Seppälä, T., S. Lesonen, P. Iikkanen & S. D'hondt (toim.) 2022.
Kieli, muutos ja yhteiskunta – Language, Change and Society. AFinLAn vuosikirja 2022.
Suomen soveltavan kielitieteen yhdistyksen julkaisuja n:o 79. Jyväskylä. s. 370–394.

Anna von Zansen[1], **Heini Kallio**[2], **Milla Sneck**[1],
Mikko Kuronen[2], **Ari Huhta**[2] **& Raili Hildén**[1]
[1]Helsingin yliopisto, [2]Jyväskylän yliopisto

Ihmisarvioijien näkemyksiä suullisen kielitaidon automaattisesta arvioinnista, digitaalisesta arviointiprosessista sekä puhesuorituksista arvioitavista ulottuvuuksista

This study investigated human raters' perceptions of automated assessment of oral language skills. The raters (n = 37) participated in three assessment rounds organized by the DigiTala research project using Moodle and Zoom. The raters assessed Finnish and Swedish learners' speech samples using one holistic and five analytical rating scales created in the project. After the assessment, the raters responded to a questionnaire that included Likert-scale and open-ended questions. Numerical responses were analyzed with descriptive statistics, open responses with content analysis. The raters think that automated scoring could support human rating. The assessment rounds were carried out successfully. The selected dimensions proved to be essential parts of the speaking performances. The results will benefit those working on automated assessment and oral language assessment.

Keywords: automated assessment, language assessment, oral language skills
Asiasanat: automaattinen arviointi, kielitaidon arviointi, suullinen kielitaito

1 Johdanto

Digitaaliset kielikokeet ovat nykypäivää muun muassa ylioppilastutkinnossa, ja lähes jokaisessa oppilaitoksessa on jonkinlainen digitaalinen oppimis- ja opetusalusta. Tekniikka on mullistanut kielitaidon testaamisen ja harjoittelun, mutta täysin automaattiset suullisen kielitaidon kokeet ovat vielä harvinaisia kansainvälisestikin. Tekoäly on kuitenkin arkipäiväistymässä monilla elämänalueilla, ja se herättää monenlaisia mielikuvia ja pohdintaa sen vaikutuksista ja eettisyydestä.

Artikkelissa kartoitamme ja analysoimme ihmisarvioijien näkemyksiä digitaalisesti toteutetusta arviointiprosessista, jossa keräsimme ihmisarvioijilta arvioita suomen ja ruotsin puhenäytteisiin. Tutkimme myös arviointikierroksille osallistuneiden näkemyksiä suullisen kielitaidon automaattisesta arvioinnista. Kolmas tutkimuskysymyksemme (ks. luku 2) liittyy tehtävänannon ja puheen ulottuvuuksiin, joita puhesuorituksista arvioidaan.

Tutkimus on osa DigiTala-tutkimushanketta (Kautonen & von Zansen 2020), jossa kehitetään puheentunnistusta, automaattista puheen arviointia sekä automaattista palautetta suomen ja ruotsin oppijoille (von Zansen ym. arvioitavana). Automatiikka lisää mahdollisuuksia puheen itsenäiseen harjoitteluun (von Zansen ym. arvioitavana). Suurissa ja tärkeissä (*high-stakes*) kokeissa se voi toimia ihmisarvioinnin tukena, vähentää arvioijien työtaakkaa sekä lisätä arvioinnin objektiivisuutta. Ihmisen tekemään arvioon voivat vaikuttaa esimerkiksi erilaiset testattavan yksilölliset tai kielelliset piirteet, kuten aksentteihin liittyvät ennakkokäsitykset (Halonen ym. 2020).

Nykynäkemyksen mukaan arviointiin osallistuvien monien toimijoiden välinen tiedonkulku ja dialogi ovat tärkeitä laadukkaalle arvioinnille. Näitä osallisia (*stakeholders*) ovat niin laatijat ja käyttäjät kuin arvioitavatkin (Hidri 2020), ja kaikkien osapuolten näkemykset on syytä huomioida jo arviointivälineiden kehittelyvaiheessa. Arvioijien tai edes opettajien käsityksiä automaattisesta suullisen kielitaidon arvioinnista ei tietojemme mukaan ole aikaisemmin tutkittu. Tutkimustiedon avulla voimme kehittää automaattisen arvioinnin menetelmiä oikeudenmukaisiksi ja käyttökelpoisiksi.

2 Tutkimuksen taustaa ja tutkimuskysymykset

Tämän luvun alaluvut käsittelevät ensin suullisen kielitaidon arviointia (luku 2.1) sekä puheen automaattista arviointia (luku 2.2), minkä jälkeen luvussa 2.3 esitetään tutkimuskysymykset.

2.1 Suullisen kielitaidon arvioinnin lähtökohtia

Lukion opetussuunnitelmassa (Opetushallitus 2019) ja työelämän kielitaitovaatimuksissa puhumisen viestinnällinen asema tunnistetaan hyvin. Lukion päätteeksi suoritettavan ylioppilastutkinnon kielikokeissa ei kuitenkaan mitata suullista kielitaitoa käytännön kysymysten takia (Vaarala ym. 2021), vaikka puhumisen osakokeesta on keskusteltu vuosikymmeniä. Arvioinnin osittainen automatisointi tarjoaa kestävän ratkaisun siihen, että lukion päättökoe vastaisi paremmin opetussuunnitelman viestinnällistä kielitaitokäsitystä.

Automaattinen arviointi toisi lisää resursseja myös lukio-opintojen aikana tapahtuvaan puhumisen harjoitteluun, sillä opiskelijat saisivat puhesuorituksestaan palautetta ajasta ja paikasta riippumatta. Tällä hetkellä suullisen kielitaidon arvioinnin välineitä lukiossa ovat opettajan laatimat ja oppimateriaalista löytyvät tehtävät, Opetushallituksen laatimat suullisen kurssin koetehtävät sekä kieliprofiili (Inha ym. 2021). Näillä ei kuitenkaan ole samaa vaikuttavuutta kuin koulutusasteiden päättöarvosanoilla tai tutkinnoilla.

Suullisen kielitaidon arviointia hankaloittaa sen erityisluonne verrattuna kirjalliseen tuottamiseen. Puhuminen on tilannesidonnainen ja sosiaalinen taito, jonka arviointi on monimutkaista. Arvioijien yksilölliset mieltymykset ja tulkinnat vaikuttavat esimerkiksi taitotason määrittämiseen (Harsch & Hartig 2015). Arviointi ei ole yhteismitallista, koska arviointia tekevät saattavat kiinnittää huomiota kielen eri piirteisiin, minkä lisäksi heidän taustansa tai käsityksensä puhumisen taidosta saattavat olla erilaiset (Han 2022). Arvioijien päätöksentekoa perinteisessä puheen arvioinnissa on tutkittu runsaasti (Fan & Yan 2020). Painopiste on kuitenkin ollut arviointien keskinäisen vastaavuuden tilastollisessa tarkastelussa, vaikka validiuden rakentumisessa oleellista on myös se, miten arvioijat käyttävät asteikkoja ja perustelevat tulkintansa sekä se, miten arviointiympäristö ja prosessit tukevat validiutta (Knoch & Chapelle 2018).

Jotta arviointi olisi oikeudenmukaista, sen tulee olla tarkoituksenmukaista (*meaningful*), virheetöntä (*reliable*), yhdenvertaista (*absence of bias*), käyttäjien tarpeiden mukaista (*accommodation*), vastuullisesti toteutettua (*administration*) ja kohdentua sellaisiin henkilöihin, joilla on ollut mahdollisuus hankkia arvioitava taito esimerkiksi opiskelemalla sitä kieltä ja oppimäärää, jonka osaamista vaaditaan (*opportunity to learn*) (Kunnan 2018: 96). Tämä tutkimus kohdentuu arvioijien näkemyksiin. Näkemykset voivat liittyä mainittuihin validiusnäkökulmiin, mutta painotamme tässä tutkimuksessa kuitenkin tarkoituksenmukaisuutta ja toimeenpanoa.

Tarkoituksenmukaisuuteen kuuluu, että koetehtävät ja arviointitilanne vastaavat riittävästi sekä opiskeltua että oppilaitoksen ulkopuolista kielenkäyttöä eli koetehtävien pitää olla näissä suhteissa autenttisia (Kunnan 2018). Koulukokeissa pedagoginen autenttisuus on varsin helppo toteuttaa, jos opettaja ja kokeen laatija ovat sama henkilö. Koulutusjärjestelmän ulkopuolisten kielikokeiden (esim. Yleiset

kielitutkinnot ja Valtionhallinnon kielitutkinnot) tehtävien tulisi muistuttaa arjen tai työelämän kielenkäyttötehtäviä. Lisäksi arvioitava käsite eli suullinen kielitaito pitää määritellä siten, että se kattaa oleelliset taidot ja käyttöyhteydet, mutta ei sisällä asiaankuulumattomia virhelähteitä (*bias*) kuten puhujan kulttuuritausta tai henkilökohtaiset ominaisuudet (Bachman & Palmer 1996: 112). Mitä ratkaisevampia päätöksiä arvioinnin perusteella tehdään, sitä tärkeämpää on arvioinnin reliaabelius eli se, että arvioinnin toimeenpano ja arviointitulos sekä kaikki siihen vaikuttavat tekijät pysyvät mahdollisimman yhdenmukaisina arviointikerrasta toiseen.

