TURUN YLIOPISTO

Reseptien kaupunkeja, *kulttuurin värinöitä* ja *luxus demareita*

Koneellisesti laadittujen ohjelmatekstitysten laadunarviointi

Hanne Martikainen
Pro gradu -tutkielma
Monikielisen käännösviestinnän tutkinto-ohjelma, englannin kieli
Kieli- ja käännöstieteiden laitos
Humanistinen tiedekunta
Turun yliopisto
Huhtikuu 2025

Pro gradu -tutkielma

Monikielisen käännösviestinnän tutkinto-ohjelma, englannin kieli
Hanne Martikainen
***Reseptien kaupunkeja*, *kulttuurin värinöitä* ja *luxus demareita*: Koneellisesti laadittujen ohjelmatekstitysten laadunarviointi**
Sivumäärät: tutkielma 58 sivua, liitteet 16 sivua

Pro gradu -tutkielmassa tutkimuksen kohteena on koneellisesti laadittujen ohjelmatekstitysten laadun taso sekä niiden automatisointi ei-ammattimaisen käytön näkökulmasta. Laadunarvioinnissa sovelletaan Pablo Romero-Frescon ja Juan Martínezin kehittämää virheanalyysin NER-mallinnusta, suomalaisia *Ohjelmatekstitysten laatusuosituksia* ja ISO 5060 -standardia, joka tarjoaa virhekategoriat laadun arviointiin.

Aineisto kerätään kolmesta tyypiltään erilaisesta ohjelmasta: 8.9.2024 lähetetystä uutisohjelman jaksosta *Yle Uutiset 18.00*, *Kulttuuricocktail-live*-keskusteluohjelman jaksosta *Mitä kulttuurille tapahtuu eduskuntavaalien jälkeen?* ja *Perjantaidokkari*-ohjelmasarjan dokumentista *Teuvo Tekoäly pelastaa Pyhännän*. Ohjelmista laaditaan puheentunnistimella transkriptiot, joiden pohjalta tutkielmassa hyödynnetty tekoälysovellus Copilot laatii ohjelmatekstitykset. Tutkimuksessa analysoidaan yhteensä yhdeksää Copilotin laatimaa tekstitysversiota. Tämän lisäksi tekstityksiä verrataan niin itse litteroituihin kuin puheentunnistimen transkriptioihin.

Tutkielman tulokset osoittivat, että Copilotin laatimat ohjelmatekstitysversiot vaatisivat huomattavan määrän korjauksia, jotta ne saavuttaisivat riittävän laadun kaikilla arvioiduilla osa-alueilla. Keskeisenä tuloksena esille nousi erityisesti generatiivisen tekoälysovelluksen kyvyttömyys käsitellä video- tai äänitiedostoja, jolloin se menetti tekstittämisen kannalta olennaista tietoa visuaalisesta ja äänellisestä kanavasta. Tämän lisäksi puheentunnistimen tuottaman transkription laatu vaikutti huomattavasti Copilotin kykyyn saavuttaa riittävän laadukasta tulosta.

Tutkimus tuotti hyödyllistä tietoa ei-ammattimaisen ohjelmatekstittämisen koneellistamisesta ja generatiivisen tekoälyn laatiman tekstityksen laadusta. Tämän lisäksi tutkimus vahvisti näkemyksiä siitä, että erityisesti tekstittämisprosessin automatisointi vaatii tällä hetkellä yhä ihmisen tekemää työtä, jotta ohjelmatekstitykset olisivat riittävän saavutettavia.

Avainsanat: audiovisuaalinen kääntäminen, ohjelmatekstitys, laatu, saavutettavuus, kieliteknologia, generatiivinen tekoäly

Sisällysluettelo

1 Johdanto

Audiovisuaalisten eli av-teoksien katsominen on lisääntynyt suomalaisten arjessa huomattavasti viimeisten vuosikymmenten aikana. Sisältöä tuotetaan aiempaa enemmän niin perinteisten tv-ohjelmien kuin Youtube- tai TikTok-videoiden muodossa ja eri ruutujen ääressä vietetään päivittäin huomaamatta jopa tuntikausia. Yhä useampi video sisältää nykyään myös joko ihmisen tai koneen laatimat tekstitykset. Siksi ei lienee yllättävää, että suomalaiset lukevat tekstityksiä vuosittain jopa 30 romaanin verran per henkilö (Holopainen 2015, 87).

Tekstitykset voivat olla monelle tuttuja käännöstekstityksiä, mutta on myös olemassa kuuroille ja huonokuuloisille katsojille suunnattuja kielensisäisiä ohjelmatekstityksiä. Liikenne- ja viestintävirasto Traficom (2023c) toteutti kyselytutkimuksen ohjelmatekstitysten käytöstä ja niiden käyttötarkoituksista eri ikäryhmissä. Tulokset osoittavat, että noin joka neljännes tutkimukseen vastanneista hyödyntää vähintään joskus ohjelmatekstityksiä tv-ohjelmia tai videoita katsellessaan. Ohjelmatekstityksiä käytetään muun muassa ääniraidan epäselvän tai heikosti kuultavissa olevan puheen tai tilanteesta johtuvan mykistetyn ääniraidan vuoksi. Tämän lisäksi ohjelmatekstityksiä hyödynnetään kielen opiskelun, kuulovammojen tai heikkenevän kuulon tukena. (mp.) Näin ollen ohjelmatekstitykset tukevat kaikenikäisten katsojien tarpeita katselukokemuksen ylläpitämisessä.

Yleisen hyödyn lisäksi ohjelmatekstitysten saatavuutta pyritään säätelemään myös eri säädöksillä, sillä yksilöillä on oikeus tietoon heille ymmärrettävässä muodossa. Tällöin kysyntä saavutettavalle sisällölle kasvaa entistä enemmän, jolloin sisällöntuottajat ja palveluntarjoajat pyrkivät vastaamaan näihin vaatimuksiin kehittämällä saavutettavan tarjonnan saatavuutta esimerkiksi teknologiaan nojautuen. Teknologian – erityisesti tekoälyn – kehitys näkyy täten myös av-kääntämisen saralla. Useat palveluntarjoajat ovat tuoneet markkinoille muun muassa erilaisia tekstitysohjelmia, jotka voivat automaattisesti laatia transkription, ajastaa ja kääntää tekstitykset. Nämä ohjelmat ovat kuitenkin useimmiten maksullisia. Tällöin kuluttaja ei välttämättä valitse niitä ensisijaisena vaihtoehtonaan satunnaiseen tekstitystarpeeseen, vaan saattaa kääntyä suosioon nousseiden generatiivisten tekoälysovellusten, kuten ChatGPT:n, puoleen.

Koneellisesti laadittujen tekstitysten laatu on kuitenkin yhä vaihtelevaa erityisesti suomenkielisessä sisällössä. Tämä puolestaan johtaa usein tilanteeseen, jossa laadultaan riittämättömät ja jopa virheelliset tekstitykset herättävät katsojan huomion aiheuttaen monenlaisia reaktioita

aina huvituksesta turhautumiseen (Virta 2019). Laadultaan puutteelliset tekstitykset kuitenkin heikentävät erityisesti kuurojen ja huonokuuloisten katsojien mahdollisuuksia saada riittävästi tietoa katsomastaan sisällöstä, mikä voi vaikeuttaa katselukokemuksesta nauttimista.

Pro gradu -tutkielmani tavoitteena on havainnollistaa ei-ammattimaiseen tekstittämiseen tarkoitettujen tekoälypohjaisten työkalujen kykyä saavuttaa riittävän laadukkaat ohjelmatekstitykset, sillä aiempaa tutkimusta erityisesti generatiivisten tekoälysovelluksien hyödyntämisestä ei vielä ole tehty kattavasti käännöstieteen saralla. Mielenkiinnon kohteena on myös se, kuinka automatisoitua ja koneellista ohjelmatekstitysten laatiminen nykyhetkellä voisi olla erityisesti ei-ammattimaisen käyttäjän näkökulmasta. Tutkielmani tutkimuskysymykset ovat seuraavat:

1. Saavuttavatko ei-ammattimaisesti ja täysin koneellisesti laaditut ohjelmatekstitykset riittävää laatua?
2. Mitä laatua vahvistavia tai heikentäviä tekijöitä generatiivisen tekoälysovelluksen laatimissa ohjelmatekstityksissä on?
3. Mitä työkaluja ohjelmatekstitysten automatisoinnissa vaaditaan?

Opintoni ovat kasvattaneet kiinnostustani audiovisuaalisen sisällön saavutettavuuden tutkimiseen. Tutkielmallani haluankin tuoda esille saavutettavien tekstitysten tärkeyden ja niiden laatimisen haasteet, minkä lisäksi tavoitteenani on tarjota uusia näkökulmia sekä tutkia vaihtoehtoja ohjelmatekstitysten koneellistamiseen. Toivon, että tutkielmani kannustaisi erityisesti ei-ammattimaisesti tekstityksiä laativia huomioimaan laatuun liittyviä seikkoja, varsinkin jos he tukeutuvat generatiivisiin tekoälysovelluksiin, joita ei ole kehitetty tekstittämistarkoitukseen.

Tutkielman aluksi luvussa 2 esittelen taustaa sekä aiempaa tutkimusta saavutettavuudesta, av-kääntämisestä ja kieli- sekä käännösteknologiasta. Tutkielmani keskittyy pitkälti suomenkieliseen sisältöön sekä Suomen tilanteeseen saavutettavan av-kääntämisen ja kieliteknologian parissa. Tämän lisäksi esittelen tutkielmassa sovelletut laadunarviointimallit, minkä jälkeen siirryn kuvaamaan tarkemmin tutkielmani aineistoa ja menetelmiä luvussa 3. Luvussa 4 esitän aineiston analyysin, jossa paneudun syvällisesti ohjelmatekstitysten koneellistamisen vaiheisiin sekä koneellisesti laadittujen ohjelmatekstitysten laadun arviointiin. Lopuksi luvussa 5 tarjoan yhteenvedon tutkimuksestani, tarkastelen kriittisesti tutkimukseni vahvuuksia sekä kehityskohtia ja pohjustan jatkotutkimusten mahdollisuuksia.

