TIETO- JA SÄHKÖTEKNIIKAN TIEDEKUNTA

Valtteri Kuosmanen

KOULUTUSDATAN VAIKUTUS VIRTUAALIASSISTENTIN SUORIUTUMISEEN

Kandidaatintyö
Tietotekniikan tutkinto-ohjelma
Joulukuu 2021

Kuosmanen V. (2021) Koulutusdatan vaikutus virtuaaliassistentin suoriutumiseen. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 40 s.

TIIVISTELMÄ

Eri yritysten henkilökohtaiset puheentunnistukseen perustuvat avustajat eli virtuaaliassistentit ja niitä hyödyntävät laitteet ovat kasvattaneet suosiotaan huomattavasti viime vuosien aikana. Applen Sirin, Amazonin Alexan, sekä Google Assistantin kaltaiset avustajat voivat esimerkiksi mahdollistaa hälytysten asettamisen ilman tarpeetonta valikkojen läpi selaamista. Lisäksi suurien yritysten assistentit kykenevät hyödyntämään valtavaa verkossa olevan tiedon määrää, parantaen niiden kykyjä vastata arkipäiväisiin kysymyksiin. Puheentunnistusratkaisuja on myös otettu käyttöön erilaisissa työtehtävissä, joissa työntekijä ei pysty käyttämään käsiään käyttöliittymien navigoimiseen.

Tässä työssä esitellään syväoppimiskoulutuksella toteutettu virtuaalinen assistentti, jonka ominaisuudet perustuvat puheentunnistukseen, luonnollisen kielen käsittelyyn, sekä verkon haravoimiseen. Tärkeimpänä tavoitteena oli toteuttaa puheentunnistukseen perustuva avustaja, joka pystyisi tarjoamaan suomeksi toimintoja, jotka vastaisivat kaupallisten englanninkielisten assistenttien asettamaa tasoa. Lisäksi haluttiin testata ylimääräisen datan lisäämistä puheentunnistukselle tarkoitettuun pakettiin ja arvioida saadun mallin suorituskykyä verrattuna yhdellä tietoaineistolla koulutettuun malliin. Nämä tavoitteet onnistuttiin saavuttamaan. Huolimatta suomenkielisen koulutusdatan vähäisestä määrästä suhteutettuna esimerkiksi englannin-, tai kiinankieliseen koulutusdataan, toteutettu puheentunnistusmalli suoriutui komentojen tulkitsemisessa yllättävän hyvin jo heti koulutuksen jälkeen, jolloin se sai tulkittavista komennoista keskimäärin 46% oikein. Mallin virheellisiä tulkintoja paikattiin onnistuneesti toteuttamalla komennot taulukoina, joissa oikein tulkittujen sanojen lisäksi oli mallin yleisimpiä virheellisiä arvauksia, sekä mahdollisia synonyymejä. Tätä menetelmää käyttämällä malli onnistui saamaan komentolauseista keskimäärin 77% oikein. Lisäksi yhdyssanavirheitä paikattiin hyödyntämällä projektille kehitettyä automaattista korjausta. Automaattinen ratkaisu korjasi 19% kaikista ylimääräisistä välilyönneistä johtuvista virheistä. Sekadatalla koulutettu malli olisi myöskin soveltunut käytettäväksi, mutta malli ei suoriutunut yhtä hyvin kuin Common Voice:n datalla toteutettu puheentunnistusmalli. Näin osoitettiin tarve Lahjoita puhetta -kampanjan ja Common Voice:n kaltaisille puheentunnistukselle suunitelluille puheaineistoille. Virtuaalisessa assistentissa käytetyt tiedostot ovat julkisesti saatavilla.

Avainsanat: syväoppiminen, koneoppiminen, NLP, tiedonharavointi

Kuosmanen V. (2021) The Effect of Training Data on the Performance of a Virtual Assistant. University of Oulu, Degree Programme in Computer Science and Engineering, 40 p.

ABSTRACT

Various intelligent personal assistants or virtual assistants and devices supporting them have risen in popularity in the past years. Assistants like Apple's Siri, Amazon's Alexa and Google Assistant can for example be used to set up alarms without needlessly going through multiple sets of menus. In addition the assistants developed by Big Tech are able to utilize the enormous amounts of data available online, improving their ability to answer everyday questions from the users. Speech recognition solutions have also been utilized in some commercial and industrial applications. This is useful in situations where the person working is unable to use their hands to navigate interfaces.

In this thesis, a Finnish virtual assistant is presented. This virtual assistant has been implemented using deep learning and its functions are based on speech recognition, natural language processing and web scraping. The primary goal was to implement a speech recognition-based assistant that could execute tasks in Finnish, in a way that is comparable to the bar set by commercial virtual assistants in English. The secondary goal was to test adding additional data to the speech recognition dataset and evaluate its performance compared to the language model which was trained with only a single dataset. In the end both of these goals were met. Despite the small amount of Finnish training data compared to the amount of data for languages such as English or Chinese, the implemented speech recognition model's ability to interpret commands right after training surpassed expectations, by managing to correctly interpret 46% of the given commands. Mistakes made by the model were rectified successfully by implementing the commands as arrays, where in addition to the intended command words, the most common misinterpretations and potential synonyms were also included. Using this method the model was able to interpret 77% of the commands correctly. Compound word errors were also fixed using an automatic solution developed for this project. This automatic solution was able to fix 19% of the errors caused by incorrect use of spaces. The model that was trained using mixed data could have also been used for the implementation, but it did not perform as well as the speech recognition model trained only with Common Voice data. This demonstrated the need for Finnish datasets similar to those of the Lahjoita puhetta -campaign and Common Voice, which are specifically intended for speech recognition. The files used in the virtual assistant are publicly available.

Keywords: deep learning, machine learning, NLP, web scraping

SISÄLLYSLUETTELO

ALKULAUSE

Työ toteutettiin osana Oulun yliopiston Sulautettujen ohjelmistojen projektia. Kiitos kaikille, jotka ottivat osaa virtuaalisen assistentin testaukseen.

Oulussa 9. joulukuuta 2021

Valtteri Kuosmanen

LYHENTEIDEN JA MERKKIEN SELITYKSET

AI	artificial intelligence, tekoäly
ANN	artificial neural network, keinotekoinen neuroverkko
ASR	automated speech recognition, automaattinen puheentunnistus
CER	character error rate, merkkivirhesuhde
CPP	cocktail party problem, cocktailkutsuongelma
GUI	graphical user interface, graafinen käyttöliittymä
HCI	human computer interaction, ihmisen ja tietokoneen vuorovaikutus
IPA	intelligent personal assistant, älykäs henkilökohtainen assistentti
IVA	intelligent virtual assistant, älykäs virtuaaliassistentti
MLP	multilayer perceptron, monikerroksinen perseptroniverkko
MPL	Mozilla public license, Mozillan vapaa ohjelmistolisenssi
NLP	natural language processing, luonnollisen kielen käsittely
NN	neural network, neuroverkko
RNN	recurrent neural network, takaisinkytketty neuroverkko
TTS	text to speech, teksti puheeksi
VUI	voice user interface, äänikäyttöliittymä
WER	word error rate, sanavirhesuhde

arg	funktion argumentti
b	vakiotermi
f	funktio
I	syöte
max	maksimiarvo
O	ulostulo
P	todennäköisyys
U, V, W	vektorin paino
W	sanajono
w	painovektori
X	akustinen observaatio
x	syötevektori

1. JOHDANTO

Vaikka perinteiset käyttöliittymät ovatkin luotettavin ratkaisu useimmille laitteille ja käyttötilanteille, puheentunnistuksen lisääminen vaihtoehtoiseksi syötetavaksi nostaa mahdollisten käyttötilanteiden ja käyttäjien määrää. Esimerkiksi nuori lapsi, joka ei vielä pysty käyttämään tietokonetta, pystyy kuitenkin soittamaan haluamansa kappaleen kommunikoimalla virtuaaliassistentin kanssa [1]. Näistä syistä virtuaaliassistenttien implementoinnista eri älylaitteille olisi potentiaalisesti paljon hyötyä.

Virtuaaliassistentteihin liittyy kuitenkin turvallisuusuhkia ja tämä joissakin tapauksissa saa ihmiset välttelemään niiden käyttämistä [2]. Tämänhetkisistä virtuaaliassistenteista suosituimmat ovat suurten yritysten kehittämiä [3] ja ne käyttävät useimmissa tilanteissa pilvilaskentaa puheentunnistukseen [4, 5, 6]. Siis sen lisäksi, että puheentunnistuksen toteuttaminen itse laitteella mahdollistaisi assistentin toiminnan osittain myös ilman verkkoyhteyttä, se voisi vedota myös ihmisiin, jotka kokevat turvallisuusuhat liian suuriksi.

Suomessa virtuaalisten assistenttien käyttäjäkuntaa rajoittaa myös se, että Applen, Googlen, Amazonin ja Microsoftin assistenteista vain Applen Siri tukee suomen kieltä ja tälläkään ei ole kaikkia ominaisuuksia englanninkielisestä versiosta [7]. Vaikka voitaisiinkin olettaa, että suomen kieltä alettaisiin tukemaan tulevaisuudessa, olisi suomenkielisen virtuaalisen assistentin toteuttamisesta kuitenkin hyötyä tällä hetkellä.

2. PUHEENTUNNISTUS JA VIRTUAALIASSISTENTIT

Älykäs henkilökohtainen assistentti (intelligent personal assistant, IPA), tai virtuaaliassistentti (intelligent virtual assistant, IVA) on ratkaisu, jossa koneen ja ihmisen välinen vuorovaikutus (human computer interaction, HCI) tapahtuu äänikomennoilla [8]. Tämäntyyppiset puheeseen pohjautuvat käyttöliittymät (voice user interface, VUI) voivat tehdä tietotekniikan käyttämisestä helpompaa esimerkiksi sellaisissa tapauksissa, jossa käyttäjällä ei ole kokemusta perinteisemmistä käyttöliittymistä [9]. Myös älykodit hyötyvät puheohjauksesta [10], tehden esimerkiksi valojen hallitsemisesta käytännöllisempää. Näistä syistä johtuen virtuaaliassistentit ja puheohjaus laajemmin ovat tärkeitä ihmisten ja tietokoneiden välisen vuorovaikutuksen osa-alueita.

2.1. Tekoäly

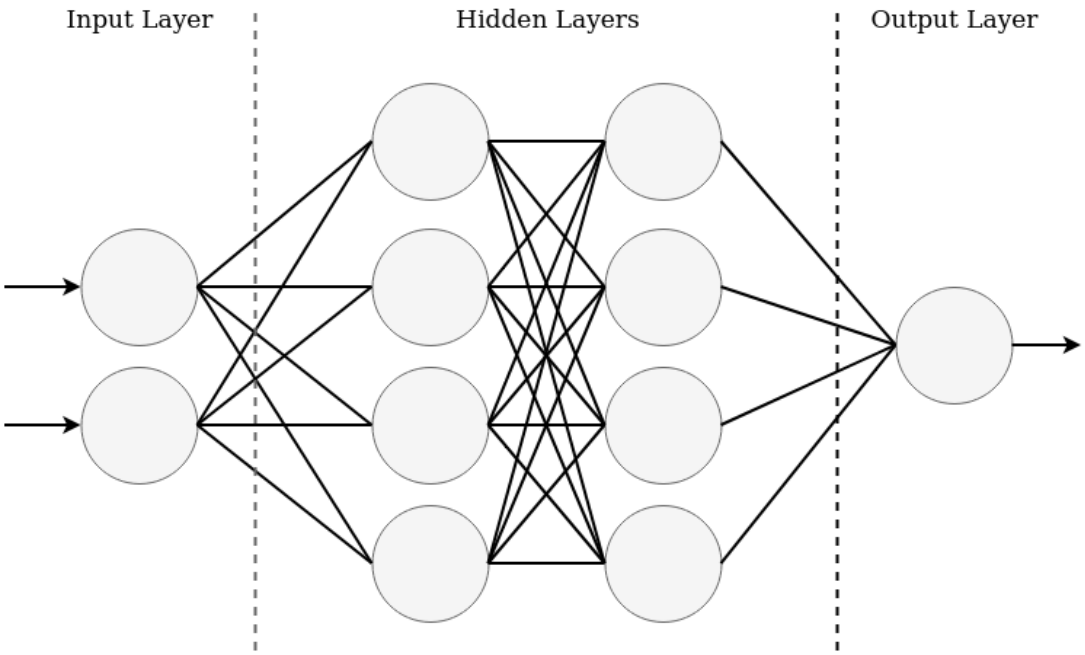
Tekoälyllä viitataan tieteen ja erityisesti tietotekniikan alueeseen, jossa pyritään kehittämään älykkäitä koneita [11]. Älykkäiden koneiden juuret ovat historiallisissa legendoissa, kuten antiikin kreikkalaisten Talos-myytissä [12] ja juutalaisen taruston Golemissa [13], sekä nuoremmissa tieteisfiktion esimerkeissä, kuten Čapek-veljesten fiktiivisissä automaattisissa ihmisissä [14] ja Isaac Asimovin robotiikan laeissa [15]. Tekoäly tieteellisessä kontekstissa alkoi saamaan muotoansa vuonna 1950, kun nykyään tietotekniikan isänäkin tunnettu Alan Turing käsitteli julkaisemassaan artikkelissa koneiden älykkyyttä ja sen testaamista. Tästä syntyi tähänkin päivään asti käytetty Turingin testi, jossa koneen älykkyyttä arvioidaan sen mukaan, kuinka uskottavasti se pystyy esittämään ihmistä keskustelussa [16, 15]. Tekoälyn varsinaisena aloituspisteenä pidetään kuitenkin yleisesti vuoden 1956 Dartmouth Summer Research Project on Artificial Intelligence -työpajaa, jossa eri alojen tutkijat yhdessä pyrkivät luomaan yhteistä tieteenalaa, joka keskittyisi jäljittelemään ihmisen älykkyyttä koneellisesti [17, 18]. Tämä työpaja antoi myös ensimmäisen kerran nimen tekoälylle (artificial intelligence, AI).

