rate of extralinguistic intervals per minute
Urate	rate of utterances per minute
Brate	rate of breaks between utterances per minute
SPratio	relative proportion of short silent pauses in a sample
LSPratio	relative proportion of long silent pauses in a sample
FPratio	relative proportion of filled pauses in a sample
CRratio	relative proportion of corrections and repetitions in a sample
ELratio	relative proportion of extralinguistic intervals in a sample
meanSP	mean duration of short silent pauses in seconds
meanLSP	mean duration of long silent pauses in seconds
meanFP	mean duration of filled pauses in seconds
meanCR	mean duration of corrections and repetitions in seconds
meanEL	mean duration of extralinguistic intervals in seconds
meanU	mean duration of utterances in seconds
meanB	mean duration of breaks between utterances in seconds

2.2 Data Analysis

The first research question was whether the selected phonetic fluency measures could predict human ratings for L2 Finnish fluency and proficiency. To study this effect, linear regression was used, similarly to previous L2 fluency research by Kallio et al. (2022) and by Bosker et al. (2013). Different linear regression models were compared to find the fluency measures that best explain the variance in the ratings, as well as to answer the second research question of how the phonetic fluency measures differ in predicting human ratings for fluency and proficiency. The fluency and proficiency ratings were treated as mean ratings calculated for each speech sample, as some samples were rated by several assessors.

The amount of corrections, repetitions, and extralinguistic phenomena was very low in the data. The frequency of CRs in the speech samples ranged from zero to three with a mean of 0.64 per sample. The frequency range for ELs was from zero to two, with a mean of 0.13 per sample. Because of this, these measures were discarded from the analyses.

Phonetic fluency measures that represented absolute durations or frequencies were not included in the analyses, as the speech samples had a lot of variety in the total durations. Instead, different fluency measures were represented by rates, ratios, and means, as described in Section 2.1.3.

2.2.1 Inter-Rater Reliability

Rater agreement was measured with intraclass correlation coefficient (ICC) and a subset of the rating data. The amount of individual ratings was 277 in the total data used in the present study, and the subset included 224 of them, with 14 ratings for 16 speech samples each. From the different forms of ICC, two-way mixed-effects model was chosen, since all 14 raters from the complete rating data were also included in the subset that was used for studying the agreement. Both single and average types were considered for the ICC: although most analyses in this work rely on the mean ratings for each speaker, some speakers in the complete data were only rated by one assessor, adding a lot of weight to individual assessors' ratings. For the ICC definition, both consistency and absolute agreement were used. Values below 0.5 were considered poor, between 0.5 and 0.75 moderate, between 0.75 and 0.9 good, and above 0.9 excellent (Koo & Li, 2016). R package irr was used for calculating the coefficients (Gamer et al., 2019).

2.2.2 Linearity Between Phonetic Fluency Measures and Human Ratings

Linear regression was used for investigating whether there was a statistically significant connection between the ratings and the phonetic fluency measures. This connection was studied first with simple linear regression models, using each phonetic fluency measure as an independent variable and mean fluency or mean proficiency rating as a dependent variable, and then with multiple linear regression models using combinations of phonetic fluency measures as independent variables. The linear models were fitted using R (R Core Team, 2021). Different regression models were compared to each other using analysis of variance (ANOVA).

As many of the used fluency measures were dependent of each other (for instance, pause durations affected speech rate), they were only combined in one model if they were not highly correlated with each other. Pearson's r was used for measuring correlation in R (R Core Team, 2021), and correlation coefficients between –0.39 and 0.39 were considered weak enough to combine two variables in the same model (Schober et al., 2018).

3 Results

3.1 Inter-Rater Reliability

Inter-rater reliability was measured separately for fluency and proficiency ratings. The obtained intraclass correlation coefficients are listed in Table 3. The inter-rater agreement ICC value for the single type was 0.38 for fluency ratings and 0.47 for proficiency ratings, indicating poor reliability for both. For consistency, the single type ICC was 0.50 for fluency ratings and 0.60 for proficiency ratings, indicating moderate reliability for both. The average type ICCs were considerably higher: for fluency ratings, a coefficient of 0.90 and for proficiency ratings, a coefficient of 0.92 was obtained for absolute agreement. For consistency, a coefficient of 0.93 was obtained for fluency and 0.95 for proficiency ratings. ICCs above 0.90 are considered excellent (Koo & Li, 2016). The coefficients were expected to be lower for the single type, as this describes the differences between individual ratings for the same samples, and it is very unlikely that human raters would assess all speech samples identically. Additionally, this measure does not provide information on the scope of the disagreement – for instance, the difference between fluency ratings 1 and 4 has a similar effect on the single type ICC as the difference between ratings 3 and 4, although in the latter case, the raters are much closer to agreement on the grade. The average type coefficients express the agreement between individual ratings and the mean ratings. As these are all excellent (> 0.90), using mean ratings when analyzing this data is considered very reliable. All correlation coefficients were slightly higher for proficiency ratings than for fluency.

Table 3. Intraclass correlation coefficients (ICC) for fluency and proficiency ratings. In the ICC types, “A” refers to absolute agreement, “C” to consistency, number 1 to single type, and number 14 to average type. Reliability below 0.5 is considered poor, between 0.5 and 0.75 moderate, between 0.75 and 0.9 good, and above 0.9 excellent (Koo & Li, 2016).

ICC type	Fluency ratings	Proficiency ratings
ICC(A,1)	0.38	0.47
ICC(C,1)	0.50	0.60
ICC(A,14)	0.90	0.92
ICC(C,14)	0.93	0.95

3.2 Simple Linear Regression

Simple linear regression models with each phonetic fluency measure as the independent variable and ratings for mean fluency or mean proficiency as the dependent variable were compared. Results from simple linear regression models are presented in Table 4, with *t* values, adjusted R^2 values, and significance codes for *p* values below 0.05 for each model. The adjusted R^2 expresses how well the independent variable explains variation inside the dependent variable (mean fluency or mean proficiency ratings). Descriptions for each phonetic fluency measure are listed in Table 2 under Section 2.1.3.

Table 4. Results from simple linear regression models. For each model, mean fluency or mean proficiency was treated as the dependent variable and one phonetic fluency measure as the independent variable. Negative adjusted R^2 values were rounded to zero. Significance codes: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.

Phonetic fluency measure	Mean fluency		Mean proficiency	
	t value	Adjusted R^2	*t* value	Adjusted R^2
speech rate	8.77 ***	0.594	5.00 ***	0.316
articulation rate	4.28 ***	0.250	3.72 ***	0.198
SPrate	2.18 *	0.067	2.39 *	0.083
LSPrate	–1.26	0.011	0.20	0
FPrate	–1.00	0	–0.78	0
SPratio	1.82	0.042	2.58 *	0.098
LSPratio	–6.39 ***	0.434	–3.84 ***	0.209
FPratio	–1.93	0.050	–1.70	0.035
Urate	–1.02	0.001	0.26	0
Brate	–1.12	0.005	0.44	0
meanSP	0.74	0	1.83	0.043
meanLSP	–4.40 ***	0.261	–3.99 ***	0.223
meanFP	–1.34	0.015	–1.57	0.027
meanU	3.41 **	0.169	1.22	0.009
meanB	–4.45 ***	0.266	–3.94 ***	0.219

Speech rate achieved the highest adjusted R^2 for both mean fluency and mean proficiency when compared to all other fluency measures. The regression coefficients were positive, which indicates that a higher speech rate resulted in higher ratings for both assessed dimensions.