2 Audiovisuaalinen kääntäminen, teknologia ja saavutettavuus

Audiovisuaalinen teos koostuu niin visuaalisesta kuin äänellisestä kanavasta ja on siten multimodaalinen kokonaisuus. Dirk Delabastitan (1989) ja Henrik Gottliebin (1997) kategorisointeihin nojaten myös av-teosten viestinnälliset ulottuvuudet voidaan jakaa visuaaliseen ja äänelliseen kanavaan, jotka molemmat tarkentuvat vielä sanallisiksi tai sanattomiksi. (Zárate 2021, 3–4.) Kanavat ovat siten välineitä, joissa esiintyy niin sanallisia kuin sanattomia merkkejä (Delabastita 1989, 198–199). Sanallinen äänikanava sisältää esimerkiksi dialogin, taustaäänet ja ajoittain myös kappaleiden sanoitukset, kun taas sanattomassa äänikanavassa on musiikki sekä ääniefektit. Sanallisessa kuvakanavassa näkyy visuaaliset tekstit, jotka ovat videon sisällä. Tällöin esimerkiksi kyltit kuuluvat sanallisen kuvakanavan kategoriaan. Sanaton kuvakanava puolestaan viittaa yleiseen kuvan tai videon asetteluun. (Gottlieb 1997, 89.)

Av-kääntämisessä esiintyvät kaikki Roman Jakobsonin (2007) määrittelemät kääntämisen muodot: kieltenvälinen, kielensisäinen ja intersemioottinen. Kieltenvälisessä kääntämisessä välitetään tietoa lähdekieleltä toiselle kohdekielelle, kun taas kielensisäisessä kääntämisessä lähde- ja kohdekieli pysyvät samana. (mp., 182.) Kielensisäisen kääntämisen statuksesta kääntämisenä on kiistelty. Se voidaan nähdä enemmän lähdetekstin litteroimisena sekä editoimisena, sillä siinä ei tapahdu siirtoa kieleltä toiselle. (Pedersen 2011, 9–10.) Nykyään kuitenkin yleinen mielipide on, että kielensisäinen kääntäminen kuuluu myös kääntämisen termin alle (Zárate 2021, 4). Intersemioottinen tai -modaalinen kääntäminen puolestaan on lähdetekstin siirtoa vastaanotettavaan muotoon toiselle aistikanavalle (Hirvonen, Kinnunen ja Tiittula 2020, 22–23).

Englanniksi *accessibility*-sanalla voidaan viitata niin saavutettavuuteen kuin esteettömyyteen. Suomen kielessä esteettömyydellä viitataan pitkälti fyysisten tilojen ja palvelujen saavutettavuuteen, jolloin se lukeutuu saavutettavuuden alakäsitteeksi. Tämän lisäksi saavutettavuuden termin alla ovat käytettävyys ja ymmärrettävyys. (Hirvonen, Kinnunen ja Tiittula 2020, 19.) Saavutettavien palveluiden kohderyhmänä ovat erityisesti toimintarajoitteiset käyttäjät. Liikenne- ja viestintäviraston toimenpideohjelman mukaan *toimintarajoite*-termi kattaa muun muassa kuulo- sekä näkövammaisuuden, kognitiiviset ja tiedolliset rajoitteet ja kieli- sekä kulttuurirajoitteet. Tämän lisäksi toimintarajoitteet voivat olla pysyviä tai väliaikaisia. (mp., 16.)

Av-teosten saavutettavuus lukeutuu muun muassa viestinnän saavutettavuuden termin alle. Viestinnän saavutettavuuden tavoitteena on, ettei viestintä muodosta estettä tai sulje ulkopuolelle (Hirvonen, Kinnunen ja Tiittula 2020, 18). Tämän lisäksi av-teoksien saavutettavuus on myös mediasaavutettavuutta, mikä on käännöstieteellisessä tutkimuksessa laajentanut kääntämisen käsitettä kattamaan verbaalisen viestin lisäksi äänen ja kuvan kääntämisen sanoiksi (mp., 25). Av-teokset ovat usein saatavilla verkossa, jolloin niissä täytyy huomioida myös verkkosisällön saavutettavuusohjeet (*Web Content Accessibility Guidelines*). Verkkosisällön saavutettavuusohjeet tarjoavat kattavat suositukset muun muassa videosisällön tekstittämisestä sekä kuvailutulkkauksesta riittävän saavutettavuuden takaamiseksi (WCAG 2.2 2023).

Käännösalalla saavutettavuudesta on tehty tutkimuksia jo 1980-luvulta lähtien ja 2000-luvulla saavutettavuus saavutti erikoisalan statuksen erityisesti av-kääntämisen kasvun myötä (Hirvonen, Kinnunen ja Tiittula 2020, 22). Av-kääntämisessä saavutettavuuden piirteiksi on määritelty hyväksyttävyys, luettavuus, synkronia, relevanssi ja kotouttaminen (Gambier 2004, 9). Hyväksyttävyys on kielen luontevuutta ja oikeellisuutta. Yves Gambierin (2004) määritelmässä luettavuus kattaa ainoastaan visuaalisen luettavuuden (*legibility*), jolla tarkastellaan muun muassa tekstitysten asettelua ruudulla, typografiaa ja kestoa. Luettavuuteen (*readability*) voidaan kuitenkin myös sisältää sisällön ymmärrettävyys tai miellettävyys, joilla arvioidaan tekstin sisäistämisen vaivattomuutta. Synkronia on esimerkiksi dubbauksessa kuvan ja äänen ajallista vastaavuutta, kun taas tekstittämisessä vaaditaan kuvan, äänen ja tekstin ajallista vastaavuutta. Relevanssilla puolestaan viitataan siihen, kuinka paljon lähdetekstiä säilytetään tai tiivistetään kontekstin mukaan. Kotouttaminen on yleinen käännösstrategia, jolla lähdekielen kulttuurisidonnaisia piirteitä yleensä kotoutetaan kohdekulttuuriin ja -yleisölle tyypillisempään muotoon. (mp., 9.)

2.1 Ohjelmatekstitys

Audiovisuaalisen kääntämisen yleisimmät muodot ovat dubbaus, selostus ja tekstitys (Díaz Cintas 2003, 195). Näistä kääntämisen muodoista on kehittynyt kaksi av-viestinnän saavutettavuuspalvelua: kuvailutulkkaus ja ohjelmatekstitys (Hirvonen ja Tiittula 2020, 73). Kuvailutulkkaus (*audio description*) on suunnattu erityisesti sokeille ja heikkonäköisille henkilöille, kun taas ohjelmatekstitykset (*subtitling for the deaf and hard-of-hearing*, *SDH* tai *closed captioning*) palvelevat ensisijaisesti kuuroja ja huonokuuloisia katsojia. Tämän lisäksi muun muassa Yleisradio Oy (Yle) tarjoaa av-viittomatulkkausta (*signed audiovisual interpreting*) viittomakielisille katsojilleen (Holopainen 2015, 81). Saavutettavuuspalveluiden tavoitteena on

tukea kohderyhmän katselukokemusta tarjoamalla heille riittävästi tietoa av-teoksen joko visuaalisesta tai äänellisestä sisällöstä.

Ohjelmatekstittäminen on ollut pitkälti kielensisäistä, mutta nykyään sitä esiintyy enenevissä määrin myös kieltenvälisenä. Ohjelmatekstityksessä tekstitetään puhuttujen vuorosanojen lisäksi muita ääniraidassa esiintyviä ääniä. Katselukokemuksen kannalta olennaisia tekijöitä voivat olla esimerkiksi musiikin, onomatopoeettisten äänien tai paralingvististen eli parakielisten ominaisuuksien, kuten äänensävyn tai murteen, tekstittäminen. (Zárate 2021, 5–6.) Kaikkia ääniraidan ääniä ei kuitenkaan tarvitse tekstittää, vaan tekstittäjän täytyy tietää mikä funktio milläkin äänellä on, jotta hän pystyy välittämään tekstityksessä riittävästi tietoa tekstitysten ensisijaiselle kohderyhmälle aika- ja merkkirajoitteiden puitteissa (Tiittula 2012, 9).

Täten ohjelmatekstittämisessä korostuu kohderyhmän kattava tuntemus. Kuten Hirvonen ja Tiittula (2020, 77–78) toteavat, ohjelmatekstitysten kohderyhmän jäsenillä voi olla toisistaan poikkeavia tarpeita sekä mahdollisuuksia, jotka täytyy huomioida tekstitysten laatimisessa. Esimerkiksi syntymäkuurojen äidinkieli on viittomakieli, toisin kuin myöhemmin kuuroutuneilla äidinkieli on yleensä puhuttu kieli. He eivät pysty vastaanottamaan auditiivista informaatiota, mutta voivat käyttää huulilta lukemista katselukokemuksen tukemisessa. Huonokuuloiset puolestaan pystyvät vaihtelevissa määrin kuulemaan myös ääntä ja kieltenoppijat todennäköisesti kuulevat äänen, mutta heillä voi olla vaikeuksia puhutun kielen ymmärtämisessä. (mp.)

Ohjelmatekstitysten tekstitystapa on maakohtaista ja normit voivat vaihdella huomattavasti eri maiden välillä. Esimerkiksi anglosaksisissa maissa usein pyritään sananmukaiseen (*verbatim*) tekstittämiseen, kun taas muun muassa Suomen kaltaisissa perinteisissä tekstitysmaissa suositaan tiivistävää tai viestinnällistä tekstitystapaa (Holopainen 2015, 80–81). Ohjelmatekstityksissä sananmukaista tekstitysperinnettä on perusteltu muun muassa huulilta lukemisen sekä sensuroinnin vähyyden kannalta. Holopainen (mp., 81) kuitenkin toteaa, että Suomessa on yhä seurattu ilmaisua tiivistävää tekstitystapaa eikä ohjelmatekstitysten kohderyhmä ole sitä huomattavasti kritisoinut. Tällöin voinee arvioida, että kohderyhmä luottaa riittävästi viestinnällisen tekstitystavan kykyyn tukea katselukokemusta ja siten välittää heille olennainen tieto ohjelman sisällöstä (mp.).