2.2. Neuroverkot

Tietotekniikassa neuroverkot (neural network, NN), eli keinotekoiset neuroverkot (artificial neural network, ANN) ovat solmuista (node), eli keinotekoisista neuroneista koostuvia laskennallisia järjestelmiä, jotka toimintatavaltaan perustuvat biologisten aivojen neuroverkkoihin [19, 20]. Vaikka keinotekoisiin neuroverkkoihin keskittyvä tutkimustyö alkoikin jo 1940-luvulla Warren McCullochin ja Walter Pittsin toimesta [21], ei alalla tapahtunut paljoa kehitystä ennen vuotta 1958, jolloin Frank Rosenblatt esitteli havaitsijan, eli perseptronin [22, 23]. Perseptronia voidaan pitää minimaalisena neuroverkkona, joka muodostuu yhdestä neuronista ja syötekerroksesta [24]. Perseptronin aktivointifunktio voidaan esittää kaavan 1

$$f(x) = \begin{cases} 1, & \text{jos } w \cdot x + b > 0 \\ 0, & \text{jos } w \cdot x + b \leq 0 \end{cases} \tag{1}$$

avulla, jossa $w \cdot x$ on paino- ja syötevektorien pistetulo ja b on vakiotermi (bias) [25, 24]. Kun perseptroneista muodostetaan verkko, jossa on useampia tasoja, puhutaan monikerroksisista perseptroniverkoista (multilayer perceptron, MLP), kuten kuvassa 1. Näitä monikerroksisia perseptroniverkkoja kutsutaan monesti "tavallisiksi" neuroverkoiksi [26]. Neuroverkot soveltuvat hyvin muun muassa kuvioiden tunnistamista ja luokittelua vaativiin tehtäviin ja niiden käyttäminen näihin tarkoituksiin on yleistynyt viime vuosien aikana [27].

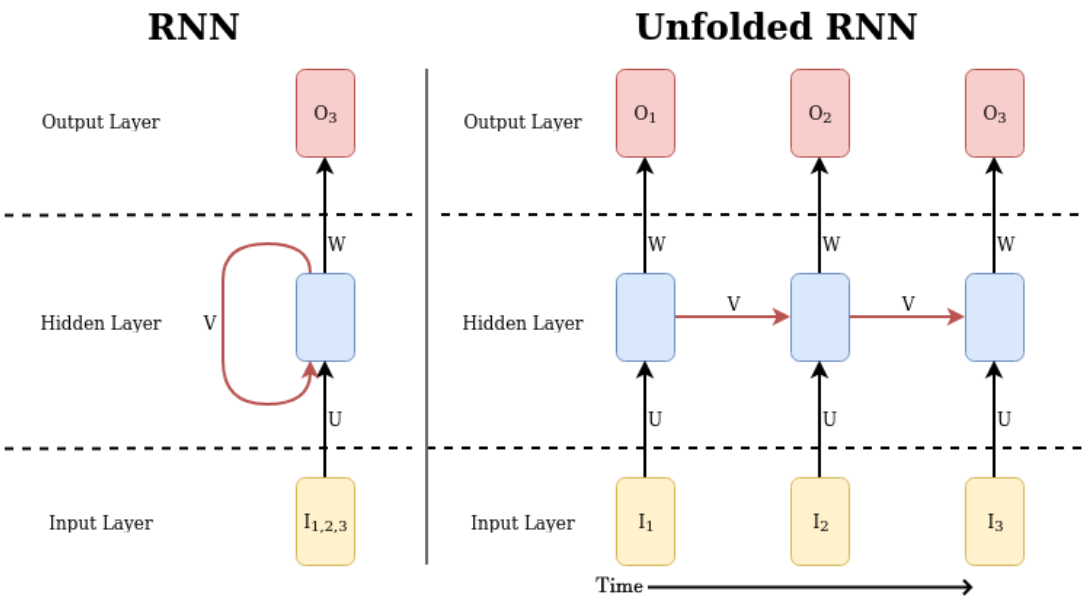


Kuva 1. Esimerkki monikerroksisesta perseptroniverkosta kahdella piilokerroksella. Kuva on luotu tätä kandidaatintyötä varten.

2.2.1. Takaisinkytketyt neuroverkot

Tavallisten neuroverkkojen lisäksi on olemassa myös takaisinkytkettyjä neuroverkkoja (recurrent neural networks, RNN). Takaisinkytketyt neuroverkot eroavat tavallisista siinä, että neuroverkon solmujen ulostuloja käytetään seuraavien kerrosten solmujen syötteinä. Tämä mahdollistaa sen, että neuroverkko pystyy ottamaan aikaisemmat ulostulot huomioon seuraavan ulostulon tuottamisessa. [28]

Takaisinkytketyt neuroverkot on siis suunniteltu sekvenssien tunnistamista varten [27]. Tämä tekee niistä soveltuvia tekstin ja puheen tunnistamiseen ja niiden käyttäminen näihin tarkoituksiin onkin yleistä [29, 30, 31]. Kuvassa 2 havainnollistetaan takaisinkytketyn neuroverkon toimintaa.



Kuva 2. Vasemmalla RNN, jolle annetaan kolmiosainen syöte $I_{1,2,3}$ ja joka antaa ulostulon O_3. Oikealla sama RNN avattuna. W, V ja U vastaavat vektorien painoja. Takaisinkytkentää havainnollistetaan punaisilla viivoilla. Kuva on luotu tätä kandidaatintyötä varten.

2.3. Siirto-oppiminen

Siirto-oppimisella (transfer learning) tarkoitetaan sitä, kun uuden suoritettavan tehtävän oppimista pyritään tehostamaan hyödyntämällä toisen tehtävän suoritukseen liittyvää tietoa. Tällä tavoin on mahdollista tehdä koulutuksesta tehokkaampaa, kuin mitä se olisi ilman aikaisemman tiedon hyödyntämistä, tuoden näin koneoppimisen lähemmäksi tapaa, jolla ihmisetkin oppivat. [32]

Siirto-oppiminen sopii hyvin muun muassa kohteiden tunnistukseen liittyviin tehtäviin [33, 34]. Eli kun neuroverkko on koulutettu tunnistamaan asia X on mahdollista kouluttaa neuroverkko tunnistamaan myös asia Y ilman, että koulutusta joudutaan aloittamaan täysin alusta.

2.4. Automaattinen puheentunnistus

Automaattinen puheentunnistus (automatic speech recognition, ASR) viittaa teknologioihin, jotka mahdollistavat ihmisten ja tietokoneiden välisen kommunikoimisen puheen kautta [35]. Automaattisessa puheentunnistuksessa ihmisen tuottama puhe muutetaan tekstimuotoon ja tätä voidaan käyttää hyväksi esimerkiksi tekstitysten luomiseen. Virtuaaliassistenttien tapauksessa ASR on olennainen ominaisuus, koska käyttäjän antama komento tunnistetaan tekstimuotoa käsittelemällä.

Käytännössä ASR-ratkaisu siis ottaa syötteenä akustisen aaltomuodon ja antaa ulostulona sanajonon. Kun syötteenä on akustinen observaatio $X = X_1, X_2 \dots X_n$

on tavoitteena tuottaa sanajono $W = W_1, W_2 \dots W_n$, maksimoiden todennäköisyyden $P(W|X)$ yhtälön 2

$$W = arg\ max\ P(W|X) = arg\ max\ \frac{P(W)P(X|W)}{P(X)} \quad (2)$$

mukaisesti. [36]

Jotta aaltomuodosta on mahdollista poimia tietoa, on suoritettava piirteiden poiminta (feature extraction). Kun puhesignaalia tarkastellaan noin 5-100 millisekunnin pituisella aikavälillä, ei signaalin ominaisuuksissa ole yleensä havaittavissa merkittäviä muutoksia. Kun tällä aikavälillä kuitenkin havaitaan muutos, tiedetään kyseessä olevan muutos lausutuissa äänteissä. Hyödyntämällä tätä lyhyillä aikaväleillä amplitudispektristä saatua informaatiota on mahdollista poimia signaalista foneemeja, eli kielen pienimpiä merkityksiä erottavia äänteitä. [37]

2.5. Puheentunnistuksen haasteet

Vaikka viimeisen kahdenkymmenen vuoden aikana onkin tapahtunut suuria muutoksia automaattisen puheentunnistuksen suoriutumisessa [38, 39] liittyy puheentunnistukseen kuitenkin vielä haasteita. Mahdollisia ASR:n suoriutumista rajoittavia tekijöitä ovat muun muassa murteiden ja aksenttien huomiotta jättäminen koulutusaineistossa, haastavat äänitysolosuhteet, sekä koulutusaineiston heikko määrä kielille, joilla on vähemmän puhujia [40, 41, 42].

2.5.1. Cocktailkutsuongelma

Cocktailkutsuilmiö (cocktail party effect) tarkoittaa ihmisten kykyä keskittyä yhden tietyn henkilön puheeseen, metelistä ja suuresta taustalla puhuvien ihmisten määrästä huolimatta [43]. Tämä ilmiö määriteltiin ensimmäisen kerran vuonna 1954 Colin Cherryn julkaisemassa tutkimuksessa, jossa hän käsitteli eri tekijöiden vaikutuksista ihmisen puheentunnistukseen [44]. Vaikka tämäntyyppiset häiriötekijät eivät aiheutakaan terveelle ihmiselle juurikaan ongelmia, coktailkutsuongelma (cocktail party problem, CPP) on kuitenkin yhä suuri haaste ASR-ratkaisuille [45, 41].

Jo ennen syväoppimisen hyödyntämistä ongelmalle oli koitettu useita eri ratkaisuja, jotka ovat jaettavissa yksikanavaisiin järjestelmiin (single-channel system) ja monikanavaisiin järjestelmiin (multi-channel system). Järjestelmien kanavien määrä kertoo äänityksessä käytettävien mikrofonien lukumäärän. [46]

Syväoppimisen nousu on johtanut kehitykseen yksikanavaisiin järjestelmiin perustuvissa ratkaisuissa [47, 48]. Monikanavaisten järjestelmien vahvuutena kuitenkin säilyy keilantaminen, jossa useaa mikrofonia käyttämällä on mahdollista asettaa eri suunnille joko vahvistavaa, tai vaimentavaa interferenssiä [41, 46].

Virtuaaliassistenttien yhteydessä käytettävä syötelaite vaikuttaa kuitenkin siihen, minkä tyyppisiä äänenpaikantamiseen liittyviä ratkaisuja on mahdollista käyttää. Esimerkiksi Amazonin älykotilaitteet pystyvät päättelemään, mikä käytössä olevista laitteista on käyttäjäänsä lähimpänä [49], mikä tarkoittaisi, että tämäntyyppisissä tilanteissa voisi olla mahdollista hyödyntää monikanavaisten järjestelmien ratkaisuja. Toisaalta yleensä puhelimien ja kannettavien tietokoneiden äänityslaitteissa ei

ole huomioitu useampia mahdollisia äänen suuntia, jolloin ollaan rajoitettuja yksikanavaisiin ratkaisuihin.