Tietoa siitä, miten arvioijat kokevat automaattisesti ja digitaalisesti toteutettavan arviointiprosessin, on vielä niukasti (automaattisen puhumisen arvioinnin reliabiliteetista ks. Khabbazbashi & Galaczi 2020). Automaattisen arvioinnin ja ihmisarvioinnin tulosten yhtäpitävyyttä on tutkittu enemmän kuin itse arviointiprosessia ja arvioijien käsityksiä siitä (Han ym. 2020; Ouyang ym. 2021). Tieto ihmisarvioijien kokemuksista digitaalisessa arviointiympäristössä auttaa parantamaan tehtäviä, asteikkoja ja digitaalisen arvioinnin oikeudenmukaisuutta.

2.2 Puheen automaattisen arvioinnin nykytilanne ja tulevaisuuden näkymät

Suullisen kielitaidon automaattisen arvioinnin perustana on automaattinen puheentunnistus eli puhesignaalin koneellinen muuntaminen tekstiksi. Puheentunnistusmenetelmien nopea kehitys on mahdollistanut muun muassa puheohjattavien laitteiden ja sovellusten tulon osaksi arkielämäämme. Automaattisia suullisen kielitaidon arviointijärjestelmiä ei kuitenkaan ole vielä yleisesti käytössä kotimaisille kielille. Sen sijaan englannin kielessä on jo täysin automatisoituja suullisen kielitaidon testejä (Educational Testing Service 2014; Pearson 2017). Näissä järjestelmissä kielenoppija vastaa koekysymyksiin tietokoneella ja palvelimella akustinen puhesignaali pilkotaan numeerisiksi piirteiksi, jotka kuvaavat esimerkiksi ääntämistä, sujuvuutta sekä kieliopin tai sanaston hallintaa. Nämä suullisen kielitaidon eri osa-alueita kuvaavat piirteet yhdistetään ja niistä muodostetaan arvio tutkimuksiin perustuvien matemaattisten mallien ja kaavojen avulla. Mallien pohjana on aina ihmisasiantuntijoiden tekemiä arvioita eritasoisten kielenoppijoiden puheesta, ja arviointialgoritmi vertaa näiden ihmisten arvioimien puhenäytteiden piirteitä automaattisesti testattavien puhenäytteiden vastaaviin piirteisiin. (Loukina & Yoon 2019.)

Kotimaisten kielten automaattista arviointia on jarruttanut sopivien aineistojen puute. Puheen automaattisen arvioinnin kehittäminen vaatii suurta määrää eritasoisten kielenoppijoiden tallennettua puhetta ja ihmisasiantuntijoiden antamia arvioita siitä. Koska puheentunnistinten akustiset mallit pohjautuvat äidinkieltään puhuvien henkilöiden näytteisiin, tunnistin pitää mukauttaa kielenoppijan puheeseen (Ylinen & Kurimo 2017). Lisäksi kielenoppijan puhe on usein epäsujuvaa, ja siinä voi olla monenlaisia ääntämis- ja kielioppivirheitä, jotka hankaloittavat sanojen tunnis-

tamista. DigiTala-hankkeessa kerätään automaattisen suullisen kielitaidon arviointiin sopivaa tietokantaa ja tutkitaan kielenoppijoiden puheelle tyypillisiä puheen piirteitä.

Jotta automaattinen arviointialgoritmi olisi luotettava ja tarkka, tarvitaan tutkimuksia selvittämään, mitkä puheen mitattavat ominaisuudet eli parametrit ovat tärkeitä kielitaidon arvioinnissa. Osa parametreista on kieliriippuvaisia, kuten esimerkiksi ääntämiseen ja painotukseen liittyvät piirteet. Toisaalta esimerkiksi puheen sujuvuutta kuvaavien parametrien on todettu toimivan kielitaidon mittarina useilla eri kohdekielillä (ks. esim. Préfontaine ym. 2016; Kallio ym. 2017; Kang & Johnson 2018). DigiTala-hankkeen alustavat tulokset osoittavat, että muiden kielten automaattisessa arvioinnissa käytetyt puheen sujuvuuden parametrit ennustavat hyvin myös suomenoppijoiden suullista taitotasoa (Kallio ym. 2022). Kotimaisten kielten, suomen ja ruotsin, suulliseen kielitaitoon vaikuttavia puheen piirteitä on kuitenkin tutkittu melko vähän. Suomalaisten ruotsinoppijoiden puheesta on tutkittu prosodisia piirteitä (Kautonen 2019; Kautonen & Kuronen 2021; Kallio ym. 2022), sujuvuutta (Kallio ym. 2017; Kallio ym. 2022) sekä rakenteita (Kautonen & Kuronen, 2021). Myös suomalaisten puhuman ruotsin ääntämisen automaattista arviointia on tutkittu (Toivanen 2016). Toivasen kuvaamassa kokeessa automaattinen luokittelu eri taitotasoihin kuuden parametrin avulla oli tarkkaa, mutta puheaineisto oli pieni ja koostui lyhyistä (10 s) näytteistä.

Nykyisillä suullisen kielitaidon automaattisilla arviointityökaluilla ei voida arvioida kaikkia puheen piirteitä. Esimerkiksi vuorovaikutuksen onnistumista kuvaavia, automaattisesti mitattavia parametreja ei vielä käytetä laajalti. Tämän takia esimerkiksi rajaamalla tehtävävalikoimaa mahdollistetaan puheen piirteiden täsmällisempi mittaaminen ja siten parannetaan automaattisen arvioinnin tarkkuutta. Samalla kuitenkin osa suullisen kielitaidon osa-alueista voi jäädä arvioinnin ulkopuolelle.

Koska kielikokeet vaikuttavat opetuksen sisältöön (*washback*, Bachman & Palmer 1996; *impact*, Zhang ym. 2020: 22), on tärkeää tietää, missä määrin automaattinen arviointi mittaa kielitaitoa puutteellisesti. Puheentunnistimen on helpompi tunnistaa lukupuhetta tai sellaista puhetta, jonka sisältö on ennakoitavissa. Tästä syystä monissa automaattisissa arviointijärjestelmissä tehtävävalikoima keskittyy ääneen luettaviin tai mallin mukaan toistettaviin lauseisiin, visuaalisten ärsykkeiden kuvailuun ja mielipiteen ilmaisemiseen annetusta aiheesta (Zechner & Evanini 2020). Lisäksi jokaiselle tehtävälle on yleensä oma arviointimallinsa, joka huomioi sen, mitä tehtävän vastaukselta odotetaan. Esimerkiksi vuorovaikutuksen onnistumista automaattisissa testeissä ei kuitenkaan mitata. Joidenkin tutkijoiden mielestä tämä ei ole ongelma, koska spontaania puhetta tuottavat monologitehtävät vaativat samoja perustavanlaatuisia psykolingvistisiä ja kielellisiä taitoja, joita tarvitaan monimutkaisemmissakin kommunikointitilanteissa (Bernstein ym. 2010).

2.3 Tutkimuskysymykset

Artikkelissa analysoimme ihmisarvioijien näkemyksiä suullisen kielitaidon automaattisesta arvioinnista. Lisäksi tutkimme arvioijien käsityksiä automaattiseen arviointiin liittyvien menetelmien ja kehitysvaiheiden toimeenpanosta. Pyrimme vastaamaan seuraaviin tutkimuskysymyksiin:

1. Mitä näkemyksiä arviointiin osallistuneilla on suullisen kielitaidon digitaalisesta arviointiprosessista, johon he osallistuivat?
2. Mitä näkemyksiä arvioijilla on suullisen kielitaidon automaattisesta arvioinnista?
3. Mitä tehtävänannon ja puheen ulottuvuuksia puhesuorituksista tulisi arvioijien mielestä arvioida?

Kaikki kysymykset tuottavat tietoa Kunnanin (2018) validiuskehyksen eri alueiden toteutumisesta, kun taas kolmas kysymys kohdentuu erityisesti puhumisen käsitteen tarkoituksenmukaisuuteen.

Näkemykset tarkoittavat tässä tutkimuksessa arvioijilta saatua palautetta, heidän mielipiteitään ja suhtautumistaan. Arviointiprosessi pitää sisällään arviointiin annetut ohjeistukset ja kriteerien käytön, arvioijien koulutuksen sekä Moodlen käytön arviointiympäristönä. Tehtävänannon ja puheen ulottuvuudet tarkoittavat osa-alueita, joihin arviointi kohdistuu. Puhenäytteestä arvioitavat ulottuvuudet vaikuttavat tehtävänlaadintaan, arviointikriteereihin sekä mitattavan käsitteen määrittelyyn.

3 Aineisto ja menetelmät

Tutkimushanke järjesti talven 2020 ja kesän 2021 välillä kolme arviointikierrosta, joihin osallistui yhteensä 37 arvioijaa. Tämän artikkelin aineistona käytetään näiden arviointikierrosten jälkeen kerättyjä kyselyvastauksia. Tässä luvussa esitellään ensin luvussa 3.1 arvioijat, luvussa 3.2 tutkimusasetelma sekä luvussa 3.3 aineiston analyysimenetelmät.

3.1 Arvioijat

Arvioijien valitsemisessa käytettiin harkinnanvaraista näytettä, eli arvioijiksi rekrytoitiin suomen ja ruotsin kielten asiantuntijoita, joilla oli aiempaa kielitaidon arviointikokemusta Yleisissä kielitutkinnoista (YKI) tai Ylioppilastutkinnosta. Lisäksi tutkijat rekrytoivat arvioijia aiemman tutkimushankkeen ruotsin arviointeihin osallistuneista lukio-opettajista sekä henkilökohtaisten verkostojensa kautta. Ulkopuolisten arvioijien lisäksi projektitiimistä yhteensä viisi tutkijaa osallistui arviointikierroksiin. Arvioijat koulutettiin tehtävään ja projektin ulkopuolisille arvioijille maksettiin yhtä työpäivää vastaava korvaus.