Articulation rate was also a significant predictor of mean fluency and mean proficiency ratings, but according to the adjusted R^2, speech rate explains a larger amount of variance in human ratings. The linear relationships between speech rate and mean fluency ratings as well as speech rate and mean proficiency ratings are presented in Figure 2, where each data point represents one speech sample. A similar plot for articulation rate as the independent variable is presented in Figure 3.

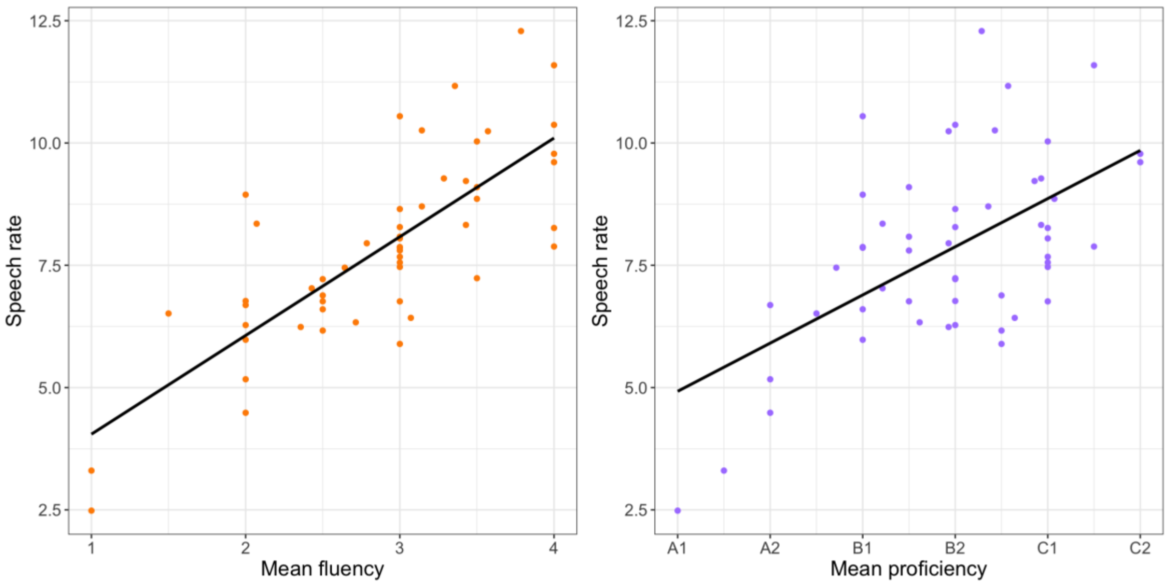


Figure 2. Linear trends between speech rate and mean ratings for fluency and proficiency.

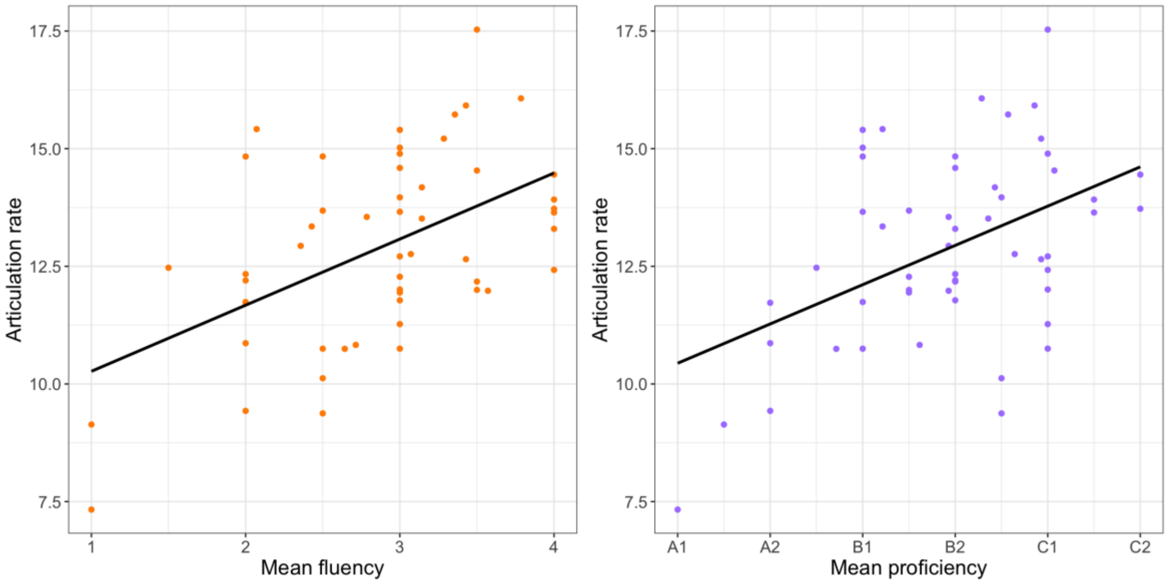


Figure 3. Linear trends between articulation rate and mean ratings for fluency and proficiency.

The portion of long silent pauses (LSPratio) and the mean duration of long silent pauses (meanLSP) in a sample had statistically significant effects on fluency and proficiency ratings. The regression coefficients for these were all negative, which implies a negative connection between the variables: a larger portion or a higher mean duration of long silent pauses in a speech sample resulted in lower ratings. The rate of long silent pauses (LSPrate) remained statistically insignificant. For the portion of short silent pauses (SPratio), only the connection with mean proficiency was statistically significant. Interestingly, the connection was positive, which means that a greater portion of short silences resulted in higher proficiency ratings. The mean duration of short silent pauses (meanSP) was not statistically significant in predicting the ratings, but the rate of short silent pauses (SPrate) had a statistically significant, positive effect on both assessed dimensions. The portion of filled pauses (FPratio), the mean duration of filled pauses (meanFP), and the rate of filled pauses (FPrate) all remained statistically insignificant in predicting the ratings. The pause-related measure with the highest adjusted R^2 was LSPratio for mean fluency, and meanLSP for mean proficiency. The linear connections between LSPratio and mean ratings for fluency and proficiency are presented in Figure 4, and similar plots for meanLSP as the independent variable are in Figure 5.

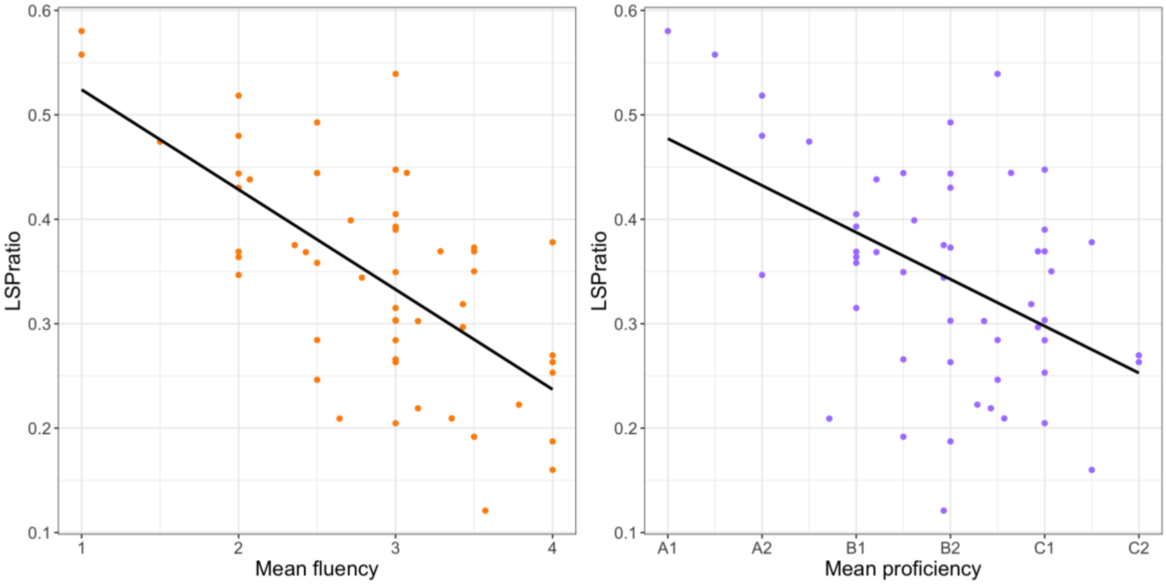


Figure 4. Linear trends between LSPratio and mean ratings for fluency and proficiency.