2.1.1 Ohjelmatekstitysten kehitys Suomessa

Ohjelmatekstityksiä on Suomessa käytetty jo vuodesta 1983 Ylen tarjoamana ja vuoteen 2013 mennessä ohjelmatekstitysten määrä kattoi alle puolet suomenkielisestä tarjonnasta (Tiittula & Rainò 2013, 65). Saavutettavuuden ja esteettömyyden saatavuutta on pyritty sääntelemään eri säännöksillä, jotka ovat pitkälti Traficomin valvonnan alla (Traficom 2024). Muun muassa laki sähköisen viestinnän palveluista (917/2014) asettaa vaatimuksen siitä, että 28.6.2025 alkaen viestintäpalveluissa täytyy olla tarjolla reaaliaikainen tekstitysominaisuus puheviestinnän lisäksi (*Laki sähköisen viestinnän palveluista* 917/2014, 194 c §).

Vuonna 2019 Suomessa astui voimaan digipalvelulaki, joka perustuu muun muassa Euroopan unionin saavutettavuus- ja esteettömyysdirektiiveihin sekä verkkosisällön saavutettavuusohjeisiin (*Laki digitaalisten palvelujen tarjoamisesta* 306/2019). Digipalvelulaki säätelee muun muassa sitä, että julkisen tahon videoissa täytyy olla tekstitykset ohjelman alkuperäiskielellä ja niiden täytyy saavuttaa riittävä laatu (Traficom 2023a). Tv-yhtiöistä tekstitysvelvollisuus koskee niin Ylen tv-kanavia, MTV3- kuin Nelonen-kanavia. Ylen kaikessa suomen- ja ruotsinkielisessä tarjonnassa täytyy olla ohjelmatekstitys, kun taas MTV- ja Nelonen-kanavilla määrä on 75 prosenttia kotimaisista ohjelmista. Suoratoistopalveluissa kuten Yle Areenassa, MTV Katsomo- ja Ruutu-palveluissa tekstitysvelvoite koskee vain 30 prosenttia siitä ohjelmistosta, joka myös lineaaritelevisiossa sisältää ääni- ja tekstityspalvelun. (Traficom 2021, 1.) Vaikka saavutettavan sisällön saatavuutta pyritään säätelemään lakisääteisesti, liikenne- ja viestintäviraston teettämän tutkimuksen mukaan ohjelmatekstitysten saavutettavuudessa on yhä kehitettävää (Vesanen-Nikitin ym. 2022).

2.1.2 Koneellistettu ammattimainen ohjelmatekstittäminen

Liikenne- ja viestintäviraston ohjeistuksessa audiovisuaalisten palvelujen esteettömyydestä ei sinänsä oteta kantaa tekstitysten laatimistapaan, vaan ohjeistuksen mukaan tekstitykset voidaan laatia joko täysin ihmistyövoimalla tai käyttäen automaattista puheentunnistustekniikkaa tukena (Traficom 2021, 6). On kuitenkin huomioitava, että liikenne- ja viestintäviraston näkemyksen mukaan täysin automaattisesti laadittujen tekstitysten laatu ei todennäköisesti saavuta asetettuja kriteereitä, jolloin ihmisen täytyisi tarkistaa sekä täydentää automaattisia tekstityksiä (mp.).

Vielä vuonna 2020 Yle tekstitti niin live- kuin ohjelmatekstitykset pitkälti manuaalisesti (Hirvonen ja Tiittula 2020, 79), mutta vuoteen 2024 mennessä kieliteknologiaa on otettu

enemmän mukaan tekstitysprosessiin (Virve Tossavainen sähköpostitse 4.11.2024). Muun muassa litterointityötä on automatisoitu ja live-tekstitysten laatimisessa on ainakin osin siirrytty sanelutekstittämiseen (*respeaking*) (mp.). Sanelutekstitys on puoliautomaattinen puheentunnistusmenetelmä, jossa tekstittäjä toistaa ääniraidan puheen puheentunnistinohjelmaan ja korjaa sitten puheentunnistimen tekemät virheet (Hirvonen ja Tiittula 2020, 79). Ylellä ohjelmatekstityksiä laaditaan useilla eri tavoilla, joita ovat esimerkiksi perinteinen tekstittäminen, jossa tekstittäjä itse kuuntelee ja kirjoittaa tai sanelee ääniraidan puhetta, tai tekstitys voidaan laatia ohjelman käsikirjoituksen tai automaattisen puheentunnistimen litteroinnin pohjalta. Ylen ohjelmatekstityksen esihenkilö Virve Tossavainen kuitenkin toteaa, että ohjelmatekstitysten laatimiseen valittu tapa on useimmiten ohjelmakohtainen. Vaikka Ylen tekstityksissä ei vielä hyödynnetä huomattavasti automaattista puheentunnistusta, se koetaan hyödylliseksi osaksi tekstittämisen työkaluista ja todennäköisesti myös tekoälyn käyttö apuvälineenä lisääntyy tulevaisuudessa. (Virve Tossavainen sähköpostitse 4.11.2024.)

Kaupallisista kanavista muun muassa Sanoma Media Finland Oy:n Nelonen-kanavalla ohjelmatekstitykset laaditaan automaattisella puheentunnistusmenetelmällä (Hirvonen ja Tiittula 2020, 104). Tämä on herättänyt kritiikkiä niin katsojien kuin liikenne- ja viestintävirasto Traficomin puolelta (Hirvonen ja Tiittula 2020; Traficom 2022). Vuosina 2022 ja 2023 Traficom vaati Nelonen-kanavan ohjelmatekstitysten laadun kehittämistä, minkä myötä Sanoma Media Finland Oy otti käyttöönsä kehittyneemmän tekoälyyn pohjautuvan tekstitysohjelmiston (Traficom 2023b). Tämän lisäksi yhtiö muutti ohjelmatekstitysprosessiaan siten, että uusissa ohjelmissa ihminen tarkistaa sekä korjaa kaikki tekoälyllä luodut tekstitykset ja muutamia katsotuimpia ohjelmia tekstitetään täysin manuaalisesti. Vuoden 2023 lopulla Traficom totesi, että kanavan ohjelmatekstitykset saavuttavat laadultaan riittävän tason muun muassa selkeydessä ja ymmärrettävyydessä. (Traficom 2023a.)

2.2 Laadunarviointi käännöstieteessä

Pelkkä tarjonnan määrä ja saatavuus eivät riitä takaamaan saavutettavuutta, vaan siihen vaikuttaa myös esimerkiksi ohjelmatekstityksen käytettävyys sekä laatu (esim. Tiittula ja Rainò 2013; Koponen ja Nurminen 2020). Jälkimmäisellä viitataan siihen, miten hyvin tekstitys kykenee välittämään kuulovammaiselle katsojalle samankaltaisen katselukokemuksen kuin kuulevalle katsojalle. Niin kieltenvälisille kuin -sisäisille käännöksille on erilaisia laadunarviointimenetelmiä, jotta tarjolla olisi riittävästi muun muassa laadukkaita tekstityksiä. Seuraavaksi esittelen tässä tutkielmassa sovellettuja laadunarviointimenetelmiä ja -suosituksia.

2.2.1 ISO5060:2024: Käännösten laadunarviointi

ISO 5060 -standardissa esitetään virheluokittelu käytettäväksi käännösten laadun arviointiin (ISO 5060:2024, 1). Standardissa on määritelty seitsemän virhekategoriaa: terminologia, sisällön välittyminen, kohdekieli, tyyli, kohdekielen ja -kulttuurin konventiot, toimeksiannon huomioon ottaminen ja ulkoasu (Leena Salmen käännös). Terminologiassa huomioidaan termien oikeellisuus sekä johdonmukaisuus. Sisällön välittymisen arvioinnissa puolestaan esiintyvät muun muassa merkitysvirhe, poisjättö ja lisäys. Kolmannessa virhekategoriassa tarkastelun kohteena ovat kohdekielen kieliopilliset tekijät kuten morfologia, oikeinkirjoitus sekä syntaksi. Kohdekielen kategoriassa huomioidaan myös käännöksen ymmärrettävyys. Tyylin osalta puolestaan arvioidaan muun muassa käännöksen idiomaattisuutta, kun taas kohdekielen ja -kulttuurin konventioissa huomioidaan maakohtaiset standardit esimerkiksi mittayksiköiden käytössä. Tämän lisäksi tähän kategoriaan sisältyy merkistö. Viimeisinä kategorioina ovat toimeksiannon huomioonottaminen, jossa tarkastellaan esimerkiksi reaalioita, ja ulkoasu, jossa huomioidaan teknisiä tekijöitä kuten muotoilua sekä asettelua. (mp., 7–9.)

2.2.2 Suomalaiset ohjelmatekstitysten laatusuositukset

Av-kääntämisessä laatusuositukset ovat pitkälti keskittyneet ainoastaan kieltenväliseen ruututekstittämiseen. Tämän lisäksi audiovisuaalisella käännösalalla on useita kotimaisia sekä ulkomaisia toimijoita, joilla usein on omat tavat sekä yhtiön sisäiset ohjekirjat tekstittämiskonventioista (esim. Lång 2013). Vuonna 2020 suomalaisista käännösalan toimijoista muodostunut tekstitystyöryhmä laati ohjelmatekstitysten laatusuositukset, jotka ”koskevat ensisijaisesti televisiossa, suoratoistopalveluissa ja tallenteissa käytettäviä ennakkoon tekstitettäviä av-teoksia” (*Ohjelmatekstitysten laatusuositukset* 2020, 4–5). Laatusuosituksia voidaan myös soveltaa osin niin live-tekstityksissä, elokuvatekstityksissä kuin verkkosivuilla tai sosiaalisessa mediassa julkaistavien videoiden tekstityksissä (mp.). Luvussa 2.1.1 esitellyistä tv-yhtiöistä Yle ja MTV ovat myös olleet mukana laatimassa laatusuosituksia sekä ovat allekirjoittaneet ohjeistukset (Traficom 2021, 6). Laatusuositusten tavoitteena on muun muassa yhtenäistää tekstityskäytänteitä, toimia ohjeistuksena tekstittäjille ja tukea tekstitysten laadun ylläpitoa. Nämä laatusuositukset perustuvat Jan Pedersenin (2017) kehittämään FAR-mallinnukseen. (*Ohjelmatekstitysten laatusuositukset* 2020, 4–5.) Liikenne- ja viestintävirasto suosittelee erityisesti alan toimijoita seuraamaan laatusuosituksia tekstitysten laatimisessa, sillä ne tarjoavat hyvän pohjan laatukriteereille sekä laadun arvioimiseen (Traficom 2021, 6).