Lukiolaisten ruotsi -arviointikierrokselle osallistui 18 ruotsin asiantuntijaa, joista neljä oli tutkijoina DigiTala-projektissa ja loput olivat ammatiltaan ruotsin opettajia tai kielitaidon arvioijia. Arvioijien äidinkielet olivat suomi (n=14) tai ruotsi (n=5), ja yksi arvioijista oli kaksikielinen.

Suomenkielistä aineistoa arvioi yhteensä 22 arvioijaa, joista neljä oli DigiTala-projektin tutkijoita ja loput suomen kielen opettajia eri oppilaitoksissa sekä YKI-arvioijia. Myös suomen arvioijilla oli kokemusta suullisen kielitaidon arvioinnista, kaikilla vähintään vuosi. Arvioijien äidinkielet olivat suomi (n=19) ja venäjä (n=3). YKI-puhujien suomi -arviointikierrokselle osallistui 20 arvioijaa. Lukiolaisten suomi -arviointikierrokselle osallistui 14 arvioijaa, joista 12 oli samoja henkilöitä kuin ensimmäisellä suomen kierroksella. Yleiskuva arviointikierroksista esitellään seuraavan luvun taulukossa 1.

3.2 Tutkimusasetelma

Arvioinnit ja kyselyvastaukset kerättiin Moodlessa (versio 3.8.3). Arvioijille järjestettiin Zoomissa koulutus, jossa käytiin läpi arviointiin liittyvät ohjeet ja kriteerit (von Zansen 2022a) sekä esiteltiin taitotasoja konkretisoivat maamerkkinäytteet. Arvioijille annetut ohjeet käsittelivät muun muassa kriteerien ja maamerkkien käyttöä, rauhallista kuunteluympäristöä, taukojen pitämisen tärkeyttä sekä kuuntelukertojen määrää.

Taulukossa 1 esitellään arviointikierroksiin liittyvät tehtävät, käytetyt kriteerit ja kyselylomakkeet. Kuten taulukosta 1 on nähtävissä, arviointikierroksilla arvioitiin eri tehtäviä. Lisäksi arviointiasteikkoihin (von Zansen 2022a) ja kyselylomakkeisiin (von Zansen 2022b, 2022c, 2022d) tehtiin arviointikierrosten välillä saadun palautteen perusteella pieniä muutoksia.

TAULUKKO 1. Yleiskuva kolmesta arviointikierroksesta.

Arviointi-kierros	**Puhenäytteet**	**Puhetehtävät**	**Arviointi**	**Kysely**
1. Lukiolaisten ruotsi (marraskuu 2020)	2025 näytettä 181 puhujaa 7,1 h puhetta keskimäärin 12,7 s	4 eri tehtävää 22 osatehtävää Lyhyitä reagointitehtäviä	18 arvioijaa 4134 arviointia 5x analyyttinen (skaala 0–3), holistinen (alle A1–C1)	14 kysymystä (6 avointa)
2.YKI -puhujien suomi (toukokuu 2021)	401 näytettä 204 puhujaa 12,5 h puhetta keskimäärin 101,9 s	4 kertomis-tehtävää	20 arvioijaa 1240 arviointia 5x analyyttinen (0–3/4), holistinen (alle A1–C2)	10 kysymystä (3 avointa)
3. Lukiolaisten suomi (kesäkuu 2021)	882 näytettä 62 puhujaa 4,3 h puhetta keskimäärin 17,4 s	8 eri tehtävää 26 osatehtävää Ääneenluku, lyhyt reagointi, pidempi monologi	14 arvioijaa 1660 arviointia 5x analyyttinen (0–3/4), holistinen (alle A1–C2)	9 kysymystä (2 avointa)

Tutkimushankkeen laatimat arviointikriteerit koostuvat holistisesta taitotasoasteikosta sekä viidestä analyyttisestä kriteeristä (von Zansen 2022a). Arviointikriteerit laadittiin vastaamaan lukion opetussuunnitelman kielitaitokäsitystä ja oppimistavoitteita. Arviointikriteerien laadinnassa hyödynnettiin etenkin aiemman opetussuunnitelman (Opetushallitus 2003) kielitaidon tasojen kuvausasteikkoa, joka sopii nykyisten opetussuunnitelmien taitotasoasteikkoa paremmin DigiTala-hankkeen analyyttisiin ja teknisiin lähtökohtiin. DigiTala-tutkimushankkeessa puhenäytteille määritetään taitotaso, jonka lisäksi puhenäytteistä arvioidaan seuraavia ulottuvuuksia: tehtävänannon täyttymistä, sujuvuutta, ääntämistä, ilmauksen laajuutta sekä rakenteiden tarkkuutta (Kautonen & von Zansen 2020).

Ääneenlukutehtävässä arvioitiin ainoastaan sujuvuutta ja ääntämistä, mutta muissa puhenäytteissä arvioijat määrittivät kunkin näytteen taitotason ja antoivat sille viisi analyyttistä arviota. Holistinen taitotasoarvio perustui taitotasokuvauksiin ja taitotasoja kuvaaviin maamerkkinäytteisiin. Analyyttisiä arvioita ei linkitetty taitotasoarvioihin, vaan puhenäytteestä arvioitiin kukin analyyttinen ulottuvuus itsenäisenä ominaisuutena.

Arvioitavat ruotsin näytteet kerättiin vuonna 2015 aiemmassa tutkimushankkeessa. Puhujat opiskelivat ruotsia toisena kotimaisena kielenä lukiossa. Arvioitavat näytteet sisälsivät vastauksia neljään eri tehtävään, jotka edellyttivät lyhyttä reagointia (esimerkiksi "vastaa kysymykseen" tai "mitä sanot ruotsiksi, kun...") tai pidem-

pää puhetuotosta (esimerkiksi "kuvaile retkipäivän säätä"). Tehtävänannot sisälsivät tekstiä, kuvia ja äänitteitä.

Arvioitavat YKI-puhujien suomen näytteet olivat peräisin suomen kielen keskitason (n=164) ja ylimmän tason (n=40) kokeeseen osallistuneilta puhujilta, joilta oli saatu lupa aineiston tutkimuskäyttöön. Arvioitavat puhenäytteet olivat neljään kertomistehtävään liittyviä suorituksia (tavoitekesto 1–1,5 minuuttia). Tehtäväkohtainen ohjeistus oli tekstimuodossa, ja tehtävissä oli lisäksi ohjaavia apukysymyksiä.

Lukiolaisten suomen näytteet kerättiin Toinen kotimainen kieli (finska) -oppimäärää sekä Suomi toisena kielenä ja kirjallisuus (S2) -oppimäärää opiskelevilta lukiolaisilta. Tehtävät sisälsivät ääneenlukutehtävän ("lue lauseet ääneen") lisäksi lyhyitä reagointitehtäviä (esimerkiksi "vastaa kuulemiisi kysymyksiin" tai "vastaa kommentteihin webinaarissa") ja pidempiä kertomistehtäviä (esimerkiksi "kerro tärkeästä paikasta" tai "kerro mitä näet kuvassa"). Tehtävänannot sisälsivät tekstiä, kuvia, äänitteitä ja videota.

Arvioitavien puhenäytteiden määrä vaihteli arviointikierrosten ja arvioijien välillä sen mukaan, minkä pituisia puhenäytteitä oli arvioitavina. Puhenäytteet pyrittiin jakamaan arvioijille niin, että niiden arvioimiseen kuluisi enintään noin seitsemän tuntia arvioijaa kohden. Arvioinnit toteutettiin Moodle-tenttinä. Kukin arvioija sai tehdä arvioinnit omassa tahdissaan. Moodlessa oli mahdollisuus kysyä ja keskustella arviointikriteereistä. Lisäksi projektin tutkija vastasi arvioijien kysymyksiin sähköpostitse ja puhelimitse.

Arviointien jälkeen arvioijat vastasivat kyselyyn Moodlessa. Kyselylomakkeen kysymykset koskivat puhetehtävien soveltuvuutta valittujen piirteiden arvioimiseen, arviointikriteerien, -ohjeiden ja -ympäristön selkeyttä ja tarkoituksenmukaisuutta sekä arvioijien näkökulmia automaattisesta arvioinnista. Kysely sisälsi Likert-asteikollisia ja avoimia kysymyksiä, joihin arvioijat vastasivat kirjallisesti.

Kyselylomaketta testattiin ennen käyttöönottoa hankkeen tutkijoilla sekä muutamalla projektin ulkopuolisella henkilöllä. Testaamisella haluttiin varmistaa, että lomakkeen kysymykset ymmärretään oikein ja ne mittaavat haluttuja asioita. Kyselylomakkeen avulla haluttiin kerätä palautetta käytetyistä puhetehtävistä, arviointityökalusta ja arviointikriteereistä. Lisäksi haluttiin kerätä arvioijien näkemyksiä puhumisen arvioinnista sekä tietoa heidän suhtautumisestaan automaattiseen arviointiin. Laadinnassa kiinnitettiin huomiota hyviin käytänteisiin, jotka liittyvät tieteellisen kyselyn laatimiseen (esim. selkeys, pituus, kieli, ks. Yhteiskuntatieteellinen tietoarkisto 2022). Tämän lisäksi tutkijat hyödynsivät kyselylomakkeen laadinnassa aiempaa kokemustaan ja teoreettista ymmärrystään kielitaidon arvioinnista. Kyselylomakkeeseen tehtiin arviointikierrosten välillä pieniä muutoksia, jos havaittiin, että jotkin osiot ovat epäselviä tai tarpeettomia (esimerkiksi kysymysten yhteydessä olevia avoimia kenttiä vähennettiin). Kyselylomakkeiden (von Zansen 2022b, 2022c, 2022d) sisältö pysyi kuitenkin olennaisilta osin samana arviointikierrosten välillä.