2.5.2. Suomenkielinen puheentunnistus

Kaikista maailman kielistä vain äärimmäisen pienelle osalle löytyy ASR-ratkaisuja [50]. Tämän huomioon ottaen suomenkielisen puheentunnistuksen tila maailman mittakaavassa on siis hyvä. Suomenkielisten ASR-ratkaisujen suoriutuminen on kuitenkin useissa tapauksessa maailmankieliä heikompaa [51, 52] ja esimerkiksi tämänhetkisistä suosituista virtuaaliassistenteista vain Applen Siri tukee puhuttua suomen kieltä [53].

Suomenkielisten ja erityisesti kuluttajakäyttöön tarkoitettujen ASR-ratkaisujen nykyiseen tilaan vaikuttavat useat tekijät. On esimerkiksi mahdollista, että suurien yritysten suomenkielisten palveluiden puute tai heikompi laatu voi johtua muun muassa pienestä markkina-alueesta johtuvasta kehitysmotiivien puutteesta [52].

Mahdollinen vaikuttava tekijä on myös suomenkielisen koulutusaineiston määrä. Käyttäen esimerkkinä joukkoistamisella kasattua vapaaehtoisilta kerättyä Common Voice ASR-aineistoa

Taulukko 1. Esimerkkejä Common Voice Corpus 7.0:n kielistä

Kieli	Aineiston koko
Englanti	65 Gt
Saksa	26 Gt
Ranska	21 Gt
Espanja	17 Gt
Italia	8 Gt
Kymri / Wales	4 Gt
Arabia	3 Gt
Kiina (Hong Kong)	3 Gt
Kiina (Taiwan)	2 Gt
Kiina (Kiina)	2 Gt
Viro	1 Gt
Ruotsi	1004 Mt
Suomi	256 Mt

voidaan taulukosta 1 huomata, että suomenkielisen aineiston määrä on huomattavasti maailmankielien aineistojen määrää vähäisempää ja on jäljessä jopa viron- ja ruotsinkielisistä aineistoista [54]. On toki tärkeää huomioida, että näiden kielien ASR-ratkaisujen taso, tai olemassa olevien koulutusaineistojen määrä ei ole suoraan pääteltävissä Common Voice:n kielijakaumasta. Esimerkiksi siniittisten kielien aineistojen yhteenlaskettu koko on 7 Gt, jääden jälkeen jopa italiasta, siitä huolimatta, että muun muassa mandariinikiinalle on runsaasti dataa

ja mandariinikiinan ASR-ratkaisujen kehittämiseksi on tehty paljon tutkimustyötä [55, 56, 57, 58]. Tämä on vain yksi tapa demonstroida laadukkaan suomenkielisen datan vähäistä määrää, suhteutettuna enemmän puhuttuihin kieliin.

Suomenkielisen puheentunnistuksen haasteisiin liittyy myös sen epätavalliset ominaisuudet verrattuna esimerkiksi englannin ja saksan kieliin. Muun muassa suomi on agglutinatiivinen kieli, eli sanojen vartaloon voidaan liittää useita affikseja [59]. Suomi on myös morfologisesti englantia rikkaampi kieli [60], eli suomen kielessä sanojen taivutuksella ilmaistaan asioita, joita esimerkiksi englannin kielessä kuvailtaisiin käyttämällä lauserakenteita [61]. Tämä on huomioitu myös ASR-ratkaisuissa. Esimerkiksi muuttamalla syötteen käyttämät yksiköt kokonaisista sanoista sanaa alemmiksi yksiköiksi (subword unit), kuten tavuiksi tai foneemeiksi [60].

Suomen kielen ominaisuuksiin liittyy kuitenkin myös vahvuuksia ASR:n kannalta. Esimerkiksi sanojen lausumisen säännöllisyys suhteutettuna kirjoitettuun muotoon mahdollistaa paremman sanaa alempien yksiköiden hyödyntämisen [62]. On myös huomioitava, että koska suomen kielen morfologisista ominaisuuksista johtuen on mahdollista ilmaista yhdellä sanalla käsitteitä, jotka esimerkiksi englannin kielessä vaatisivat useita sanoja, on ymmärrettävää, että suomenkielisten ASR-ratkaisujen sanavirheiden määrät ovat suurempia [63]. Suomenkielinen puheentunnistusmalli voi siis tuottaa hyödyllisiä tuloksia, vaikka se jäisikin sanavirheiden määrää tarkastellessa esimerkiksi suurien yritysten englanninkielisten ratkaisujen jälkeen.

3. VIRTUAALIASSISTENTIN TOTEUTUS

Toteuttavan virtuaaliassistentin on ymmärrettävä ennaltamäärätyt ääneenlausutut suomenkieliset komennot. Tämän lisäksi assistentin on yleisesti ymmärrettävä suomen kieltä tarpeeksi hyvin, jotta käyttäjän on mahdollista hyödyntää ennalta määrittelemättömiä sanoja ja lauseita, esimerkiksi kysymysten yhteydessä. Assistenttia tullaan käyttämään tietokoneella, joten on suunniteltava graafinen käyttöliittymä, jotta käyttäjän on mahdollista kommunikoida assistentin kanssa käytännöllisesti.

3.1. Puheentunnistusmallin koulutus

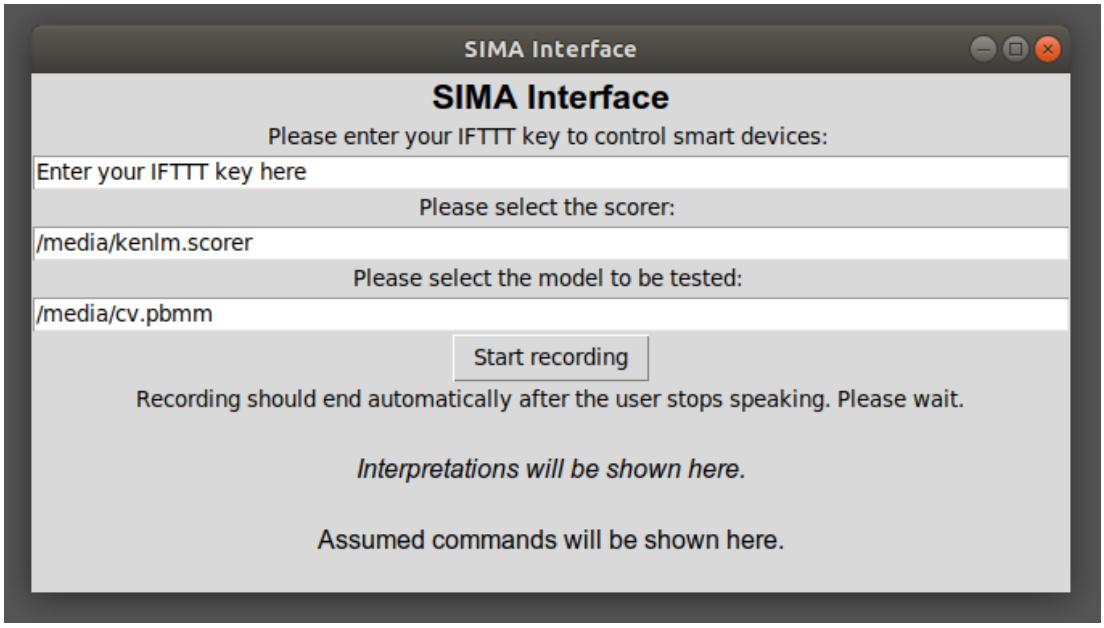
Puheentunnistusmallin koulutuksessa käytettiin Mozillan avoimen lähdekoodin DeepSpeech -projektia suomenkielisellä Common Voice Corpus 7.0 -datalla [54]. Koulutuksessa käytettiin myös Ylen vuosien 2011-2018 suomenkielisestä sekoitetusta uutisarkistosta [64] koottua pisteyttäjää (scorer), joka auttaa parantamaan tulkinnan laatua ennakoimalla mitkä sanat seuraavat todennäköisesti toisiaan. Toteutuksessa käytettiin DeepSpeech:n omaa MPL-2.0-lisenssillä (Mozilla public license, Mozillan vapaa ohjelmistolisenssi) julkaistua, valmiiksi koulutettua, englanninkielistä akustista mallia [65], josta tiputettiin kaksi päällimmäistä tasoa, jonka jälkeen se koulutettiin uudelleen suomenkielisellä datalla siirto-oppimista hyödyntäen.

Koulutus suoritettiin myös käyttämällä sekoitettua dataa, jossa koulutus-, testaus- ja validointiaineistojen määrät kaksinkertaistettiin lisäämällä näytteitä eduskunnan istunnoista vuosilta 2008-2020 kootusta puheentunnistuskorpuksesta [66]. Tämän korpuksen transkriptiot ovat toteutettu puheentunnistusta käyttämällä, mutta tästä huolimatta niiden laatu on hyvä lausujan puhuessa normaalilla puhetahdilla. Aineisto valittiin sen suuren koon, tasaisen äänenlaadun ja puhujien korkean määrän lisäksi siksi, että aineiston jäsentely teki sen käyttöönottamisen koulutuksessa helpoksi. Näin on mahdollista vertailla miten suurikokoisempi, sekä auditiivisilta ominaisuuksiltaan moninaisempi koulutusdata suoriutuu verrattuna Common Voice:n 256 megatavun aineistoon.

3.2. Käyttöliittymä ja toiminnot

Toteutettu virtuaaliassistentti käyttää graafista käyttöliittymää, jonka avulla käyttäjän on mahdollista vaihdella eri puheentunnistusmallien ja pisteyttäjien välillä, sekä kokeilla ennalta määrättyjä komentolauseita. Toteutus käyttää myös verkkosivustojen haravointiin suunniteltuja ohjelmointirajapintoja, mahdollistaen tiedonhaun muun muassa Wikipediasta [67], sekä käyttäjän valitsemasta hakukoneesta. Käyttöliittymän tekstimuotoisten vastausten lisäksi assistentti pystyy myös vastaamaan puhutulla suomen kielellä Googlen TTS-ohjelmistorajapintaa käyttämällä [68, 69].

Älylaitteiden kanssa kommunikoiminen toteutettiin IFTTT-palvelun kautta. Assistentti lähettää haluttua komentoa vastaavan http-pyynnön IFTTT:lle, joka puolestaan välittää halutun tehtävän varsinaiselle laitteelle. Tällä tavoin assistentti pystyy helposti kommunikoimaan laitteiden kanssa, jotka muuten kuuluisivat lukuisiin eri ohjelmistoekosysteemeihin.



Kuva 3. Virtuaaliassistentin graafinen käyttöliittymä

Komentolauseet on toteutettu merkkijonojen, tai listojen sijasta kaksiulotteisilla taulukoilla, joissa on täydellisesti tulkittujen sanojen lisäksi myös mallin yleisimmät virhesanat. Nämä virhesanat kerättiin testaamalla komentolausetta kunnes koottuja virhesanoja oli kymmenen. Mikäli komennon alustavan testaamisen aikana ei jollekin sanalle saatu kymmentä virhettä, saatettiin näitä lisätä, jos virhetulkintoja löytyi kehityksen myöhemmässä vaiheessa. Kuvasta 3 näkee virtuaaliassistentin käyttöliittymän.

3.3. Komennon nauhoitus ja tulkinta

Käyttäjän painettua nappia aloitetaan puheen nauhoittaminen. Tässä käytetään PyAudiota [70], PortAudion [71] rajapintaa. Äänitys toteutetaan taulukon 2 mukaisilla parametreilla, joilla saadaan aikaseksi tulkittava wav-tiedosto. SoX-komentorivityökalu [72] konvertoi wav-tiedoston automaattisesti DeepSpeech:n ymmärtämään muotoon. Sekunnin hiljaisuuden jälkeen nauhoitus lopetetaan ja äänitys tallennetaan väliaikaisesti.

Tämän jälkeen seuraava funktio hakee tämän äänitiedoston ja suorittaa sille tulkinnan valittua mallia ja pisteyttäjää käyttämällä. Sitten tämä tulkinta tallennetaan väliaikaisesti tekstitiedostolle. Seuraavaksi displayIntep-funktio näyttää saadun tulkinnan käyttöliittymässä.

3.3.1. NLP-komennot

Seuraavaksi assistentti siirtyy mahdollisten komentojen valintaan, jossa arvioidaan mitä toimintoa käyttäjä todennäköisimmin tarkoittaa. Testin alussa funktio arvioi

Taulukko 2. Pyaudion äänityksessä käytetyt parametrit

Parametri	Arvo
Formaatti	pyaudio.paInt16
Kanavien määrä	2
SHORT NORMALIZE	(1.0/32768.0)
Lohko (Chunk)	1024
Näytetaajuus	48000
Näytteen leveys	2
Kohinaportti	40

onko kyseessä luonnollisen kielen käsittelyä vaativa komento. Esimerkiksi komennon ollessa "mitä on tekoäly" on pääteltävä, että kyseessä on tiedonhakukomento hakusanalla "tekoäly". Nämä NLP-komennot vaativat tarkempia sanajärjestyksiä ja tästä syystä niiden tarkistaminen suoritetaan erillään muiden komentovaihtoehtojen pisteytyksestä.