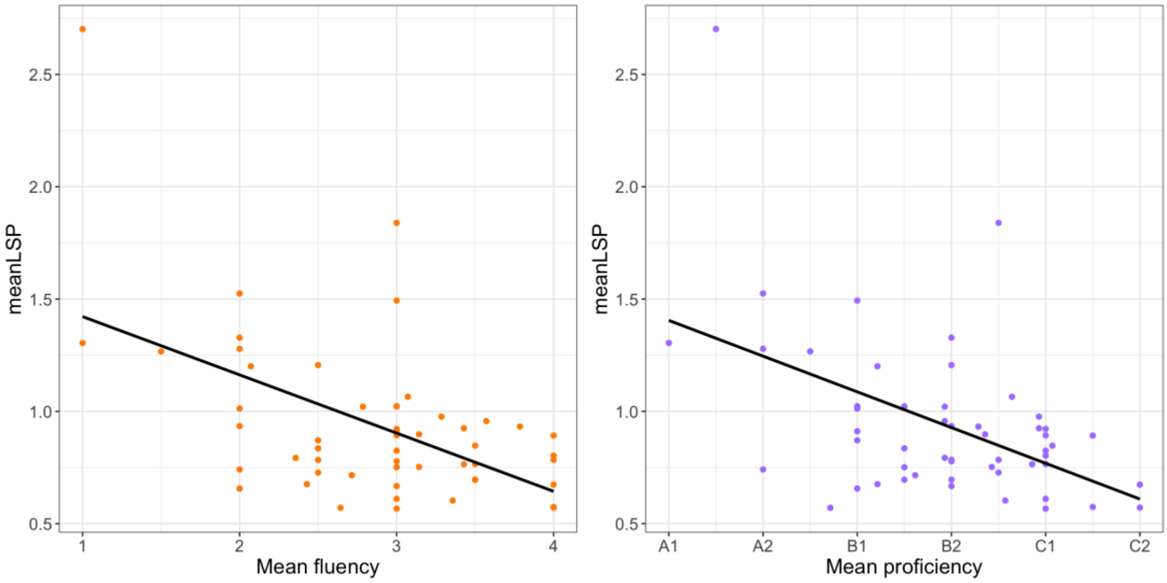


Figure 5. Linear trends between meanLSP and mean ratings for fluency and proficiency.

The mean duration of a break interval (meanB) was a highly significant predictor of both fluency and proficiency ratings. The connections were negative, which indicates that a higher mean duration of break intervals resulted in lower ratings for both fluency and proficiency. These connections are plotted in Figure 6. The mean duration of an utterance interval (meanU) was only significant for fluency ratings. The effects of Urate and Brate remained statistically insignificant.

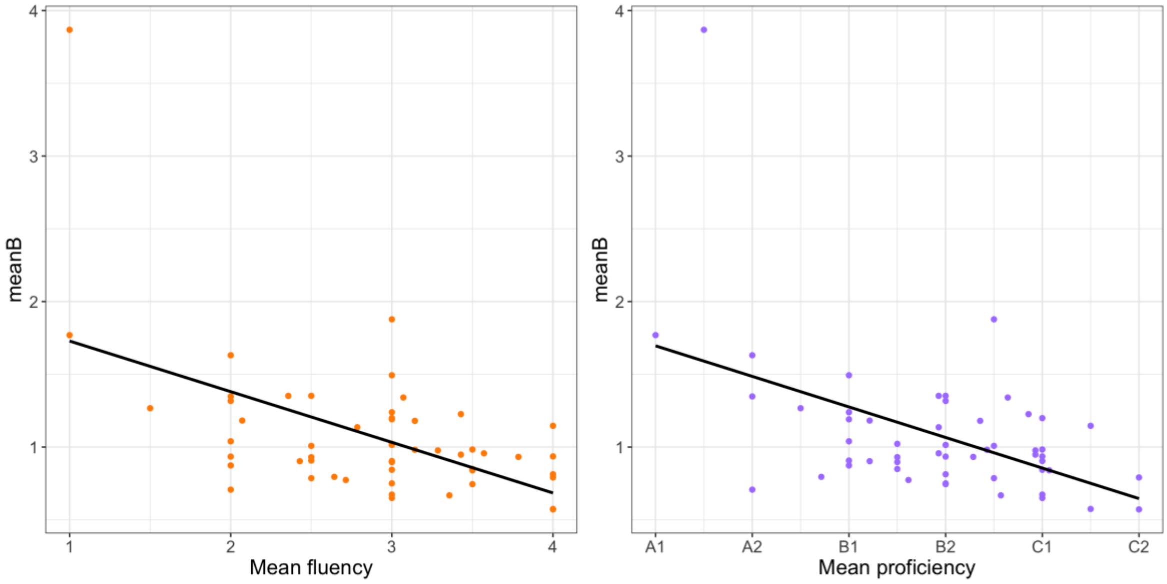


Figure 6. Linear trends between meanB and mean ratings for fluency and proficiency.

In Figures 2–6, ratings were treated as continuous variables. As speech rate had the highest adjusted R^2, its effect on ratings of fluency and proficiency was also examined by treating the ratings as discrete variables. For this, the mean ratings were divided into categories: for instance, mean fluency ratings between 2.50 and 3.49 were rounded to 3. Figure 7 presents the amounts of speech samples for each fluency and proficiency level category. It is noteworthy that for level 1 fluency, there were only two speech samples, and for level A1 proficiency, only one.

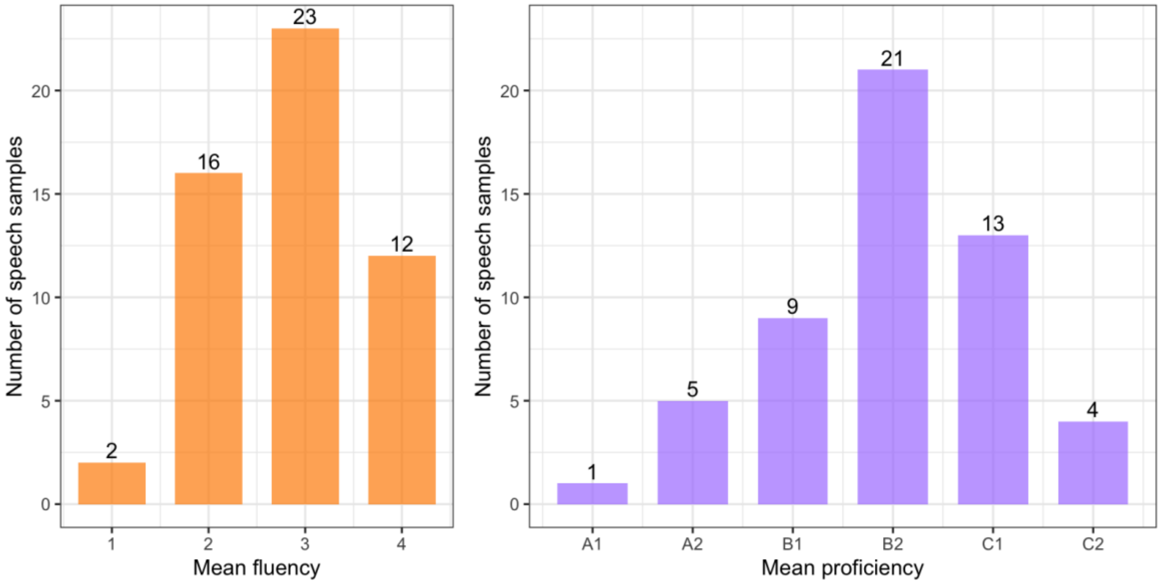


Figure 7. The amounts of speech samples in the mean fluency and mean proficiency categories.