3.3 Aineiston analyysimenetelmät

Ennen aineiston analyysia kyselyvastaukset vietiin Moodlesta taulukkolaskentaohjelmaan (Excel 2016), josta poimimme avovastaukset tekstitiedostoihin. Kahden kysymyksen kohdalla ruotsin kyselyvastauksissa (ks. von Zansen 2022b kysymykset 9 ja 10) oli käytetty suppeampaa 4-portaista skaalaa, joten nämä vastaukset koodattiin vastaamaan suomen arvioijien kyselyjen (von Zansen 2022c, 2022d) 5-portaista skaalaa vastausvaihtoehtoja muuttamatta. Osa arvioijista osallistui useampaan kuin yhteen arviointikierrokseen (ks. luku 3.1). Arviointikierroksia yhdistävissä analyyseissa heidän osaltaan aineistoon sisällytettiin ainoastaan arvioijan ensimmäiset kyselyvastaukset. Likert-asteikolliset kyselyvastaukset analysoitiin tilastollisella kuvailevalla analyysillä Excel 2016 -ohjelmalla. Vastauksista laskettiin frekvenssit, keskiarvot ja keskihajonnat. Tämän lisäksi vastausten jakaumia esitetään ositettujen palkkikaavioiden avulla.

Avoimet vastaukset analysoitiin sisällönanalyysillä (Vuori n.d.) käyttäen Atlas.ti -ohjelmaa (versio 9). Aineiston koodaus oli iteratiivinen prosessi, johon osallistui kaksi tutkijaa. He järjestivät aineistoa vuorotellen ja keskustelivat aineiston luokittelusta ja tehdyistä havainnoista videopuheluiden välityksellä. Aineiston alustava analyysi eteni aineistolähtöisesti (Vuori n.d.). Tämän vaiheen ja tutkimuskirjallisuuden perusteella muotoiltiin tutkimuskysymykset. Sen jälkeen avoimia kyselyvastauksia ryhmiteltiin Atlas.ti:n avulla neljään pääluokkaan (kriteerit, tehtävät, prosessi, automaattinen arviointi), jotka sisälsivät myös alaluokkia (esim. holistinen arviointi, analyyttinen arviointi, Moodle arviointiympäristönä). Alustavaan analyysiin verrattuna toinen koodauskierros eteni teoriavetoisemmin (Vuori n.d.), sillä tulkinta rakentui aiemman suullisen kielitaidon arviointiin liittyvän tutkimustiedon (ks. luku 2) varaan. Tämän vaiheen jälkeen aineistoa tarkasteltiin myös arviointikierroksittain, jolloin huomiota kiinnitettiin osa-aineistojen välisiin eroihin ja yhteneväisyyksiin (Vuori n.d.). Lopulta sisällönanalyysi eteni tutkimuskysymyksittäin: arvioijien vastauksia koostettiin ja jäsennettiin kuhunkin aihepiiriin (arviointiprosessi, automaattinen arviointi, arvioitavat ulottuvuudet) liittyen. Artikkeliin valitut suorat lainaukset arvioijien kirjallisista huomioista poimittiin numeeristen analyysien valmistuttua. Määrällisen ja laadullisen aineiston perusteella muodostettiin kokonaiskuva tutkittavasta ilmiöstä (ks. luku 2.3).

4 Tulokset

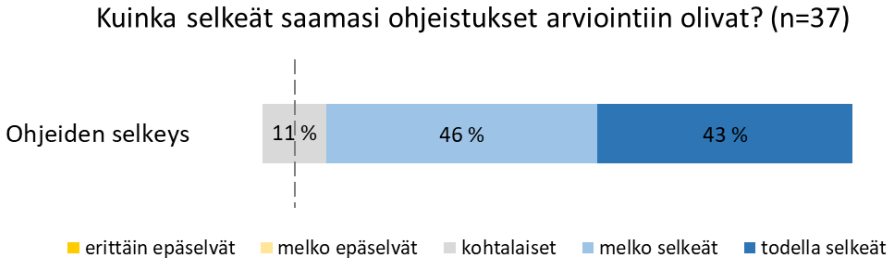
Seuraavissa alaluvuissa esittelemme tutkimustulokset tutkimuskysymyksien mukaisessa järjestyksessä. Esittelemme arvioijien näkemyksiä arviointiprosessista (luku 4.1) ja suullisen kielitaidon automaattisesta arvioinnista (luku 4.2). Luvussa 4.3 ker-

romme tehtävänannon ja puheen ulottuvuuksista, joita arvioijien mielestä tulisi arvioida puhesuorituksissa.

4.1 Näkemykset digitaalisesta arviointiprosessista

Tässä luvussa käsittelemme ensimmäistä tutkimuskysymystä. Esittelemme arvioijien kyselyvastauksia koskien konkreettista arviointiprosessia eli arviointiohjeita ja -koulutusta, kriteerien käytön omaksumista sekä Moodlea arviointiympäristönä.

Arviointiin annetut ohjeet liittyvät keskeisesti arviointiprosessiin. Kuten kuviosta 1 näkyy, valtaosa arvioijista koki arviointiin saamansa ohjeistukset joko selkeiksi tai melko selkeiksi (ka. 4,3). Toisaalta Lukiolaisten ruotsi -arviointikierroksella lyhyiden näytteiden arvioitavuus ja muiden kielten käyttö kohdekielen rinnalla mietityttivät arvioijia (lainaukset 1 ja 2).



KUVIO 1. Arvioijien kokemus ohjeiden selkeydestä.

(1) Olisiko pitänyt tarkemmin määritellä ”Ei voi arvioida”-tapaukset. Esim. yhden sanan vastaukset ”Ja” ja ”Nej”, jotka kylläkin vastaavat tehtävänantoon osittain.

(2) – – Myös suomenkielisten sanojen käytöstä olisimme voineet linjata jotain. Monissa näytteissä esiintyi suomea välissä, viesti ei välity. Eli silloin on ehkä turhanaikaista ja mahdotontakin alkaa pohtia lauseprosodiaa.

Digitaaliset työkalut ja tietoliikenneyhteydet mahdollistivat arviointiprosessin toteuttamisen täysin etänä. Zoomissa järjestetty koulutus ja tallenteen jakaminen koettiin hyödylliseksi (lainaus 3), sillä käytetyt arviointikriteerit ja tehtävätyypit olivat arvioijille uusia. Analyyttisten kriteerien käyttö oli arvioijille vierasta, sillä he ovat tottuneempia arvioimaan puhetta laajempina kokonaisuuksina. Toisille uusien arviointikriteerien omaksuminen oli vaivattomampaa (lainaus 4), toisille taas haastavampaa, mikä näkyi arvioinnin hitautena (lainaukset 5 ja 6). Maamerkkien lisäksi arvioijat kaipasivat lisää esimerkkejä, jotka tukisivat analyyttisten kriteerien käyttämistä (lainaus 7).

(3) Koulutustilaisuus oli hyvä (katsoin myös videon jälkeenpäin).

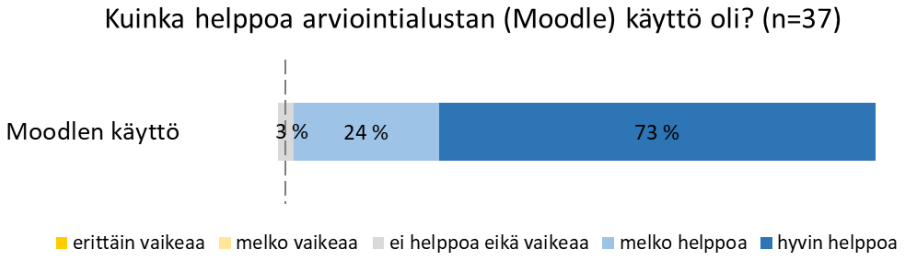
(4) Oli mukava, että taitotasojen lisäksi oli muitakin arviointikriteerejä

(5) – – Vaikka YKIssä olen tottunut tekemään sekä analyyttista että holistista arviota, tämä oli paljon hankalampaa, kun piti samanaikaisesti arvioida niin monia asioita. Etenin arvioinneissa varsin hitaasti.

(6) – – Det tog också mycket tid med nödvändiga pauser, eller så kunde man inte vara så noggrann som önskat..

(7) Kuitenkin huomasin näitä tehdessä että analyyttiset kriteerit eivät olleetkaan minulle niin selviä. Niistä olisi voinut olla enemmän erilaisia esimerkkejä.

Arvioinnit kerättiin Moodlessa, joka oli osalle uusi ympäristö. Kuten kuviosta 2 nähdään, Moodlen käyttö oli arvioijille pääsääntöisesti helppoa (ka. 4,7), vaikkakin sanallisessa palautteessa arvioijat toivat esiin epäkohtia (lainaukset 8–10). Häiritsevin ongelma liittyi ohjelmointivirheeseen, jonka takia arviointinäkymä varoitti puutteellisesta vastauksesta. Lisäksi arvioijat kokivat Moodlessa toimimiseen vaaditut siirtymät turhauttavaksi. Sen sijaan äänitteiden välillä liikkumisen mahdollistanut navigointinäkymä koettiin hyödylliseksi. Vaikeuksista huolimatta arvioijat suhtautuivat myönteisesti Moodle-tentin käyttöön arviointiympäristönä (lainaukset 11 ja 12).



KUVIO 2. Arvioijien kokemus arviointialustan käytöstä.

(8) Päätettyäni tentin, kaikissa luki puutteellinen vastaus. Palautin silti - toivottavasti olin tehnyt oikein. Mitä pisteet tarkoittivat? Sain 54% oikein?