Useimmat luonnollisen kielen käsittelyä vaativista komennoista koskevat tiedonhakuun liittyviä toimintoja. Alustavasti toteutettu virtuaaliassistentti pyrkii haravoimaan (web scraping) tarvittavan tiedon Wikipediasta, mutta mikäli sopivaa osumaa ei löydy, suoritetaan haravoiminen verkkohausta. Tämä vastaa myös teknologiajättien virtuaaliassistenttien toimintaa. Esimerkiksi Amazonin Alexa, sekä Microsoftin Cortana käyttävät kummatkin Bing-hakukonetta, kun taas Google Assistant käyttää luonnollisesti Googlea. Näin toteutettu assistentti pystyy tarjoamaan käyttäjälleen toiminnallisuuksia, jotka ovat verrattavissa viimeisintä tekniikkaa edustaviin ratkaisuihin.

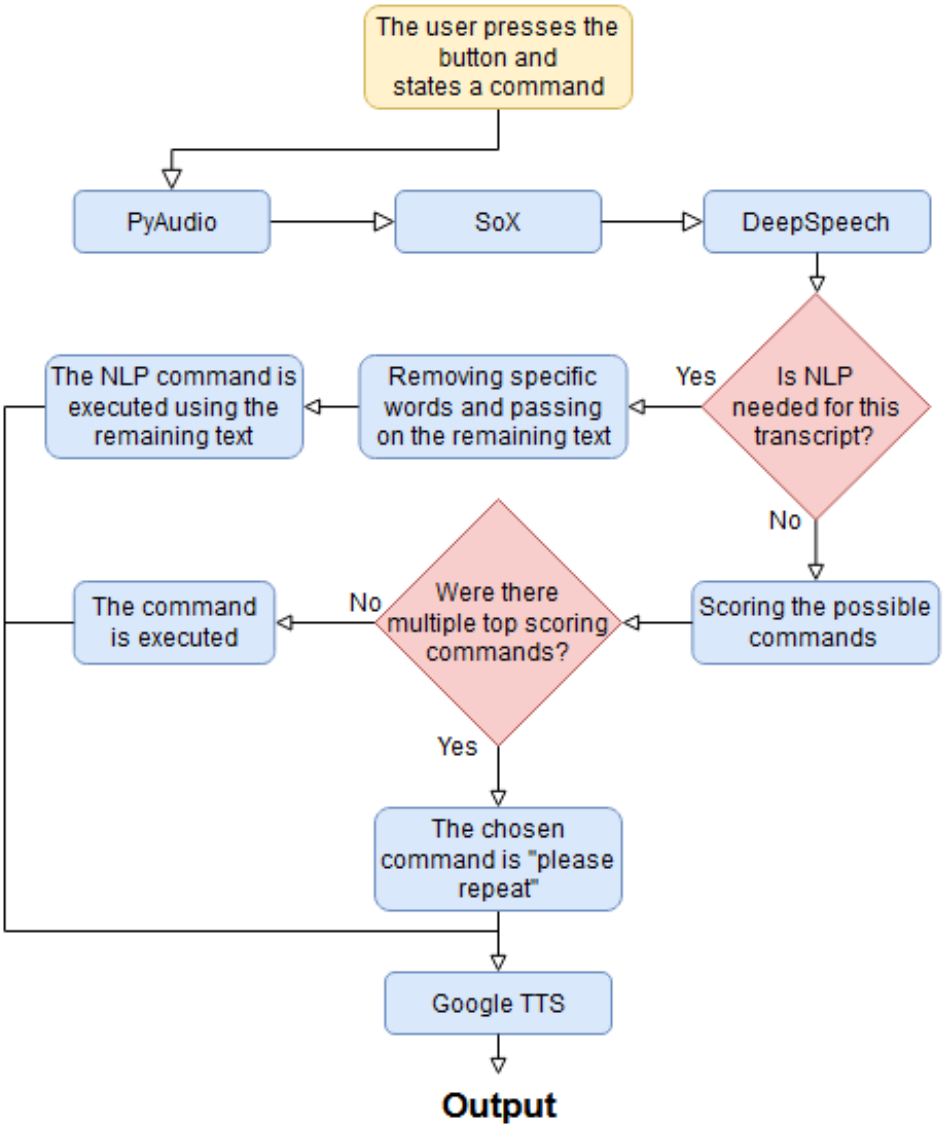
3.3.2. Komentojen pisteytys

Mikäli NLP-komennon ehdot eivät täyty, siirrytään muiden komentojen pisteyttämiseen. Komennot kuten “aseta valot punaisiksi” eivät vaadi lauseen parsimista ja tästä syystä on mahdollista hyväksyä myös epätavallisempia sanajärjestyksiä käyttäviä syötteitä, kuten “valot punaisiksi aseta”. Tämän tyyppisissä komennoissa ei ole myöskään välttämätöntä vaatia syötteeltä kaikkia komentoon kuuluvia sanoja. Esimerkiksi syöte “valot punaisiksi” sisältää olennaisen informaation puuttuvasta sanasta huolimatta.

Pisteytyksessä syötelause jaetaan listaksi sanoja ja näitä sanoja verrataan jokaista komentoa vastaavaan taulukkoon. Mikäli listalla oleva sana löytyy taulukon riviltä, tämä rivi tiputetaan pisteytyksestä ja komento saa yhden pisteen. Lopullinen komennon saama pisteytys saadaan jakamalla kerättyjen pisteiden määrä taulukon rivien, eli komennon sanojen määrällä. Mikäli taulukossa on enemmän kuin yksi rivi ja

taulukosta löytyi vain yksi oikea rivi, asetetaan komennon lopulliseksi pistemääräksi nolla. Tällä tavalla vältytään siltä, että monisanainen komento voisi aktivoitua vain yhden oikean syötesanan perusteella.

Pisteytyksen lopussa vertaillaan komentojen saamia pisteitä ja valitaan parhaiten sopiva toiminto. Mikäli useampi komento saa saman pistemäärän, mitään toimintoa ei suoriteta, käyttäjää ilmoitetaan virheestä ja pyydetään uutta syötettä. Pisteytyksen jälkeen seuraavalle funktiolle kerrotaan toteutettava toiminto mahdollisine parametreineen. Tämä funktio lopulta suorittaa halutun tehtävän. Seuraavaksi speakLine-funktio generoi määrätystä vastauksesta äänitiedoston, toistaa sen käyttäjälle ja sitten poistaa äänitiedoston. Lopulta onClick-funktio päättyy ja käyttäjä voi aloittaa uuden komennon nauhoittamisen painamalla nappia. Kuvassa 4 havainnollistetaan virtuaaliassistentin toimintaa. Virtuaalisessa assistentissa käytetyt tiedostot ovat saatavilla GitHubista: https://github.com/vkuosman/SIMA-IPA-Interface



Kuva 4. Virtuaaliassistentin toimintaa havainnollistava kaavio. Kuva on luotu tätä kandidaatintyötä varten.

4. KIELIMALLIEN TESTAUS JA SUORITUSKYVYN KEHITTÄMINEN

Toteutetulle assistentille koulutettiin kaksi erillistä kielimallia, joista tarkoituksena on valita parhaiten suoriutuva. Lisäksi kielimallin toimintaa tehostavien toimintojen suoriutumista on testattava, jotta voidaan olla varmoja siitä, mitkä toiminnot on viisainta ottaa käyttöön varsinaisessa assistentissa.

4.1. Koulutusaineistojen suoriutuminen ja vertailu

Ensimmäisen puheentunnistusmallin kouluttamisessa käytettiin pelkästään Mozillan Common Voice Corpus 7.0:n suomenkielistä dataa. Tässä tietoaineistossa on 12 tuntia puhetta, josta 6 tuntia on validoitua, eli puhujan lisäksi erillisen ihmisen tarkistamaa. Aineistossa olevien puhujien määrä on 139.

Toisen puheentunnistusmallin koulutuksessa aikaisempaan tietoaineistoon lisättiin Aalto yliopiston Finnish Parliament ASR korpus, joka on kasattu eduskunnan täysistuntojen äänitteistä ja pöytäkirjoista. Käytetyssä tietoaineiston osassa puhujien määrä on 58. Tästä osasta otetaan yksi näytepari yhtä Common Voice:n näyteparia kohden, kaksinkertaistaen koulutusdatan määrän.

Ensimmäisen koulutuksen yhteydessä epookin koulutusvaiheen kestoaika oli noin 8 minuuttia askelmäärällä 2104, kun taas validointivaiheen kesto oli noin 2,5 minuuttia askelmäärällä 1464. Toisen puheentunnistusmallin koulutuksessa puolestaan epookin koulutusvaiheen kestoaika oli noin 24 minuuttia askelmäärällä 4209 ja validointivaihe kesti noin 7,5 minuuttia askelmäärällä 2929. Kummankin koulutuksen askelmäärät määräytyivät aineiston koon mukaisesti. Kummankin mallin koulutukseen käytettiin noin neljä tuntia.

Taulukko 3. Koulutettujen puheentunnistusmallien saamat testitulokset

Mallin käyttämä koulutainesto	Testauksessa käytetty aineisto	WER	CER	Loss
Common Voice Corpus 7.0	Common Voice Corpus 7.0	0,536	0,223	48,601
Common Voice Corpus 7.0 + Parliament ASR Corpus	Common Voice Corpus 7.0 + Parliament ASR Corpus	0,541	0,290	110,849

4.1.1. CommonVoiceCorpus 7.0 -aineistolla toteutettu malli

Ensimmäisen mallin lopulliset testitulokset näkyvät taulukossa 3. Taulukossa näytetään mallien sanavirhesuhteet (word error rate, WER), merkkivirhesuhteet (character error rate, CER), sekä hukka (loss), joka kertoo kuinka paljon mallin antama tulkinta eroaa todellisesta syötteestä. Puhtaalla Common Voice -korpuksen aineistolla koulutetun ja testatun mallin sanavirhesuhteeksi saatiin 0,536. Verrattaessa esimerkiksi Googlen 4,9% ja Microsoftin 5,1% WER-tuloksiin [39, 73], suoriutuu koulutettu malli suhteellisen heikosti. On kuitenkin huomioitava, että Googlen ja Microsoftin tulokset tulevat englanninkielisestä puheesta. Tulokset ovat kuitenkin riittävät assistentin toteutukseen, sillä mallin tarvitsee vain valita syötteen perusteelta haluttu komento, eikä varsinaisen tulkinnan tarvitse olla täydellinen.

Taulukko 4. Näytteitä Common Voice Corpus 7.0 -mallin vahvimmista tulkinnoista

	Lause 1	Lause 2
Alkuperäinen lause	"olen itse henkilökohtaisesti käynyt kuussa katsomassa rakennusprojektia"	"äänestin luonnollisesti mietinnön hyväksymisen puolesta"
Mallin tulkinta	"olen itse henkilökohtaisesti käynyt kuussa katsomassa rakennusprojektia"	"äänestin luonnollisesti mietinnön hyväksymisen puolesta"
Tulkinnan WER	0,000	0,000
Tulkinnan CER	0,000	0,000

Taulukon 4 tuloksista voidaan nähdä, että vaikka malli ei saavutakaan tekniikan nykytilaa edustavien englanninkielisten mallien tasoa, onnistuu se siitä huolimatta tuottamaan täydellisiä tulkintoja suhteellisen haastavillekkin lauseille. Kaikkien komentolauseiden tapauksissa ei siis välttämättä tarvitsisi ottaa huomioon virheellisiä tulkintoja, mutta taulukon 4 näytteet vastaavat vain mallin parhaimpia suorituksia.

Taulukko 5. Näytteitä Common Voice Corpus 7.0 -mallin mediaanitulkinnoista

	Lause 1	Lause 2
Alkuperäinen lause	"semmoiselta se toisistakin tuntui"	"eikä kuule kanan laulua"
Mallin tulkinta	"semmoisella se toisestakin tuntui"	"eikä kuulekaan laulua"
Tulkinnan WER	0,500	0,500
Tulkinnan CER	0,061	0,087

Mediaanisuoritusta vastaavat lauseet taulukossa 5 olivat virheellisiä, mutta useimmissa tapauksissa alkuperäisen lauseen olennaisin informaatio oli kuitenkin ymmärrettävissä. Tästä huolimatta mediaanituloksista on jo havaittavissa lauseita, joiden tulkinnoissa on liiallisesti lauseen ymmärrettävyyten vaikuttavia virheitä.

