Toivola, M., P. Lintunen & L. M. Heikkola (toim.) 2024. Puheen tutkimuksen uusia suuntia – Aineistona vapaasti tuotettu puhe. New directions in speech research – Freely produced speech as data. AFinLA-teema / n:o 17, 91–116.

Anna von Zansen
Helsingin yliopisto

Heini Kallio
Tampereen yliopisto

DigiTala – Moodle-sovellus suullisen kielitaidon automaattiseen arviointiin

Nostot

- DigiTala kehitti puheentunnistukseen ja koneoppimiseen perustuvan työkalun suomen ja ruotsin oppijoille.
- Työkalu antaa automaattista palautetta myös spontaaneista puhenäytteistä.
- Automaattinen palaute antaa tietoa puhujan kehittämiskohteista sekä konkretisoi taitotasoja.
- Artikkelissa esittelemme työkalun kehittämisvaiheet ja ominaisuudet.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Abstract

This article presents an online tool developed by the DigiTala research project in 2022. The tool is the first one to provide comprehensive automated assessment of spontaneous L2 Finnish and Swedish speech. The Moodle-based tool records and analyzes L2 learners' speech automatically and provides holistic as well as analytic feedback to learners. In addition to self-regulated learning purposes, the tool can be used to support human assessment in high-stakes tests and language courses. We discuss the benefits and weaknesses of the tool and automated speaking assessment in general from the perspectives of low-resourced languages as well as developers and users.

Keywords: automated assessment, spontaneous speech, L2 skills, agile software development
Asiasanat: automaattinen arviointi, spontaani puhe, suullinen kielitaito, ketterä ohjelmistokehitys

1 Johdanto

Suullista kielitaitoa tarvitaan yhteiskunnassa ja yhteisöissä toimimiseen. Kielenoppijat tarvitsevat paljon tilaisuuksia käyttää kieltä, jotta he saavat valmiuksia toimia kohdekielellä arkisissa kielenkäyttötilanteissa. Lisäksi oppijat saattavat tarvita harjoitusta ääntämiseen, sujuvuuteen tai ilmausten laajuuteen liittyen voidakseen itsevarmasti käyttää kieltä esimerkiksi työelämässä. Erilaisissa elämäntilanteissa kielenoppijoilla ei kuitenkaan aina ole tarvittavaa mahdollisuutta, motivaatiota tai uskallusta hakeutua aitoihin vuorovaikutustilanteisiin harjoittelemaan kohdekieltä. Lisäksi vaikka oppija hakeutuisi kielikurssille, oppituntien ja saatavilla olevan ohjauksen määrä ei välttämättä riitä suullisen kielitaidon kehittämiseen toivotussa määrin.

Puheteknologia mahdollistaa ajasta ja paikasta riippumattoman suullisen kielitaidon harjoittelun. Vuorovaikutustaitojen kehittymisen kannalta etenkin spontaanin puheen[1] harjoittelu on tärkeää. Automaattisten puheen arviointimenetelmien kehittäjät ovat kuitenkin suosineet lukupuhetta (*read-aloud*) sen ennustettavuuden kannalta (Evanini & Zechner 2020). Puheentunnistimet suoriutuvat paremmin luetun tai imitoidun kuin spontaanin puheen tunnistamisesta, ja saman sisältöiset puhenäytteet mahdollistavat esimerkiksi tarkempien ääntämismallien kehittämisen (Bernstein ym. 1990). 2000-luvulla tutkijoiden huomio laajeni spontaanin puheen automaattiseen arviointiin. Tällöin tutkimus alkoi puheen sujuvuuden mittaamisesta (Cucchiarini ym. 2002) ja laajeni lopulta useiden kielitaitoa eri näkökulmista mittaavien piirteiden tarkasteluun (Xi ym. 2008). Spontaanin puheen automaattista arviointia on kehitetty kuitenkin lähinnä suurille kielille, kuten englannille, jossa on tarjolla paljon puheaineistoa niin äidinkielisiltä kuin kielenoppijoiltakin.

1 Käytämme termiä *spontaani puhe* kuvaamaan ns. vapaata puhetta, jossa puhuja tuottaa sisältöä itse ääneen lukemisen tai perässä toistamisen sijaan.

DigiTala-tutkimushankkeen (Kautonen & von Zansen 2020) Moodle-sovellus (von Zansen ym. 2022a) on ensimmäinen kotimaisten kielten automaattiseen arviointiin kehitetty työkalu, joka kattaa useita kielitaidon osa-alueita ja arvioi myös spontaaneja puhenäytteitä. Automaattiseen puheen tunnistukseen (*automatic speech recognition*), arviointiin (*automated assessment*) ja palautteeseen (*automated feedback*) perustuvan työkalun kohderyhmänä ovat suomen ja ruotsin oppijat, jotka haluavat harjoitella puheen tuottamista ja ääntämistä osana kieltenopetusta tai itsenäisesti. Suomi ja suomenruotsi ovat maailmanlaajuisesti tarkasteltuna pieniä kieliä, joissa sekä äidinkielisiä puhujia että kielten oppijoita on suhteellisen vähän. Puheteknologian näkökulmasta nämä kielet ovat vähäresurssisia: soveltuvaa aineistoa kuten kielenoppijoiden spontaania puhetta on rajallisesti saatavilla. Teknologisia sovelluksia kehittämällä edistämme näiden pienten kielten opiskelua.

Puhumisen harjoittelun lisäksi tutkimushankkeen kehittämää puheen automaattista arviointia voitaisiin hyödyntää yksilön tulevaisuuden kannalta tärkeissä (*high-stakes*) kielikokeissa, kuten yleisten kielitutkintojen ja ylioppilastutkinnon kielikokeissa. Yleiset kielitutkinnot toimivat portinvartijana (*gate-keeper*) esimerkiksi Suomen kansalaisuutta hakiessa. Ylioppilastutkinnon arvosanoja puolestaan hyödynnetään muun muassa jatko-opintoihin hakeutuessa. Työkalua voidaan käyttää tärkeisiin kokeisiin valmistautumisen lisäksi ihmisarvioinnin tukena. Yleisiin kielitutkintoihin kuuluu puhumisen osakoe, mutta ylioppilastutkinnon kielikokeista se edelleen puuttuu ratkaisemattomien käytännön kysymysten takia (Vaarala ym. 2021).

Tämän artikkelin tavoitteena on luoda katsaus puheen automaattista arviointia koskevaan aiempaan tutkimukseen, dokumentoida työkalun kehittämisvaiheet sekä tarjota läpinäkyvä ja ymmärrettävä kuvaus DigiTalan kehittämästä arviointityökalusta. Automaattisen arviointityökalun toimintaperiaatteiden kuvaaminen on välttämätöntä, jotta sen toimivuutta voidaan arvioida niin pedagogisesti kuin teknisestikin. Artikkelissa esittelemme DigiTalan Moodle-sovelluksen tutkimusperustaisen taustan, sovelluksen kehitysvaiheet sekä itse sovelluksen toiminnot. Pohdimme myös kielenoppijan puheen automaattisen arvioinnin vahvuuksia ja heikkouksia vähäresurssisten kielten sekä kehittäjien ja käyttäjien näkökulmista.

2 Tausta

Seuraavissa alaluvuissa esittelemme puheen automaattisen arviointiin liittyviä lähtökohtia (luku 2.1) sekä aiempaa tutkimusta (luku 2.2).

2.1 Puheen automaattisen arvioinnin lähtökohtia

Suullisen kielitaidon automaattisen arvioinnin tärkein komponentti on tarkka kohdekielinen puheentunnistin. Tunnistin muuntaa puheen tekstiksi, josta voidaan laskea automaattisesti erilaisia ääntämisen laatuun, kielioppiin ja sanastoon liittyviä piirteitä (Hsieh ym. 2020; Yoon ym. 2020). Lisäksi suoraan puhesignaalista voidaan mitata puheen sujuvuutta laskemalla mm. puhenopeutta ja taukojen ja epäröintien määrää tai kestoa (Hsieh ym. 2020). Kielenoppijan puheen tunnistaminen on kuitenkin hankalampaa kuin äidinkielisen puhujan, koska kielenoppijan puhe on usein epäsujuvaa ja siinä voi esiintyä erilaisia kielioppi-, sanasto- ja ääntämisvirheitä. Puheentunnistinten akustiset mallit pohjautuvat äidinkielisten puhujien näytteisiin, mutta automaattista arviointia varten tunnistin täytyy mukauttaa kielenoppijan puheeseen opettamalla sitä eritasoisten oppijoiden puheella (Ylinen & Kurimo 2017).

Automaattisten arviointimallien pohjana käytetään ihmisten tekemiä arvioita eritasoisten kielenoppijoiden puheesta. Arviointialgoritmit vertaavat ihmisten arvioimien puhenäytteiden laskennallisia piirteitä testattavien puhenäytteiden piirteisiin. (Loukina & Yoon 2020.) Jokaiselle arvioitavalle suullisen kielitaidon ulottuvuudelle kehitetään oma arviointimallinsa. Esimerkiksi ETS:n (English Testing Services) kehittämässä automaattisessa arviointijärjestelmässä on holistisen taitotason lisäksi kolme arvioitavaa ulottuvuutta: ulosanti (ääntäminen ja puheen sujuvuus), kielenkäyttö (kieliopin tarkkuus ja sanaston laajuus) sekä aiheenmuodostus (sisällön kattavuus ja koherenssi) (Loukina & Yoon 2020).

DigiTalan työkalussa arvioitavat puheen ulottuvuudet ovat taitotason lisäksi tehtävänannon täyttyminen, sujuvuus, ääntäminen ja ilmauksen laajuus (ks. von Zansen ym. 2022c). Tutkimukseen (ks. luku 2.2) ja koneoppimiseen pohjautuva arviointimalli huomioi sen, mitä tehtävän vastaukselta odotetaan sekä määrittää, mitä puheen piirteitä käytetään kyseisen ulottuvuuden arvioimiseen ja miten kutakin piirrettä painotetaan arvioinnissa.

Useimmiten arviointimallin kehittämiseen on käytetty ohjattua oppimista: puheaineisto, josta algoritmi oppii, on ihmisten esikäsittelemää (Evanini & Zechner 2020). Aineisto voi olla nimikoitua tai siitä on muodostettu erilaisia funktioita, joiden avulla arviointimalli päättelee, miten uusia puhenäytteitä tulisi arvioida. Opetusaineistoon voidaan esimerkiksi luokitella sana- ja lausepainoja, joiden toteutumista arviointimalli vertaa arvioitavan puhenäytteen ja opetusaineiston kesken (Hsieh ym. 2020). Puheen piirteiden laskentaan käytetään kuitenkin enenevässä määrin ohjaamatonta oppimista, jolloin algoritmi oppii aineistosta itsenäisesti ilman ihmisen väliintuloa (Al-Ghezi ym. 2023).

Oikeudenmukaisuuden ja luotettavuuden kannalta automaattisen arvioinnin läpinäkyvyys (*transparency*) on tärkeää (Evanini & Zechner 2020; von Zansen & Heijala 2023). Ohjattuun oppimiseen perustuvien arviointimallien toimintaperiaatteet tiedetään, jolloin työkalun käyttäjälle voidaan tarjota tietoa siitä, mihin automaattinen

arvio perustuu. Tämä esimerkiksi mahdollistaa analyyttisen automaattisen palautteen näyttämisen oppijalle (ks. von Zansen & Heijala 2023; von Zansen & Huhta 2022).