(9) Unohdin vastata tähän kyselyyn, eli tämä ei ole virtaviivainen alusta, siirtymät ongelmallisia. Lisäksi arviointinäkymässä oli bugi, joka väitti vastauksia puuttuvan. Tuli epävarma olo, tallentuvatko kaikki arvioinnit. Kriteerit olivat hyvin esillä, samoin tehtävänannot ja äänitteiden kuunteleminen sujui ongelmitta. Oli hyödyllistä päästä liikkumaan vastausten välillä. – –

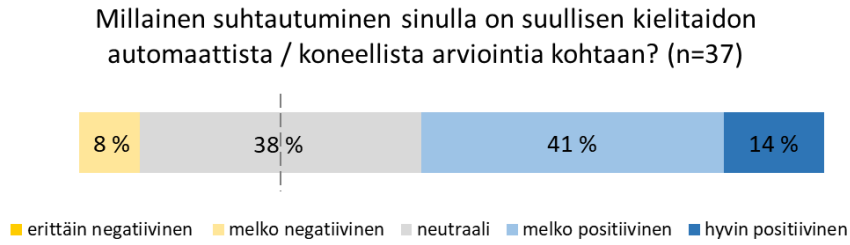
(10) Hiukan kömpelöhän tuo Moodle on. Valittaa keskeneräisistä vastauksista ym. En uskaltanut painaa Palauta ja lopeta -nappia, ja arvioinnit saattoivat jäädä roikkumaan puolivalmiiseen tilaan.

(11) – – [Moodlessa] Tentti toimi kuitenkin hyvin ja arviointi oli teknisesti sujuvaa ja helppoa.

(12) – – Moodle on erinomainen alusta tähänkin. – –

4.2 Suhtautuminen automaattiseen arviointiin

Tässä alaluvussa tarkastelemme toista tutkimuskysymystä. Pyysimme arvioijia kertomaan heidän suhtautumisestaan automaattiseen arviointiin, ja kuviossa 3 esitetään arvioijien vastaukset viisiportaisella mielipideasteikolla. Enemmistö vastaajista suhtautui automaattiseen arviointiin positiivisesti tai neutraalisti (ka. 3,6). Sanallisissa vastauksissa arvioijat kommentoivat, että automaattinen arviointi sopisi heidän mielestään ihmisarvioinnin tueksi (lainaus 13) tai lisäksi, josta esimerkkeinä itsenäinen harjoittelu (lainaus 14), lähtötason kartoitus (lainaus 15) ja adaptiiviset eli mukautuvat kielitestit (lainaus 16).



KUVIO 3. Arvioijien suhtautuminen automaattiseen arviointiin.

(13) Automaattinen kielitaidon arviointi sopii mielestäni parhaiten tukemaan ihmisen tekemää arviointia. Kuitenkin uskon, että mahdollisuuksia on valtavasti.

(14) itsenäiseen harjoitteluun, esitestiksi jonka perusteella esim valitaan sopiva testitehtävä

(15) massatestauksiin esimerkiksi alkukartoituksen ja ohjauksen apuna

(16) Ihmisen apuna, alustava taitotasoarvio, adaptiivisessa kielikokeessa sopivantasoisten tehtävien löytäminen jne

Arvioijat näkivät automaattisen arvioinnin myös mahdollisuutena parantaa arvioinnin luotettavuutta ja käytettävyyttä (lainaukset 17–19).

(17) ihmisarvioinnin tukena konearviointi on erittäin hyvä asia, sillä voi lisätä arvioinnin luotettavuutta esim. ääntämisen, sanaston laajuuden, sujuvuuden jne. kannalta

(18) – – Tämä on tärkeä ja kannatan vahvasti suullista arviointia ja ymmärrän hyvin, että resurssipulan takia se onnistuu suuremmissa määrin ainoastaan koneellisesti.

(19) Kone ei väsy eikä tee satunnaisvirheitä ja kohtelee kaikkia suorituksia samalla tavalla. Tasolle B1 saakka hyvin soveltuva ratkaisu isojen suorittajamäärien arviointiin.

Vastauksissaan arvioijat toivat kuitenkin esiin myös automaattiseen arviointiin liittyviä haasteita ja rajoituksia sekä peräänkuuluttivat tutkimustietoa (lainaukset 20 ja 21). Joillakin vastaajilla oli jo jonkin verran kokemusta puheentunnistukseen perustuvista työkaluista (lainaukset 22 ja 23). Heitä mietitytti se, miten automaattinen arviointi soveltuu spontaanin puheen arviointiin sekä miten erilaiset puhujat huomioidaan (lainaukset 21 ja 22).

(20) Miten kone voi korvata ihmisen tässä tapauksessa? Tästä olisi hyvä saada enemmän tutkimustietoa.

(21) – – Tämä on varmasti väistämätön kehitys tulevaisuudessa, mutta en voi olla ajattelematta, että tässä teen itseäni ja kielenopettaja kollegoitani työttömäksi (vähän kärjistäen :-)).. Lisäksi en usko että kone pystyy milloinkaan täydellisesti arvioimaan spontaania puhetta?

(22) – – Ollessani mukana XX:n hankkeessa, jossa kehitettiin peliä, jossa edettiin puheentunnistuksen avulla, nousi esiin monia ongelmia mm. paikallismurteen tunnistamisessa ja persoonallisen tyylin hyväksymisessä. – –

(23) – – Olen ollut hankkeessa, jossa puheen ymmärtämistä kehitettiin ja vaikeaa se on. Avoimin mielin suhtaudun koneelliseen arviointiin.

Lisäksi arvioijat miettivät, mihin heidän tulisi kiinnittää huomiota, kun he työskentelevät automaattisen arvioinnin kehittämistä varten – he arvelivat, että automaattisessa arvioinnissa keskeistä on kiinnittää huomiota yksityiskohtiin (lainaukset 24 ja 25). Arvioijat pohtivat automaattisen arvioinnin reunaehtoja ja omaa rooliaan sen kehittämisessä (lainaus 26).

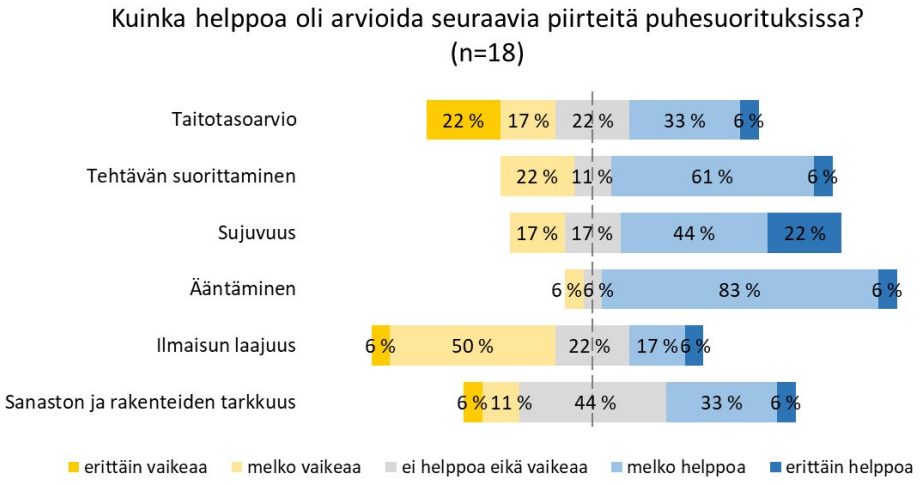
(24) – – Mietin, täytyykö koneellisesti arvioidussa puheessa olla tarkempi painotuksissa, intonaatiossa ja äänteissä kun ihmisten välisessä vuorovaikutuksessa, jossa kuuntelija saattaa silti ymmärtää, vaikka niissä olisikin virheitä.

(25) – – Tavallaan pitäisi tässä arvioitsijana ajatella, kuinka meidän antamamme vastaus vaikuttaa tekoälyn arviointitaitoon, millä tulee eroteltua nyanssit ja sanavalinnat.

(26) – – Kokemus tästä arvioinnista oli se, että lyhyitä puheenvuoroja on vaikea arvioidakaikilla kriteereillä. Tässä oli monta tehtävää, joihin ei tarvinnut tuottaa pitkää puheenvuoroa. Koneellisessa arvioinnissa niin kaiketi täytyy ollakin. Jäljelle jää usein kysymyksiä: Olinko oikeudenmukainen hitaalle puhujalle? Vaadinko turhan täydellistä vastausta kysymykseen, josta selviytyy kunnialla lyhyemmälläkin vastauksella.

4.3 Tehtävänannon ja puheen ulottuvuuksien arviointi

Kolmas tutkimuskysymyksemme käsittelee tehtävänannon ja puheen ulottuvuuksia, joita arvioijien mielestä tulisi puhesuorituksissa arvioida. Aloitamme esittelemällä arviointikierroksittain arvioijien näkemyksiä arvioinnin helppoudesta. Lukiolaisten ruotsi -arviointikierroksella arvioijilla (n=18) oli haasteita etenkin taitotason määrittelyssä (ka. 2,8, kuvio 4) ja arvioitaessa ilmaisun laajuutta (ka. 2,7). Lukiolaisten ruotsi -arviointikierroksella vaikeuksia aiheuttivat erityisesti puhenäytteiden lyhyys ja tehtävissä edellytetyn sanaston vaikeus (lainaus 27). Lisäksi arvioijat toivoivat selkeämpää tehtäväkohtaista ohjeistusta kriteerien käyttöön sekä tarkempaa määrittelyä tehtävänannon täyttymiselle – jälleen puhenäytteiden lyhyyteen viitaten (lainaukset 28 ja 29).