Taulukko 6. Näytteitä Common Voice Corpus 7.0 -mallin heikoimmista tulkinnoista

	Lause 1	Lause 2	Lause 3
Alkuperäinen lause	"mysteerimiehen matka"	"hauskaa pikkujoulua"	"selässä"
Mallin tulkinta	"myös termiä hän matka"	"hauskat pikku jolla"	"siellä saa"
Tulkinnan WER	1,500	1,500	2,000
Tulkinnan CER	0,350	0,211	0,714

Yllättäen vaikka 1,5-2,0 WER:n tulokset taulukossa 6 eroavatkin selkeästi oikeista vastauksista, on kuitenkin useimmissa tapauksissa helppoa hahmottaa miksi kyseinen lause on tuottanut ongelmia. Esimerkiksi “mysteerimiehen” ja “myös termiä hän” muistuttavat foneettisesti paljon toisiaan ja lausumisen nopeudesta ja selkeydestä riippuen tämä voisi aiheuttaa virheellisen tulkinnan ihmisellekin.

4.1.2. Usean lähteen aineistolla toteutettu malli

Taulukosta 3 voidaan nähdä, että sekakoulutettu malli suoriutui yhtä pitkän koulutusajan jälkeen heikommin kuin Common Voice:n datalla koulutettu malli, kun askelten määrää oli nostettu lisätyn datan mukaisesti.

Taulukko 7. Näytteitä sekadatamallin vahvimmista tulkinnoista

	Lause 1	Lause 2
Alkuperäinen lause	"arvoisa puhemies hyvät kollegat miksi meille nyt esitettävät ehdotukset ovat niin tärkeitä"	"lapsen oikeuksien sopimuksen mukaisesti suomella on velvollisuus toimia lasten pelastamiseksi"
Mallin tulkinta	"arvoisa puhemies hyvät kollegat miksi meille nyt esitettävät ehdotukset ovat niin tärkeitä"	"lapsen oikeuksien sopimuksen mukaisesti suomella on velvollisuus toimia lasten pelastamiseksi"
Tulkinnan WER	0,000	0,000
Tulkinnan CER	0,000	0,000

Kuten ensimmäinenkin malli, myös sekadatalla koulutettu malli onnistuu tuottamaan täydellisiä tulkintoja haastavistakin lauseista. Taulukosta 7 voidaan nähdä, että toinen malli onnistui saamaan täydellisiä vastauksia myös pidemmille testilauseille, mikä saattaa johtua täysistuntoaineistoon kuuluvista huomattavasti pidemmistä koulutuslauseista.

Taulukko 8. Näytteitä sekadatamallin mediaanitulkinnoista

	Lause 1	Lause 2
Alkuperäinen lause	"mutta aika jännä että kun mennään sijoitusvakuutusten puolelle näiden kuorien puolelle niin siellä sisällä saa kyllä tehdä kaikenlaista kauppaa ja voittoa ja se ei sitten kuulu verottajalle ollenkaan"	"ravintolayrittäjä tai laivaristeilyjä järjestävä yrittäjä syyllistyy laittomuuteen jos myy esimerkiksi naisille tässä tapauksessa lippuja puoleen hintaan esimerkiksi siten että risteilylle tai ravintolaan pääsee kaksi naista yhden lipun hinnalla"
Mallin tulkinta	"mutta aika jännä että kun mennään tänne sijoitusvakuutuksen puolelleen kuoren puolelle siellä sisällä sitä kyllä tehdä kaikenlaista kaavoittaa ja sitten kuljettajalle olleen"	"ravintolan yrittäjä tai risteilylaivat järjestävä yrittää ylityötunteja naisille esimerkiksi naisille tässä tapauksessa lippuja puoleen hintaan esimerkiksi siten että telttakatoksessa yhden lipun"
Tulkinnan WER	0,536	0,536
Tulkinnan CER	0,240	0,318

Mediaanitason saavuttaneista tulkinnoista taulukossa 8 voidaan kuitenkin nähdä, että mallilla on yhä ongelmia täysistuntodatan pidempien lauseiden kanssa.

On toisaalta tärkeää huomioida, että kummankin testausmateriaalin lauseen kohdalla on otettu vapauksia puheen kirjaamisessa. Esimerkiksi lauseesta 1 on tiputettu sana "tänne", joka löytyy kuitenkin mallin tekemästä tulkinnasta. Lauseessa 2 puolestaan puhuja sekoaa sanoissaan ja sanoo sanan "risteily" ennen sanaa "laivaristeilyjä", selittäen mallin arvauksen "risteilylaivat".

Taulukko 9. Näytteitä sekadatamallin heikoimmista tulkinnoista

	Lause 1	Lause 2
Alkuperäinen lause	"hauskaa pikkujoulua"	"poikki heinäkarheiden"
Mallin tulkinta	"hauska pikku joulua"	"kaikki heinä kahden"
Tulkinnan WER	1,500	1,500
Tulkinnan CER	0,105	0,286

Taulukon 9 huonoiten suoriutuneet tulkinnat muistuttavat kuitenkin hyvin paljon ensimmäisen mallin tuloksia. Erityisesti syötesanan “selässä” virheellinen tulkinta “siellä saa” on täsmälleen sama kuin ensimmäisellä mallilla. Yhtäläisyydet mallien säännöllisesti tekemissä virheissä voisi mahdollistaa kummankin mallin käyttämisen, vaikka oletetut virhetulkinnat oltaisiinkin määritelty vain ensimmäistä mallia käyttämällä.

4.2. Automaattinen yhdyssanavirheiden korjaaminen

Suomen kielen ominaisuuksista johtuen suomenkieliset puheentunnistusmallit saavat yleensä heikompia tuloksia WER-testeissä, kuin esimerkiksi englanninkieliset mallit [63]. Yksi vaikuttavista tekijöistä on yhdyssanojen suuri määrä [74]. Vaikka tuotettu tulkinta olisi täysin ymmärrettävissä, yksikin yhdyssanavirhe voi heikentää merkittävästi tulkinnan WER-tulosta. Esimerkiksi tulkinta "missä jalka pallo" antaa WER-tulokseksi 1,000. Korjaamalla yhdyssanavirheitä mallin suoriutumisen arvioiminen WER-testien perusteella olisi siis luotettavampaa. Myös toteutetun virtuaalisen assistentin tiedonhakukomennot voisivat toimia paremmin, jos yhdyssanavirheiden määrää onnistuttaisiin vähentämään.

Ongelman ratkaisemiseksi kehitettiin automaattinen tarkistin, joka pyrkii korjaamaan mahdollisimman monta yhdyssanavirhettä. Koska koulutukseen ja testaukseen käytettävästä aineistosta on poistettu väliviivat, ratkaisu ei pyri tuottamaan väliviivallisia yhdyssanoja, koska niiden onnistumista on mahdotonta arvioida nykyisellä aineistolla. Sanojen testaamiseen kerättiin 61803 sanaa sisältävä aineisto, joka koottiin Wikisanakirjan suomenkielisten yhdyssanojen listasta [75]. Tarkistimen testaamiseen otettiin Common Voice:n tietoainestosta 5000 näytettä, joita ei oltu käytetty mallin koulutuksessa. Common Voice Corpus 7.0 -malli tuotti tulkinnat kaikille näytteille ja tarkistin kävi läpi kaikki nämä tulkinnat. Jokaisesta vierekkäin esiintyvistä sanoista muodostettiin yhdistelmäsanat. Suomen kielen morfologiseen analyysiin suunnitellulla Voikolla [76] luotiin lista kaikista yhdistelmäsanoista, joille Voikko ei löytänyt korjausta. Seuraavaaksi muodostettiin lista sanoista, jotka Voikko onnistui muuttamaan perusmuotoon. Mikäli yhdistelmäsana löytyi Wikisanakirjasta kootusta yhdyssanalistasta, käytettiin yhdistelmäsanaa tulkinnan korjaamiseen.

4.2.1. Korjauksen arviointi

Jotta automaattisen korjauksen suoriutumista voitaisiin arvioida kunnolla, on korjausratkaisua verrattava täydellisiin korjauksiin. Kaikkien oikein suoritettujen yhdyssanavirheiden korjausten saamista ei ole käytännöllistä suorittaa käsin, käytettyjen näytteiden suuresta määrästä johtuen. Oikein korjattujen tulkintojen saamista ei voi myöskään automatisoida, joten ihannetuloksina käytetään tulkintoja, jotka on korjattu näytteiden litteroituja lauseita hyödyntämällä. Mikäli yhdistelmäsana löytyy alkuperäisestä lauseesta, sitä käytetään tulkinnan korjaamisen. On tärkeää huomioida, että vaikka tulkinnasta löytyisi kaksi sanaa, joista olisi muodostettava yhdyssana, yhdyssanaa ei kuitenkaan muodosteta, jos sitä ei löydy alkuperäisestä lauseesta. Testauksesta saadut tulokset näkyvät taulukossa 10.

Taulukko 10. Yhdyssanavirheiden korjauksen tulokset

	Ilman korjausta	Automaattinen korjaus	Ihannetulos
WER	0,472	0,470	0,460
Muutettujen tulkintojen määrä	0 / 5000	81 / 5000	185 / 5000

Toteutetulla automaattisella korjauksella saatu muutos tuloksessa oli noin 19% ihannetuloksen muutoksesta. On hyvä huomioida, että ihannetulosta on mahdotonta saavuttaa tietämättä alkuperäisiä litteroituja lauseita. Muokattuja lauseita tarkastellessa havaittiin, että automaattinen korjaus oli tuottanut kieliopillisesti korrekteja yhdyssanoja, joilla oli kuitenkin ollut negatiivinen vaikutus WER-tuloksiin, koska tulkinta oli ollut muutenkin virheellinen. Esimerkiksi sanojen "ihmisen" ja "aluksi" yhdistämisestä saatu "ihmisenaluksi" alensi tulkinnan tulosta, koska todelliset sanat olivat "ihmisen" ja "alkukotiin". Ihannetuloksen korjauksissa vastaavia virheitä ei ollut mahdollista tapahtua.

Ihannetuloksen korjauksissa oltiin myös yhdistetty kieliopillisesti korrekteja perussanoja toisiksi perussanoiksi. Esimerkiksi "osasivat" ja "pahan" oltiin yhdistetty, koska alkuperäisessä lauseessa esiintyi sana "osasivatpahan". Tämäntyyppisiä muutoksia ei tehty toteutettussa automaattisessa korjauksessa, koska käytetyssä tarkistuslistassa oli ainoastaan yhdyssanoja. Perussanoja ei kuitenkaan ole viisasta lisätä tarkistuslistaan, koska tällöin myös oikein tulkitut sanat voitaisiin yhdistää virheellisesti.

Ihannetuloksen korjauksista löytyy kuitenkin lauseita, jotka on mahdollista muodostaa myös automaattista korjausta hyödyntämällä. Näissä tapauksissa yhdyssanat ovat olleet harvinaisempia, eikä niitä ole löytynyt tarkistuslistasta. Tämänkaltaisia sanoja olivat esimerkiksi "toimielinuudistus" ja "lainsäädäntöpäätöslauselma". Tämä toi esille myös mielenkiintoisia aukkoja tarkistusanastossa. Esimerkiksi sanat "ajanjumala" ja "tyttöjoukko" eivät löydy, vaikka sanat "sodanjumala" ja "miesjoukko" löytyvät.

4.3. Toivotun komennon valinta pisteytyksellä

Koska puheentunnistusmalli ei tuota aina täydellisiä tuloksia, oli toteuttava pisteytyssysteemi, jota käyttämällä assistentilla olisi mahdollista päätellä, mitä toimintoa käyttäjä tarkoittaa. Tulkintavirheiden korjaamisen lisäksi toteuttavan toiminnon valitseminen pisteytyksellä mahdollistaa myös sen, että käyttäjän ei ole muistettava täsmälleen oikeaa lausetta toiminnon aktivoimiseksi. Komentolauseet toteutettiin taulukoina, joissa oli komennon perusmuodon lisäksi otettu huomioon synonyymit ja mallin tekemät yleisimmät virheet. Ennen uusien toimintojen lisäämistä oli kuitenkin suoritettava testauksia, jotta voitaisiin olla varmoja pisteytyksen tuomista hyödyistä.