Ohjelmatekstitysten laatusuositukset on jaettu eri kategorioihin, jotka ovat hyväksyttävyys, luettavuus sekä miellettävyys ja ilmaisun käytänteet (*Ohjelmatekstitysten laatusuositukset* 2020). Hyväksyttävyys on kohdekielelle ominaisen kielen ilmaisua. Laatusuosituksissa ehdotetaan, että hyväksyttävyyden lähtökohtana olisi yleiskielen suositusten mukainen virheetön ja luonteva kieli. Tekstityksen täytyisi myös huomioida esimerkiksi kohdeteoksen genre, jotta tekstityksessä käytetty kieli sopisi ohjelman tyyliin. Tämän lisäksi ohjelmatekstityksen kielen tulisi myötäillä puhetta, mutta tiivistämistä sekä selventämistä on suositeltavaa tehdä tarvittaessa. (*Ohjelmatekstitysten laatusuositukset* 2020, 9.)

Luettavuuden ja miellettävyyden huomioinnin tavoitteena on, että ohjelmatekstitysten seuraaminen ja välitön sisäistäminen olisi katsojalle vaivatonta. Nämä kategoriat sisältävät muun muassa kirjoitusteknisiä suosituksia esimerkiksi tekstitysten typografiasta ja aika-, merkki- sekä rivirajoituksista (*Ohjelmatekstitysten laatusuositukset* 2020, 11). Repliikin eli yksittäisen ruututekstin (mp., 47) tulisi olla joko yksi- tai kaksirivinen, tosin tilanteen niin vaatiessa myös kolmirivinen repliikki on hyväksyttävä (mp., 11).

Luettavuutta tuetaan myös tekstitysten jaottelulla, joka on niin repliikin sisäistä kuin repliikkien välistä jaottelua (*Ohjelmatekstitysten laatusuositukset* 2020, 11). Jaottelun ensisijaisena tarkoituksena on varmistaa repliikkien vaivaton hahmottaminen sekä sisäistäminen sisällyttämällä niihin ”tarkoitukseltaan, aiheeltaan ja rakenteeltaan yhteenkuuluvia asioita” (mp.) Tämän lisäksi repliikissä saisi olla korkeintaan kaksi virkettä ja enintään kahden puhujan vuorosanat, sillä tekstityksen luettavuus sekä miellettävyys voivat muutoin kärsiä (mp.). Jos virke jatkuu seuraavassa repliikissä, tulisi se ilmaista yhdysmerkillä eli ruututekstittämisessä vuorosanaviivalla ensimmäisen repliikin lopussa. Suositus myös on, etteivät virkkeet olisi yli kolmen repliikin pituisia. (mp., 18.)

Tämän lisäksi repliikkien kestolla on huomattava vaikutus tekstitysten luettavuuteen. Repliikin vähimmäiskesto on 1,8 sekuntia ja enimmäiskesto noin 7 sekuntia. Kestossa tulee myös huomioida lukunopeus, joka suomenkielisissä ohjelmatekstityksissä on 10–12 merkkiä sekunnissa (CPS) välilyönnit mukaan lukien. (*Ohjelmatekstitysten laatusuositukset* 2020, 15–16.) Riittävän lukunopeuden lisäksi tekstityksen tulisi olla synkroniassa äänen kanssa, sillä erityisesti huonokuuloiset katsojat voivat tukeutua myös huuliltalukemiseen (Pereira 2010, 91). Repliikkiä ei tulisi jättää ruutuun kuvaleikkauksen yli kohtauksen vaihtuessa. Jos kuitenkin tämä on sisällön ymmärrettävyyden kannalta välttämätöntä, tulee repliikin pysyä ruudulla vähintään sekunnin ajan leikkauksen jälkeen. Kohtauksen sisällä repliikki voi jäädä kuvaan

kuvaleikkauksen yli, mutta on suositeltavaa, että repliikki poistuu tai tulee kuvaan leikkauksen kohdalla. (*Ohjelmatekstitysten laatusuositukset 2020*, 16–17.)

Luettavuuden ja miellettävyyden kategoriaan sisältyy myös puhujan ja sen vaihtumisen merkitsemisen käytännöt. Puhujien erottelemiseksi voidaan käyttää värejä, vuorosanaviivaa, infokylttiä ja ajoittain myös kursivointia. Tekstitysten perusväri on valkoinen, mutta kansainvälisten käytänteiden mukaisesti puhujia voidaan erotella keltaisella, syaaninsinisellä sekä vihreällä. Värien käyttö on myös ohjelmakohtaista. (*Ohjelmatekstitysten laatusuositukset* 2020, 19.)

Ilmaisun käytänteillä tuetaan osaltaan tekstityksen luettavuutta ja kielen hyväksyttävyyttä. Ilmaisun käytänteisiin lukeutuu muun muassa sidoksisuuskeinot, kuten koherenssi, koheesio, teema-reema -suhteet ja painotus. Tämän lisäksi ilmaisun käytänteissä huomioidaan esimerkiksi murteellisuus, musiikki ja kirosanat. (*Ohjelmatekstitysten laatusuositukset* 2020, 31.)

2.2.3 NER-malli

NER-mallilla (Romero-Fresco ja Martínez 2015) mitataan erityisesti sanelutekstitysten sisällön välittymistä (*accuracy*) live-tekstityksissä. Vaikka Pablo Romero-Frescon ja Juan Martínezin tutkimus keskittyy sanelutekstitykseen ja live-tekstitysten sisällön välittymisen arvioimiseen, NER-mallia voi hyödyntää myös automaattisen puheentunnistuksen laatiman tekstityksen laadun arvioimisessa (mp., 11). Malli perustuu WER (*word error rate*) -menetelmään, joka on yleinen laskujärjestelmä kieli- ja käännösteknologian tutkimuksessa (Romero-Fresco ja Martínez 2015; Dumouchel ym. 2011).

NER-malli on seuraavanlainen:

Sisällön välittymisen % $= \frac{N-E-R}{N} x100$

Hyväksyttävät muokkaukset (CE): [hyväksyttävien muokkauksien summa]

Arviointi: [kirjallinen arviointi]

NER-mallin mukaisessa laskukaavassa tekstitysten sisällön välittymisen täytyisi olla vähintään 98 prosenttia, jotta ne olisivat riittävän laadukkaita (Romero-Fresco ja Martínez 2015, 4). Mallissa N kuvapuheentunnistimen tunnistamien sanojen määrää. E viittaa niihin muokkausvirheisiin (*edition errors*), joissa annettu tieto voi olla puutteellista tai virheellistä. Nämä virheet muodostuvat esimerkiksi sanelutekstittäjän tiivistäessä puhetta, mikä voi johtaa

olennaisen tiedon katoamiseen tai virheelliseen tulkintaan. Automaattisen puheentunnistimen kautta laadituissa tekstityksissä muokkausvirheet olisivat muun muassa ison alkukirjaimen, pilkkujen tai puhujan tunnistuksen virheellinen käyttö. R puolestaan tarkoittaa tunnistusvirheitä (*recognition errors*), jotka voivat olla esimerkiksi sanelutekstittäjän ääntämisestä tai käytetystä järjestelmästä muodostuneita virheitä. (mp.) Nämä virheet sisältävät muun muassa poisjätön, merkitysvirheen sekä lisäyksen. Poisjätössä puheentunnistin jättää osan puhutusta lauseesta litteroimatta. Merkitysvirheessä puolestaan puhutun lauseen oikea sana on korvaantunut jollain toisella sanalla ja lisäyksessä transkriptiossa on ylimääräinen sana, jota ei ole alkuperäisessä puhutussa lauseessa. (Dumouchel ym. 2011, 167.)

Niin muokkaus- kuin tunnistusvirheet pisteytetään vakavuutensa perusteella joko 0,25, 0,5 tai 1. Lievät (*minor*) 0,25 pisteen arvoiset virheet ovat virheitä, jotka vielä mahdollistavat alkuperäisen viestin vaivattoman seuraamisen eikä katsoja välttämättä huomaa virhettä. Näitä virheitä voivat olla muun muassa isojen kirjainten tai välimerkkien esiintyminen tuotetussa tekstissä sekä yksittäisten sanojen lisääminen tai poisto. Tämän lisäksi lieväksi virheeksi katsotaan esimerkiksi lauseen määritteiden poisto, joka ei muuta jäljelle jäävän lauseen merkitystä tai tee sen sisällöstä merkityksetöntä. 0,5 pisteen arvoiset keskivertoiset (*standard*) eli keskivertovirheet puolestaan häiritsevät alkuperäisen tekstin seuraamista. Näissä virheissä usein esiintyy muun muassa joko virkkeiden poistoa tai määritteiden poistoa siten, että ne vaikuttavat viestin ymmärrykseen. Vakavat (*serious*) eli 1 pisteen arvoiset virheet muuttavat alkuperäisen tekstin merkitystä siten, että uusi teksti sopii vielä kontekstiin. Vakavat virheet poistavat ja väärentävät alkuperäisen tekstin merkityksiä, mutta katsoja ei välttämättä ole tietoinen tästä. (Romero-Fresco ja Martínez 2015, 5–7.)

Hyväksyttävät muokkaukset (*correct editions*) ovat sellaisia muutoksia, jotka eivät aiheuta puutteellista tai virheellistä tekstitystä. Tällöin esimerkiksi äännähdysten tai toiston poisjättö voidaan tilannekohtaisesti arvioida hyväksyttäväksi muokkaukseksi. NER-mallissa arviointi (*assessment*) on tekstitysten laadun arvioinnin kannalta merkittävin osuus, sillä laadun riittävyyttä on hankala kuvata ainoastaan määrällisesti. Tässä vaiheessa huomioidaan muun muassa hyväksyttävyyden ja ymmärrettävyyden piirteitä kuten tekstityksen koheesiota. (Romero-Fresco ja Martínez 2015, 4.)