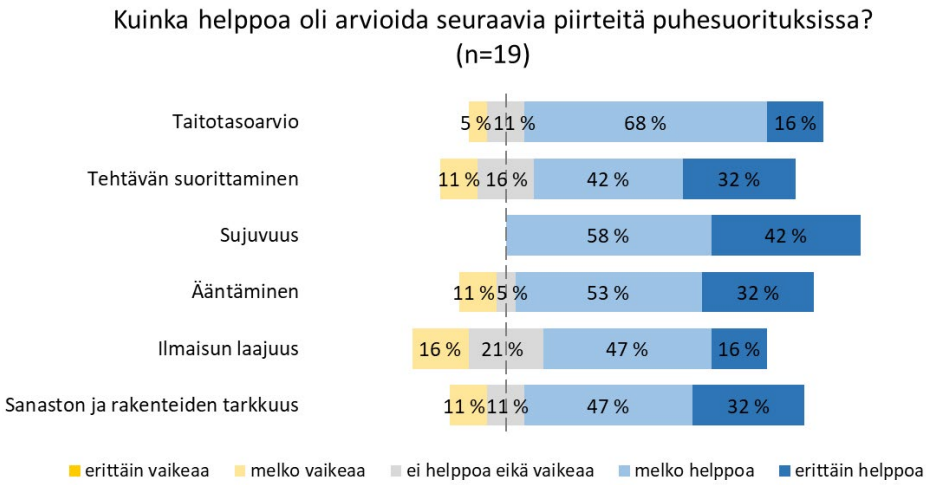
KUVIO 4. Arvioijien kokemus tehtävänannon ja puheen ulottuvuuksien arvioinnista Lukiolaisten ruotsi -kierroksella.

(27) – – taitotasosta on minusta suht mahdotonta antaa arvioita näin lyhyen vastauksen perusteella; yleensä se vaatii vähintään 10 min kuuntelua, jolloin suoritus usein paranee kun alun jännittyneisyys poistuu. – – Tehtävät olivat mielestäni aika vaikeita. En halunnut kauheasti sakottaa siitä että joku ei muistanut sanoja ”dimma/dimmigt” tai tack för «senast”. ”Ledig” oli myös vaikea monelle. Hankalaa kun koko lause/suoritus monesti jää yhdestä sanasta kiinni.

(28) – – joidenkin oppilaiden/suoritusten kohdalla kriteerit olivat mahdottomia käyttää. Ne eivät sopineet tilanteeseen lainkaan. Esim. ääntäminen, virheettömyys, jos oppilas sanoi vain tack. – – Voinnin kyselytehtävä: jos oppilas ei tehnyt vastakysymystä, oliko tehtävänannon noudattaminen puutteellista jne. ?

(29) – – Pohdin myös sitä, miten arvioida suorituksia reagointitehtäviin ts. milloin tehtävänanto täyttyy. – –

YKI-puhujien suomi -arviointikierroksella valittujen ulottuvuuksien arvioiminen oli arvioijien (n=19) mielestä helpompaa, vaikkakin 16 % vastaajista koki ilmaisun laajuuden ”melko vaikeaksi” arvioida (kuvio 5). Esimerkiksi yhdelle arvioijalle (lainaus 30) oli epäselvää, tulisiko ilmaisun laajuus suhteuttaa tehtävänantoon vai taitotasoon, sillä analyyttisiä kriteerejä ei linkitetty taitotasokuvauksiin. Kaikkien arviointikierrokselle osallistuneiden mielestä oli vähintään ”melko helppoa” (kuvio 5) arvioida sujuvuutta (ka. 4,4). Pääosin helppoa oli myös sijoittaa suoritukset taitotasoille (ka. 4) sekä arvioida sanaston ja rakenteiden tarkkuutta (ka. 4), vaikka puhujan yksittäisten puhenäytteiden arviointi olikin vieraampaa kuin puhujan kokonaisvaltaisempi arviointi (lainaus 31). Lukiolaisten ruotsin näytteisiin verrattuna arvioitavat YKI-puhujien suomen näytteet olivat pidempiä, minkä uskomme helpottavan eri ulottuvuuksien arvioimista puhesuorituksista.

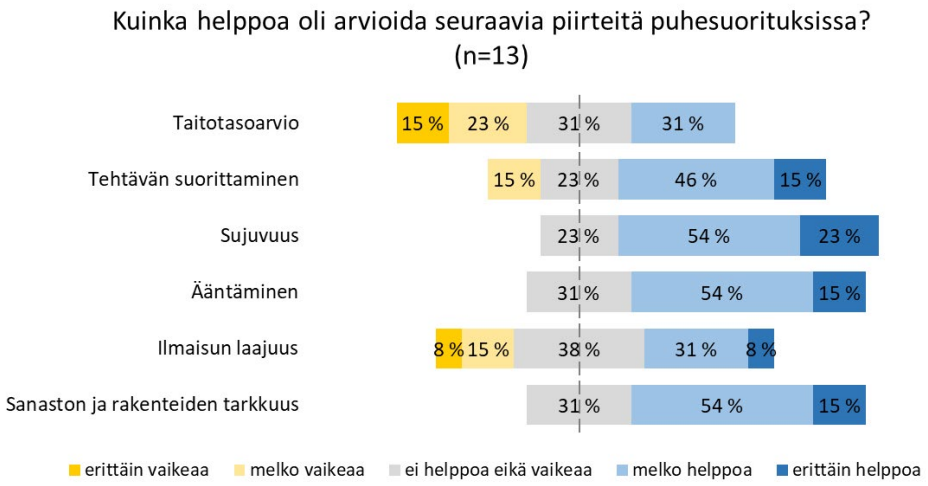


KUVIO 5. Arvioijien kokemus tehtävänannon ja puheen ulottuvuuksien arvioinnista YKI-puhujien suomi -kierroksella.

(30) Arviointikriteereissä ilmaisun laajuus -kohdassa heräsi kysymys siitä, pitääkö laajuus käsittää taitotason mukaisesti. Toisin sanoen laajuus A-tasolla on eri asia kuin B-tasolla.

(31) Taitotason ja ilmaisun laajuuden arvioiminen oli hankalaa siksi, että on tottunut saamaan yhdeltä suorittajalta kokonaisen puhumisen osakokeen (n. 20-25 min) verran puhetta. – – monologin puhuminen on hankalaa monille ylimmänkin tason osallistujillekin. – –

Lukiolaisten suomi -arviointikierroksella selkeästi haastavimmiksi osoittautuivat taitotason määrittäminen (ka. 2,7) ja ilmaisun laajuuden (ka. 3,1) arvioiminen, kun taas sujuvuuden (ka. 4), ääntämisen (ka. 3,8) ja tarkkuuden (ka. 3,7) arviointi koettiin neutraaliksi tai helpoksi (kuvio 6). Vastaajista (n=13) 15 % koki ”melko vaikeaksi” arvioida, oliko tehtävänanto täyttynyt riittävästi (kuvio 6), ja tehtävänannon täyttymiseen toivottiin tarkempaa ohjeistusta (lainaus 32). Kuten Lukiolaisten ruotsi -kierroksella, myös Lukiolaisten suomi -kierroksen arvioijat kommentoivat, että taitotason arvioiminen oli hankalaa näytteiden lyhyyden takia (lainaukset 33 ja 34).



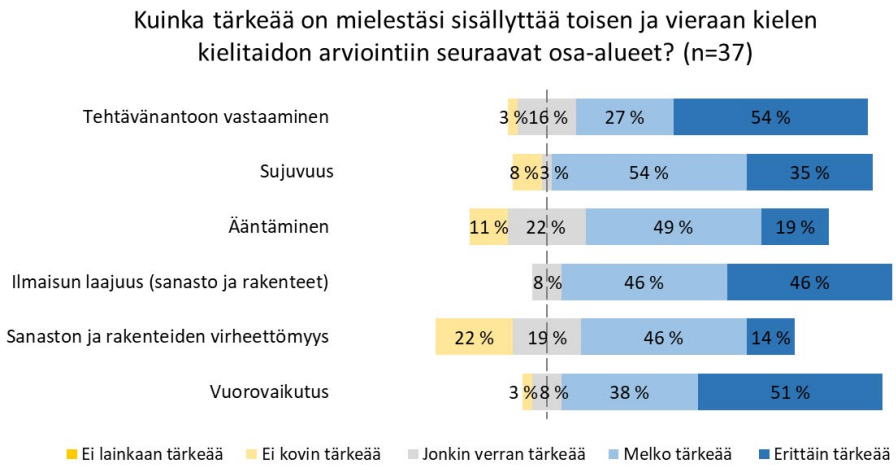
KUVIO 6. Arvioijien kokemus tehtävänannon ja puheen ulottuvuuksien arvioinnista Lukiolaisten suomi -kierroksella.

(32) – – Haasteena tässä aineistossa oli erittäin lyhyiden lauseiden arviointi (ääneenluku), joidenkin tehtävien konkreettisuus (> miten päästä ylimmän tason suoritukseen), erittäin suppeat näytteet (muutama sekunti, kun ohjeissa pyydettiin 30 s). Pohdin myös sitä, miten arvioida suorituksia reagointitehtäviin ts. milloin tehtävänanto täyttyy. Yleensä arvioijilla on tästä jokin ohje (ts. mitä vähintään vaaditaan). – –

(33) Puhetehtävät eivät elisitoineet riitävästi puhetta, jotta taitotason arvioiminen olisi itseltäni onnistunut.

(34) – – Jotkut näytteistä olivat aivan liian lyhyitä kunnollisen arvioinnin tekemiseen. Yleensä tarvitsen noin 1-2 minuuttia puhetta taitotason arviointiin. Yksittäinen lause ei anna kunnon kuvaa tasosta ja oma sisäinen arviointikoneisto meni vähän sekaisin... – –

Lisäksi kolmanteen tutkimuskysymykseen liittyen kysyimme, kuinka tärkeinä arvioijat pitävät tiettyjen ulottuvuuksien sisällyttämistä kielitaidon arviointiin. Kuten kuviosta 7 käy ilmi, arvioijat (n=37) pitävät tärkeänä etenkin vuorovaikutuksen (ka. 4,4), tehtävänannon täyttymisen (ka. 4,3), ilmaisun laajuuden (ka. 4,4) sekä sujuvuuden (ka. 4,2) arviointia. Annetuista ulottuvuuksista vähiten tärkeiksi koettiin ääntäminen (ka. 3,8) ja sanaston ja rakenteiden virheettömyys (ka. 3,5).