Testaukseen osallistui kehittäjän lisäksi neljä vapaaehtoista henkilöä. Koehenkilölle annettiin lista ääneen lausuttavia komentoja, jotka kaikki toistettiin kolmesti. Äänitykset suoritettiin koehenkilöiden valitsemissa sisätiloissa ja ennen testausta kaikilta koehenkilöiltä varmistettiin, että he suostuivat vapaaehtoisesti tietojen luovuttamiseen. Testauksen aikana kirjattiin ylös assistentin tuottamat tulkinnat, sekä aktivoituiko toiminto vai ei. Tulkinnan ollessa identtinen komennon perusmuodon kanssa tiedetään, että komento olisi onnistunut myös ilman pisteytystä. Koehenkilöiden saamat tulokset, sekä ikä- ja sukupuolijakaumat näkyvät taulukossa 11. Tarkemmat testitulokset löytyvät liitteestä 1.

4.3.1. Pisteytyksen testitulokset

Taulukko 11. Tietoa koehenkilöistä. Alaikäiseltä koehenkilöltä vaadittiin huoltajan hyväksyntä. Oletetut virheet valikoitu kehittäjän ääninäytteillä. Koehenkilön 3 äänityksessä oli lievää taustahälyä.

	Ikä	Sukupuoli	Käyttänyt aikaisemmin virtuaalista assistenttia	Tulos ilman pisteytystä	Tulos pisteytyksellä
Kehittäjä Koehenkilö 0	24	mies	kyllä	21 / 30	28,5 / 30
Koehenkilö 1	54	mies	kyllä	13,5 / 30	23 / 30
Koehenkilö 2	17	nainen	kyllä	12,5 / 30	20 / 30
Koehenkilö 3	22	nainen	kyllä	11,5 / 30	25 / 30
Koehenkilö 4	51	nainen	ei	10,5 / 30	18,5 / 30

Taulukon 11 testituloksista voimme nähdä, että komennon valitseminen toteutetulla pisteytyksellä paransi kaikkien testihenkilöiden tulkintoja. Saaduista tuloksista voidaan päätellä, että puheentunnistusmallin säännöllisesti tuottamilla virheillä on merkittävästi päällekkäisyyksiä puhujien välillä. Tämä mahdollistaa toimivien komentotaulukoiden muodostamisen yhdeltä ihmiseltä saatujen säännöllisten virhetulkintojen perusteella. Komentotaulukoiden sanastot hyötyisivät kuitenkin useammilta ihmisiltä kerätyistä säännöllisistä virhetulkinnoista. Esimerkiksi testihenkilön 2 tapauksessa sana "paljonko" sai testauksen aikana kahdesti tulkinnan "palmian".

4.3.2. Mahdollisten väärien positiivisten testaaminen

Koska virtuaaliassistentti ei vaadi komennolta tiettyjä sanoja ennaltamäärätyssä sanajärjestyksessä, on mahdollista, että pisteytys voi tuottaa komennolle väärän positiivisen. Väärien positiivisten esiintymistodennäköisyyden arvioimiseksi koehenkilöitä 1, 2, 3 ja 4 pyydettiin keksimään ja lausumaan lauseita, sekä komentoja,

joihin he toivoisivat virtuaalisen assistentin reagoivan. Jokainen koehenkilö testasi kymmenen lausetta, eli assistenttille syötettiin yhteensä 40 äänitettä. Tämä testi suoritettiin ennen pisteytyksen testaamista, jotta koehenkilöt eivät voisi tietää, mitä toimintoja assistenttiin on toteutettu.

Testauksen yhteydessä vain yksi lause tuotti väärään positiivisen. Koehenkilön 1 lause "mikä maa on lentopallon maailmanmestari" sai tulkinnan "mikä maa on lentopallon maailman näistä". Tämä tulkinta aktivoi toiminnon, jonka komentolauseen perusmuoto on "mikä on elämän tarkoitus". Tämä toiminto toteutettiin varhaisessa vaiheessa kehitystyötä, eikä se ole enää aktivoitavissa perusmuotoa käyttämällä. Komennon perusmuodon käyttäminen assistentin tämänhetkisessä tilassa johtaa tiedonhakukomentoon. Koska tämänhetkinen tiedonhakukomento hyväksyy vain tietyllä tavalla muotoillut kysymykset, "mikä maa on lentopallon maailman näistä" -syötettä ei tulkita tiedonhakukomennoksi. Koska syötteestä löytyvät sanat "mikä" ja "on" on lähimpänä oleva komentolause "mikä on elämän tarkoitus".

Vaikka testauksessa saatiin vain yksi väärä positiivinen, voidaan päätellä, että assistentin nykyisessä tilassa virheelliset tiedonhakukomennot voivat helposti aktivoida elämän tarkoitusta kysyvän toiminnon. Helpoin tapa ratkaista ongelma olisi muuttaa "mikä on elämän tarkoitus" -komento vain muotoon "elämän tarkoitus". Komento oltiin alunperin suunniteltu pelkäksi pääsiäismunaksi (easter egg), eli piilotetuksi viestiksi, joten komennon muokkaamisella ei olisi negatiivista vaikutusta virtuaaliassistentin toimintaan.

5. JATKOKEHITYS

Vaikka virtuaaliassistenttin ja sen käyttämän puheetunnistusmallin toteuttamisessa onnistuttiin, löytyy assistentista ja puheetunnistusmallista kuitenkin vielä mahdollisia jatkokehityksen kohteita.

5.1. Virtuaaliassistentin kehittäminen

Siitä huolimatta, että virtuaaliassistentti kykeneekin useisiin samoihin toimintoihin kuin suurien yritysten assistentit, osa komennoista vaatii tällä hetkellä liian spesifejä lauserakenteita. Assistentti tuottaa siis halutun vastauksen esimerkiksi lauseella "mikä on suomen itsenäisyyspäivän päivämäärä", mutta ei lauseella "milloin on suomen itsenäisyyspäivä". Toiminnot soveltuvat siis käyttäjälle, joka on tietoinen assistentin rajoitteista, mutta uudelle käyttäjälle virtuaaliassistentin käyttäminen voi vaatia hieman opettelua. Ongelmaa voisi koittaa ratkaista lisäämällä uusia lausetaulukoita ja ylimääräisiä NLP-komentoihin liittyviä tarkastuksia.

Assistentti ei myöskään tällä hetkellä suoriudu useimpien englanninkielisten sanojen tunnistamisessa. Vaikka kyseessä onkin suomenkielinen virtuaaliassistentti, olisi englannin kielen sanaston osaaminen hyödyksi muun muassa soitettavan musiikin valitsemisessa. Tällä hetkellä assistentti hyväksyy esimerkiksi kysymyksen "ketkä ovat metallican jäsenet?", mutta kysymystä "ketkä ovat rainbow:n jäsenet?" assistentti ei onnistu tulkitsemaan oikein.

Mahdollinen ratkaisu ongelmaan olisi tulkita käyttäjän antamat lauseet suomenkielisen mallin lisäksi, myös englanninkielisellä mallilla. Jos verkkoharavointi ei palauta tulosta suomenkielisen mallin tekemällä tulkinnalla, tehtäisiin englanninkielisestä tulkinnasta lista. Käyttämällä esimerkkinä englanninkielisen Google Assistantin tulkintaa edellä mainitulle rainbow-bändin jäsenistä tiedustelevalle kysymykselle, tulkinta voisi olla esimerkiksi "get cabbage rainbow me a sentence". Tämän listan perusteella voidaan suorittaa useampia verkkohakuja, kuten "get", "get cabbage"... "rainbow", "rainbow me" ja niin edelleen. Näistä hauista verkkoharavointi saisi tuloksen vain sanalla "rainbow", jolloin yksinkertaisimmillaan olisi mahdollista korvata suomenkielisestä tulkinnasta "ketkä" ja "ovat" -sanojen jälkeiset sanat, tuottaen lopulliseksi hauksi "ketkä ovat rainbow?". Tämä ei täysin vastaisi käyttäjän antamaa lausetta, mutta tuottaisi kuitenkin hyödyllisiä hakutuloksia. Olisi myös mahdollista testata hienostuneempia ratkaisuja, jotka mahdollisesti pystyisivät säilyttämään syötteen lauserakenteen.

Assistentin tiedonhakutoimintoja voisi parantaa myös hyödyntämällä kehitettyä automaattista yhdyssanavirheiden korjausta. Esimerkiksi oikean tiedon haravoiminen Wikipediasta onnistuisi suuremmalla todennäköisyydellä, mikäli yhdyssanavirheiden määrää vähennettäisiin. Harvinaissempien yhdyssanojen korjaamiseksi olisi myös täydennettävä olemassaolevaa yhdyssanojen tarkistuslistaa.

5.2. ASR:n kehittäminen

Koulutettu puheentunnistusmalli soveltuu yksinkertaisen virtuaaliassistentin toimintoihin, mutta monimutkaisempien chatbot-ominaisuuksien toteuttaminen tämänhetkisellä mallilla ei todennäköisesti tuottaisi luonnolliselta vaikuttavia keskusteluja. ASR:n kehittämistä olisi siis jatkettava, mikäli vastaavia hienostuneempia toimintoja haluttaisiin implementoida assistenttiin.

Paremmin suoriutuvan mallin toteuttaminen vaatisi enemmän koulutusdataa. Vaikka testauksessa sekadatalla kouluttaminen ei tuottanut parempia tuloksia, ei se kuitenkaan tarkoita, etteikö koulutus voisi hyötyä uusien aineistojen sovittamisesta osaksi nykyistä koulutusaineistoa. Esimerkiksi Lahjoita puhetta -kampanjan keräämä koulutusaineisto sopisi todennäköisesti hyvin tähän käyttötarkoitukseen. On myös todennäköistä, että Common Voice -korpusta tullaan päivittämään vielä tulevaisuudessa, jolloin myös on mahdollista suorittaa koulutus käyttäen suurempaa aineistoa.

6. YHTEENVETO

Tavoitteena olleet koneoppimisella koulutettu suomenkielinen puheentunnistusmalli ja tätä hyödyntävä virtuaalinen assistentti toteutettiin onnistuneesti. Työn alussa Common Voice:n puheaineistolla koulutetun kielimallin suoriutuminen ei vaikuttanut lupaavalta, mutta käyttämällä Mozillan tarjoamaa englanninkielistä puheentunnistusmallia oli mahdollista hyödyntää siirto-oppimista, tuottaen lopulta hyviä tuloksia. Yksinkertaisen graafisen käyttöliittymän toteuttaminen Ubuntulla ei tuottanut ongelmia. Toteutus mahdollistaa muun muassa lamppujen hallinnan ja tiedonhaun ilman, että käyttäjä joutuu kaivamaan puhelintaan esille.

Verkon haravointi tuottaa yksinkertaisesta toteutuksestaan huolimatta hyviä tuloksia. Kuten myös esimerkiksi Google Assistant, toteutettu virtuaaliassistentti pystyy haravoimaan ääneen luettavan vastauksen useimpiin yksinkertaisiin, oikein tulkittuihin kysymyksiin ja vähintään näyttämään hakukoneella saadut tulokset kysymyksen ollessa haastavampi. Assistentin kyky vastata kysymyksiin on kuitenkin hyvin riippuvainen valitusta hakukoneesta, sillä toteutus ei saa tietoa Wikipediasta, mikäli ei haun nimistä artikkelia ole olemassa.

Parhaiten toimiva malli koulutettiin pelkällä Mozillan Common Voice:n puheainestolla, eikä lisätyllä täysistunnoista saadulla datalla ollut positiivista vaikutusta sen suoriutumiseen. Tämä johtuu todennäköisesti puheaineiston heikommasta laadusta. Puhe on äänitetty Eduskuntatalon istuntosalissa, joka eroaa huomattavasti akustisilta ominaisuuksiltaan sekä asumiseen tarkoitetuista huoneista, että ulkotiloista, joissa virtuaaliassistenttia todennäköisimmin käytetään. Täysistunnosta nauhoitetusta datasta löytyy myös paljon äänitteitä joissa on muun muassa sanoissa sekoilua, epäselvää lausumista, sekä tavallisesta puheesta eroavaa sanojen painotusta ja nostettua ääntä. Lisäksi aineston transkriptioista löytyy myös virheitä. Erityisesti silloin kun henkilö puhuu luonnottoman nopeasti, esimerkiksi lakitekstejä ääneen luettaessa.

Puheentunnistusmallin tuottamia virheitä kompensoitiin onnistuneesti automaattisella yhdyssanavirheiden korjaamisella, sekä hyödyntämällä komentotaulukoita suoritettavan toiminnon valitsemisessa. Automaattinen yhdyssanavirheiden korjaus löysi 19% ihanteellisten korjausten löytämistä virheistä. Komentotaulukoita ja pisteytystä hyödyntävän toiminnonvalinnan suoriutumista testattiin useilla puhujilla. Tuloksista voitiin todeta ratkaisun parantaneen tuloksia kaikilla koehenkilöillä. Testauksessa esiintyi vain yksi väärä positiivinen, joka aiheutui virheellisen toiminnon komentotaulukosta.