Puheen automaattiseen arviointiin liittyy kuitenkin rajoitteita. Automaattiset arviointityökalut mahdollistavat nykyään erityyppisiä puhumisen tehtäviä, mutta tehtävät ovat yleensä rajattuja, koska työkalut on säädetty tiettyä tarkoitusta varten (Ylinen & Kurimo 2017). Rajoitetuimmat tehtävät sisältävän ääneen lukua tai lyhyitä toistoja mallin perässä, ja näistä suorituksista voidaan arvioida vain ääntämistä ja sujuvuutta (von Zansen ym. 2022c). Avoimemmista tuottamistehtävistä, kuten kuvasta kertomisesta tai annetusta aiheesta puhumisesta, voidaan arvioida ääntämisen ja sujuvuuden lisäksi myös kieliopin tarkkuutta ja ilmaisun laajuutta. Spontaanit puhetehtävät ovat lähempänä autenttisia kielenkäyttötilanteita kuin ääneen lukeminen. Kaikkia suullisen kielitaidon ulottuvuuksia ei kuitenkaan päästä vielä arvioimaan automaattisesti: nykyisistä työkaluista puuttuvat vuorovaikutteiset tehtävät (kielenoppijoiden suullisen vuorovaikutuksen automaattiseen arviointiin tähtäävästä tutkimushankkeesta ks. von Zansen 2023).

2.2 Aiempi tutkimus puheen automaattisesta arvioinnista

Luomme tässä luvussa katsauksen aiempaan sekä DigiTala-hankkeessa tehtyyn monitieteiseen tutkimukseen. Esittelemme ensin kielididaktista ja foneettista tutkimusta, jonka jälkeen esittelemme hankkeen kieliteknologista tutkimusta.

2.2.1 Kielididaktinen tutkimus

Puheen automaattiseen arviointiin liittyvä kielididaktinen tutkimus ulottuu ensinnäkin mitattavaan käsitteeseen (*construct*), tehtävänlaadintaan (*task design*) sekä tavoiteltavaan puhesuoritukseen liittyviin tarkasteluihin (Luoma 2004; Fulcher 2014). Toiseksi automaattisten arviointimallien kehitys perustuu yleensä ihmisten tekemiin arviointeihin (Evanini & Zechner 2020), joten arviointien laadun ja skaalojen toimivuuden (ks. tarkemmin luku 3.4) tutkiminen on keskeistä. Lisäksi käytettävien arviointikriteerien tulee olla linjassa puheesta mitattavien piirteiden (*speech feature*, Evanini & Zechner 2020, ks. tarkemmin foneettista tutkimusta koskeva alaluku 2.2.2) kanssa. Kolmas keskeinen aihe kielididaktisessa tutkimuksessa liittyy automaattisen arvioinnin käyttöön. Arviointityökalujen käytettävyyttä ja hyödyllisyyttä tutkitaan esimerkiksi kartoittamalla kielenoppijoiden, opettajien ja kielitaidon arvioijien näkemyksiä. Neljäntenä kielididaktisen tutkimuksen aiheena nostamme esiin puheen harjoittelua ja kielen oppimista tukevan automaattisen palautteen (Gu & Davis 2020).

Hankkeen kielididaktinen tutkimus on käsitellyt muun muassa ihmisarviointien laatua ja arviointiin osallistuvien (*stakeholders*) käsityksiä puheen automaattisesta arvioinnista. Osallistujien suhtautuminen puhetehtävien tai arvioinnin suorittamiseen tietokoneella on ollut positiivista, vaikka se on ollut heille pääosin uutta. Laatimamme

arviointikriteerit ovat osoittautuneet toimiviksi, ja kouluttamamme ihmisarvioijat ovat suoriutuneet arviointitehtävästä hyvin (von Zansen & Huhta 2022). Niin kouluttamiemme ihmisarvioijien (von Zansen ym. 2022c), lukiolaisten (von Zansen ym. 2022b) ja aikuisopiskelijoiden (von Zansen & Hilden 2022) kuin opettajienkin (von Zansen & Heijala 2023) tutkimuksen aineistonkeruun yhteydessä ilmaisemat näkemykset ovat tuoneet esiin alan kirjallisuudessakin (ks. esim. Evanini & Zechner 2020) esiintyviä mahdollisuuksia ja uhkia automaattiseen arviointiin liittyen.

Automaattisen arvioinnin kiistaton hyöty on mahdollisuus harjoitella puhumista ajasta ja paikasta riippumatta. Lisäksi kone arvioi kaikki puhujat väsymättä ja yhtenevillä kriteereillä. Automaattiseen arviointiin liittyvät huolet puolestaan liittyvät arvioinnin oikeudenmukaisuuteen: tunnistaako kone puheen oppijan erilaisesta aksentista tai puheeseen liittyvästä vaikeudesta huolimatta? (von Zansen ym. 2022b.) Tiettyjen ulottuvuuksien, kuten sujuvuuden ja taitotason, arvioiminen liian lyhyistä puhenäytteistä on puolestaan haastavaa niin koneelle kuin ihmisarvioijille (von Zansen ym. 2022c). Moodle on toiminut hyvin suoritusympäristönä kyseisille kohderyhmille, vaikkakin sen käyttö edellyttää jonkin verran teknisiä valmiuksia (ks. von Zansen & Hilden 2022).

Haastattelemamme opettajat käyttäisivät Moodle-työkalua esimerkiksi kotitehtävissä, suullisissa kokeissa ja osana lukiossa laadittavaa kieliprofiilia. Lisäksi muotoilemamme automaattinen palaute on heidän arvionsa mukaan oppijoille ymmärrettävää ja hyödyllistä, vaikkakin opettajan apua saatetaan tarvita palautteen välittämisessä näille. Arviointikriteerien pohjalta laaditut palauteväittämät eivät kuitenkaan sisällä kaikkia puhumisen ulottuvuuksia – opettajat lisäisivät esimerkiksi kehonkielen, sisällön ja rakenteiden virheettömyyden. (von Zansen & Heijala 2023.)

Työkalun yksi merkittävä hyöty liittyy taitotasoihin, jotka pohjautuvat Eurooppalaiseen viitekehykseen ja ovat toistakymmentä vuotta kuuluneet suomalaisiin kielten opetussuunnitelmiin (Opetushallitus 2003, 2019). Opetussuunnitelmista huolimatta taitotasoja ei ole kattavasti hyödynnetty kielten opetuksessa (ks. Härmälä & Marjanen 2023: 122), joten työkalun antama holistinen palaute konkretisoisi taitotasoja niin opiskelijoille kuin opettajille. Konkretisoimalla puheesta mitattavia piirteitä työkalu voisi yhdenmukaistaa arviointia. (von Zansen & Heijala 2023.)

2.2.2 Foneettinen tutkimus

Puheen automaattiseen arviointiin liittyvä foneettinen tutkimus on usein taustatutkimusta, jossa verrataan puheesta mitattavien piirteiden ominaisuuksia ihmisten tekemiin arvioihin kielitaidosta. Tutkimuksista saatava tieto on tärkeää piirrepohjaisten arviointimallien kehittämisessä. Lisäksi tutkimusten avulla voidaan päätellä, mitä piirteitä niin sanottu *black box* -malli (ks. esim. Zechner 2020) voisi käyttää puhenäytteiden arviointiin, kun halutaan tarjota kielenoppijalle sanallista palautetta numeerisen arvion lisäksi. Ihmisarvioiden ja puheen piirteiden yhteyttä on tutkittu suhteellisen paljon; etenkin englannin oppijoiden puhetta on tutkittu niin äänteiden

tuottamisen (esim. Munro & Derwing 1995; Isaacs & Trofimovich 2012) kuin puheen sujuvuudenkin näkökulmista (esim. Derwing ym. 2009; Tavakoli 2011; Kahng 2018; Kallio ym. 2022c; Peltonen & Lintunen 2022).

Sujuvuustutkimuksissa on todettu, että alkeisoppijoiden puhe on usein hitaampaa ja taukoja esiintyy useammin kuin edistyneempien kielenoppijoiden tai äidinkielisten puheessa. Ääntämisen arvioinnin tutkimuksissa korostuu virheettömyyden sijasta ymmärrettävyys, ja yksittäisten äännetason virheiden sijaan tutkijat ovat havainneet prosodisten piirteiden, kuten sana- ja lausepainotuksen, tauotuksen ja nopeuden, vaikuttavan voimakkaammin puheen ymmärrettävyyteen (Anderson-Hsieh ym. 1992; Isaacs & Trofimovich 2012; Kang 2012).

Hankkeessa kieltenoppijoiden puheen tutkimus on keskittynyt etenkin spontaaneihin puhunnoksiin. Ruotsinoppijoiden monologipuhetta tutkineet Kautonen ja Kuronen (2021) havaitsivat, että sanojen ääntäminen kehittyy lineaarisesti suhteessa taitotasoon, mutta lausetason prosodiassa esiintyy enemmän virheitä B2-tasolla kuin A1–B1-tasojen puhujilla. Tämä on todennäköisesti yhteydessä muun kielitaidon kehittymiseen: B2–C1-tasoilla puhujat tuottavat enemmän puhetta ja monimutkaisempia rakenteita, jolloin lauseprosodiaa on vaikeampi hallita. Tutkimustuloksia voidaan automaattisessa arvioinnissa hyödyntää siinä, miten puheen eri piirteitä painotetaan milläkin taitotasolla: kun tiedetään, että tietynlaiset prosodiset virheet yleistyvät muun kielitaidon kasvaessa, suhteutetaan näiden piirteiden esiintyminen esimerkiksi vastauksen syntaktiseen syvyyteen ja sanavaraston laajuuteen.

Sana- ja lausepainojen vaihtelu muodostaa puheelle sen ominaisen rytmin, mutta painotuksia tuotetaan eri kielissä eri tavoin. Puherytmejä on vertailtu eri kielissä sekä äidinkielisten ja kieltenoppijoiden välillä erilaisilla tavu- tai äännekestoon perustuvilla parametreilla (White & Mattys 2007). Ruotsin sanapainoa tuotetaan ensisijaisesti tavukestojen avulla, ja painon tuottotavassa on enemmän haasteita ja suurempaa vaihtelua alemman taitotason kielenoppijoilla kuin edistyneemmillä kielenoppijoilla (Kallio ym. 2020, 2021). Sanapainoon liittyvistä piirteistä DigiTalan suomenruotsiin keskittyvä arviointityökalu mittaakin tavukestojen suhteita. Painotusten objektiivinen mittaaminen osana automaattista arviointia voi kuitenkin olla vaikeaa, jos kielenoppija yrittää puhua esimerkiksi riikinruotsia tai norjaa, joiden painotustavat eroavat suomenruotsista, tai jos puheen intonaatio on muuten poikkeavaa (Kallio ym. 2021).