KUVIO 7. Arvioijien kokemus tehtävänannon ja puheen ulottuvuuksien tärkeydestä.

Valmiiksi annettujen kielitaidon ulottuvuuksien (kuvio 7) lisäksi arvioijilla oli mahdollisuus nimetä avoimeen kenttään, mitä muita ulottuvuuksia tulisi sisällyttää suullisen kielitaidon arviointiin. Tilanteeseen sopivuus eli rekisteri mainittiin kaikilla arviointikierroksilla (lainaukset 35 ja 36) ja Lukiolaisten suomi -kierroksen arvioinneissa tuotiin esiin lisäksi idiomaattisuus (lainaus 37). Avointen vastausten perusteella voidaan todeta, että tutkimushankkeen ennakkoon määrittelemät ulottuvuudet kattavat hyvin ne suullisen kielitaidon osa-alueet, joiden arviointia arvioijatkin pitävät keskeisinä.

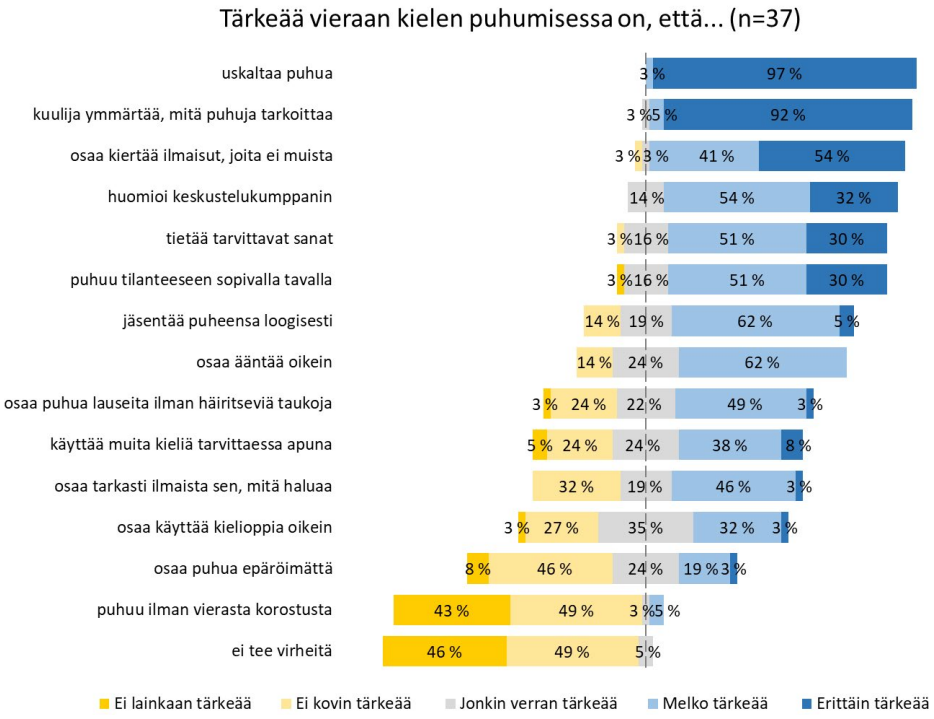
(35) riippuu tilanteesta, näiden lisäksi esiintymistä, nonverbaalista viestintää, kohteliaisuutta, tilanteeseen sopivuutta jne harjoitellaan ja on hyvä arvioidakin.

(36) – – Asiointitilanteissa puhelimessa (ystävyyskoulu) olisin halunnut ottaa huomioon myös rekisterin, mutta se ei tuntunut kuuluvan mihinkään kriteeriin. – –

(37) – – puheen idiomaattisuus ja käsitteellisyys arvioinnin kohteena korostuvat ylemmillä taitotasoilla (B2 - C2). – –

Lähestyimme puhesuorituksista arvioitavia ulottuvuuksia (tutkimuskysymys 3) myös vapaammin muotoilluilla väittämillä, joiden avulla kartoitimme, mitä arvioijat pitävät tärkeänä puhumisessa. Kuvio 8 osoittaa, ettei kaikkia arvioijien tärkeänä pitämiä asioita – kuten vuorovaikutukseen liittyvää keskustelukumppanin huomioimista tai ymmärretyksi tulemista – ole välttämättä mahdollista mitata automaattisesti. Kuviosta näemme myös, että vieraat aksentit, epäröinti ja virheet kuuluvat arvioijien

mielestä puheeseen. Arvioijien mielipiteitä jakaneet väittämät liittyvät muun muassa tarkkuuteen, taukoihin ja monikieliseen kompetenssiin (kuvio 8).



KUVIO 8. Väittämät puhumisen käsitteestä.

5 Pohdinta

Tutkimuksemme tavoitteena oli kartoittaa ja analysoida ihmisarvioijien näkemyksiä digitaalisesti toteutetusta arviointiprosessista ja tuottaa tietoa oikeudenmukaisen arvioinnin keskeisistä piirteistä (Kunnan 2018). Lähtökohtaisesti kaikilla arvioitavilla oli ollut mahdollisuus hankkia puhetehtävien edellyttämä suullinen kielitaito. Muiden osa-alueiden toteutumisesta kartoitus antoi runsaasti uutta ja hyödyllistä tietoa.

Arviointiprosessin toimeenpano oli arvioijien mielestä varsin onnistunut ja he suhtautuivat automaattiseen arviointiin enimmäkseen positiivisesti tai neutraalisti. Havaitsimme eroja siinä, kuinka hyvin arvioijat tuntevat automaattisten arviointityökalujen toimintaa (puhutun englannin osalta esim. Educational Testing Service 2014; Pearson 2017). Heidän näkemyksensä vahvistivat aiemmassakin tutkimuksessa esille tulleita tuloksia siitä, että automaattinen arviointi soveltuu etenkin

ihmisarvioinnin tueksi (hybridiarvioinnista Zechner & Evanini 2020) ja vaikutukseltaan vähäisiin (*low-stakes*) kielikokeisiin (Zhang ym. 2020). Kuten muiden osallisten (*stakeholders*) automaattiseen arviointiin liittyvästä käsitystutkimuksesta tiedetään, myös arvioijat pitivät reliaabeliutta ja käytettävyyttä automaattisen arvioinnin vahvuutena. Tutkimuksen osallistujat olivat huolissaan erilaisten puhujien oikeudenmukaisesta arvioinnista, mikä on tullut esiin myös opiskelijoiden käsitystutkimuksessa (von Zansen ym. arvioitavana; puheentunnistimen mukauttamisesta kielenoppijan puheeseen Ylinen & Kurimo 2017). Käytetyn validiusmallin (Kunnan 2018) näkökulmasta kysymys on arvioinnin mukautuvuudesta ja yhdenvertaisuudesta.

Arvioinnin tarkoituksenmukaisuuden ydin on suullisen kielitaidon käsitteen määrittely, mikä puolestaan ohjaa tehtävien ja arviointikriteerien laadintaa. Puhesuorituksista arvioitavien ulottuvuuksien määrittely puolestaan vaikuttaa automaattisen arvioinnin kehittämiseen (Zechner & Evanini 2020). Tässä tutkimuksessa osa tehtävistä ei mahdollistanut riittävän laajan puhesuorituksen antamista. YKI-puhujien suomi -arviointikierroksella, jossa puhenäytteet olivat pidempiä, kaikki ulottuvuudet koettiin helpommiksi arvioida kuin muilla kierroksilla. Lyhyet, kontrolloidut puhenäytteet mahdollistavat kuitenkin tarkempien arviointialgoritmien kehittämisen kuin pitkät, vapaata puhetta sisältävät näytteet, joiden sisältöä on vaikeampi ennustaa. Syväoppimista hyödyntävillä tekoälymenetelmillä (Al-Ghezi ym. arvioitavana) on saatu lupaavia tuloksia myös pidempien puhenäytteiden arvioinnissa. Tällaisen "mustan laatikon" toiminta on kuitenkin läpinäkymätöntä, eli arviointia ei pystytä perustelemaan tietyillä puhenäytteestä löytyvillä piirteillä.

Sujuvuutta ja ääntämistä pidettiin helppoina arvioida. Nämä ulottuvuudet ovat myös muita puheen ulottuvuuksia helpompia arvioida automaattisesti, sillä esim. sujuvuuden automaattiseen arviointiin käytetyt parametrit ovat suhteellisen kieliriippumattomia ja niiden yhteydestä ihmisten tekemiin arvioihin on vahvaa tutkimusnäyttöä (ks. esim. Préfontaine ym. 2016; Kang & Johnson 2018; Kallio ym. 2021; Kautonen & Kuronen 2021). Keskustelukumppanin huomioiminen ylsi kolmanneksi tärkeimmäksi asiaksi (kuvio 8). Vuorovaikutusta ei kuitenkaan huomioitu puhesuorituksista arvioitavissa ulottuvuuksissa, sillä sen automaattinen arviointi on haastavaa ja rajoittaa näin tehtävien autenttisuutta (koetehtävien ja muiden kielenkäyttötilanteiden vastaavuudesta Kunnan 2018). Ehkä tulevaisuudessa tekoäly voidaan opettaa arvioimaan myös vuorovaikutusta ja strategista osaamista.