Boxplots were created to visually examine the differences in speech rate between the discrete fluency and proficiency levels, and they are presented in Figure 8. The visualization reveals that for mean fluency, different levels are clearly separated from each other in terms of speech rate, whereas for mean proficiency, speech rates on levels B1–C1 appear to be relatively close to each other. Similar plots were also created for the variable meanB, as it was a statistically significant predictor for both fluency and proficiency ratings, and a relatively new measure in L2 fluency research. Figure 9 shows that for mean fluency, the values for meanB on levels 2–4 are close to each other. For mean proficiency, meanB for levels B1–C1 are again close to each other, whereas the meanB on level A2 is higher and on level C2 lower.

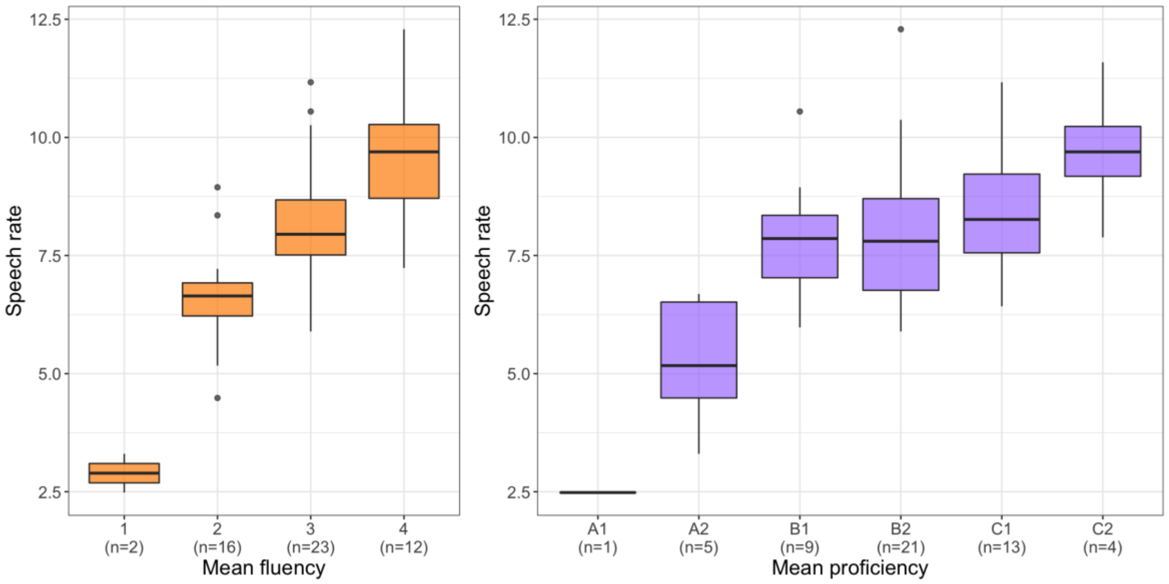


Figure 8. Boxplots for speech rate with mean fluency and mean proficiency.

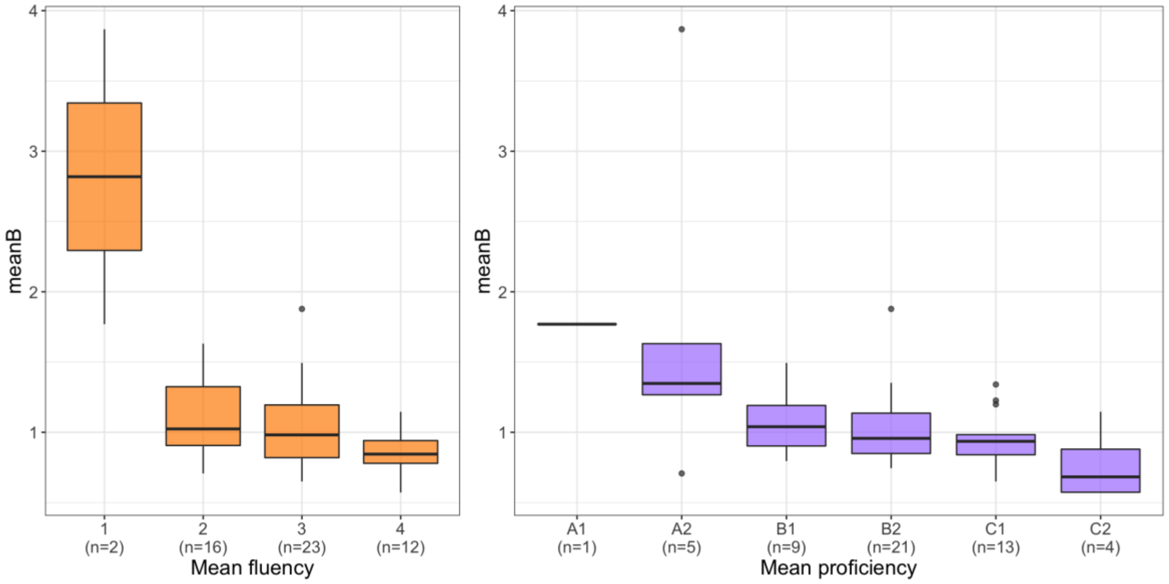


Figure 9. Boxplots for meanB with mean fluency and mean proficiency.

3.3 Multiple Linear Regression

Multiple linear regression (MLR) models were created by treating two or more phonetic fluency measures as independent variables and mean fluency or mean proficiency as the dependent variable. Only measures that had low correlation coefficients were combined in the same MLR model. Correlations between individual variables were measured with Pearson's r, and correlation coefficients between –0.39 and 0.39 were considered low (Schober et al., 2018). The statistical difference between two models was measured with ANOVA.

A comparison of simple linear regression and MLR models with mean fluency as the dependent variable indicated that combining speech rate with other fluency measures did not significantly improve the simple linear regression model with speech rate. However, a model combining articulation rate with certain fluency measures significantly improved the simple model with articulation rate. Table 5 presents four best models for both mean fluency and mean proficiency ratings according to the adjusted R^2 and t values. The highest adjusted R^2 was achieved with model fluency-1 where articulation rate and the portion of long silent pauses (LSPratio) were treated as independent variables. This indicates that a faster articulation rate and a smaller portion of long silent pauses led to higher fluency ratings. The adjusted R^2 of fluency-1 is even higher than that of the best simple linear regression model where speech rate was treated as the independent variable, but according to ANOVA, the difference between these two models is not statistically significant. As shown by fluency-2 and fluency-3, a model combining articulation rate and the mean duration of utterance intervals (meanU) with the mean duration of long silent pauses (meanLSP) or the mean duration of break intervals (meanB) as independent variables predicted the fluency ratings well. A relatively high adjusted R^2 was also achieved without measures of speech rate or articulation rate with the model fluency-4, where meanLSP and meanU were treated as independent variables; a speech sample with longer, uninterrupted utterances without too long silences resulted in higher fluency ratings.