Puhenopeudella, epäröintien määrällä ja tavukestojen hajonnalla puolestaan on tilastollinen yhteys ruotsinoppijoiden spontaanin puheen sujuvuuden arviointiin (Kallio ym. 2023). Tulokset ovat linjassa kansainvälisten tutkimusten kanssa, joissa samankaltaisten piirteiden on todettu vaikuttavan puheen sujuvuuteen (Derwing ym. 2009; Tavakoli 2011; Kahng 2018). Suomenkielisten, A2-tason ruotsinoppijoiden ja suomenruotsalaisten puheen sujuvuutta verranneet Peltonen ja Lintunen (2022) havaitsivat myös selkeän eron L1- ja L2-ryhmien välillä niin puhenopeudessa kuin useissa epäsujuvuutta kuvaavissa piirteissä. Puhenopeutta ja erilaisia epäsujuvuuksia mitataan myös DigiTalan arviointityökalussa.

Suomenoppijoillakin puhenopeus on merkittävä sujuvuuden mittari, mutta vielä paremmin ihmisten antamia arvioita ennustaa artikulaationopeus yhdistettynä erityyppisiin taukoihin, kuten puhunnosten välisten katkosten kestoon, hiljaisten taukojen esiintymistiheyteen ja epäröintien kestoon (Kallio ym. 2022b; Koivusalo 2022). Artikulaationopeus kertoo yksittäisten sanojen tuottonopeuden ilman taukoja, puhenopeus taas koko puhunnoksen tuottonopeuden taukoineen. Tulokset taukojen esiintymistiheydestä tukevat aiempaa tutkimusta, jossa suomenoppijoiden todettiin pitävän ääneen lukiessaan lyhyitä taukoja huomattavasti useammin kuin äidinkielisten puhujien (Toivola ym. 2009). Näitä piirteitä käytetään myös DigiTalan arviointityökalussa (Al-Ghezi ym. 2023). Lisäksi Kallio ym. (2022b) havaitsivat, että etenkin C1–C2-tasoille arvioiduilla suomenoppijoilla esiintyi puheessa narinaa, mutta sellainen oli yleisesti niin harvinaista, ettei sen perusteella voi luotettavasti erottaa eritasoisia puhujia toisistaan.

Kallio ym. (2022a) tutkivat taukojen sijainnin vaikutusta arvioituun kielitaitotasoon ja spontaanin puheen sujuvuuteen suomenoppijoilla. Tulosten perusteella kieliopillisten lausekkeiden sisälle jäävillä tauoilla ja epäröinneillä on merkittävä vaikutus sekä suomen taitotason että sujuvuuden arviointiin. Tulokset ovat linjassa aiempien tutkimusten kanssa (Kahng 2018; Tavakoli 2011). Taukojen sijaintiin perustuvilla parametreilla voitaisiin myös parantaa automaattisen arvioinnin tarkkuutta (Kallio & Kuronen 2023). DigiTalassa etsittiin tapoja integroida taukojen sijaintiparametreja arviointityökaluun; yksi tällainen menetelmä voisi perustua puheen automaattisen syntaktisen jäsennyksen ja akustisten taukojen mittaamisen yhdistämiseen.

2.2.3 Kieliteknologinen tutkimus

Puheen automaattisen arvioinnin teknologinen tutkimus alkoi 1990-luvulla ja keskittyi ääntämisen laadun (Bernstein ym. 1990), puheen sujuvuuden (Cucchiarini ym. 1997) ja intonaation arviointiin (Eskenazi 1996). Näiden tutkimusten kohde oli vielä rajattu kielenoppijan lukemaan tai perässä toistamaan puheeseen, jota algoritmi vertasi puhujan samansisältöiseen tuotokseen. Teknologian nopea kehitys ja laskentatehon kasvu on mahdollistanut yhä monimutkaisempien algoritmien ja sitä myötä tarkempien puheentunnistimien ja monipuolisempien arviointimallien kehittämisen (ks. esim. Loukina & Yoon 2020).

Hankkeen puhe- ja kieliteknologinen tutkimus on keskittynyt automaattisen puheentunnistimen (Al-Ghezi ym. 2021; Getman 2021a) sekä automaattisen arvioinnin ja palautteen kehittämiseen (Getman 2021b; Al-Ghezi ym. 2023). Al-Ghezi ym. (2021) käyttivät Meta AI -yhtiön julkaisemaa, avoimen lähdekoodin wav2vec-algoritmia ruotsinoppijoiden puheentunnistimen kehittämiseen. Menetelmä on itseohjattu: kone oppii tunnistamaan puhetta suoraan suuresta, monikielisestä aineistosta ilman ihmisen väliintuloa. Itseohjatun alustavan opetuksen jälkeen tunnistin optimoitiin ohjatusti pienellä, ruotsinoppijoiden puheesta koostuvalla aineistolla. Getman (2021a) tutki

maisterintutkielmassaan tällaisten itseohjattujen tunnistimien soveltuvuutta suomenruotsin, suomen ja saksan oppijoiden puheentunnistukseen ja havaitsi niiden olevan tehokkaita työkaluja vähäresurssiseen toisen kielen oppijoiden puheentunnistukseen. DigiTalassa kehitetyt puheentunnistimet ovat tietääksemme ensimmäiset suomen ja suomenruotsin oppijoiden puheeseen adaptoidut tunnistimet.

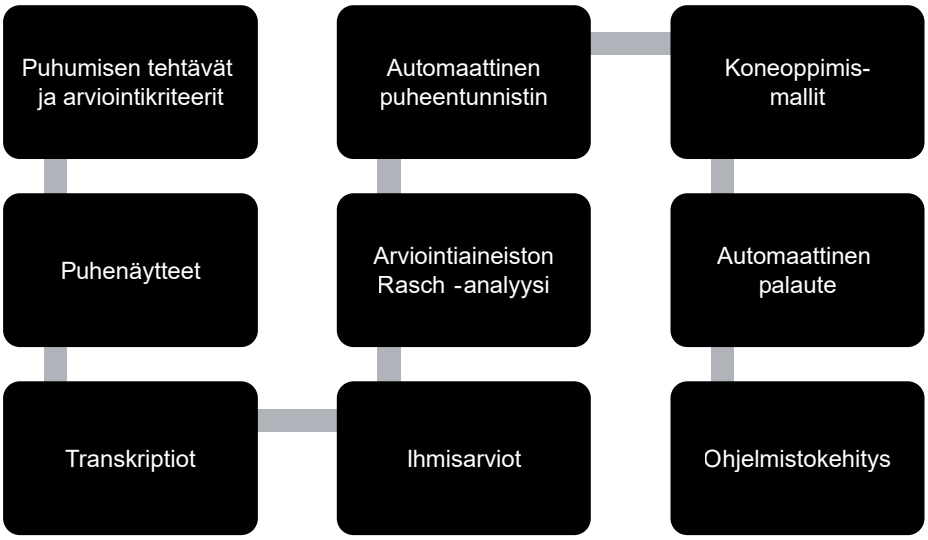
Getman on myös kehittänyt työkalua, joka antaa automaattista palautetta ruotsin kielen sanastosta ja kieliopista (Getman 2021b). Palautetyökalu paikantaa tekstistä virheitä sisältäviä sanoja tai sanajonoja ja antaa käyttäjälle korjausehdotuksia. Työkalua voi käyttää myös suullisen kielitaidon arviointiin, kun puhe on ensin muutettu tekstiksi automaattisen puheentunnistimen avulla.

Voskoboinik ym. (2023) puolestaan tutkivat eri menetelmiä automaattiseen tehtävän suorittamisen arviointiin. Testatuista menetelmistä alustavasti opetettu suuri kielimalli suoriutui sekä suomen- että ruotsinoppijoiden vastausten arvioinnista tarkemmin kuin sanojen yleisyyteen perustuva vektorimalli. Molemmat menetelmät perustuvat tuotetun ja odotetun sisällön samankaltaisuuden mittaamiselle.

Tuorein tutkimuksemme (Al-Ghezi ym. 2023) raportoi ensimmäisen, useita suullisen kielitaidon osa-alueita kattavan automaattisen arviointityökalun eri moduulien kehitysprosessit, sekä testaa ja pohtii niiden toimivuutta. Pääsääntöisesti koneen antama arvio poikkesi vähemmän ihmisen antamasta arviosta (*human-machine agreement*) kuin ihmisten arvioinnit toisistaan (*human-human agreement*), mutta sekä mallien että puheentunnistinten tarkkuus vaihteli jonkin verran. Sekä suomen- että ruotsinkielisen puheentunnistuksen tarkkuuteen vaikuttivat todennäköisesti mm. ihmisen tekemän transkription tarkkuus, puhenäytteen tekninen laatu ja puhujan puhenopeus.

3 Työkalun kehittäminen

Tässä luvussa esittelemme DigiTalan Moodle-sovelluksen (von Zansen ym. 2022a) kehittämisvaiheet. Kyseessä on ensimmäinen kotimaisten kielten automattiseen arviointiin kehitetty työkalu, joka kattaa useita kielitaidon osa-alueita. Koska vastaavaa ei ole tehty aiemmin, puuttui myös sopiva tietokanta työkalun kehittämistä varten. Projektissa on suunniteltu ja kerätty laaja tutkimus- ja kehitysaineisto, joka sisältää suomen ja ruotsin oppijoille suunnattuja suullisen kielitaidon tehtäviä (von Zansen 2022e, 2022f, 2022g), arviointikriteerit (von Zansen 2022d), puhenäytteet litteraatteineen, asiantuntija-arvioinnit (ks. von Zansen ym. 2022c), käyttäjätutkimusten kyselylomakkeet (von Zansen 2022a, 2022b, 2022c), puheentunnistimet, koneoppimismallit sekä tietovaraston Moodle-sovelluksen lataamiseen, käyttöön ja kehittämiseen (von Zansen ym. 2022a; Alanen ym. 2022). Esittelemme kuviossa 1 työkalun kehittämisvaiheet.



KUVIO 1. Työkalun kehittämisvaiheet.

Seuraavat alaluvut etenevät kuviota 1 mukailevassa järjestyksessä. Kuvaamme ensin puhumisen tehtävät (3.1), arviointikriteerit (3.2), puheaineistot (3.3) sekä ihmisarviointien keräämisen ja analyysin (3.4). Tämän jälkeen esittelemme puheentunnistimen (3.5) sekä koneoppimismallit (3.6). Lopuksi käsittelemme automaattista palautetta (3.7) sekä käyttöliittymän (3.8) ja ohjelmiston (3.9) kehitystä.

3.1 Puhumisen tehtävät

Aineistonkeruun suunnittelu käynnistyi mitattavan puhumisen käsitteen määrittelemisellä etenkin Glenn Fulcherin luokittelun mukaisesti (ks. Fulcher 2014 puhumisen käsitteen määrittelystä ja puhetehtävien ominaisuuksista). Kohderyhmän ja kielenkäyttötilanteiden lisäksi otimme huomioon puheen automaattiseen arviointiin ja tietokoneella suoritettavaan puhumisen kokeeseen liittyvät rajoitukset. Käytännössä puhumisen tehtävät suunniteltiin siten, että ne tuottaisivat mahdollisimman monipuolisia puhenäytteitä kielitaidon kattavaa arviointia varten, mutta olisivat puheentunnistimelle riittävän helppoja ja laskentatehon kannalta sopivan mittaisia. Lisäksi tehtävien suunnittelussa huomioitiin niiden pedagoginen mielekkyys muun muassa luomalla useita tehtäväkokonaisuuksia eritasoisille kielenoppijoille.