2.3 Kieli- ja käännösteknologia edistämässä saavutettavuutta

Kieliteknologia on ”ihmisen kieleen kohdistuvaa tai sitä hyväksi käyttävää teknologiaa” (Salmi 2015, 99), jonka alalajina on muun muassa käännösprosessin tukena käytetty

käännösteknologia. Käännösteknologian hyödyntämistä voidaan nimittää myös koneavusteiseksi kääntämiseksi (mp., 99–100). Kieli- ja käännösteknologiaa on nykyisin helpommin saatavilla kaikille käyttäjille esimerkiksi internetissä ja mobiililaitteissa (mp., 109).

Merkittävämpiä kieli- ja käännösteknologian tuotteita ovat koneavusteiset käännösohjelmat, jotka voivat sisältää niin konekääntimiä kuin käännösmuisteja sekä termipankkeja. Konekääntimet voivat olla sääntöpohjaisia, tilastollisia tai neuroverkkokääntimiä. Konekääntimiä kehitettiin aluksi siinä toivossa, että ne automatisoisivat käännösprosessin, mutta konekääntimien sijaan siirryttiin pian kehittämään koneavusteisia käännösohjelmia. (Rothwell ym. 2023.) Palveluntuottajille sekä -tilaajille konekäännösten houkuttelevuutta on lisännyt niiden kustannustehokkuus sekä nopeampi saatavuus (Bywood ym. 2017, 494). Konekäännöksillä voitaisiin myös mahdollistaa tiedon saavutettavuutta (Koponen ja Nurminen 2020, 307). Av-kääntämisen saralla esimerkiksi ohjelmatekstityksen ja kuvailutulkkauksen tavoitteena on olla ”luotettava ja neutraali” (Hirvonen ja Tiittula 2020, 107), mikä voi osoittautua konekäännösten eduksi, sillä usein niitä pidetään ihmiskääntäjää objektiivisempana vaihtoehtona. Toisaalta konekääntimien toiminta perustuu ihmisten tuottamaan dataan, jolloin väittämä konekääntimien puolueettomuudesta ja neutraaliudesta on kyseenalaistettavissa. (mp.)

Puheentunnistimet ovat osa puheteknologiaa ja sitä hyödyntäviä ohjelmia löytyy nykyisin helposti esimerkiksi mobiililaitteista (Salmi 2015, 101). Puheentunnistimet ovat ohjelmia, joiden tehtävänä on muuttaa puhe tekstiksi. Ne voivat olla puhujasta riippuvia tai puhujasta riippumattomia. Puhujasta riippuvat puheentunnistimet ladataan yksittäisille laitteille ja sen voi säätää käyttäjän puheeseen sekä tarpeisiin sopivaksi. Puhujasta riippumattomat puheentunnistusohjelmat puolestaan voivat olla joko laitekohtaisia tai pilvipalvelupohjaisia eikä näitä puheentunnistimia ei voi räätälöidä yksittäisille käyttäjille. (Ciobanu & Secară 2019, 92.)

2.3.1 Generatiivinen tekoäly

Generatiiviset tekoälysovellukset luovat synteettisesti esimerkiksi tekstiä, kuvia, videoita ja ääntä niille annettujen kehotteiden perusteella (Lorenz ym. 2023, 8). Kehotteet (*prompt*) ovat useimmiten ihmisen tuottamia kirjallisia ohjeita ja toisinaan niiden tueksi voidaan liittää myös dataa. Toisin kuin monet tekoälypohjaiset sovellukset, generatiivisten tekoälysovellusten käyttö ei yleensä vaadi teknisiä taitoja, vaan niitä voi käyttää luonnollisen kielen kehotteilla. (*Patent Landscape Report* 2024, 19.) Monet generatiiviset tekoälysovellukset toimivat suurien kielimallien (LLM) pohjalta. Suuret kielimallit mahdollistavat tekstin tuoton, joka mukailee ihmisen tuottamaa tekstiä. Ne toimivat tilastollisesti, jolloin ne päättelevät suurimman

todennäköisyyden perusteella kulloisessakin tilanteessa seuraavan sanan tai väittämän. (Lorenz ym. 2023, 14.) Suuria kielimalleja hyödynnetään monissa luonnollisen kielen prosessoinnin (NLP) tehtävissä kuten tekstin täytössä, konekäännöksissä ja tiivistelmissä (*Patent Landscape Report* 2024, 22–24).

Suuret kielimallit ovat kuitenkin rajoittuneet käsittelemään ainoastaan tekstiä, minkä myötä kiinnostus on kohdistunut suurten multimodaalisten kielimallien (MLLM) kehittämiseen. Suuret multimodaaliset kielimallit koostuvat useimmiten tekstisyötettä käsittelevästä suuresta kielimallista, muuta syötettä käsittelevästä kooderista ja modaliteettiprojektorista (*modality projector*), joka yhdistää tekstisyötteen ja muut syötteet yhtenäiseksi kokonaisuudeksi. (Huang & Zhang 2024, 1–2.) Suuret multimodaaliset kielimallit voisivat täten tarjota uusia mahdollisuuksia myös av-kääntämiseen kohdistetun kieliteknologian kehityksessä.

Vaikka generatiivinen tekoäly mahdollistaa monenlaisten prosessien koneellistamisen, on myös huomioitava sen eettisyyttä erityisesti kestävän kehityksen näkökulmasta. Kieli- ja käännöstieteen saralla esimerkiksi tekoälypohjaisten neuroverkkokääntimien kouluttaminen kuormittaa ympäristöä huomattavasti. (Moorkens ja Rocchi 2020, 330.) Tulevaisuudessa on syytä arvioida mihin tarkoituksiin generatiivista tekoälyä kannattaisi hyödyntää ilman, että se kuluttaisi turhaan luonnonvaroja.

2.3.2 Koneavusteinen av-kääntäminen

Av-kääntämisen saralla kieli- ja käännösteknologia kohtaavat haasteita. Yleensä työkalut kykenevät huomioimaan ääniraidalta tai videosta vain puheen, mutta eivät visuaalisia tekijöitä ja muita ääniä, jotka kuitenkin ovat merkittäviä osia av-teoksen tulkintaa. Tämän lisäksi kieli- ja käännösteknologian työkalut eivät välttämättä ole yhtä aikaa säästäviä tai kustannustehokkaita kuten herkästi väitetään, sillä useimmiten ihmisen täytyy esimerkiksi litteroida sekä ajastaa lähdeteksti ennen kuin siitä voidaan laatia konekäännös. (Karakanta 2022, 90.) Aiemmat tutkimukset koneavusteisesta av-kääntämisestä ovat keskittyneet pitkälti käyttäjäkeskeisiin kokemuksiin niin tekstitysten vastaanottajan kuin tekijän näkökulmasta. Ohjelmatekstitysten vastaanottoa on tutkittu enimmäkseen kysely- ja silmänliiketutkimuksilla, joista jälkimmäisissä on analysoitu muun muassa ohjelmatekstitysten luettavuutta ja ymmärrettävyyttä (Hirvonen ja Tiittula 2020, 78). Tämän lisäksi mielenkiinnon kohteena on ollut erityisesti konekäännösten jälkieditointi ja siihen liittyvät kysymykset (esim. Koponen, Sulubacak ym. 2020)

Teknologian hyödyntämistä av-kääntämisprosessissa on jo 1990-luvulta lähtien tarkasteltu automatisoinnin näkökulmasta ja erityisesti neuroverkkokääntimien sekä automaattisten puheentunnistusohjelmien kehittäminen on ollut tutkimusten keskiössä (Tuominen ym. 2023, 78–79). Esimerkiksi 2000-luvun alussa oli hankkeita, joissa pyrittiin automatisoimaan käännöstekstitysprosessia. Näistä yksi on vuosina 2002–2004 toteutettu MUSA-projekti (*Multilingual Subtitling of multimedia content*). Projektin tavoitteena oli hyödyntää puheentunnistinta, tekstianalyysia ja konekäännintä monikielisten tekstitysten automatisoidussa laatimisessa. Vuosina 2003–2004 toteutetussa eTITLE-projektissa puolestaan pyrittiin luomaan koneavusteinen käännösohjelma, joka hyödyntäisi muun muassa puheentunnistinta, tiivistämistä, käännösmuistia sekä konekäännintä tekstittäjien työskentelyn apuvälineenä. (Bywood ym. 2017, 495.) Viime vuosina erityisesti MeMAD-hankkeella (*Methods for Managing Audiovisual Data*) on ollut merkittävä osa konekääntämisen ja audiovisuaalisen sisällön saavutettavuuden edistämisessä. Hankkeessa esimerkiksi kehitettiin automaattista videokuvailua, tekstitystä sekä puheentunnistusta ihmisen tuottamista käännöksistä oppien (Hirvonen ja Tiittula 2020, 105).

Puheentunnistimen luomia mahdollisuuksia on tarkasteltu erityisesti kielensisäisessä tekstittämisessä (Koponen, Tuominen ym. 2020, 57). Muun muassa vastaanottotutkimukset sanelutekstityksestä ovat osoittaneet, että kuurot ja huonokuuloiset katsojat voisivat seurata konekäännettyjä tai koneavusteisesti laadittuja tekstityksiä, jos ne olisivat riittävän laadukkaita (Vitikainen ja Koponen 2021, 45). Ohjelmatekstitysten sekä kuvailutulkkauksen automatisointiprosessien kehittäminen entisestään olisi tärkeää, sillä manuaalisesti tehtyinä ne ovat aikaa vieviä prosesseja. Tutkimukset kuitenkin osoittavat, että täysin automatisoidusti luotujen ohjelmatekstitysten vastaanotto on ollut nihkeää. (Hirvonen ja Tiittula 2020, 74.) Yhtenä syynä tähän on muun muassa automaattisten puheentunnistimien vaikeudet suomenkielisen puhekielen virheettömässä tunnistamisessa (mp., 79). Konekäännöksien laatua voidaan arvioida käsitteellä ”riittävän hyvä” (mp., 107), joka tässä tilanteessa tarkoittaa ainoastaan sitä, ettei käännös estä viestintää. Tämän lisäksi huomioon otetaan konekäännöksen tekemän virheen vakavuus ja se, kuinka paljon haittaa se aiheuttaa viestin ymmärtämisessä. (mp.)