For MLR models with mean proficiency as the dependent variable, the highest adjusted R^2 was achieved with the model proficiency-1 where speech rate and the mean duration of short silent pauses (meanSP) were combined as independent variables. The model proficiency-3 combined speech rate with the portion of short silent pauses (SPratio). The effects of meanSP and SPratio were positive, which indicates that the higher the mean duration or the portion of short silent pauses, the higher the proficiency rating. Articulation rate, portion of short silent pauses (SPratio), and the mean duration of break intervals (meanB) were treated as independent variables in the model proficiency-2. However, articulation rate and meanB were more important in the model than SPratio, which only had a barely significant effect statistically. Combining speech rate with the rate of long silent pauses per minute (LSPrate) as independent variables in the model proficiency-4 resulted in relatively high adjusted R^2, although in the simple linear regression models, LSPrate did not have a statistically significant effect on the ratings on its own.

Table 5. Results from MLR models. The table presents four best multiple linear regression models for both mean fluency and mean proficiency according to the adjusted R^2 and *t* values. Individual MLR models are separated with a line. Models from fluency-1 to fluency-4 had mean fluency as the dependent variable, and models from proficiency-1 to proficiency-4 had mean proficiency as the dependent variable. Each independent variable's *t* value is reported separately. Significance codes: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, and . for $p < 0.1$.

MLR model	Phonetic fluency measures	*t* value	Adjusted R^2
fluency-1	articulation rate	4.80 ***	0.604
	LSPratio	–6.84 ***	
fluency-2	articulation rate	4.45 ***	0.583
	meanLSP	–4.83 ***	
	meanU	4.33 ***	
fluency-3	articulation rate	4.20 ***	0.568
	meanU	4.19 ***	
	meanB	–4.56 ***	
fluency-4	meanLSP	–4.88 ***	0.426
	meanU	3.96 ***	
proficiency-1	speech rate	5.31 ***	0.376
	meanSP	2.43 *	
proficiency-2	articulation rate	3.43 **	0.370
	SPratio	1.80 .	
	meanB	–2.63 *	
proficiency-3	speech rate	4.68 ***	0.361
	SPratio	2.14 *	
proficiency-4	speech rate	5.58 ***	0.359
	LSPrate	2.11 *	

4 Discussion

In this thesis, the effect of phonetic fluency measures on human ratings was studied using linear regression. The analyzed speech samples contained L2 Finnish spoken by 53 high school students, and their fluency and proficiency levels were assessed by 14 expert raters. The aim was to find out whether the selected phonetic fluency measures can predict human ratings for L2 Finnish fluency and proficiency, and if they can, how the different measures differ in predicting the ratings. Fluency research on L2 Finnish is still limited, and this thesis is the first that studied the phonetic fluency of high school students' L2 Finnish. The findings expand previous knowledge on phonetic L2 fluency and can benefit both language learners and teachers, as well as developers of automatic assessment of L2 speech.

In the following discussion, the results from comparisons between simple and multiple linear regression models are analyzed in relation to the research questions, and the strengths and limitations of this research are reviewed.

4.1 Phonetic Fluency Measures Contributing to Perceived Fluency and Proficiency

The effect of phonetic fluency measures on human ratings was studied using simple and multiple linear regression models. After a comparison of simple linear regression models where each phonetic fluency measure was treated as the independent variable, the following fluency measures were found statistically significant predictors of both fluency and proficiency ratings: articulation rate, speech rate, portion of long silent pauses (LSPratio), mean duration of long silent pauses (meanLSP), mean duration of breaks between utterances (meanB), and rate of short silent pauses per minute (SPrate). The mean duration of utterance (meanU) was only significant for fluency ratings, whereas the portion of short silent pauses (SPratio) was only significant for proficiency ratings.

Speech rate was the best predictor for both fluency and proficiency ratings among the simple linear regression models. This was expected, as faster rate has many times proven to be related to fluent speech (Kallio et al., 2017; Kormos & Dénes, 2004; Cucchiarini et al., 2002; Kallio et al., 2022; Préfontaine et al., 2016). Unlike with articulation rate, calculating speech rate takes pauses and other disfluencies into consideration: more disfluencies in the speech sample results in lower speech rate, even if the speech segments themselves were spoken rapidly. With speech rate as good predictor in fluency ratings, it might even be the pauses and disfluencies that actually produce this outcome, as combining speech rate with other phonetic fluency measures did not improve the model. Articulation rate also proved to be one of the best predictors of fluency and proficiency ratings, which is again a familiar outcome from previous fluency research (Kallio et al., 2017; Kallio et al., 2022; Préfontaine et al., 2016). Combining articulation rate in multiple linear regression models with other fluency measures, such as LSPratio, improved the simple linear regression models significantly. This suggests that the fluency ratings are affected by a combination of different phonetic fluency measures, and that the raters ground their assessments on this combination, although some phonetic fluency measures might be more important on their own than others. This finding is also supported by the significance of meanB in predicting fluency and proficiency ratings, as it takes both pauses and repair fluency measures into account. Combining articulation rate with meanB as predictors was significant for both fluency and proficiency ratings. Previous fluency research has also found the mean duration of utterance break as a significant predictor of fluency and proficiency (Kallio et al., 2022).

Long silent pauses were also strongly connected with the ratings, as LSPratio and meanLSP were highly significant in the simple linear regression models for both fluency and proficiency ratings. This supports previous research where the mean duration of silent pauses above 250 ms predicted fluency ratings well (Préfontaine et al., 2016; Bosker et al., 2013). In the research by Kallio et al. (2017), only the portion of silent pauses above 1,000 ms was found significant in predicting fluency ratings, and not silent pauses between 200 and 1,000 ms. In this thesis, silences above 250 ms were not divided into categories, and thus similar conclusions cannot be made. A multiple linear regression model combining meanLSP and meanU as predictors created one of the best models in predicting the fluency level. For LSPrate, the outcome was not significant. One reason to this might be that LSPratio and meanLSP depend on the overall durations of long silent pauses, and LSPrate on the frequency – a sample with a few silences that last for many seconds might appear less fluent than a sample with many silences that last for only 250 ms each. However, combining LSPrate with speech rate in a multiple linear regression model, the effect was statistically significant when predicting proficiency ratings.

Interestingly, the rate of short silent pauses per minute (SPrate) had a small yet significant positive effect on fluency and proficiency ratings. The effect of the portion of short silent pauses (SPratio) on proficiency ratings was positive as well. In other words, having brief silences in a speech sample resulted in higher ratings for fluency and proficiency. In a previous study concerning L2 Swedish, the portion of short silences (50–200 ms) had a negative effect on fluency ratings in read speech, and in spontaneous speech, the effect was not statistically significant (Kallio et al., 2017). There may be different expectations when it comes to prosody in different types of speech. Spontaneous speech requires more content planning and is prone to more disfluencies, whereas in read speech, the speaker knows what to say, which usually results in more fluent speech. Among spontaneous speech samples, fluent ones might have a lower mean duration of silences overall, which could decrease the amount of long pauses and increase the amount of brief pauses. The speech samples of Kallio et al. (2017) contained quite brief spontaneous responses, whereas in this research, the task required longer responses. Thus, it is possible that when listening to a longer, spontaneous monologue, brief silences might even be expected. These findings indicate that shorter silences can increase fluency, and this is important to take into consideration when developing automated assessment for spoken L2 – depending on the task type and the expected duration of a response, the automated assessment criteria could vary if the aim is to provide human-like assessment.

In the simple linear regression models, measures concerning filled pauses (FPratio, FPrate, and meanFP) were not significant in predicting the human ratings. Filled pauses were characterized by Cucchiarini et al. (2002) as “secondary variables”, since they are not as common as silent pauses in speech. Based on the linear regression results, the effect of filled pauses is clearly different from that of silent pauses. Filled pauses seemed to appear several times together with silent pauses – perhaps their effect was more significant in the break intervals, but this is not clear from the analyses.