B1/B2-tasojen puhetehtävät laadittiin vastaamaan lukion opetussuunnitelman (Opetushallitus 2019) sisältöjä ja A1/A2-tasojen puhetehtävät yliopiston alkeiskurssien sisältöjä (ks. von Zansen & Hilden 2022; von Zansen ym. 2022b). Tehtäviä testattiin

opettajien ja kielenoppijoiden kanssa, ja niihin tehtiin parannuksia saadun palautteen perusteella esimerkiksi tehtävänantoja muokkaamalla.

Tehtävät sisältävät ääneen lukua, lyhyitä reagointeja kuvan, tekstin tai äänitteen pohjalta sekä pidempiä tuottamistehtäviä, esimerkiksi kuvasta tai annetusta aiheesta kertomista (esim. kerro päivästäsi tai sinulle tärkeästä paikasta). Näistä spontaaniksi puheeksi lasketaan lyhyet reagoinnit sekä pidemmät tuottamistehtävät. Suomen oppijoiden puheen keräämisessä käytetyt tehtävät on julkaistu verkossa (von Zansen 2022e, 2022f, 2022g) lukuun ottamatta Yleisten kielitutkintojen puhekokeista saatua aineistoa. Ruotsin A1/A2-tasojen puhetehtävät on julkaistu suomen tehtävien tapaan verkossa (von Zansen & Tarvainen-Li 2024). Ruotsin B1-tason suulliset tehtävät puolestaan suunniteltiin aiemmassa projektissa (Karhila ym. 2016), eikä niitä ole julkaistu.

3.2 Arviointikriteerit

Tehtävänlaadinnan rinnalla tutkimushankkeessa laadittiin holistiset ja analyyttiset arviointikriteerit (von Zansen 2022d), jotka suunniteltiin vastaamaan viimeaikaisten lukion opetussuunnitelmien kielitaitokäsitystä (puhumisen arviointikriteereistä tarkemmin Luoma 2004 ja Fulcher 2014). Kriteerien laadinnassa käytettiin pohjana ensisijaisesti vuoden 2003 opetussuunnitelman kielitaidon tasojen kuvausasteikkoja (Opetushallitus 2003), koska ne sopivat nykyisen opetussuunnitelman asteikkoa paremmin hankkeen analyyttisiin ja teknisiin tavoitteisiin. Aiemmat kuvausasteikot eivät ole ristiriidassa nykyisen asteikon kanssa: Opetushallitus kannustaa hyödyntämään niitä esimerkiksi arvioinnissa. Tämä on mahdollista, koska molemmat asteikot ovat paikallisia sovellutuksia Eurooppalaisesta viitekehyksestä. Automaattisesta arvioinnista ks. tarkemmin alaluku 3.6.

Holistinen kriteeristö koostuu kuvauksista taitotasoille asteikolla alle A1–C2. Analyyttiset kriteerit pitivät sisällään kuvaukset viidelle analyyttiselle suullisen kielitaidon osa-alueelle (asteikolla 0–3/4): tehtävän suorittaminen, sujuvuus, ääntäminen, ilmaisun laajuus sekä sanaston ja kieliopin tarkkuus. Ääneenlukutehtävissä arvioitiin vain puheen sujuvuutta ja ääntämistä. Tuottamistehtävissä arvioijat määrittivät puhenäytteille taitotason ja antoivat sille arvion analyyttisistä osa-alueista. Analyyttiset osa-alueet arvioitiin itsenäisinä, taitotasosta riippumattomina ominaisuuksina. Kriteerejä testattiin projektitiimin kielitaidon arvioinnin asiantuntijoiden kanssa ennen niiden käyttöönottoa, ja ne havaittiin toimiviksi.

3.3 Puheaineistot

DigiTala-hanke on kerännyt kieltenoppijoiden puhetta yllä kuvatuilla tehtäväkokonaisuuksilla sekä lukioista että korkeakouluista. Osallistujat suorittivat puhetehtävät Moodle-tenttinä joko luokkahuoneympäristössä tai kotonaan osana etäopetusta. Osa

osallistujista käytti ulkoista kuulokemikrofonia puhevastaustensa äänittämiseen, osa käytti kannettavan tietokoneen sisäistä mikrofonia ja osa älypuhelimen tai tabletin mikrofonia. Vähintään yksi hankkeen tutkija ohjeisti osallistujat puhetehtävien tekoon ja auttoi teknisissä ongelmissa. Puhenäytteiden keruu on dokumentoitu tarkemmin osatutkimuksissa (Al-Ghezi ym. 2023; von Zansen ym. 2022b; von Zansen & Hilden 2022). B1-tason ruotsin näytteet kerättiin aiemmassa projektissa (Karhila 2016).

Artikkelia kirjoittaessa projektissa on kerätty ja litteroitu (ks. myös luku 3.5) yhteensä 4442 puhenäytettä (noin 19 h puhetta) 325 suomenoppijalta ja 8834 näytettä (noin 12,5 h puhetta) 301 ruotsinoppijalta. Näiden lisäksi saimme puheaineistoa Yleisten Kielitutkintojen (YKI) suomen (356 litteroitua näytettä, noin 8,5 h) ja ruotsin (12 litteroitua näytettä, noin 0,3 h) puhekokeista.

3.4 Ihmisarviointien kerääminen ja analyysi

Varsinaisten puhenäytteiden ja transkriptioiden lisäksi automaattisen arvioinnin kehitys on edellyttänyt arviointien keräämistä koulutetuilta ihmisarvioijilta. Arvioijien kouluttamisella pyrimme lisäämään arvioijien ymmärrystä käytettävistä arviointikriteereistä sekä vähentämään arviointivirheitä. Arvioijat saattavat esimerkiksi tulkita annettuja kriteerejä eri tavoin (lisää arviointiin vaikuttavista tekijöistä Fulcher 2014, laajempi katsaus puhumisen arviointitutkimukseen Fan & Yan 2020).

Tutkimushanke on vuosina 2020–2023 järjestänyt yhteensä viisi arviointikierrosta, joihin on osallistunut 39 arvioijaa (ks. arviointiprosessista tarkemmin von Zansen ym. 2022c). Arvioijiksi rekrytoitiin suomen ja ruotsin kielten asiantuntijoita, joilla oli aiempaa kielitaidon arviointikokemusta yleisistä kielitutkinnoista tai ylioppilastutkinnosta. Lisäksi arvioijia rekrytoitiin aiemman tutkimushankkeen ruotsin arviointeihin osallistuneista lukio-opettajista sekä hankkeen tutkijoiden henkilökohtaisten verkostojen kautta. Ulkopuolisten arvioijien lisäksi yhteensä viisi tutkijaa projektitiimistä osallistui arviointikierroksiin.

Arvioijat koulutettiin tehtävään Zoomissa, ja projektin ulkopuolisille arvioijille maksettiin kultakin arviointikierrokselta yhtä työpäivää vastaava korvaus. Arvioinnit ja kyselyvastaukset kerättiin Moodlessa (versio 3.8.3). Koulutuksessa käytiin läpi arviointiin liittyvät ohjeet ja kriteerit (von Zansen 2022d) sekä esiteltiin taitotasoja konkretisoivat maamerkkinäytteet.

Ihmisarvioijien välillä on tunnetusti tiukkuuseroja, joiden tasaamiseksi ordinaaliasteikollinen arviointiaineisto muutettiin lineaarisiksi ja alkuperäisiä pisteitä paikkansapitävämmiksi arvoiksi Rasch-analyysin avulla (*Many-facet Rasch measurement*, ks. Boone ym. 2014, hankkeen aiemmista tutkimuksista von Zansen & Huhta 2022). Lisäksi puheaineistosta karsittiin tässä vaiheessa ongelmallisiksi havaitut näytteet, jotka olivat saaneet jostakin kriteeristä arvosanan 0. Näitä kullekin puhenäytteelle lasket-

tuja raakapisteitä reilumpia arvoja (*fair averages*) hyödynnettiin arviointimoduulien opettamisessa, mihin palaamme puheentunnistimen esittelyn jälkeen.

3.5 Puheentunnistin

Automaattisen puheen arvioinnin tärkein osa on tarkka puheentunnistin, joka muuntaa kielenoppijan puheen tekstiksi. DigiTalan suomen ja ruotsin puheentunnistimet ovat niin sanottuja itseohjattuja (*self-supervised*) päästä-päähän-tunnistimia (*end-to-end,* ks. esim. Baevski ym. 2020; Al-Ghezi ym. 2021; Getman 2021a). Nämä puheentunnistimet opetettiin alustavasti suurella, litteroimattomalla monikielisellä aineistolla, joka sisälsi puhetta 23 eri kielestä. Itseohjattu opetus tarkoittaa, että kone oppii suoraan puheaineistosta ilman ihmisen väliintuloa (ks. luku 2.2.3).

Ruotsin puheentunnistin opetettiin lisäksi ruotsinkielisellä, litteroimattomalla puheaineistolla. Suomen kielestä vastaavaa laajaa aineistoa ei ollut saatavilla. Alustavan opetuksen jälkeen tunnistin hienosäädettiin DigiTala-projektissa kerätyllä, litteroidulla eritasoisten suomen- ja ruotsinoppijoiden puheella (ks. luku 3.3). Litteraatteihin merkittiin muun muassa ääntövirheet ja epäröinnit helpottamaan puheentunnistimen työtä. Näin kehitetyt puheentunnistimet ovat tarkempia kuin pelkästään kohdekielisellä puheaineistolla opetetut, kun kohdekielistä aineistoa on suhteellisen vähän saatavilla. (Baevski ym. 2020; Al-Ghezi ym. 2021; Getman 2021a.)

3.6 Koneoppimismallit

Koneoppimisen menetelmillä pyritään ennustamaan ihmisten antamia arviointeja. Kutsumme näitä koneoppimismalleja arviointimalleiksi. Ihmisarvioijilta saatu arviointiaineisto kategorisoitiin *fair average* -arvoja pyöristämällä, ja jos johonkin arvosanakategoriaan jäi liian vähän näytteitä, sitä ei sisällytetty arviointimoduulien opettamiseen. Holistinen taitotason arviointimalli opetettiin ohjatusti syviä neuroverkkoja hyödyntäen. Kyseessä on päästä-päähän -malli, jossa ihmisten antamaa taitotasoarvioita ennustetaan suoraan puhesignaalista saatavien akustisten ja kielellisten piirteiden pohjalta (Chen ym. 2018; Al-Ghezi ym. 2023). Samoja mitattavia piirteitä käytetään valikoiden myös analyyttisissä arviointimalleissa. Arviointimallit painottavat eri piirteitä eri tavoin riippuen paitsi arvioitavasta kielitaidon alueesta, myös kielestä ja opetusaineistosta.

Tehtävän suorittamisen arviointiin käytettiin itseohjattua koneoppimista, joka laskee semanttisen etäisyyden puhetuotoksen ja tehtävänannon välillä (ks. esim. Wang ym. 2020; Voskoboinik ym. 2023). Tämän arviointimallin tehtävänä on toimia portinvartijana, jos kokelas pelkästään lukee tehtävänannon ääneen tai ei vastaa tehtävänantoon.