Mahdollisia jatkokehityksen kohteita on esimerkiksi englanninkielisen mallin lisääminen erilaisten lainasanojen, sekä brändien ja yhtyeiden nimien tunnistamiseksi. Lisäksi uudempien tietoaineistojen hyödyntäminen puheentunnistusmallin koulutuksessa voisi parantaa virtuaalisen assistentin suorituskykyä.

7. VIITTEET

[1] Campagna G., Ramesh R., Xu S., Fischer M. & Lam M.S. (2017) Almond: The architecture of an open, crowdsourced, privacy-preserving, programmable virtual assistant. Teoksessa: Proceedings of the 26th International Conference on World Wide Web, ss. 341–350.

[2] Lau J., Zimmerman B. & Schaub F. (2018) Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. Proceedings of the ACM on Human-Computer Interaction 2, ss. 1–31.

[3] Singh H. (2019), Apple's Siri the Most Popular Virtual Assistant With a 35 Percent Market Share: Report. URL: https://gadgets.ndtv.com/smart-home/news/apple-siri-leads-virtual-assistant-market-share-alexa-google-assistant-cortana-report-2151436. [Verkossa; 10-9-2021].

[4] Microsoft Support (2021), Cortana and privacy. URL: https://support.microsoft.com/en-us/topic/cortana-and-privacy-47e5856e-3680-d930-22e1-71ec6cdde231. [Verkossa; 10-9-2021].

[5] Levy A. (2016), Amazon brings alexa voice control to cloud computing. URL: https://www.cnbc.com/2016/11/30/amazon-brings-alexa-voice-control-to-cloud-computing.html. [Verkossa; 10-9-2021].

[6] Sterling G. (2019), Google assistant moves from the cloud to the phone, now 10x faster. URL: https://searchengineland.com/google-assistant-moves-from-the-cloud-to-the-phone-now-10x-faster-316556. [Verkossa; 10-9-2021].

[7] Apple (2021), iOS ja iPadOS ominaisuuksien saatavuus. URL: https://www.apple.com/fi/ios/feature-availability/#siri. [Verkossa; 10-9-2021].

[8] Chung H. & Lee S. (2018) Intelligent virtual assistant knows your life. arXiv preprint arXiv:1803.00466 .

[9] Kowalski J., Jaskulska A., Skorupska K., Abramczuk K., Biele C., Kopeć W. & Marasek K. (2019) Older adults and voice interaction: A pilot study with google home. Teoksessa: Extended Abstracts of the 2019 CHI Conference on human factors in computing systems, ss. 1–6.

[10] Sen S., Chakrabarty S., Toshniwal R. & Bhaumik A. (2015) Design of an intelligent voice controlled home automation system. International Journal of Computer Applications 121.

[11] McCarthy J. (2007) What is artificial intelligence? .

[12] Mayor A. (2020) Gods and robots: myths, machines, and ancient dreams of technology. Princeton University Press.

[13] Hutton D. (2011) The quest for artificial intelligence: A history of ideas and achievements. Kybernetes .

[14] Hockstein N.G., Gourin C., Faust R. & Terris D.J. (2007) A history of robots: from science fiction to surgical robotics. Journal of robotic surgery 1, ss. 113–118.

[15] Haenlein M. & Kaplan A. (2019) A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. California management review 61, ss. 5–14.

[16] Muggleton S. (2014) Alan turing and the development of artificial intelligence. AI communications 27, ss. 3–10.

[17] Moor J. (2006) The dartmouth college artificial intelligence conference: The next fifty years. Ai Magazine 27, ss. 87–87.

[18] Solomonoff R.J. (1985) The time scale of artificial intelligence: Reflections on social effects. Human Systems Management 5, ss. 149–153.

[19] Bullinaria J.A. (2004) Introduction to neural networks. University of Birmingham, UK .

[20] He T., Fan Y., Qian Y., Tan T. & Yu K. (2014) Reshaping deep neural network for fast decoding by node-pruning. Teoksessa: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, ss. 245–249.

[21] McCulloch W.S. & Pitts W. (1943) A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5, ss. 115–133.

[22] Rosenblatt F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 65, s. 386.

[23] Abdi H. (1994) A neural network primer. Journal of Biological Systems 2, ss. 247–281.

[24] Tuominen H. (n.d), Tietotekniikan ajankohtaisia teemoja -kurssi: Johdatus tekoälyn taustalla olevaan matematiikkaan, Informaatioteknologian tiedekunta, Jyväskylän Yliopisto. URL: https://tim.jyu.fi/view/143092#DKUvbnUuGytQ. [Verkossa; 15-9-2021].

[25] Kainen P., Krková V. & Vogt A. (2000) Best approximation by heaviside perceptron networks. Neural networks 13, ss. 695–697.

[26] Friedman J., Hastie T., Tibshirani R. et al. (2001) The elements of statistical learning, nide 1. Springer series in statistics New York.

[27] Abiodun O.I., Jantan A., Omolara A.E., Dada K.V., Mohamed N.A. & Arshad H. (2018) State-of-the-art in artificial neural network applications: A survey. Heliyon 4, s. e00938.

[28] Heininen S. (2020) Pro gradu -tutkielma: Syväoppivat neuroverkot ja niiden sovellukset, matematiikan ja tilastotieteen laitos, turun yliopisto .

[29] Graves A., Mohamed A.r. & Hinton G. (2013) Speech recognition with deep recurrent neural networks. Teoksessa: 2013 IEEE international conference on acoustics, speech and signal processing, Ieee, ss. 6645–6649.

[30] Sak H., Senior A. & Beaufays F. (2014) Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv preprint arXiv:1402.1128 .

[31] Mikolov T. & Zweig G. (2012) Context dependent recurrent neural network language model. Teoksessa: 2012 IEEE Spoken Language Technology Workshop (SLT), IEEE, ss. 234–239.

[32] Torrey L. & Shavlik J. (2010) Transfer learning. Teoksessa: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI global, ss. 242–264.

[33] Abd Almisreb A., Jamil N. & Din N.M. (2018) Utilizing alexnet deep transfer learning for ear recognition. Teoksessa: 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), IEEE, ss. 1–5.

[34] Alexandre L.A. (2016) 3d object recognition using convolutional neural networks with transfer learning between input channels. Teoksessa: Intelligent Autonomous Systems 13, Springer, ss. 889–898.

[35] Yu D. & Deng L. (2016) Automatic Speech Recognition. Springer.

[36] Radha V. & Vimala C. (2012) A review on speech recognition challenges and approaches. doaj. org 2, ss. 1–7.

[37] Shrawankar U. & Thakare V.M. (2013) Techniques for feature extraction in speech recognition system: A comparative study. arXiv preprint arXiv:1305.1145 .

[38] Munteanu C., Penn G., Baecker R. & Zhang Y. (2006) Automatic speech recognition for webcasts: how good is good enough and what to do when it isn't. Teoksessa: Proceedings of the 8th international conference on Multimodal interfaces, ss. 39–42.

[39] Protalinski E. (2017), Google's speech recognition technology now has a 4.9% word error rate. URL: https://venturebeat.com/2017/05/17/googles-speech-recognition-technology-now-has-a-4-9-word-error-rate/. [Verkossa; 14-9-2021].

[40] Sheng L.M.A. & Edmund M.W.X. (2017) Deep learning approach to accent classification. CS229 .

[41] Qian Y.m., Weng C., Chang X.k., Wang S. & Yu D. (2018) Past review, current progress, and challenges ahead on the cocktail party problem. Frontiers of Information Technology & Electronic Engineering 19, ss. 40–63.

[42] Thai B., Jimerson R., Ptucha R. & Prud'hommeaux E. (2020) Fully convolutional asr for less-resourced endangered languages. Teoksessa: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), ss. 126–130.

[43] Arons B. (1992) A review of the cocktail party effect. Journal of the American Voice I/O Society 12, ss. 35–50.

[44] Cherry E.C. (1953) Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America 25, ss. 975–979.

[45] Chang X., Qian Y., Yu K. & Watanabe S. (2019) End-to-end monaural multi-speaker asr system without pretraining. Teoksessa: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, ss. 6256–6260.

[46] Chen Z., Li J., Xiao X., Yoshioka T., Wang H., Wang Z. & Gong Y. (2017) Cracking the cocktail party problem by multi-beam deep attractor network. Teoksessa: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, ss. 437–444.

[47] Hershey J.R., Chen Z., Le Roux J. & Watanabe S. (2016) Deep clustering: Discriminative embeddings for segmentation and separation. Teoksessa: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, ss. 31–35.

[48] Chen Z., Luo Y. & Mesgarani N. (2017) Deep attractor network for single-microphone speaker separation. Teoksessa: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, ss. 246–250.

[49] Moscaritolo A. (2018), Amazon helps alexa know which device you're talking to. URL: https://uk.pcmag.com/speakers/116596/amazon-helps-alexa-know-which-device-youre-talking-to. [Verkossa; 16-9-2021].

[50] Barnard E., Davel M. & Van Heerden C. (2009) Asr corpus design for resource-scarce languages. ISCA.

[51] Caballero D., Araya R., Kronholm H., Viiri J., Mansikkaniemi A., Lehesvuori S., Virtanen T. & Kurimo M. (2017) Asr in classroom today: automatic visualization of conceptual network in science classrooms. Teoksessa: European Conference on Technology Enhanced Learning, Springer, ss. 541–544.

[52] Krogerus M. (2020), Savolainen, kainuulainen ja oululainen yrittivät puhua sirille, mutta turhaan – katso videolta, kuinka vaikeita murteet ovat tekoälylle. URL: https://yle.fi/uutiset/3-11490240. [Verkossa; 16-9-2021].

[53] Summalinguae (2021), Language support in voice assistants compared. URL: https://summalinguae.com/language-technology/language-support-voice-assistants-compared/. [Verkossa; 16-9-2021].

[54] Mozilla Foundation (2021), Common voice corpus 7.0. URL: https://commonvoice.mozilla.org/en/datasets. [Verkossa; 12-9-2021].

[55] Zhou S., Dong L., Xu S. & Xu B. (2018) Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese. arXiv preprint arXiv:1804.10752 .

[56] Fu S.W., Lee C. & Clubb O.L. (1996) A survey on chinese speech recognition. Communications of COLIPS 6, ss. 1–17.

[57] Wang D., Zhang X. & Zhang Z. (2015), Thchs-30 : A free chinese speech corpus. URL: http://arxiv.org/abs/1512.01882.

[58] Magic Data Technology Co., Ltd. (2019), Magicdata mandarin chinese read speech corpus. URL: http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101.

[59] Hakkani-Tür D.Z., Oflazer K. & Tür G. (2002) Statistical morphological disambiguation for agglutinative languages. Computers and the Humanities 36, ss. 381–410.

[60] Jain A., Rouhe A., Grönroos S.A., Kurimo M. et al. (2020) Finnish asr with deep transformer models. Teoksessa: Interspeech, ss. 3630–3634.

[61] Dahl Ö. (2008) Kuinka eksoottinen kieli suomi on? Virittäjä 112, ss. 545–545.

[62] Smit P., Virpioja S. & Kurimo M. (2021) Advances in subword-based hmm-dnn speech recognition across languages. Computer Speech & Language 66, s. 101158.

[63] Kurimo M. (2008) Puheentunnistus. Puhe ja kieli , ss. 73–83.

[64] Westerlund H. & Dieckmann U. (2021), Yle Finnish News Archive 2011-2018, scrambled, VRT. URL: https://metashare.csc.fi/repository/browse/yle-finnish-news-archive-2011-2018-scrambled-vrt/7f48ccc0438511eaa8e7005056be118eefd767c621f64753aa4e0b5f439be593/. [Verkossa; 12-9-2021].

[65] Mozilla Foundation (2020), Deepspeech 0.9.3. URL: https://github.com/mozilla/DeepSpeech/releases/tag/v0.9.3. [Verkossa; 12-9-2021].

[66] Axelson E. & Dieckmann U. (2021), Aalto Finnish Parliament ASR Corpus 2008-2020. URL: https://metashare.csc.fi/repository/browse/aalto-finnish-parliament-asr-corpus-2008-2020/da99e872b88511eb9cdefa163ec5ae3ed97d37d535fb4ab0a83b82a805704243/. [Verkossa; 12-9-2021].