Of all compared linear regression models, speech rate (adj. R^2 = 0.594) and a combination of articulation rate and LSPratio (adj. R^2 = 0.604) explained the variance in fluency ratings the best. For proficiency ratings, the best predictors were speech rate (adj. R^2 = 0.316) and a combination of speech rate and meanLSP (adj. R^2 = 0.376). However, the fluency measure combinations in multiple linear regression models for proficiency ratings presented in this work did not produce significantly different results. The adjusted R^2 values remained lower for proficiency than for fluency ratings, which suggests that there were other important aspects that affected the perceived proficiency level. It is very likely that, for instance, the use of grammar and vocabulary affected the perceived proficiency. For fluency ratings, the relatively high adjusted R^2 values indicate that most of the variance in fluency ratings can be explained with speech rate or a combination of multiple phonetic fluency measures.

The effects of speech rate and the mean duration of breaks between utterances (meanB) on fluency and proficiency ratings were also examined by treating the ratings as discrete variables instead of continuous (Figures 8 & 9). The created plots were interesting: There was a clear rising trend for speech rate for the fluency levels from 1 to 4. For proficiency, speech rate was relatively similar for levels B1–C1, whereas levels A1 and A2 had clearly lower and level C2 clearly higher rate. This implies that while other language skills related to proficiency develop through different levels, speech rate might remain stationary. For instance, Tavakoli et al. (2020) have found that speed, including speech rate, can be used for distinguishing between proficiency levels apart from levels B2 and C1, where they found no statistical difference. The measure meanB was examined separately since its effect on fluency and proficiency is less recognized in L2 fluency research. For the values of meanB, the trend was descending from lower levels to higher ones, although the values seemed to be closer to each other in general than for speech rate. For fluency ratings, meanB was a lot higher on level 1 than on the others. As there were only two speech samples for this level, reliable conclusions cannot be made, although meanB can be expected to be the highest on level 1. For the proficiency ratings, meanB was again very similar for levels B1–C1, as with speech rate, while meanB on level A2 was clearly higher and on level C2 clearly lower. The preceding observations were merely based on the plotted figures, and the differences between the levels were not tested statistically. However, these notes raise interest towards the evolution of different areas of language proficiency, and further research could be conducted similarly to Tavakoli et al. (2020) when it comes to more specific differences between individual fluency and proficiency levels.

4.2 Limitations and Future Research

The size of the data brought some inevitable limitations to this research. The total duration of the 53 analyzed speech samples was approximately 40 minutes. Although the speakers were instructed to continue speaking for one minute, some samples were less than 20 seconds in duration. This might have affected the amount of corrections and repetitions in the samples, as the frequencies were so low that measures regarding specifically them (CRrate, CRratio, and meanCR) were excluded from the analyses. With more speech data, it is likely that the different types of disfluencies would have been represented better, and the contribution of corrections and repetitions to human assessments could have been studied as well. On the other hand, it might be that these types of disfluencies are simply not that common in high school students' spontaneous speech in L2 Finnish. The effect of corrections and repetitions on fluency ratings has typically been weaker than the effect of speed or breakdown phenomena (Cucchiarini et al., 2002; Bosker et al., 2013), and it might have more significance in read speech than in spontaneous speech where disfluencies are better tolerated overall (see Kallio et al., 2017). In this research, corrections and repetitions were defined as partial words – they could have included complete words as well, which would have increased their frequency. When it comes to automatic methods, identifying corrections and repetitions can be difficult since they are simply less frequent in speech than, for instance, pauses (Hsieh et al., 2020: 106). Moreover, it is not always clear which parts in speech are disfluencies; words that seem partial can also be spoken language in Finnish, and not disfluencies, and repeating complete words can be a way to show emphasis.

The speakers had participated in an examination containing eight spoken tasks. The analyzed task was chosen for this research since it elicited spontaneous speech, and the responses were among the longest. Some samples were still relatively short, and according to the assessors, evaluating spoken language skills, especially overall proficiency, is very difficult with brief samples (von Zansen et al., 2022). Nonetheless, the inter-rater agreement for the analyzed task was relatively high, and overall slightly higher for proficiency ratings than for fluency. Some of the raters were likely more used to assessing a speaker's proficiency level than using specific analytic rating scales, as for fluency (von Zansen et al., 2022). Differences between individual raters were not examined in detail. One of the raters was a non-native speaker of Finnish, and it would have been interesting to see whether they assessed the samples differently from the native Finnish speakers.

The analyses did not consider the locations of pauses, which would have been useful to study since placing longer pauses in the middle of a clause might affect perceived fluency (Skehan, 2009). Simple and multiple linear regression were used for statistical analyses, but the interactions between the phonetic fluency measures were not examined – this could be studied further.

Although the overall representation of different fluency (1–4) and proficiency levels (A1–C2) was good among the speech samples, the amount of samples in the lowest levels was small (only two samples were rated as 1 for fluency and A1 for proficiency), and conclusions based on statistical analyses are not as reliable for these than for levels with over 20 samples. However, observations regarding speech rate and meanB over discrete fluency and proficiency levels revealed that there might be interesting patterns in the evolution of specific phonetic fluency measures through the different levels. In further research, this could be studied more.

5 Conclusions

In this thesis, high school students’ spontaneous speech samples in L2 Finnish were analyzed using acoustic measurements. The effects of different phonetic fluency measures on human ratings for fluency and overall proficiency were studied to find out which features in speech are the most connected with the perception of fluency and proficiency. The results showed that there were several phonetic fluency measures that were able to predict human ratings for fluency and proficiency in L2 Finnish, and the best individual predictors were related to speed, long silent pauses, and breaks between utterances. However, the perception of fluency and proficiency is usually affected by a combination of multiple fluency features in speech.

Automatic tools are developed to assess spoken language proficiency more efficiently and reliably. Automatically calculated phonetic fluency measures have a relatively strong connection with human ratings for fluency and proficiency. By expanding previous knowledge on phonetic fluency in L2 Finnish, the findings of this thesis also advance the development of automated proficiency assessment of spoken L2 Finnish.

References

Boersma, P, & Weenink, D. (2022). Praat: doing phonetics by computer. [Computer program]. Version 6.2.12. http://www.praat.org/

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

Brown, A. (2013). Uses of language assessments. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing Ltd.

Campione, E. & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Proceedings of Speech Prosody 2002.*

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*, 2862–2873. https://doi.org/10.1121/1.1471894

de Jong, N. H. & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In *Proceedings of DiSS 2013,* 17–20.

de Wet, F., Van der Walt, C., & Niesler, T. R. (2009). Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication, 51*(10), 864–874.

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

Dowle, M. & Srinivasan, A. (2021). data.table: Extension of `data.frame`. R package version 1.14.2. https://CRAN.R-project.org/package=data.table

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

Fontan, L., Le Coz, M., & Detey, S. (2018). Automatically measuring L2 speech fluency without the need of ASR: a proof-of-concept study with Japanese learners of French. In *Proceedings of Interspeech 2018,* 2544–2548.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Gass, S. M. & Selinker, L. (2008). Second language acquisition: An introductory course. Routledge.

Hsieh, C.-N., Zechner, K., & Xi, X. (2020). Features measuring fluency and pronunciation. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 101–122). Routledge. https://doi.org/10.4324/9781315165103-7

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly, 15*(3), 273–293. https://doi.org/10.1080/15434303.2018.1472264

Kallio, H., Suviranta, R., Kuronen, M., & von Zansen, A. (2022). Creaky voice and utterance fluency measures in predicting perceived fluency and oral proficiency of spontaneous L2 Finnish. In *Proceedings of Speech Prosody 2022*.