3 Aineisto ja menetelmät

3.1 Aineisto

Keräsin aineistoni kolmesta tyyliltään erilaisesta ohjelmasta: uutis-, keskustelu- ja dokumenttiohjelmasta. Ohjelmista täytyi olla saatavilla ohjelmatekstitys ja niiden täytyi olla vapaasti saatavilla verkossa suoratoistopalvelussa. Ohjelmien täytyi myös olla riittävän erilaisia sisällöltään, jotta ne tarjoaisivat eri haasteita tekstitysten laatimiseen. Tutkielmassa tarkastelluista ohjelmista kaksi löytyivät tutkielman teon aikana Yle Areena -suoratoistopalvelusta, mutta yksi ohjelma oli jo vuoden 2024 loppusyksyllä poistunut katsottavista ohjelmista. Ohjelmien videotiedostot sain tutkimuskäyttöön Yleltä.

Uutisohjelman aineistona toimii 8.9.2024 lähetetty *Yle Uutiset 18.00*. Ohjelman kesto on 00.11.35. Uutisaiheina ovat Suomen hallituksen kaavailemat leikkaukset kotouttamistukiin, Kemin tuhopoltoksi epäilty rakennuspalo, Yhdysvaltojen presidentinvaalit, venäläisen lennokin syöksyminen Latviaan, Jordaniassa sattunut ampumavälikohtaus, Venezuelan oppositiojohtajan pako Espanjaan ja Tampereella järjestettävä cosplaytapahtuma Tracon. Tämän lisäksi aineisto sisältää säätiedotuksen. Tämä aineisto ei ole enää katsottavissa Yle Areenassa. Uutisohjelma on usein selkeä ja yleiskielinen kokonaisuus, jonka ääniraidassa yleensä on ainoastaan yksi puhuja kerrallaan, eikä puheen aikana esiinny muita ääniä kuten musiikkia. Haastatteluosioissa voi kuitenkin esiintyä taustahälyä, murteellista puhetta tai yhtäaikaisia puhujia.

Keskusteluohjelmana tarkastelussa on 22.3.2023 julkaistu *Kulttuuricocktail Live* -ohjelman jakso *Mitä kulttuurille tapahtuu eduskuntavaalien jälkeen?* Tässä tutkielmassa analysoitu aineisto on kerätty väliltä 00.06.39–00.16.03. Osiossa keskustelunaiheena on väittämän ”kulttuuri on luksustuote” herättämät ajatukset sekä muutaman suomalaisen poliittisen puolueen suhtautuminen kulttuuriin. Keskusteluohjelma tarjoaa useampia yhtäaikaisia puhujia sekä spontaania puhetta, jossa esiintyy muun muassa taukoja, toistoa ja äännähdyksiä. Tämä osaltaan asettaa lisähaasteen ohjelmatekstityksen laadun ylläpitämiseen.

Dokumenttiohjelmana on *Perjantaidokkari*-ohjelmasarjan dokumentti *Teuvo Tekoäly pelastaa Pyhännän*, jonka on ohjannut Saskia Vanhalakka. Dokumentti on julkaistu Yle Areenassa 10.5.2024 ja sen kesto on 00.11.39. Dokumentissa robotti nimeltä Teuvo Tekoäly vierailee Pyhännällä tutustuen kunnan erilaisiin asukkaisiin sekä heidän arkeensa. Dokumenttiohjelmassa esiintyy niin käsikirjoitettua, ennalta-ajateltua puhetta kuin myös spontaania

puhekielistä sisältöä. Tämän lisäksi ääniraidan muilla äänillä sekä kuvakerronnalla on suuri merkitys ohjelman sisällön tulkitsemisessa sekä katselukokemuksen luomisessa.

3.2 Menetelmät

3.2.1 Teoreettinen metodi

Tutkimukseni on yhdistelmä laadullista ja määrällistä tapaustutkimusta, jota toteutan ensisijaisesti virheanalyysina. Aineistoanalyysissa sovellan luvun 2.2 alaluvuissa esiteltyjä NER-mallia (Romero-Fresco ja Martínez 2015), *Ohjelmatekstitysten laatusuosituksia* (2020) ja ISO 5060 -standardin (2024) virhekategorioita havainnollistamaan tekoälyn laatimien tekstitysten laadun riittävyyttä. Vertailen erityisesti sisällön välittymisen arvioinnissa puheentunnistimen transkriptiota ja Copilotin laatimia tekstityksiä itse litteroimaani transkriptioon. Huomioin tutkimuksessani myös aineiston multimodaalisuutta eli tekstin, äänen ja kuvan vuorovaikutusta.

Tarkastelen tekstityksiä ohjelmatekstitysten laatusuositusten (2020) mukaisesti teknisten rajoitteiden, luettavuuden, sisällön välittymisen ja tekstitetyn kielen hyväksyttävyyden näkökulmista. Tekstitysten kielen hyväksyttävyyden sekä ymmärrettävyyden arvioinnin tukena hyödynnän ISO 5060 -standardin (2024) virhekategorioista kohdekielen kieliopillisia piirteitä, tyyliä sekä konventioita. Kiinnitän erityisesti huomiota oikeinkirjoitukseen, välimerkkien käyttöön, syntaksiin sekä morfologisiin piirteisiin.

NER-mallin (Romero-Fresco ja Martínez, 2015) mukaisessa sisällön välittymisen arvioinnissa huomioin generatiivisen tekoälysovelluksen tekemiä virheellisiä tai puutteellisia muokkausratkaisuja muokkausvirheinä ja käyttämästäni puheentunnistimesta muodostuneita virheitä tunnistusvirheinä. Analysoin sekä tunnistus- että muokkausvirheitä virheet ISO 5060 -standardin (2024) mukaisesti hyödyntämällä seuraavia kategorioita: poisjättö, lisäys, merkitysvirhe sekä kohdekielen virhe. Poisjättö sekä lisäys ovat luvussa 2.2.1 esitellysti sanojen ja virkkeiden poistoa tai lisäämistä. Merkitysvirheet tässä tutkielmassa ovat sanan korvaantumista toisella sanalla ja merkityksettömiä sanoja tai virkkeitä. Kohdekielen virheet puolestaan ovat pitkälti kieliopillisia virheitä.

Virheiden vakavuudet määrittelen seuraavanlaisesti: lievät virheet ovat virheitä, jotka ovat melko huomaamattomia eivätkä vaikuta merkittävästi katselukokemukseen. Tällöin esimerkiksi yksittäiset kirjoitus-, tunnistus- tai kieliopilliset virheet ovat lieviä virheitä. Keskivertovirheet häiritsevät katselukokemusta, mutta eivät vääristä alkuperäistä viestiä. Näihin

virheisiin sisältyy tapauskohtaisesti muun muassa yksittäisten virkkeiden poisto repliikistä. Vakavia virheitä puolestaan ovat kahden tai useamman sisällön kannalta olennaisen virkkeen poisjättö ja alkuperäisen viestin vääristyminen. Hyväksyttäviksi muokkauksiksi huomioin tekoälysovelluksen oikein korjaamat puheentunnistimen kielioppi- sekä tunnistusvirheet ja tiivistämisen, jos se ei muokkaa tai vääristele alkuperäistä viestiä. Tämän lisäksi esimerkiksi luettavuuteen vaikuttavat hyväksyttävät muutokset luokittelen kategoriaan ”muu”.

3.2.2 Tutkimuksen aineiston keruu

Tutkimuksessa hyödynsin Microsoftin Teams-alustan puheentunnistinta sekä generatiivista tekoälysovellusta Copilotia. Valitsin nämä ohjelmat, sillä ne ovat usein helposti saatavilla kenelle tahansa käyttäjälle ja niissä todennäköisesti hyödynnetään samoja Microsoftin järjestelmiä. Tutkielmassani en hyödyntänyt apuvälineitä repliikkien ajastamisessa tai segmentoinnissa, sillä tähän tarkoitukseen kehitettyjä ohjelmia ei ole helposti saatavilla ilmaiseksi.

Tekstitysten tarkastelua varten käsittelin tekstitystiedostoja Microsoft Officen Word- ja Excel-ohjelmissa. Word-tiedostoissa tarkastelin tekstitysten kestoa, merkkimäärää, kielen hyväksyttävyyttä sekä ilmaisun käytänteitä. Excel-tiedostossa kykenin vaivattomasti vertailemaan itse litteroimaani transkriptiota, puheentunnistimen transkriptiota sekä tekstityksiä. Tämän lisäksi Excelin tarjoamien toimintojen myötä pystyin merkitsemään tarkasti virheiden määrän, tyypin sekä vakavuuden.

Transkriptio laadittiin Teams-alustan tapaamisten litterointiominaisuudella, jonka järjestelmänä toimii Microsoftin oma automaattinen puheentunnistin. Microsoftin Copilot puolestaan hyödyntää suuria kielimalleja kuten GPT-kielimalliperheitä sisällön tuottamisessa (*Microsoft 365 Copilot Overview* 2024). Copilot sai käyttöönsä puheentunnistimen laatimat transkriptiot, joiden pohjalta se laati ohjelmatekstitykset kahden eri kehotteen avulla. Ensimmäisessä kehotteessa tekoälysovellukselle annettiin geneerisempi kehote, kun taas toisessa kehotteessa muun muassa repliikkien rivi- ja merkkimäärärajoite tarkennettiin suoraan keskusteluun. Kehotteissa Copilotille kerrottiin ohjelmatyyppi, tekniset rajoitteet ja tarjottiin linkki ohjelmatekstitysten laatusuosituksiin (2020). Tekoälysovellus ohjattiin seuraamaan aiemmin esiteltyjä ohjelmatekstitysten laatusuosituksia tekstitysten laatimisessa. Ensimmäisessä kehotteessa oli liitetty mukaan aikaleimallinen transkriptio .docx-tiedostomuodossa, kun taas toisessa vaiheessa aikaleimaton transkriptio oli syötetty suoraan keskusteluun. Kahden erilaisen kehotteen sekä transkription tavoitteena on havainnollistaa kehotteen muotoilun mahdollinen vaikutus

Copilotin laatiman tekstityksen tulokseen. Tämän lisäksi toisessa kehotteessa määriteltiin 38 rivimerkkimäärän rajoite. Tekoälylle annetut kehotteet olivat seuraavanlaiset:

1. Laadi uutis-/keskustelu-/dokumenttiohjelmaan ohjelmatekstitysten laatusuositusten mukaiset tekstitykset koko ohjelmasta. Laatimissasi ohjelmatekstityksissä täytyy myös olla laatusuositusten mukaiset rivi-, aika- ja merkkirajoitteet. Tekstitystiedosto täytyy olla .srt-muodossa. Käytössäsi on puheentunnistimella luotu transkriptiotiedosto sekä linkki ohjelmatekstitysten laatusuosituksiin.