Analyyttiset arviointimallit ovat piirrepohjaisia (*feature-based)*, eli ne mittaavat puheesta useita kymmeniä parametreja, joihin malli perustaa arvionsa (Evanini &

Zechner 2020; Al-Ghezi ym. 2023). Sujuvuuden arviointimalli perustuu akustisesta signaalista mitattuihin parametreihin, kuten puhenopeuteen (montako tavua puhuja keskimäärin tuottaa tietyssä ajassa) ja hiljaisten ja täytettyjen taukojen määrään. Ääntämisen arviointimalli laskee esimerkiksi arvon, joka kuvaa puheentunnistimen tunnistusvarmuutta eli kuinka todennäköisesti sen tunnistama sana tai äänne vastaa kohdekielen akustista mallia. Lisäksi ääntämisen arviointimalli laskee puheintonaatioon ja -rytmiin liittyviä parametreja, kuten perustaajuuden hajontaa ja tavukestojen suhteita.

Ilmaisun laajuuden sekä sanaston ja kieliopin tarkkuuden arviointimallit puolestaan perustuvat puheentunnistimen tuottamasta tekstistä laskettuihin piirteisiin. Ilmaisun laajuutta kuvaavia parametreja ovat muun muassa näytteen pituus sanoissa, uniikkien sanojen osuus puhunnoksessa, vastauksen syntaktinen syvyys eli kuinka monimutkaisia rakenteita puhuja on keskimäärin käyttänyt ja verbien suhteellinen osuus (ks. esim. Yoon ym. 2020; Al-Ghezi ym. 2023). Koska kaikki edellä mainitut tekstipohjaiset parametrit kuvaavat ilmaisun laajuutta ennemmin kuin kieliopillista tarkkuutta (ks. esim. Yoon ym. 2020), sanaston ja kieliopin tarkkuus jätettiin pois hankkeen lopullisesta arviointisovelluksesta.

3.7 Automaattinen palaute

Oppijoille annettava automaattinen palaute on teknis-tieteellisten STEM-alojen (*science, technology, engineering and mathematics*) lisäksi yleistynyt myös kielten opetuksessa (Deeva ym. 2021). DigiTalan Moodle-sovelluksessa oppijalle näytettävä automaattinen palaute muotoiltiin työkalun ensimmäisessä versiossa sekä ihmisarvioijien käyttämien arviointikriteerien (von Zansen 2022d, ks. myös luku 3.2) että automaattisten arviointimoduulien hyödyntämien puheen piirteiden pohjalta. Palautteessa pyrimme valitsemaan puheen ulottuvuuksia, jotka ovat koneelle laskettavissa ja ihmisille ymmärrettävissä kuten ääntäminen, sujuvuus, tehtävänantoon vastaaminen ja laajuus (ks. myös von Zansen & Heijala 2023; von Zansen & Huhta 2022). Myöhemmässä vaiheessa automaattisen arvion lisäksi oppijalle voitaisiin antaa oppimista edistäviä neuvoja, kuten Gu ja Davis (2020) kuvaavat.

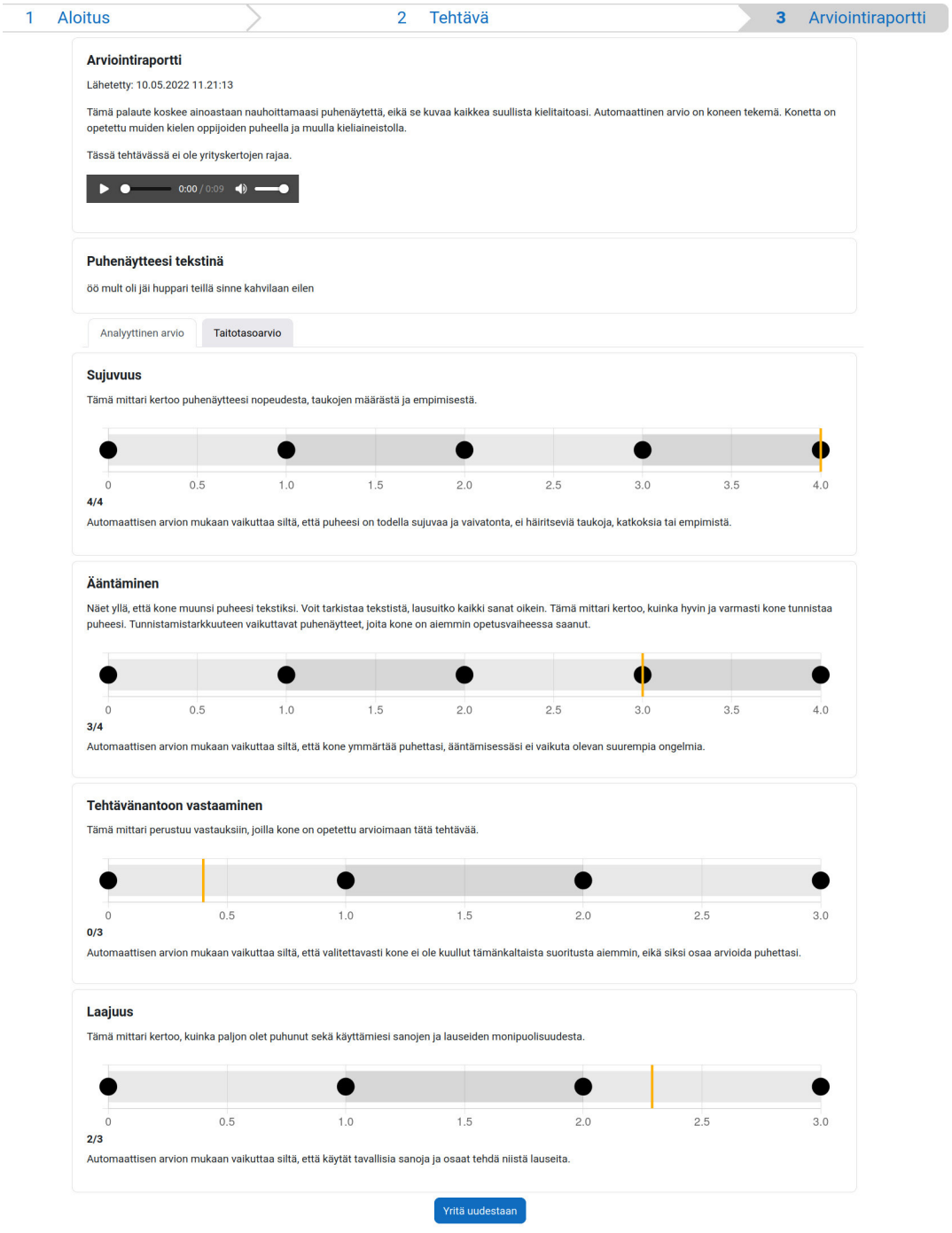
3.8 Käyttöliittymä

Käyttöliittymän ohjelmoinnista vastasi neljä Helsingin yliopiston tietojenkäsittelytieteen opiskelijaa. Heidän laatimansa käyttöohjeet kuvaavat yksityiskohtaisemmin toiminnallisuuksia ja sisältävät lisää kuvia työkalusta eri käyttäjäryhmien näkökulmista (Alanen ym. 2022), joten kuvaamme käyttöliittymän seuraavaksi pääpiirteittäin. DigiTalan verkkotyökalu (von Zansen ym. 2022a) on julkaistu vapaiden ohjelmistojen lisenssillä, mikä mahdollistaa myös käyttöliittymän jatkokehityksen. Ohjelmistokehitystä kuvaamme tarkemmin luvussa 3.9.

Käyttöliittymän avulla oppija nauhoittaa puhenäytteitä, lähettää puhenäytteet arvioitavaksi sekä vastaanottaa automaattisen arvion. Opettaja puolestaan laatii puhetehtävät sekä voi halutessaan täydentää ja kommentoida koneen oppijalle antamaa arviota. Lisäksi opettaja saa ladattua oppijoidensa puhesuoritukset ja niihin liittyvät arviot.

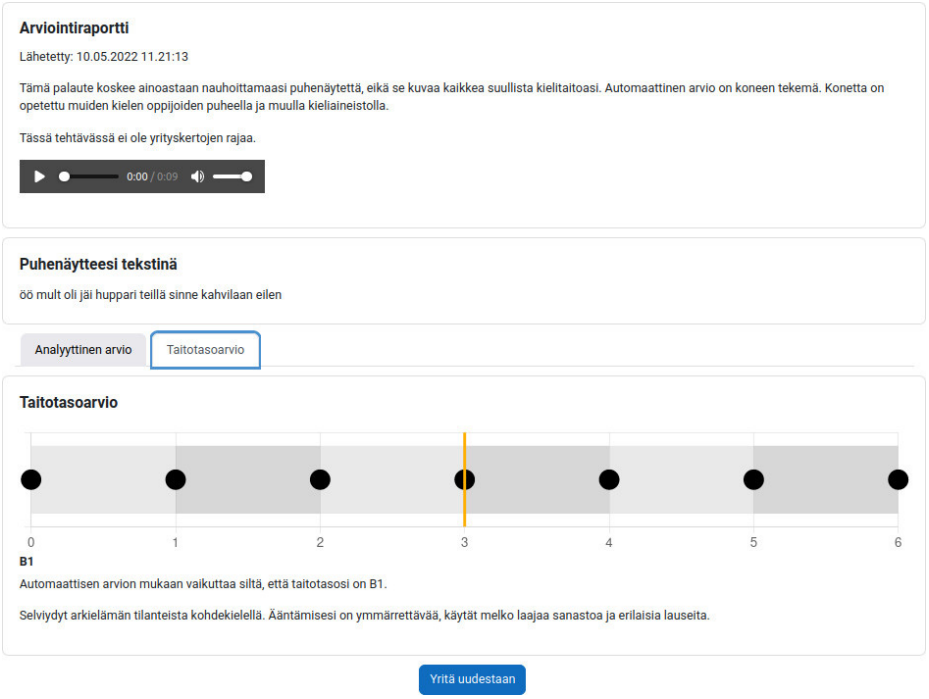
Työkalua käytetään verkkoselaimessa sijaitsevassa Moodlessa, johon kirjauduttuaan oppija näkee opettajan valmistelemat tehtävät, jotka voivat olla ääneenlukua tai puheen tuottamistehtäviä, kuten lyhyitä reagointeja tai pidempiä kertomistehtäviä. Opettaja voi helposti kopioida ja muokata laatimiaan tehtäviä määrittelemällä kohdekielen (suomi/ruotsi), tehtävänannon, aikarajan, äänityskertojen määrän sekä tehtävän sisältämän aineiston. Ääneenlukutehtävien aineistona on luettava teksti tai lause tekstinä, tuottamistehtävät puolestaan voivat sisältää myös kuvia tai äänitteitä.

Testattuaan mikrofoninsa oppija äänittää vastauksensa ja lähettää sen arvioitavaksi Aalto-yliopiston palvelimille, joissa puheentunnistin ja arviointimallit prosessoivat vastauksen. Sovellus hakee palvelimelta kullekin puhenäytteelle erilliset automaattiset arviot, joten oppija voi nauhoittaa useita puhenäytteitä ja palata tarkastelemaan automaattista palautetta (kuva 1) sen valmistuttua.