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), 193–213. https://doi.org/10.30660/afinla.73137

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. In *Proceedings of Interspeech 2016.*

Kautonen, M. & von Zansen, A. (2020). DigiTala research project: Automatic speech recognition in assessing L2 speaking. *Kieli, koulutus ja yhteiskunta*, 11(4).

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145–164.

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

Lintunen, P., Mutta, M., & Peltonen, P. (Eds.). (2019a). *Fluency in L2 learning and use.* Multilingual Matters.

Lintunen, P., Mutta, M., & Peltonen, P. (2019b). Defining fluency in L2 learning and use. In P. Lintunen, M. Mutta, & P. Peltonen (Eds.), *Fluency in L2 learning and use* (pp. 1–15). Multilingual Matters. https://doi.org/10.21832/9781788926317-003

Ministry of Education and Culture. (2017). *Gaudeamus igitur – ylioppilastutkinnon kehittäminen*. Opetus- ja kulttuuriministeriön julkaisuja 2017:16. http://urn.fi/URN:ISBN:978-952-263-462-7.

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

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team. (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. URL http://www.rstudio.com/.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia 126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal, 104*(1), 169–191. https://doi.org/10.1111/modl.12620

Toivola, M. (2011). Vieraan aksentin arviointi ja mittaaminen suomessa. Helsingin yliopisto. http://urn.fi/URN:ISBN:978-952-10-7217-8

Toivola, M., Lennes, M., & Aho, E. (2009). Speech rate and pauses in non-native Finnish. In *Proceedings of Interspeech 2009,* 1707–1710.

von Zansen, A., Kallio, H., Sneck, M., Kuronen, M., Huhta, A., & Hildén, R. (2022). *Ihmisarvioijien näkemyksiä suullisen kielitaidon automaattisesta arvioinnista, digitaalisesta arviointiprosessista sekä arvioitavista puheen ulottuvuuksista*. [Unpublished manuscript].

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40*(1), 1–29. http://www.jstatsoft.org/v40/i01/.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. http://www.jstatsoft.org/v21/i12/.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition, 27*(4), 567–595. doi:10.1017/S0272263105050254

Appendices

Appendix A. Task 2 instructions for the DigiTala oral examination in Finnish

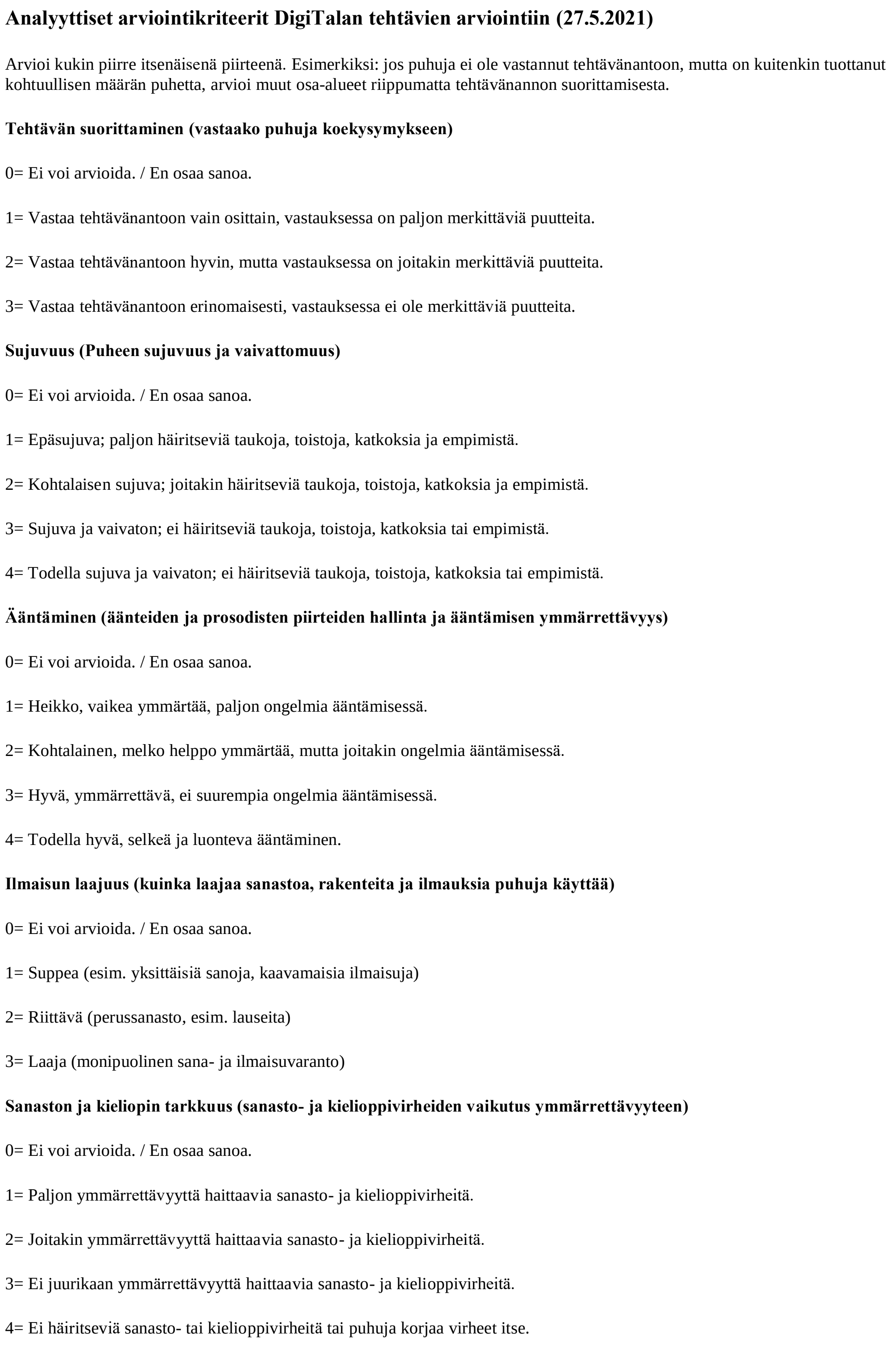
Tehtävä 2) Tärkeä paikka

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?

Appendix B. Analytic rating criteria for spoken L2 assessments in Finnish



Analyyttiset arviointikriteerit DigiTalan tehtävien arviointiin (27.5.2021)

Arvioi kukin piirre itsenäisenä piirteenä. Esimerkiksi: jos puhuja ei ole vastannut tehtävänantoon, mutta on kuitenkin tuottanut kohtuullisen määrän puhetta, arvioi muut osa-alueet riippumatta tehtävänannon suorittamisesta.

Tehtävän suorittaminen (vastaako puhuja koekysymykseen)

0= Ei voi arvioida. / En osaa sanoa.

1= Vastaa tehtävänantoon vain osittain, vastauksessa on paljon merkittäviä puutteita.

2= Vastaa tehtävänantoon hyvin, mutta vastauksessa on joitakin merkittäviä puutteita.

3= Vastaa tehtävänantoon erinomaisesti, vastauksessa ei ole merkittäviä puutteita.

Sujuvuus (Puheen sujuvuus ja vaivattomuus)

0= Ei voi arvioida. / En osaa sanoa.

1= Epäsujuva; paljon häiritseviä taukoja, toistoja, katkoksia ja empimistä.

2= Kohtalaisen sujuva; joitakin häiritseviä taukoja, toistoja, katkoksia ja empimistä.

3= Sujuva ja vaivaton; ei häiritseviä taukoja, toistoja, katkoksia tai empimistä.