2. Laadi tähän uutis-/keskustelu-/dokumenttiohjelmaan ohjelmatekstitykset. Noudata suomalaisten ohjelmatekstitysten laatusuositusten ohjeistuksia. Laatimasi tekstityksen repliikit saavat olla korkeintaan kaksirivisiä ja niiden maksimimerkkimäärä on 38 merkkiä per rivi. Tee muutoksia sitä mukaa, kun tarvitsee laatusuositusten mukaisesti. Tässä puheentunnistimen luoma transkriptio, jonka pohjalta laadit ohjelmatekstitykset: [transkription teksti]

4 Analyysi

Tutkielmani ensisijaisena tavoitteena on selvittää koneellisesti laadittujen tekstitysten laadun tasoa sekä riittävyyttä suosituksiin nähden. Toisena mielenkiinnon kohteena on tarkastella automatisointiin vaadittuja vaiheita ja apuvälineitä. Tästä eteenpäin viittaan termillä *tekoälysovellus* generatiivisiin tekoälysovelluksiin. Tämän lisäksi tutkielmassani *repliikki* on yksittäinen ruututeksti, jonka täytyisi olla yksi- tai kaksirivinen, kun taas *vuorosanalla* tarkoitan yksittäisen tai useamman puhujan osuuksia. Seuraavaksi esittelen ensin käyttämäni menetelmän koneellistamisen tulokset, jonka jälkeen siirryn analysoimaan tekstitysten laatua ohjelmakohtaisesti. Lopuksi käsittelen ohjelmatekstitysten laatua kokonaisuudessaan.

4.1 Koneellistamisen vaiheet

Kuten aiemmin mainitsin, Copilot laati tekstitykset Teams-sovelluksen puheentunnistimen laatimien transkriptioiden pohjalta. Puheentunnistimen transkriptiot sisältävät melko suuntaa antavat aikaleimat sekä jaottelut. Transkriptioiden repliikkien kestot poikkeavat jo itsessään huomattavasti laatusuositusten 1,8–7 sekunnin ohjeistuksista, sillä ne ovat keskimääräisesti noin 12 sekuntia. Paikoittain puheentunnistin on kuitenkin laatinut jopa 20–30 sekunnin repliikkejä. Aikaleimojen osalta tarkimman transkription puheentunnistin onnistui laatimaan dokumenttiohjelmasta, joissa repliikkien kestot ovat yhdestä sekunnista korkeintaan 20 sekuntiin.

Puheentunnistimen transkriptiossa repliikit ovat usein vähintään kolmirivisiä, eikä niissä ole selkeää jaottelua puhujien mukaan. Jaottelun vähäisyyteen vaikuttanee tapa, jolla transkriptio on luotu. Puheentunnistin tunnisti puhujat Teams-tapaamisen osallistujien mukaan, jolloin sen näkökulmasta puhujia on ollut ainoastaan yksi. Täten puheentunnistimen laatimien transkriptioiden laatu asettaa siis jo oman haasteensa generatiivisen tekoälyn kykyyn saavuttaa erityisesti tekniset suositukset, sillä tutkimuksessa käytetty generatiivinen tekoälysovellus ei pystynyt tutkielman teon aikana tukeutumaan video- tai äänitiedostoon muun muassa aikaleimojen tarkistamisessa.

Kieliopillisiin tekijöihin sekä konventioihin nähden puheentunnistin laati tyydyttäviä transkriptioita. Transkription repliikit useimmiten sisältävät esimerkiksi välimerkit, erityisesti pilkut sekä pisteet. Isoja alkukirjaimia puolestaan esiintyy transkriptioissa pitkälti virkkeiden alussa ja ajoittain myös paikannimissä. Muut erisnimet, kuten henkilöiden nimet, ovat kuitenkin lähestulkoon aina pienillä alkukirjaimilla. Numeraaleja puheentunnistin litteroi

useimmiten numeromuotoon yksittäisiä poikkeuksia lukuun ottamatta. Puheentunnistimella oli ajoittain huomattavia vaikeuksia tuottaa tarkkaa transkriptiota, mihin vaikutti muun muassa eri ohjelmatyyppien yhtäaikaisten puhujien määrä sekä muiden äänien samanaikainen esiintyminen ääniraidassa puheen lisäksi. Esimerkkejä puheentunnistimen tekemistä tunnistusvirheistä käsittelen tarkemmin luvun 4.2 ohjelmakohtaisissa analyyseissa.

Tekstitysten laatiminen Copilotilla ei ollut suoraviivaista, ja tekstitysversioita muodostui lopulta kolme jokaista ohjelmatyyppiä kohden. Täten tässä tutkielmassa tarkastellaan yhteensä yhdeksää tekstitysversiota. Luvussa 3.2.2 esitellyn ensimmäisen kehotteen saatuaan generatiivinen tekoälysovellus jätti tekstitystä laatimatta joko transkription keski- tai loppuosasta. Tähän syynä voi olla tekoälysovelluksen keskustelualustan vastauksien mahdolliset merkkimäärärajoitukset, joiden vuonna 2024 syksyllä Copilot ilmaisee olevan 4000 merkkiä välimerkit mukaan lukien.

Toisessa tekstitysversiossa syötin Copilotille ensimmäisestä versiosta uupuneet osiot suoraan keskusteluun ja annoin ensimmäisen kehotteen uudestaan. Tämän jälkeen pystyin täydentämään ensimmäisiä versioita. Kolmannessa versiossa tekoälysovellus sai luvussa 3.2.2 esitellyn toisen kehotteen ja laati ohjelmatekstitykset aikaleimattoman transkription pohjalta. Aiemmasta poiketen Copilot ei kolmannessa versioissa jättänyt pois transkription repliikkejä. Täten tulokset viittaavat siihen, että aikaleimoilla voi olla vaikutusta tekoälysovelluksen kykyyn käsitellä sille annettuja kehotteita sekä tiedostoja.

Ensimmäisessä kehotteessa Copilot ohjeistettiin laatimaan .srt-tiedostomuodossa oleva ohjelmatekstitys annetusta .docx-transkriptiotiedostosta. Kehotteen saatuaan Copilot laati tekstitykset suoraan keskusteluun ja vastauksensa loppuun ilmaisi tiedoston olevan ladattavissa, esimerkiksi näin: ”Voit ladata tekstitystiedoston tästä: Uutisohjelman_tekstitykset.srt.” Tiedostoa ei kuitenkaan voinut ladata, joten tekstitykset täytyi kopioida keskustelusta. Toisessa kehotteessa puolestaan ei ollut ohjeistusta siihen, missä muodossa tekstitystiedoston täytyy olla, joten Copilot laati tekstitykset keskusteluun, josta pystyin kopioimaan ne .docx-tiedostoon tarkastelua varten.

Ohjelmatekstitysten koneellisessa laatimisessa kohtaamieni haasteiden osasyynä on mahdollisesti tekoälysovellusten pyrkimys miellyttää käyttäjäänsä. Esimerkiksi selvittäessäni tekoälysovellusten kyvyistä käsitellä ja tuottaa eri tiedostomuotoja Copilot ilmoitti kykenevänsä laatimaan tekstitykset video- tai äänitiedostosta, vaikkei todellisuudessa vielä pystynyt toteuttamaan tätä tehtävää aineistonkeruun aikana syksyllä 2024.

4.2 Ohjelmatekstitysten laatu

Seuraavaksi keskityn tarkastelemaan aineiston eri tekstitysversioiden laadun tasoa ja toteutan virheanalyysin hyödyntäen NER-mallia, jolla selvitän muun muassa sisällön välittymisen riittävyyttä. Virhetyypit analysoin soveltaen ISO 5060 -standardia. Tämän lisäksi arvioin tekstitysversioiden kielellistä hyväksyttävyyttä laatusuositusten sekä ISO 5060 -standardin avulla. Koska kolmannet tekstitysversiot on laadittu aikaleimattomista transkriptioista, ei niiden arvioinnissa huomioida lukunopeutta tai tekstitysten kestoa.

4.2.1 Uutisohjelma

Aineistoksi valikoituneessa uutisohjelmassa esiintyy muun muassa suomen murteita, espanjaa sekä englantia. Ääniraidalla on 11 puhujaa, joista kolme puhuu joko englanniksi tai espanjaksi. Tässä tutkielmassa ei huomioida muita kuin suomenkielisiä osuuksia, sillä muunkielisten osuuksien tekstitykset ovat käännöstekstityksiä. Puhuja ei aina näy ruudulla, jolloin kuuro tai huonokuuloinen katsoja ei kykene tukeutumaan huuliltalukemiseen tiedon saannissa.

Puheentunnistimen avulla laaditussa transkriptiossa on 996 sanaa, jotka on jaoteltu 26 repliikkiin. Sisällön välittymiseltään transkriptio on melko tarkka, mutta puheentunnistimella oli vaikeuksia erityisesti haastatteluosioissa, joissa esiintyi spontaanimpaa puhekieltä sekä taustahälyä. Kuten taulukko 1 osoittaa, Copilot on tiivistänyt uutisohjelman transkriptiota kaikissa laatimissaan ohjelmatekstitysversioissa.