KUVA 1. Oppijan saama analyyttinen palaute tuottamistehtävästä.

Palautesivulla oppija voi kuunnella nauhoittamansa äänitteen sekä nähdä puheensa muutettuna tekstiksi. Tuottamistehtävissä automaattinen arvio koostuu eri välilehdillä näytettävästä analyyttisestä (kuva 1) ja holistisesta palautteesta (kuva 2), kun taas ääneenlukutehtävissä automaattinen palaute sisältää vain sujuvuuden ja ääntämisen arviot.



KUVA 2. Oppijan saama holistinen palaute tuottamistehtävästä.

Tuottamistehtävistä annettava analyyttinen palaute (kuva 1) sisältää arviot sujuvuudesta, ääntämisestä, tehtävänantoon vastaamisesta ja ilmausten laajuudesta asteikolla 0–3/4. Taitotasoarvio-välilehdellä puolestaan näytetään automaattinen arvio oppijan taitotasosta asteikolla alle A1–C2 (kuva 2).

3.9 Ohjelmistotuotanto

Ohjelmisto tuotettiin yhteistyössä tietojenkäsittelytieteen opiskelijoiden kanssa tammi-toukokuun 2022 aikana. Työskentelimme ketterässä ohjelmistokehityksessä (*agile software development*) käytettävän *scrum*-periaatteen mukaisesti lyhyissä 1–2 viikon kehitysjaksoissa (*sprint*), joiden aikana työkaluun kehitettiin lisää toimintoja. Kehitysjaksojen sisältö suunniteltiin yhteisissä Zoom-videoneuvottelutyökalun avulla järjestetyissä palavereissa, joissa kävimme läpi tehdyn kehitystyön, listasimme asioita

kehitysjonoon (*backlog*) ja sovimme seuraavaksi kehitettävistä toiminnoista. (Stober & Hansmann 2010.) Lisäksi käytimme yhteydenpitoon projektin ja kehittäjien välillä Teams-ohjelmaa ja sähköpostia. Ketterän kehityksen lisäksi kehitystyötä ohjasivat suomalaisen opetussuunnitelman mukainen kielitaitokäsitys (Opetushallitus 2019), saavutettavuusvaatimukset (Etelä-Suomen aluehallintovirasto 2019) sekä saatavilla oleva aiempi tutkimus (ks. luku 2.2). Kehitystyön vaatimia päätöksiä tehtäessä priorisoitiin yleisesti ottaen oppijan etua ja työkalun käyttöä tutkimustarkoituksissa.

DigiTala-hankkeessa tutkimustuotokset jaetaan mahdollisuuksien mukaan avoimen tieteen periaatteita noudattaen. Näin ollen käyttöliittymän toteutustavaksi valittiin Moodle muun muassa avoimen lähdekoodin takia. Kehitystyössä huomioitiin Moodlen omat tarkat tyyli- ja tekniikkavaatimukset (Moodle 2023). Kehitetty työkalu on julkaistu vapaiden ohjelmistojen GNU GPL -lisenssillä, joka on niin kutsuttu *copyleft*-lisenssi: jos lisensoitua koodia muokataan, myös muokattu koodi on jaettava samalla lisenssillä. Ohjelmiston dokumentaatio on saatavilla Github-verkkosivustolla (von Zansen ym. 2022a). Vapaiden ja avointen ohjelmistojen hyödyt liittyvät avoimuuteen ja yhteistyöhön – työkalua voidaan kehittää hankkeen päättymisen jälkeenkin.

4 Pohdinta

Artikkelissa esittelimme DigiTala-tutkimushankkeessa kehitetyn automaattiseen puheentunnistukseen, arviointiin ja palautteeseen perustuvan työkalun, jonka avulla suomen ja ruotsin oppijat voivat harjoitella puheen tuottamista ja ääneen lukemista. Monitieteiseen soveltavaan tutkimukseen (luku 2.2) perustuva työkalu tarjoaa kieltenoppijoille ajasta ja paikasta riippumattomia harjoittelumahdollisuuksia sekä antaa heille sekä analyyttistä palautetta että taitotasoarvion puhesuorituksesta.

DigiTalan työkalu on ensimmäinen, joka antaa analyyttista palautetta useasta kielitaidon osa-alueesta nimenomaan spontaaneista suomen- ja ruotsinkielisistä puhenäytteistä. Analyyttinen palaute tukee erilaisia oppijoita, sillä kielitaidon osa-alueet kehittyvät usein eri tahtiin (ALTE 2016: 28, 40). Työkalumme antama palaute auttaa yksilöllisten kehitystarpeiden hahmottamisessa sekä puhujan yleisen taitotason määrittämisessä.

Tulevaisuudessa automaattista arviointia voidaan toivottavasti hyödyntää kielitaidon arvioinnissa laajemmin – myös yksilöä koskevissa tärkeissä (*high-stakes*) päätöksissä, kuten kansalaisuuden, työpaikan tai jatko-opiskelupaikan hakuprosesseissa (ks. ALTE 2016; Vaarala ym. 2021). Tekoälyn hyödyntämiseen kielitaidon arvioinnissa liittyy kuitenkin uhkia ja rajoitteita, joiden takia automaattista arviointia suositellaan yleensä käytettäväksi ihmisen tekemän arvioinnin rinnalla (*hybrid approach*, ks. Evanini & Zechner 2020: 11).

Koneoppimismallit ovat esimerkiksi alttiita aineistosta kumpuaville vinoumille (*bias*). Mitä rajallisempi tekoälyn opetusaineisto on, sitä tärkeämpää on huomioida siitä mahdollisesti johtuvat vinoumat. DigiTalassa kohdekielet, suomi ja suomenruotsi, ovat vähäresurssisia ja arviointityökalun kehittämiseen on käytetty huomattavasti suppeampaa opetusaineistoa kuin esimerkiksi ETS:n tuottamissa tutkimuksissa (ks. esim. Zhang ym. 2020). Haasteet kotimaisten kielten automaattisen arvioinnin käyttämisessä ja tutkimuksessa koskevatkin mahdollisten vinoumien havaitsemista malleissa ja niiden mahdollisesti aiheuttaman syrjinnän ennaltaehkäisemistä.

Tekoälyn kontekstissa vinoumalla viitataan koneoppimismalliin tai tekoälyä hyödyntävään sovellukseen, joka kohtelee tiettyä ihmisryhmää systemaattisesti epäsuotuisammin (Ojanen ym. 2022: 14). Tällainen automaattisen arviointityökalun vinouma johtuu yleensä opetusaineiston rajallisuudesta: arviointimalli oppii arvioimaan vain sellaisia puhujia (ja sellaisia tehtäviä), joiden vastauksista opetusaineisto koostuu.

Puhumisen arvioinnissa on haasteena saada kerättyä riittävän kattava aineisto, jossa huomioitaisiin kaiken tasoiset ja eri taustoista tulevat kielenoppijat. Myös kohdekielen variantit tuovat omat haasteensa. Esimerkiksi DigiTalan ruotsin arviointimallit on opetettu aineistolla, joka sisältää pääosin kieltenoppijoiden puhumaa suomenruotsia, jolloin arviointimallit eivät välttämättä toimi yhtä hyvin oppijoille, jotka puhuvat riikinruotsia (Kallio ym. 2021). Myös osa oppijoista tiedostaa puhetyylistä johtuvan vinouman mahdollisuuden (von Zansen ym. 2022b). Lisäksi DigiTalan aineistossa puhujien kielitaitotaso oli jakautunut epätasaisesti. Tämä tarkoittaa, että työkalu arvioi luotettavammin paremmin edustetuille taitotasoille osuvia puhenäytteitä kuin niitä puhenäytteitä, joille opetusaineistosta löytyy vähemmän vertailukohteita (Al-Ghezi ym. 2023).

Puhetyylien erot korostuvat mitä pidemmistä puhunnoksista on kyse. Etenkin spontaaneissa, pitkissä puhunnoksissa voi olla paljon variaatiota, joka ei liity tai vaikuta kielitaidon arviointiin. Tällainen ”normaali” variaatio voi liittyä esimerkiksi sujuvuuspiirteisiin (ks. esim. Penttilä & Korpijaakko-Huuhka 2019; Lintunen ym. 2022). Spontaanissa puheessa esiintyy yleensä enemmän epäsujuvuuksia kuin luetussa puheessa ja niitä myös siedetään paremmin (Kallio ym. 2017). Mikäli sujuvuuden arviointimalli on opetettu suhteellisen lyhyillä ja hyvin valmistelluilla puhunnoksilla, se saattaa arvioida pidempiä puhenäytteitä turhan tiukasti. Tällaista arviointivinoumaa voidaan pienentää toisaalta aineistoa laajentamalla, toisaalta puhetehtäviä tai ohjeistuksia tarkentamalla ja siten vähentämällä tehtävän tulkinnasta johtuvaa variaatiota puhujien välillä.

Arviointivinoumat koskevat kuitenkin myös ihmisarvioijia, joiden toimintaan voidaan vaikuttaa koulutuksen ja ohjeistusten avulla sekä analysoimalla heiltä kerättyjä arviointeja (luku 3.4). Automaattinen arviointityökalu voi toimia myös koulutusvälineenä: vertaamalla omia arvioitaan koneen antamaan palautteeseen arvioija voi havaita itselleen uusia suullisen kielitaidon piirteitä tai tiedostaa arvioivansa tietynlaista puhetta tietyllä tavalla.

Yksi tekoälyn hyödyntämisen rajoite liittyy sen toimintaperiaatteiden ja käytännön soveltamisen ymmärtämiseen. Yhtäältä automaattisen arviointityökalun käyttäjien on

tärkeää ymmärtää, mihin automaattinen palaute perustuu, jotta he osaavat tulkita koneen antamaa palautetta oikein ja käyttää työkalua tarkoituksenmukaisesti. Toisaalta työkalun kehittäjillä tulee olla riittävä ymmärrys luotettavuuteen, validiteettiin ja oikeudenmukaisuuteen liittyvistä kysymyksistä. (Evanini & Zechner 2020.)

Tämän artikkelin tavoitteena oli koostaa aiempaa puheen automaattista arviointia käsittelevää tutkimusta, dokumentoida DigiTalan arviointityökalun kehittämisvaiheet sekä kuvata työkalun toimintaperiaatteet. Työkalu toteutettiin ketterän ohjelmistokehityksen (Stober & Hansmann 2010) periaatteita noudattaen avoimeen lähdekoodiin perustuvalle Moodle-oppimisalustalle. Kokemuksen perusteella suosittelemme sekä ketterää ohjelmistokehitystä että Moodlea. Kannustammekin kielten opetuksen ja kielitaidon arvioinnin parissa työskenteleviä työkalujen kehittäjiä ja materiaalien laatijoita osallistamaan monipuolisesti eri käyttäjäryhmiä suunnitteluun varhaisesta vaiheesta lähtien. Tämän avulla pyrimme palvelemaan mahdollisimman erilaisia käyttäjiä (*design for all*, Etelä-Suomen aluehallintovirasto 2019).