[67] Goldsmith J. (2014), Wikipedia python library. URL: https://pypi.org/project/wikipedia/, [Verkossa; 14-10-2021].

[68] Google LLC (2021), text-to-speech: Lifelike speech synthesis. URL: https://cloud.google.com/text-to-speech, [Verkossa; 14-10-2021].

[69] Durette P.N. (2021), gtts 2.2.3. URL: https://pypi.org/project/gTTS/, [Verkossa; 14-10-2021].

[70] Pham H. (2017), Pyaudio 0.2.11. URL: https://pypi.org/project/PyAudio/, [Verkossa; 14-10-2021].

[71] Bencina R. & Burk P. (2021), Portaudio. URL: http://www.portaudio.com/, [Verkossa; 14-10-2021].

[72] Bagwell C. e.a. (2015), Sox - sound exchange. URL: http://sox.sourceforge.net/Main/HomePage, [Verkossa; 14-10-2021].

[73] Chen H. (2021), Does Word Error Rate Matter? URL: https://www.smartaction.ai/blog/does-word-error-rate-matter/. [Verkossa; 14-9-2021].

[74] Karjalainen P. (2008), Yhdyssanavirheet merkonomiopiskelijoiden itsearvioinneissa.

[75] Wiktionary (2021), Finnish compound words. URL: https://en.wiktionary.org/wiki/Category:Finnish_compound_words, [Verkossa; 12-10-2021].

[76] Pitkänen, H (2019), Voikko, free linguistic software and data for finnish. URL: https://voikko.puimula.org/. [Verkossa; 12-10-2021].

8. LIITTEET

. Liite 1. Komentojen pisteytyksen testauksesta saadut tulokset. Ulostulojen yhdyssanat ovat kirjoitettu yhteen luettavuuden parantamiseksi. Todellisessa ulostulossa sanat ovat erillään. Osittaisessa onnistumisessa oikea toiminto on suoritettu virheellisillä syötesanoilla. Osittaisista onnistumisista annetaan 0,5 pistettä.

Testihenkilö 0 (kehittäjä)			
Syötelause	Ulostulo	Tulkinnan onnistuminen pelkän ulostulon perusteella	Tulkinnan onnistuminen oletetuilla virheillä ja pisteytyksellä
paljonko kello on	paljonko kello on	kyllä	kyllä
paljonko kello on	paljonko kello on	kyllä	kyllä
paljonko kello on	aluksellaan	ei	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	mikä sinun nimesi vaan	ei	kyllä
minun nimeni on väinämöinen	minun nimeni on vetäminen	osittainen	osittainen
minun nimeni on väinämöinen	minun nimeni on väinämöinen	kyllä	kyllä
minun nimeni on väinämöinen	minun nimeni on väinämöinen	kyllä	kyllä
kuka on sauli niinistö	kukaan sauli niinistä	ei	osittainen
kuka on sauli niinistö	kukaan sauli niinistö	ei	kyllä
kuka on sauli niinistö	kuka on sauli niinistö	kyllä	kyllä
näytä uutisia	näytä uutisia	kyllä	kyllä
näytä uutisia	näytä uutisia	kyllä	kyllä
näytä uutisia	näitä uutisia	ei	kyllä
laita valot päälle	laita valot päälle	kyllä	kyllä
laita valot päälle	laita palot päällä	ei	kyllä
laita valot päälle	laita valot päälle	kyllä	kyllä
aseta valot punaisiksi	aseta palot punaiseksi	ei	kyllä
aseta valot punaisiksi	aseita valot punaiseksi	ei	kyllä
aseta valot punaisiksi	aseta valot punaisiksi	kyllä	kyllä
mitä on suomenkieli	mitä on suomenkieli	kyllä	kyllä
mitä on suomenkieli	mitä on suomenkieli	kyllä	kyllä
mitä on suomenkieli	mitä on suomenkieli	kyllä	kyllä
mikä on kalevala	mikä on kalevala	kyllä	kyllä
mikä on kalevala	mikä on kalevala	kyllä	kyllä
mikä on kalevala	mikä on kalevala	kyllä	kyllä
sano testilause	sano testilause	kyllä	kyllä
sano testilause	sano testilause	kyllä	kyllä
sano testilause	sano estela usa	osittainen	osittainen
		21 / 30	28,5 / 30
Testihenkilö 1			
Syötelause	Ulostulo	Tulkinnan onnistuminen pelkän ulostulon perusteella	Tulkinnan onnistuminen oletetuilla virheillä ja pisteytyksellä
paljonko kello on	alan kokelaan	ei	ei
paljonko kello on	palon kukaan	ei	ei
paljonko kello on	alan kokea on	ei	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
minun nimeni on väinämöinen	minun nimeni on vain noin	osittainen	osittainen
minun nimeni on väinämöinen	minun nimeni on väinämöinen	kyllä	kyllä
minun nimeni on väinämöinen	minun nimeni on väinämöinen	kyllä	kyllä
kuka on sauli niinistö	kukaan sauli niinistö	ei	kyllä
kuka on sauli niinistö	kukaan sauli niinistö	ei	kyllä
kuka on sauli niinistö	kuka on sauli niinistö	kyllä	kyllä
näytä uutisia	näitä uutisia	ei	kyllä
näytä uutisia	näytä uutisia	kyllä	kyllä
näytä uutisia	näitä uutisia	ei	kyllä
laita valot päälle	aika valot päälle	ei	kyllä
laita valot päälle	laita valot päälle	kyllä	kyllä
laita valot päälle	laita valot alle	ei	kyllä
aseta valot punaisiksi	aseta valot punaisiksi	kyllä	kyllä
aseta valot punaisiksi	aseta valot punaisiksi	kyllä	kyllä
aseta valot punaisiksi	aseta palot punaisiksi	ei	kyllä
mitä on suomenkieli	mitä suomenkieli	ei	ei
mitä on suomenkieli	mitä suomenkieli	ei	ei
mitä on suomenkieli	mitä on suomenkieli	kyllä	kyllä
mikä on kalevala	mikään kalevala	ei	ei
mikä on kalevala	mikä on kalevala	kyllä	kyllä
mikä on kalevala	mikä on kalevala	kyllä	kyllä
sano testilause	sanatestin lause	ei	ei
sano testilause	sanot testillä se	ei	osittainen
sano testilause	sanot testilause	ei	kyllä
		13,5 / 30	23 / 30

Testihenkilö 2			
Syötelause	Ulostulo	Tulkinnan onnistuminen pelkän ulostulon perusteella	Tulkinnan onnistuminen oletetuilla virheillä ja pisteytyksellä
paljonko kello on	palmian kokeella anna	ei	ei
paljonko kello on	palmian kokeella on	ei	kyllä
paljonko kello on	alan kopilla on	ei	ei
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	mukaan nimesi anna	ei	kyllä
minun nimeni on väinämöinen	minun nimeni väinämöinen	ei	kyllä
minun nimeni on väinämöinen	minun nimeni on väinämöinen	kyllä	kyllä
minun nimeni on väinämöinen	minun nimeni on väinämöinen	kyllä	kyllä
kuka on sauli niinistö	kuka on sauli niinistö	kyllä	kyllä
kuka on sauli niinistö	kuka on solin niinistä	osittainen	osittainen
kuka on sauli niinistö	kuka on sauli niinistö	kyllä	kyllä
näytä uutisia	näitä uutisia	ei	kyllä
näytä uutisia	näitä uutisia	ei	kyllä
näytä uutisia	näitä uutisia	ei	kyllä
laita valot päälle	laitan valot päälle	ei	kyllä
laita valot päälle	laita valot päälle	kyllä	kyllä
laita valot päälle	laita valot päälle	kyllä	kyllä
aseta valot punaisiksi	aseta valot kun aiheiksi	ei	ei
aseta valot punaisiksi	aseta valot punaisiksi	kyllä	kyllä
aseta valot punaisiksi	aseta valvoituissa	ei	ei
mitä on suomenkieli	niitä on suomen pull	ei	ei
mitä on suomenkieli	mitä on suomenkieli	kyllä	kyllä
mitä on suomenkieli	mikä suomenkieli	ei	ei
mikä on kalevala	mikä on kalevala	kyllä	kyllä
mikä on kalevala	mikä on kalevala	kyllä	kyllä
mikä on kalevala	mikä kalevala	ei	ei
sano testilause	säesti lause	ei	ei
sano testilause	sanot kestilause	ei	osittainen
sano testilause	sen testilause	ei	ei
		12,5 / 30	20 / 30
Testihenkilö 3			
Syötelause	Ulostulo	Tulkinnan onnistuminen pelkän ulostulon perusteella	Tulkinnan onnistuminen oletetuilla virheillä ja pisteytyksellä
paljonko kello on	paljon kokeella on	ei	kyllä
paljonko kello on	paljon kokeella on	ei	kyllä
paljonko kello on	paljon kokeella on	ei	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
minun nimeni on väinämöinen	minun nimeni väinämöinen	ei	kyllä
minun nimeni on väinämöinen	minun nimeni väinämöinen	ei	kyllä
minun nimeni on väinämöinen	minun nimeni väinämöinen	ei	kyllä
kuka on sauli niinistö	kuka on sauli niinistö	kyllä	kyllä
kuka on sauli niinistö	kuka on salinin mistä	osittainen	osittainen
kuka on sauli niinistö	kukaan sauli niinistö	ei	kyllä
näytä uutisia	aika uutisia	ei	kyllä
näytä uutisia	malta uutisia	ei	kyllä
näytä uutisia	näitä uutisia	ei	kyllä
laita valot päälle	laita valot päälle	kyllä	kyllä
laita valot päälle	laita valot	ei	kyllä
laita valot päälle	laita maltalle	ei	ei
aseta valot punaisiksi	asiasta valot punaisiksi	ei	kyllä
aseta valot punaisiksi	aseta valot punaisiksi	kyllä	kyllä
aseta valot punaisiksi	aseta valot punaisiksi	kyllä	kyllä
mitä on suomenkieli	mitä on suomenkieli	kyllä	kyllä
mitä on suomenkieli	mitä on suomenkieli	kyllä	kyllä
mitä on suomenkieli	mitä suomenkieli	ei	ei
mikä on kalevala	mikä on kalevala	kyllä	kyllä
mikä on kalevala	mikä on kalevala	kyllä	kyllä
mikä on kalevala	mitä on kalevala	ei	kyllä
sano testilause	sankasti lause	ei	ei
sano testilause	sanat pestilause	ei	osittainen
sano testilause	anttila	ei	ei
		11,5 / 30	25 / 30

Testihenkilö 4			
Syötelause	Ulostulo	Tulkinnan onnistuminen pelkän ulostulon perusteella	Tulkinnan onnistuminen oletetuilla virheillä ja pisteytyksellä
paljonko kello on	paljon kokeella on	ei	kyllä
paljonko kello on	paljon kokeilla on	ei	kyllä
paljonko kello on	paljon kokeella on	ei	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	mikä sinun nimesi on	kyllä	kyllä
mikä sinun nimesi on	miksi nimesi on	ei	ei
minun nimeni on väinämöinen	minun nimeni on väinämöinen	kyllä	kyllä
minun nimeni on väinämöinen	minun nimeni on väinämöinen	kyllä	kyllä
minun nimeni on väinämöinen	minun nimeni on väinämöinen	kyllä	kyllä
kuka on sauli niinistö	kukaan niinistä	ei	osittainen
kuka on sauli niinistö	kukaan sauli niinistö	ei	kyllä
kuka on sauli niinistö	kuka sai niistä	ei	ei
näytä uutisia	näitä uutisia	ei	kyllä
näytä uutisia	näitä uutisia	ei	kyllä
näytä uutisia	näitä uutisia	ei	kyllä
laita valot päälle	laita valot päälle	kyllä	kyllä
laita valot päälle	laita maltalle	ei	ei
laita valot päälle	laita alapää le	ei	ei
aseta valot punaisiksi	aseta valot punaisiksi	kyllä	kyllä
aseta valot punaisiksi	aseta valuisi sin	ei	ei
aseta valot punaisiksi	aseta valot punaisiksi	kyllä	kyllä
mitä on suomenkieli	mitä suomenkieli	ei	ei
mitä on suomenkieli	mitään suomenkieli	ei	ei
mitä on suomenkieli	mitä on suomenkielen	osittainen	osittainen
mikä on kalevala	mikä kalevala	ei	ei
mikä on kalevala	mikä on kalevala	kyllä	kyllä
mikä on kalevala	mikä on kalevala	kyllä	kyllä
sano testilause	saat testilause	ei	ei
sano testilause	antti lause	ei	ei
sano testilause	ateisti lause	ei	ei
		10,5 / 30	18,5 / 30