4= Todella sujuva ja vaivaton; ei häiritseviä taukoja, toistoja, katkoksia tai empimistä.

Ääntäminen (äänteiden ja prosodisten piirteiden hallinta ja ääntämisen ymmärrettävyys)

0= Ei voi arvioida. / En osaa sanoa.

1= Heikko, vaikea ymmärtää, paljon ongelmia ääntämisessä.

2= Kohtalainen, melko helppo ymmärtää, mutta joitakin ongelmia ääntämisessä.

3= Hyvä, ymmärrettävä, ei suurempia ongelmia ääntämisessä.

4= Todella hyvä, selkeä ja luonteva ääntäminen.

Ilmaisun laajuus (kuinka laajaa sanastoa, rakenteita ja ilmauksia puhuja käyttää)

0= Ei voi arvioida. / En osaa sanoa.

1= Suppea (esim. yksittäisiä sanoja, kaavamaisia ilmaisuja)

2= Riittävä (perussanasto, esim. lauseita)

3= Laaja (monipuolinen sana- ja ilmaisuvaranto)

Sanaston ja kieliopin tarkkuus (sanasto- ja kielioppivirheiden vaikutus ymmärrettävyyteen)

0= Ei voi arvioida. / En osaa sanoa.

1= Paljon ymmärrettävyyttä haittaavia sanasto- ja kielioppivirheitä.

2= Joitakin ymmärrettävyyttä haittaavia sanasto- ja kielioppivirheitä.

3= Ei juurikaan ymmärrettävyyttä haittaavia sanasto- ja kielioppivirheitä.

4= Ei häiritseviä sanasto- tai kielioppivirheitä tai puhuja korjaa virheet itse.

Appendix C. Holistic rating criteria for spoken L2 assessments in Finnish

Taitotaso	**Kuvaus puhumisesta (A1-C1: LOPS 2003, C2: EVK 2003)**
Alle A1	• Pystyy tuottamaan vain joitakin yksittäisiä, irrallisia sanoja kohdekielellä.
EROT ALLE A1 > A1	***Osaa tuottaa hieman enemmän kuin vain yksittäisiä, irrallisia sanoja kohdekielellä.***
A1	• Osaa kertoa lyhyesti itsestään ja lähipiiristään, selviytyy kaikkein yksinkertaisimmista vuoropuheluista ja palvelutilanteista • Tauot, toistot ja katkokset ovat yleisiä • Ääntäminen voi tuottaa ymmärtämisongelmia • Osaa suppean perussanaston, perustason lauserakenteita sekä ulkoa opeteltuja ilmauksia ja fraaseja • Kielioppivirheitä esiintyy paljon vapaassa puheessa
EROT A1 > A2	***Osaa kertoa enemmän tutuista aiheista, ääntäminen pääsääntöisesti ymmärrettävää, laajempi perussanasto.***
A2	• Selviytyy yksinkertaisista sosiaalisista kohtaamisista, osaa aloittaa ja lopettaa lyhyen vuoropuhelun • Puheessa voi olla välillä sujuvaa, mutta taukoja, katkoksia ja vääriä aloituksia esiintyy paljon • Ääntäminen on ymmärrettävää, mutta satunnaisia ymmärtämisongelmia voi esiintyä ääntämisen takia • Hallitsee perussanaston ja perusrakenteita sekä joitakin idiomaattisia ilmauksia • Hallitsee yksinkertaisimman peruskieliopin, mutta virheitä voi esiintyä paljon perusrakenteissakin
EROT A2 > B1	***Viestii myös hieman vaativammissa tilanteissa pääasiassa sujuvasti, ääntämisestä ei ymmärtämisongelmia, laajempi sanasto ja rakenteita, kielioppivirheitä kuitenkin esiintyy.***
B1	• Osaa kuvailla konkreetteja aiheita, selviytyy tavallisimmista arkitilanteista, mutta ilmaisu ei välttämättä ole kovin tarkkaa • Osaa pitää yllä melko sujuvaa puhetta • Ääntäminen on ymmärrettävää, mutta ääntämisvirheitä, kohdekielelle epätyypillistä intonaatiota ja painotusta esiintyy • Käyttää melko laajaa sanastoa ja tavallisia idiomeja, erilaisia rakenteita ja lauseita • Kielioppivirheitä esiintyy, mutta ne haittaavat harvoin viestin välittymistä
EROT B1 > B2	***Ilmaisu tarkempaa myös spontaanisti, myös käsitteellisiä aiheita, tilannetaju, ääntäminen ja intonaatio luontevia, laajempi sanaston ja rakenteiden hallinta, satunnaisia kielioppivirheitä.***
B2	• Osaa ilmaista itseään varmasti, selkeästi ja kohteliaasti tilanteen vaatimalla tavalla, osaa keskustella monista asioista, mutta tarvitsee joskus kiertoilmauksia • Puhuu sujuvasti myös spontaanisti, puheessa on harvoin pidempiä taukoja tai epäröintiä • Ääntäminen on ymmärrettävää, ääntäminen ja intonaatio ovat selkeitä ja luontevia • Laajahkoa sanastoa konkreeteista ja käsitteellisistä sekä tutuista ja tuntemattomista aiheista, monipuolisia rakenteita • Kieliopin hallinta on hyvää, satunnaiset kielioppivirheet eivät vaikuta ymmärrettävyyteen, korjaa välillä ne itse
EROT B2 > C1	***Myös monimutkaisia käsitteellisiä ja yksityiskohtia sisältäviä tilanteita, puhe lähes vaivatonta, ilmaisee merkitysvivahteita ääntämisen (intonaatio ja painotus) avulla, sanaston ja rakenteiden hallinta ei rajoita ilmaisua, korjaa kielioppivirheet tarvittaessa itse.***
C1	• Osallistuu aktiivisesti monimutkaisiin käsitteellisiä ja yksityiskohtia sisältäviin tilanteisiin, selviää monenlaisesta sosiaalisesta vuorovaikutuksesta tilanteen vaatimalla tavalla • Puhe on sujuvaa, spontaania ja lähes vaivatonta • Ääntäminen on ymmärrettävää, vaihtelee intonaatiota ja hallitsee lausepainot • Sanasto ja rakenteet ovat laajat, eivätkä juuri rajoita ilmaisua • Kieliopin hallinta on hyvää, satunnaiset kielioppivirheet eivät vaikuta ymmärrettävyyteen, korjaa ne itse
EROT C1 > C2	***Puhe erittäin sujuvaa ja tyyliltään tilanteeseen sopivaa, ilmaisee hienojakin merkitysvivahteita. Kieliopin ja sanaston hallinta on varmaa lähes kaikissa tilanteissa.***
C2	• Osallistuu vaivatta kaikenlaisiin keskusteluihin tilanteen ja puhekumppanien edellyttämällä tavalla, välittää täsmällisesti hienojakin merkitysvivahteita • Puhuu sujuvasti, luontevasti ja epäröimättä myös pitkäkestoisessa puhetilanteessa • Ääntäminen on täysin ymmärrettävää, vaihtelee intonaatiota ja hallitsee lausepainot • Ilmaisu täsmällistä ja asianmukaista, merkitysvivahteetkin välittyvät, käyttää idiomaattisia tai puhekielisiä ilmauksia, sanasto ja rakenteet eivät rajoita ilmaisua • Hallitsee vaativatkin rakenteet, korjaa tarvittaessa ilmaisuaan, kiertää vaikeudet

Kussakin tasossa kuvattu 1. Kielenkäyttötilanteet 2. Sujuvuus 3. Ääntäminen, 4. Sanasto, 5. Kielioppi (19.5.2021)