Käyttäjäkokemusten tutkiminen on oleellinen osa työkalujen kehitystä, ja jatkamme sen äärellä tutkimushankkeen loppuun saakka. Oppijoille suunnattujen verkkotyökalujen käytettävyyttä voitaisiin jatkossa tutkia entistä monipuolisemmilla tutkimusmenetelmillä kuten yhdistämällä kyselyistä, havainnoinnista ja haastatteluista saatavaa tietoa käyttäjän fysiologisiin mittauksiin. Esimerkiksi katseenseurannan avulla voitaisiin tutkia, mihin kohtiin ja kuinka kauaksi aikaa kielenoppija todella kohdistaa katseensa automaattista palautetta tarkastellessaan (ks. esim. Liu & Yu 2022). Hankkeessa saadut tulokset ja kokemukset ovat lupaavia puheen automaattisen arvioinnin ja koneoppimisen hyödyntämisen kannalta siinä. Tulevaisuudessa tarvitaan kuitenkin tutkimusta liittyen puhumisen muihin osa-alueisiin, kuten nonverbaaliseen viestintään, jota ei perinteisesti ole mitattu tai eksplisiittisesti arvioitu osana suullista kielitaitoa (akateemisten suomenoppijoiden suullisen vuorovaikutuksen automaattista arviointia koskevasta Aasis-tutkimushankkeesta ks. von Zansen 2023).

Kiitokset

Suomen Akatemia rahoittaa DigiTala-tutkimushanketta (2019–2023). Kiitämme yhteistyöstä projektin tutkijoita Raili Hildeniä Helsingin yliopistosta (rahoituspäätös 322619), Mikko Kurimoa, Ragheb Al-Gheziä, Yaroslav Getmania ja Ekaterina Voskoboinikia Aalto-yliopistosta (rahoituspäätös 322625) sekä Mikko Kurosta, Maria Kautosta ja Ari Huhtaa Jyväskylän yliopistosta (rahoituspäätös 322965). Lisäksi kiitämme Moodle-sovelluksen kehittämiseen osallistuneita Helsingin yliopiston tietojenkäsittelytieteen opiskelijoita: Tuomas Alanen, Topi Harjunpää, Joona Erkkilä, Maikki Heijala. Artikkelin oikolukemisesta kiitämme tutkimusavustaja Ilona Lähteenmäkeä. Kiitämme myös kaikkia projektin aikana aineiston keruuseen, litterointiin ja arviointiin osallistuneita avustajia, arvioijia ja opettajia.

Kirjallisuus

Alanen, T., J. Erkkilä, T. Harjunpää & M. Heijala 2022. Digitala Moodle plugin user manual (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.6535377

Al-Ghezi, R., Y. Getman, A. Rouhe, R. Hildén & M. Kurimo 2021. Self-supervised end-to-end ASR for low resource L2 Swedish. Teoksessa H. Heřmanský & H. Černocký (toim.) *Proceedings of Interspeech 2021*. Brno: International Speech Communication Association (ISCA), 1429–1433. http://dx.doi.org/10.21437/Interspeech.2021–1710

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

ALTE 2016. *Kielitestien avulla osalliseksi, integroituneeksi ja kansalaiseksi. Opas päätöksenteon tueksi*. https://alte.wildapricot.org/resources/Documents/LAMI%20Booklet%20FI.pdf [luettu 22.11.2023]

Anderson-Hsieh, J., R. Johnson & K. Koehler 1992. The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentais, prosody, and syllable structure. *Language Learning*, *42* (4), 529–555. https://doi.org/10.1111/j.1467-1770.1992.tb01043.x

Baevski, A., Y. Zhou, A. Mohame, & M. Auli 2020. Wav2vec 2.0: a framework for self-supervised learning of speech representations. Teoksessa H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (toim.) *Proceedings of the 34th International conference on neural information processing systems (NIPS'20)*. Vancouver: Curran Associates Inc., 12449–12460. https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

Bernstein, J., M. Cohen, H. Murveit, D. Rtischev & M. Weintraub 1990. Automatic evaluation and training in English pronunciation. Teoksessa *First International conference on spoken language processing (ICSLP 1990)*. Kobe: Acoustical Society of Japan, 1185–1188. https://doi.org/10.21437/ICSLP.1990-313

Boone, W. J., J. R. Staver & M. S. Yale 2014. *Rasch analysis in the human sciences.* Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Chen, L., J. Tao, S. Ghaffarzadegan & Y. Qian 2018. End-to-end neural network based automated speech scoring. Teoksessa M. Hayes & H. Ko (toim.) *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. Calgary: Institute of electrical and electronics engineers, 6234–6238. https://doi.org/10.1109/ICASSP.2018.8462562

Cucchiarini, C., H. Strik & L. Boves 1997. Automatic evaluation of Dutch pronunciation by using speech recognition technology. Teoksessa *1997 IEEE Workshop on automatic speech recognition and understanding proceedings*. Santa Barbara: Institute of electrical and electronics engineers, 622–629. https://doi.org/10.1109/ASRU.1997.659144

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

Deeva, G., D. Bogdanova, E. Serral, M. Snoeck & J. De Weerdt 2021. A review of automated feedback systems for learners: classification framework, challenges and opportunities. *Computers & Education*, *162*. https://doi.org/10.1016/j.compedu.2020.104094

Derwing, T. M., M. J. Munro, R. I. Thomson & M. J. Rossiter 2009. The relationship between L2 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31* (4), 533–557. https://doi:10.1017/S0272263109990015

Eskenazi, M. 1996. Detection of foreign speakers' pronunciation errors for second language training – preliminary results. Teoksessa *Proceedings of the International congress on spoken language processing.* Philadelphia: International Speech Communication Association, 1465–1468. https://doi.org/10.1109/ICSLP.1996.607892

Etelä-Suomen Aluehallintovirasto 2019. *Yleistä saavutettavuudesta*. https://www.saavutettavuusvaatimukset.fi/yleista-saavutettavuudesta [luettu 22.11.2023]

Evanini, K. & K. Zechner 2020. Overview of automated speech scoring. Teoksessa K. Zechner & K. Evanini (toim.) *Automated speaking assessment: using language technologies to score spontaneous speech*. New York: Routledge, 3–20.

Fan, J. & X. Yan 2020. Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology*, *11*, 330–330. https://doi.org/10.3389/fpsyg.2020.00330

Fulcher, G. 2014. *Testing second language speaking*. London: Routledge.

Getman, Y. 2021a. *End-to-end low-resource automatic speech recognition for second language learners*. Pro gradu -tutkielma. Aalto-yliopisto, Sähkötekniikan korkeakoulu/ELEC. https://aaltodoc.aalto.fi/handle/123456789/110588

Getman, Y. 2021b. Automated writing support for Swedish learners. Teoksessa P. Ljunglöf, S. Dobnik & R. Johansson (toim.) *Selected contributions from the eighth Swedish language technology conference (SLTC-2020), 25-27 November 2020.* Linköping electronic conference proceedings 184. Linköping: University of Linköping, 21–26. https://doi.org/10.3384/ecp184171

Gu, L. & L. Davis 2020. Providing speech rater feature performance as feedback on spoken responses. Teoksessa K. Zechner & K. Evanini (toim.), *Automated speaking assessment: using language technologies to score spontaneous speech*. New York: Routledge, 159–175.

Hsieh, C.-N., K. Zechner & X. Xi 2020. Features measuring fluency and pronunciation. Teoksessa K. Zechner & K. Evanini (toim.) *Automated speaking assessment: using language technologies to score spontaneous speech.* New York: Routledge, 101–122.

Härmälä, M. & J. Marjanen 2023. *A-ruotsin oppimistulokset perusopetuksen päättövaiheessa 2022*. Helsinki: Kansallinen koulutuksen arviointikeskus. https://www.karvi.fi/fi/julkaisut/ruotsin-oppimistulokset-perusopetuksen-paattovaiheessa-2022

Isaacs, T., & P. Trofimovich 2012. Deconstructing comprehensibility: identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34* (3), 475–505. https://doi.org/10.1017/S0272263112000150

Kahng, J. 2018. The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39* (3), 569–591. https://doi.org/10.1017/S0142716417000534

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

Kallio, H. & M. Kuronen 2023. Revising parameters for predicting L2 speech fluency and proficiency. Teoksessa R. Skarnitzl & J. Volín (toim.) *Proceedings of the 20th International congress of phonetic sciences 2023*. Prague: Guarant International, 2452–2456. https://guarant.cz/icphs2023/221.pdf

Kallio, H., M. Kuronen & M. Kautonen 2021. Differences in acoustically determined sentence stress between native and L2 speakers of Finland Swedish. *Working papers – Lund University, Department of linguistics, General linguistics, Phonetics, 56*, 42–47. http://urn.fi/URN:NBN:fi:jyu-202111115628

Kallio, H., M. Kuronen & L. Koivusalo 2022a. The role of pause location in perceived fluency and proficiency in L2 Finnish. Teoksessa *Proceedings of the International symposium of applied phonetics 2022*. Lund: University of Lund, 22–27. https://doi.org/10.21437/ISAPh.2022-5

Kallio, H., J. Šimko, A. Huhta, R. Karhila, M. Vainio, E. Lindroos, R. Hildén & M. Kurimo 2017. Towards the phonetic basis of spoken second language assessment: temporal features as indicators of perceived proficiency level. Teoksessa M. Kuronen, P. Lintunen & T. Nieminen (toim.) *Näkökulmia toisen kielen puheeseen. Insights into second language speech.* AFinLA-e. Soveltavan kielitieteen tutkimuksia 10. Jyväskylä: Suomen soveltavan kielitieteen yhdistys AFinLA, 193–213. https://journal.fi/afinla/article/view/73137

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

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

Kallio, H., R. Suviranta, M. Kuronen & A. von Zansen 2022b. Creaky voice and utterance fluency measures in predicting fluency and oral proficiency of spontaneous L2 Finnish. Teoksessa S. Frota, M. Cruz & M. Vigário (toim.) *Proceedings of Speech prosody 2022*. Lisbon: International Speech Communication Association (ISCA), 777–781. https://doi.org/10.21437/SpeechProsody.2022-158

Kang, O. 2012. Relative impact of pronunciation features on ratings of non-native speakers' oral proficiency. Teoksessa J. Levis & K. LeVelle (toim.) *Proceedings of the 4th Pronunciation in second language learning and teaching conference.* Ames: Iowa State University, 10–15.

Karhila, R., A. Rouhe, P. Smit, A. Mansikkaniemi, H. Kallio, E. Lindroos, R. Hildén, M. Vainio & M. Kurimo 2016. Digitala: An augmented test and review process prototype for high-stakes spoken foreign language examination. Teoksessa *Proceedings of Interspeech 2016.* San Francisco: International Speech Communication Association (ISCA), 784–785. https://www.isca-archive.org/interspeech_2016/karhila16_interspeech.pdf

Kautonen, M. & M. Kuronen 2021. Kvantitativa perspektiv på L2-tal på olika färdighetsnivåer. *Folkmålsstudier*, *59*, 11–40. https://journal.fi/folkmalsstudier/article/view/112545

Kautonen, M. & A. von Zansen 2020. DigiTala research project: Automatic speech recognition in assessing L2 speaking. *Kieli, koulutus ja yhteiskunta, 11* (4). https://www.kieliverkosto.fi/fi/journals/kieli-koulutus-ja-yhteiskunta-kesakuu-2020/digitala-research-project-automatic-speech-recognition-in-assessing-l2-speaking.