Kyselyvastausten perusteella näyttää siltä, että tutkimushanke on onnistunut valitsemaan keskeisiä ulottuvuuksia automaattiseen arviointiin. Puhumisen käsite on kuitenkin laaja, eikä kaikkia arvioijien tärkeinä pitämiä ulottuvuuksia voida sisällyttää puheen automaattisiin arviointimalleihin (puheen automaattisen arvioinnin rajallisuudesta Zechner & Evanini 2020). Lopulta arvioinnin tarkoitus eli se, mitä päätöksiä suorituksen perusteella tehdään, määrittää vahvasti sitä, mitä ulottuvuuksia puhesuorituksista kannattaa arvioida. Hankkeessa kehitettävillä sovelluksilla onkin

useampia mahdollisia kohderyhmiä sekä formatiivisen että summatiivisen arvioinnin osalta.

Tällä tutkimuksella on tärkeä merkitys oikeudenmukaisen automaattisen arvioinnin kehittämisessä. Saadun palautteen avulla olemme voineet kehittää arviointimenetelmiä toimivammiksi. Onnistumisten ohella tunnistamme työssämme myös rajoituksia. Tutkimuksen kannalta mittareiden ja menetelmien muuttaminen kesken prosessin on haastavaa, koska eri välineistöllä kerätyt havainnot eivät ole täysin vertailukelpoisia keskenään. Toinen rajoitus koskee osallistujien valintaa: harkinnanvarainen näytteemme ei edusta kattavasti kielitaidon arvioijien näkemyksiä. Kolmas rajoitus koskee arviointikriteerien laatimista aiemman opetussuunnitelman (Opetushallitus 2003) yksityiskohtaisempien taitotasokuvausten pohjalta. Niitä ei kuitenkaan ole linkitetty nykyisen opetussuunnitelman (Opetushallitus 2019) taitotasojen kuvausasteikolle. Jatkotutkimusta tarvitaan muiden osallisten kuten opettajien näkemyksistä, sillä opettajat ovat keskeisessä asemassa arviointityökalujen käyttöönotossa.

Suomen Akatemia rahoittaa DigiTala-tutkimushanketta (2019–2023). Kiitämme yhteistyöstä projektin tutkijoita Helsingin yliopistosta (rahoituspäätös 322619), Aaltoyliopistosta (rahoituspäätös 322625) sekä Jyväskylän yliopistosta (rahoituspäätös 322965).

Kirjallisuus

Al-Ghezi, R., K. Voskoboinik, Y. Getman, A. von Zansen, H. Kallio, C. Akiki, M. Kuronen, A. Huhta & R. Hilden (arvioitavana). Automatic speaking assessment of spontaneous L2 Finnish and Swedish.

Bachman, L. F. & A. S. Palmer 1996. *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford: Oxford University Press.

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

Educational Testing Service 2014. *TOEFL iBT speaking section scoring guide.* https://www.ets.org/toefl/ibt/scores/understand/ [luettu 11.2.2022].

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.

Halonen, M., A. Huhta, S. Ahola, T. Hirvelä, R. Neittaanmäki, S. Ohranen & R. Ullakonoja 2020. Ensikielen tunnistamisen merkityksestä suullisen kielitaidon arvioinnissa Yleisissä kielitutkinnoissa. Teoksessa S. Grasz, T. Keisanen, F. Oloff, M. Rauniomaa, I. Rautiainen & M. Siromaa (toim.) *Menetelmällisiä käänteitä soveltavassa kielentutkimuksessa - Methodological turns in applied language studies*. AFinLA:n vuosikirja, 2020. Jyväskylä: Suomen soveltavan kielitieteen yhdistys ry, 56–70. https://doi.org/10.30661/afinlavk.89453.

Han, C. 2022. Interpreting testing and assessment: A state-of-the-art review. *Language Testing, 39* (1), 30–55. https://doi.org/10.1177/02655322211036100.

Han, C., S.-J. Chen, R.-B. Fu & Q. Fan 2020. Modeling the relationship between utterance fluency and raters' perceived fluency of consecutive interpreting. *Interpreting*, *22* (2), 211–237. https://doi.org/10.1075/intp.00040.han.

Harsch, C. & J. Hartig 2015. What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly, 12* (4), 333–362. https://doi.org/10.1080/15434303.2015.1092545.

Hidri, S. 2020. New challenges in language assessment. Teoksessa S. Hidri (toim.) *Changing Language Assessment*. London: Palgrave Macmillan Cham. https://doi.org/10.1007/978-3-030-42269-1_1.

Inha, K., A. Halvari, Y. Nummela, P. Mattila & S. Sigvart 2021. *Kieliprofiili*. https://www.oph.fi/fi/kieliprofiili [luettu 15.2.2022].

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.) *AFinLA-e: Soveltavan kielitieteen tutkimuksia, 10*, 193–213. https://doi.org/10.30660/afinla.73137.

Kallio, H., A. Suni & J. Šimko 2021. 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%2F00238309211040175.

Kallio, H., R. Suviranta, M. Kuronen & A. von Zansen 2022. Creaky voice and utterance fluency measures in predicting perceived fluency and oral proficiency of spontaneous L2 Finnish. Teoksessa S. Frota, M. Cruz, & M. Vigário (toim.), *Speech Prosody 2022: Proceedings of the 11th International Conference on Speech Prosody*, 777-781. https://doi.org/10.21437/SpeechProsody.2022-158.

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

Kautonen, M. 2019. *Finskspråkiga inlärares uttal av finlandssvenska i fritt tal på olika färdighetsnivåer*. JYU Dissertations 90. Jyväskylä: Jyväskylän yliopisto. http://urn.fi/URN:ISBN:978-951-39-7778-8.

Kautonen, M. & M. Kuronen 2021. Kvantitativa perspektiv på L2-tal på olika färdighetsnivåer. *Folkmålsstudier, 59,* 11–39. 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.

Khabbazbashi, N. & E. D. Galaczi 2020. A comparison of holistic, analytic, and part marking models in speaking assessment. *Language testing*, *37* (3), 333–360. https://doi.org/10.1177%2F0265532219898635.

Knoch, U. & C. A. Chapelle 2018. Validation of rating processes within an argument-based framework. *Language testing, 35* (4), 477–499. https://doi.org/10.1177%2F0265532217710049.

Kunnan, A. J. 2018. *Evaluating language assessments*. New York: Routledge.

Loukina, A. & S. Y. Yoon 2019. Scoring and filtering models for automated speech scoring. Teoksessa K. Zechner & K. Evanini (toim.) *Automated Speaking Assessment.* New York: Routledge, 75–98.

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

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

Ouyang, L., Q. Lv & J. Liang 2021. Coh-Metrix model-based automatic assessment of interpreting quality. Teoksessa J. Chen & C. Han (toim.) *Testing and assessment of interpreting: Recent developments in China.* Singapore: Springer Singapore, 179–200. https://doi.org/10.1007/978- 981-15-8554-8_9.

Pearson 2017. *PTE Academic score guide for test takers.* https://pearsonpte.com/wp-content/uploads/2017/08/Score-Guide.pdf [luettu 11.2.2022].

Préfontaine, Y., J. Kormos & D. E. Johnson 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%2F0265532215579530.

Toivanen, J. 2016. Ruotsin ääntämisen automaattinen arviointi – onko se mahdollista? Teoksessa *Ammatilliset ruotsin opettajat opetuksen kehittäjinä – digitalisaatio ja yhteistyö fokuksessa*. OKKA-säätiö 2016, 66–71.

Vaarala, H., S. Riuttanen, E. Kyckling & S. Karppinen 2021. *Kielivaranto. Nyt!: Monikielisyys vahvuudeksi -selvityksen (2017) seuranta*. Soveltavan kielentutkimuksen keskus, Jyväskylän yliopisto. https://www.jyu.fi/hytk/fi/laitokset/solki/tutkimus/julkaisut/pdf-julkaisut/kielivaranto-nyt-_julkaisu_sivuittain-1.pdf.

von Zansen, A., M. Sneck & R. Hildén (arvioitavana). Lukiolaisten käsitykset ja heidän antamansa palaute suullisen kielitaidon automaattisesta arvioinnista. Teoksessa R. Kantelinen, M. Kautonen & Z. Elgundi (toim.) *Kielipedagogisia näkökulmia elinikäiseen kielenoppimiseen* (tulossa). Suomen ainedidaktinen tutkimusseura.

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

von Zansen, A. 2022b. DigiTala's post-rating questionnaire for human raters (Swedish, upper secondary schools, Dec2020). Zenodo. https://doi.org/10.5281/zenodo.6469605.

von Zansen, A. 2022c. DigiTala's post-rating questionnaire for human raters (Finnish, YKI, May2021). Zenodo. https://doi.org/10.5281/zenodo.6476995.

von Zansen, A. 2022d. DigiTala's post-rating questionnaire for human raters (Finnish, upper secondary schools, Jun2021). Zenodo. https://doi.org/10.5281/zenodo.6477015.

Vuori, J. n.d. Laadullinen sisällönanalyysi. Teoksessa J. Vuori (toim.) *Laadullisen tutkimuksen verkkokäsikirja*. Tampere: Yhteiskuntatieteellinen tietoarkisto https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus [luettu 17.03.2022].

Yhteiskuntatieteellinen tietoarkisto 2022. Kyselylomakkeen laatiminen. Teoksessa *Kvantitatiivisen tutkimuksen verkkokäsikirja*. Tampere: Yhteiskuntatieteellinen tietoarkisto. https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvanti/kyselylomake/laatiminen/ [luettu 12.4.2022].

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

Zechner, K. & K. Evanini (toim.) 2020. *Automated speaking assessment: Using language technologies to score spontaneous speech.* New York: Routledge.

Zhang, M., B. Bridgeman & L. Davis 2020. Validity considerations for using automated scoring in speaking assessment. Teoksessa K. Zechner & K. Evanini (toim.) *Automated speaking assessment: Using language technologies to score spontaneous speech.* New York: Routledge, 21–31.