Koivusalo, L. 2022. *Phonetic fluency in Finnish as a second language: acoustic analysis of high school students' spontaneous speech*. Pro gradu -tutkielma, Helsingin yliopisto. http://urn.fi/URN:NBN:fi:hulib-202206152522

Lintunen, P., M. Mutta, S. Olkkonen, P. Peltonen & O. Veivo 2022. Sujuvuus ja epäsujuvuus vieraan kielen oppimisen näkökulmasta: monitahoinen ilmiö edellyttää monitieteistä tutkimusta. Teoksessa S. Loukusa, T. Hautala & A.-K. Tolonen (toim.) *Sujuvaa vai sujumatonta? Puheen ja kielen sujuvuutta tutkimassa*. Helsinki: Puheen ja kielen tutkimuksen yhdistys, 75–87.

Liu, S. & G. Yu 2022. L2 learners' engagement with automated feedback: An eye-tracking study. *Language Learning & Technology*, *26* (2), 78–105. https://doi.org/10125/73480

Loukina, A. & S. Y. Yoon 2020. Scoring and filtering models for automated speech scoring. Teoksessa K. Zechner & K. Evanini (toim.) *Automated speaking assessment: using language technologies to score spontaneous speech*. New York: Routledge, 75–98.

Luoma, S. 2004. *Assessing speaking*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

Moodle 2023. *Moodle developer documentation*. https://docs.moodle.org/dev/Main_Page [luettu 22.11.2023]

Munro, M. J. & T. M. Derwing 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45* (1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Ojanen, A., O. Sahlgren, J. Vaiste, A. Björk, J. Mikkonen, K. Kimppa, A. Laitinen & N. Oljakka 2022. *Algoritminen syrjintä ja yhdenvertaisuuden edistäminen: arviointikehikko syrjimättömälle tekoälylle*. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 2022, 54. Helsinki: Valtioneuvosto. http://urn.fi/URN:ISBN:978-952-383-404-0

Opetushallitus 2003. *Lukion opetussuunnitelman perusteet*. Helsinki: Opetushallitus.

Opetushallitus 2019. *Lukion opetussuunnitelman perusteet*. Helsinki: Opetushallitus.

Peltonen, P. & P. Lintunen 2022. Multilingual speakers' L1, L2, and L3 fluency across languages: a study of Finnish, Swedish, and English. *Nordand*, *17* (1), 48–63. https://doi.org/10.18261/nordand.17.1.4

Penttilä, N. & A. M. Korpijaakko-Huuhka 2019. Disfluencies in typical Finnish-speaking adults. *The Phonetician*, *27* (116), 28–41.

Stober, T. & U. Hansmann 2010. Overview of agile software development. Teoksessa T Stober & U. Hansmann (toim.) *Agile software development*. Heidelberg: Springer Berlin, 35–59. https://doi.org/10.1007/978-3-540-70832-2_3

Tavakoli, P. 2011. Pausing patterns: differences between L2 learners and native speakers. *ELT Journal*, *65* (1), 71–79. https://doi.org/10.1093/elt/ccq020

Toivola, M., M. Lennes & E. Aho 2009. Speech rate and pauses in non-native Finnish. Teoksessa *Proceedings of Interspeech 2009*. Brighton: International Speech Communication Association (ISCA), 1707–1710. https://doi.org/10.21437/Interspeech.2009-515

Vaarala, H., S. Riuttanen, E. Kyckling & S. Karppinen 2021. *Kielivaranto. Nyt!: monikielisyys vahvuudeksi -selvityksen (2017) seuranta*. Jyväskylä: Soveltavan kielentutkimuksen keskus, Jyväskylän yliopisto. https://jyx.jyu.fi/handle/123456789/74416

Voskoboinik, E., Y. Getman, R. Al-Ghezi, M. Kurimo & T. Grósz 2023. Automated assessment of task completion in spontaneous speech for Finnish and Finland Swedish language learners. Teoksessa D. Alfter, E. Volodina, T. François, A. Jönsson & E. Rennes (toim.) *Proceedings of the 12th Workshop on natural language processing for computer assisted language learning (NLP4CALL 2023)*. Linköping: University of Linköping, 102–110. https://doi.org/10.3384/ecp197012

Wang, X., K. Zechner & C. Hamill 2020. Targeted content feedback in spoken language learning and assessment. Teoksessa H. Meng, B. Xu & T. Zheng (toim.) *Proceedings of Interspeech 2020*. Sanghai: International Speech Communication Association (ISCA), 3850–3854. https://doi.org/10.21437/Interspeech.2020-1766

White, L. & S. L. Mattys 2007. Calibrating rhythm: first language and second language studies. *Journal of Phonetics*, *35* (4), 501–522. https://doi.org/10.1016/j.wocn.2007.02.003

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

Ylinen, S. & M. Kurimo 2017. Kielenoppiminen vauhtiin puheteknologian avulla. Teoksessa H. Savolainen, R. Vilkko & L. Vähäkylä (toim.) *Oppimisen tulevaisuus.* Helsinki: Gaudeamus, 57–69. http://hdl.handle.net/10138/309750

Yoon, S.-Y., X. Lu & K. Zechner 2020. Features measuring vocabulary and grammar. Teoksessa K. Zechner & K. Evanini (toim.) *Automated speaking assessment: using language technologies to score spontaneous speech*. New York: Routledge, 123–137.
von Zansen, A. 2022a. DigiTala's post-rating questionnaire for human raters (Finnish, upper secondary schools, Jun2021). Zenodo. https://doi.org/10.5281/zenodo.6477015
von Zansen, A. 2022b. DigiTala's post-test questionnaire for L2 Finnish learners (upper secondary schools 2021). Zenodo. https://doi.org/10.5281/zenodo.6562884
von Zansen, A. 2022c. DigiTala's pre-test consent and background information form for L2 Finnish learners (upper secondary schools 2021). Zenodo. https://doi.org/10.5281/zenodo.6562663
von Zansen, A. 2022d. DigiTala's rating criteria: Holistic and analytic scales for assessing L2 speaking. Zenodo. https://doi.org/10.5281/zenodo.6477089
von Zansen, A. 2022e. DigiTala's speaking tasks and questionnaire for L2 Finnish learners (proficiency level A). Zenodo. https://doi.org/10.5281/zenodo.6627533
von Zansen, A. 2022f. DigiTala's speaking tasks for L2 Finnish learners (proficiency level B1). Zenodo. https://doi.org/10.5281/zenodo.6562855
von Zansen, A. 2022g. DigiTala's speaking tasks for L2 Finnish learners (proficiency level B2). Zenodo. https://doi.org/10.5281/zenodo.6562865
von Zansen, A. 2023. The Aasis research project: automatically assessing spoken interaction in L2 Finnish. *Kieli, koulutus ja yhteiskunta*, *14* (7). https://www.kieliverkosto.fi/fi/journals/kieli-koulutus-ja-yhteiskunta-joulukuu-2023/the-aasis-research-project-automatically-assessing-spoken-interaction-in-l2-finnish
von Zansen, A., T. Alanen, R. Al-Ghezi, J. Erkkilä, T. Harjunpää., M. Heijala & H. Kallio 2022a. *DigiTala Moodle plugin*. https://github.com/aalto-speech/moodle-mod_digitala
von Zansen, A. & M. Heijala 2023. Miten suomen ja ruotsin opettajat käyttäisivät puheen automaattiseen arviointiin kehitettyä työkalua?. Teoksessa T. Mäkipää, R. Hilden & A. Huhta (toim.) *Kielenoppimista tukeva arviointi – Assessment for supporting language learning.* AFinLA-teema 15. Jyväskylä: Suomen soveltavan kielitieteen yhdistys AFinLA, 124–141. https://journal.fi/afinla/article/view/124822
von Zansen, A. & R. Hilden 2022. "It was cool and comfortable!" Akateemisten alkeistason S2-opiskelijoiden kokemuksia tietokoneella suoritettavasta puhumisen kokeesta. Teoksessa S. Routarinne, P. Heinonen, T. Kärki, A. Roiha. M.-L. Rönkkö & A. Korkeaniemi (toim.) *Ainedidaktiikka ajassa: laajenevat oppimisympäristöt ja eri-ikäiset oppijat.* Suomen ainedidaktisen tutkimusseuran julkaisuja 22. Turku: Turun yliopisto, Suomen ainedidaktinen tutkimusseura ry, 72–90. http://hdl.handle.net/10138/353562
von Zansen, A., R. Hilden & M. Sneck 2022b. Lukiolaisten käsitykset ja heidän antamansa palaute suullisen kielitaidon automaattisesta arvioinnista. Teoksessa R. Kantelinen, M. Kautonen & Z. Elgundi (toim.) *Linguapeda 2021*. Suomen ainedidaktisen tutkimusseuran julkaisuja 21. Joensuu: Itä-Suomen yliopisto, Suomen ainedidaktinen tutkimusseura ry, 176–205. http://hdl.handle.net/10138/352128
von Zansen, A. & A. Huhta 2022. Developing automated feedback on spoken performance: exploring the functioning of five analytic rating scales using many-facet Rasch measurement. Teoksessa J. H. Jantunen, J. Kalja-Voima, M. Laukkarinen, A. Puupponen, M. Salonen, T. Saresma, J. Tarvainen, & S. Ylönen (toim.) *Diversity of methods and materials in digital human sciences: proceedings of the Digital research data and human sciences DRDHum conference 2022, December 1–3, Jyväskylä, Finland.* Jyväskylä: Jyväskylän yliopisto, 211–229. http://urn.fi/URN:ISBN:978-951-39-9450-1

von Zansen, A., H. Kallio, M. Sneck, M. Kuronen, A. Huhta & R. Hilden 2022c. Ihmisarvioijien näkemyksiä suullisen kielitaidon automaattisesta arvioinnista, digitaalisesta arviointiprosessista sekä puhesuorituksista arvioitavista ulottuvuuksista. Teoksessa T. Seppälä, S. Lesonen, P. Iikkanen & S. D'hondt (toim.) *Kieli, muutos ja yhteiskunta – Language, change and society*. AFinLAn vuosikirja 2022. Jyväskylä: Suomen soveltavan kielitieteen yhdistys AFinLA, 370–394. https://journal.fi/afinlavk/article/view/114821

von Zansen, A. & L. L. Tarvainen-Li 2024. DigiTala's speaking tasks and questionnaire for L2 Swedish learners (proficiency level A1-A2). Zenodo. https://doi.org/10.5281/zenodo.10693570

Zechner, K. 2020. Summary and outlook on automated speech scoring. Teoksessa K. Zechner & K. Evanini (toim.) *Automated speaking assessment: using language technologies to score spontaneous speech.* New York: Routledge, 192–204.

Zhang, M., L. Yao, S. J. Haberman & N. J. Dorans 2020. Assessing scoring accuracy and assessment accuracy for spoken responses: using human and machine scores. Teoksessa K. Zechner & K. Evanini (toim.) *Automated speaking assessment: using language technologies to score spontaneous speech*. New York: Routledge, 32–58.