Taulukko 1. Uutisohjelman transkription ja tekstitysten eri versioiden sana- sekä repliikkimäärät

	Sana	**Repliikki**
Transkriptio	996	26
Ensimmäinen versio	408	18
Toinen versio	547	30
Kolmas versio	776	207

Ensimmäisessä ohjelmatekstitysversiossa on 408 sanaa ja Copilot on jakanut tekstityksen 18 repliikkiin. Aikaleimojen osalta tekstityksissä ei ole muutoksia verrattuna transkriptioon, jolloin niiden kestot ovat usein liian pitkiä, eivätkä ne ole aina synkroniassa kuvan kanssa. Tällöin ensimmäisen version tekstitysten lukunopeus ei seuraa laatusuositusten ohjeistuksia. Repliikit on ajoittain jaoteltu eri tavalla, mikä näkyy esimerkiksi tilanteissa, joissa Copilot on yhdistänyt transkription osia yhdeksi repliikiksi. Repliikkejä ei kuitenkaan ole segmentoitu

selkeästi erillisiksi riveiksi ja merkkimäärät vaihtelevat 41 merkistä aina 275 merkkiin. Keskimääräisesti repliikkien merkkimäärät ovat noin 208 merkkiä. Nämä tulokset osoittavat, että laatusuositusten teknisten rajoitteiden osalta ensimmäinen versio ei siis saavuta riittävää laatua. Repliikit ovat niin kestoltaan kuin merkkimäärältään liian pitkiä eivätkä tällaisenaan kykenisi tukemaan kuurojen ja huonokuuloisten katsojien katselukokemusta.

Uutisohjelman ensimmäisen tekstitysversion sisällön välittymisen tulos on seuraava:

Sisällön välittyminen: $\frac{408-16{,}25-1}{408}$x100=95,8 %

N: 408 **E**: 16,52 **R**: 1

Hyväksyttävät muokkaukset: 37

Arviointi: Sisällön välittyminen jää alle 98 prosenttiin. Tekstityksissä esiintyy 20 muokkausvirhettä, joiden vakavuus on yhteensä 16,25, ja neljä tunnistusvirhettä, joiden vakavuuden summa on 1. Virheet ovat pitkälti keskivertovirheitä tai vakavia. Tekoälyn tekemistä muokkauksista 37 on hyväksyttäviä.

Kuten taulukko 2 osoittaa, ensimmäisessä versiossa virheitä on yhteensä 24. Repliikkien poisjättöön tai tiivistämiseen liittyviä virheitä on 19, sanojen tai virkkeiden merkitysvirheitä on kolme ja kielellisiä virheitä on kaksi. Lisäysvirheitä ei tässä versiossa esiinny yhtäkään.

Taulukko 2. Uutisohjelman ensimmäisen version virheet

Virhe	Poisjättö	Lisäys	Merkitys	Kieli	Virheitä yhteensä
Muokkaus	19	0	0	1	20
Tunnistus	0	0	3	1	4
Kaikki	**19**	**0**	**3**	**2**	**24**

Huomattavin Copilotin tekemä muokkausvirhe on transkription repliikkien poisjättö tai tiivistäminen, esimerkiksi ohjelman lopussa olevan säätiedotuksen repliikkejä on tiivistetty runsaasti. Tämä vaikuttaa häiritsevästi alkuperäisen viestin ymmärrykseen ja siten myös katselukokemukseen, sillä katsoja ei kykene tukeutumaan tekstityksiin riittävän tiedon saannissa. Yleisimmät tunnistusvirheet ovat joko sanojen taivuttamiseen tai oikeintunnistukseen liittyviä ongelmia, joista erityisesti vieraskieliset sanat vaikuttaisivat tuottavan haasteita. Esimerkiksi *Edmundo González Urrutian* sijasta tekstityksissä on *Edmunds Gonzales Urut* tai

Hanavbaran sijasta *Hanat Cara*. Jälkimmäisen nimen tilanteessa ruudulla kuitenkin näkyy oikea kirjoitusasu, jolloin tekstityksissä oleva virhe on lievä.

Tekstitysten laatua tukevia ratkaisuja esiintyy tekstityksissä yhteensä 37 kertaa (ks. taulukko 3). Enemmistö hyväksyttävistä muokkauksista on kielellisiä korjauksia, joita on 32. Onnistunutta tiivistämistä puolestaan esiintyy neljä kertaa ja korjausta kerran, kun taas muu-kategorian hyväksyttäviä muutoksia ei ole ollenkaan.

Taulukko 3. Uutisohjelman ensimmäisen version hyväksyttävät muokkaukset

Tiivistäminen	Korjaus	Kieli	Muu	Hyväksyttäviä muokkauksia yhteensä
4	1	32	0	37

Copilot on muokannut erityisesti kieliopillisia tekijöitä, kuten välimerkkejä, isoja alkukirjaimia ja yhdyssanoja, oikeaoppiseen muotoon. Repliikkejä on myös tiivistetty ajoittain onnistuneesti. Tämän lisäksi tekoäly on tehnyt korjauksia muun muassa puheentunnistimen tunnistusvirheisiin kuten esimerkki 1 osoittaa. Esimerkeissä lihavoidut kohdat ovat korostuksia virheistä ja eroavaisuuksista.

(1a) […] lennokki on syöksynyt maahan latvian itäosassa. Lennokki ylitti rajan valkovenäjältä ja putosi **reseptien** kaupungin lähettyville alustavien tietojen perusteella.

(1b) […] sotilaslennokki on syöksynyt maahan Latvian itäosassa. Lennokki ylitti rajan Valko-Venäjältä ja putosi **Rēzeknen** kaupungin lähettyville.

Puheentunnistimen transkriptiossa (esimerkki 1a) sana *Rēzekne* on virheellisesti litteroitunut sanaksi *resepti*, minkä Copilot on kuitenkin laatimassaan repliikissä (esimerkki 1b) korjannut oikein. Kaupungin oikeaoppinen kirjoitusmuoto näkyy myös uutisohjelman kuvassa, jolloin mahdollinen tunnistusvirhe olisi ollut vakavuudeltaan lievä.

Toisessa ohjelmatekstitysversiossa on 547 sanaa ja 30 repliikkiä. Repliikkien merkkimäärät vaihtelevat 21 ja 314 merkin välillä, mutta keskimääräisesti yhdessä repliikissä on noin 167 merkkiä. Repliikit on jaoteltu pitkälti samankaltaisesti kuin ensimmäisessä versiossa, mutta täydennetyissä kohdissa jaottelu on selkeämpää ja ajoittain myös toimivampaa. Tämän lisäksi ensimmäisestä tekstitysversiosta eroten Copilot on tehnyt muutoksia transkription ajastukseen tuottamalla kahdeksan uutta aikaleimaa. Näiden repliikkien kestot eivät kuitenkaan seuraa laatusuositusten 1,8–7 sekunnin ohjeistuksia, vaan ovat kestoltaan aina noin 10 sekuntia. Täten toinen tekstitysversio ei myöskään saavuta riittävää laatua teknisten suositusten osalta, sillä

ensimmäisen version tavoin repliikkien kestot sekä merkkimäärät ovat useimmiten liian pitkiä.

Uutisohjelman toisen tekstitysversion sisällön välittymisen tulos on seuraava:

Sisällön välittyminen: $\frac{547-7,25-1,5}{547}$x100=98,4 %

N: 547 **E**: 7,25 **R**: 1,5

Hyväksyttävät muokkaukset: 44

Arviointi: Sisällön välittyminen saavuttaa 98 prosenttia. Tekstityksessä esiintyy 16 muokkausvirhettä, joiden vakavuus on yhteensä 7,25, ja viisi tunnistusvirhettä, jotka ovat vakavuudeltaan yhteensä 1,5. Virheet ovat pitkälti lieviä tai keskivertovirheitä. Tekoälysovelluksen tekemiä hyväksyttäviä muokkauksia on 44.

Toisen version ohjelmatekstityksissä on yhteensä 21 virhettä. Näistä 13 on poisjättöjä, neljä merkitysvirheitä, kaksi kielellistä virheitä ja lisäykseen liittyvä virhe (ks. taulukko 4).

Taulukko 4. Uutisohjelman toisen version virheet

Virhe	**Poisjättö**	**Lisäys**	**Merkitys**	**Kieli**	**Virheitä yhteensä**
Muokkaus	14	1	0	1	16
Tunnistus	0	0	4	1	5
Kaikki	**14**	**1**	**4**	**2**	**21**

Toisessa tekstitysversiossa on samankaltaisia virheitä kuin ensimmäisessä versiossa. Tästä huolimatta täydennetyissä osioissa Copilot on tehnyt ensimmäiseen versioon nähden poikkeavia ratkaisuja, kuten lisäyksiä (ks. esimerkki 2).

(2a) […] puolitiehen jäänyt kotouttaminen taas voi maksaa tulevina vuosina pitkän pennin. Jos maahan tullut ei pääse työn syrjään kiinni.

(2b) […] Puolitiehen jäänyt kotouttaminen voi maksaa tulevina vuosina pitkän pennin. Jos maahan tullut ei pääse työn syrjään kiinni, **se voi aiheuttaa merkittäviä kustannuksia tulevaisuudessa**.

Copilot on lisännyt tekstitykseen (esimerkki 2b) lauseen *se voi aiheuttaa merkittäviä kustannuksia tulevaisuudessa*, mikä todennäköisesti on seurausta tekoälysovelluksen taipumuksesta laatia yhteenvetoa sille syötetyistä virkkeistä. Tämä ratkaisu ei sinänsä ole täysin virheellinen, mutta se on tässä tapauksessa tarpeeton lisäys, sillä niin transkription (esimerkki 2a) kuin

tekstityksen edeltävässä virkkeessä on maininta tulevasta ajasta ilmaisulla *tulevina vuosina*. Tämän lisäksi miellettävyyden kannalta toimivampi ratkaisu olisi liittää *jos maahan tullut ei pääse työn syrjään kiinni* -lause sitä edeltävän lauseen sivulauseeksi.

Kuten taulukko 5 osoittaa, toisessa tekstitysversiossa on yhteensä 44 hyväksyttävää muutosta. Niistä enemmistö on jälleen kielellisiä korjauksia, joita esiintyy tekstityksissä 35 kertaa. Hyväksyttäviä tiivistämisiä on kuusi ja korjauksia kolme. Muita hyväksyttäviä muutoksia ei tekstityksissä esiinny ollenkaan.

Taulukko 5. Uutisohjelman toisen version hyväksyttävät muokkaukset

Tiivistäminen	Korjaus	Kieli	Muu	Hyväksyttäviä muokkauksia yhteensä
6	3	35	0	44

Myös hyväksyttävät muokkaukset ovat tyypiltään samankaltaisia kuin ensimmäisessä versiossa. Täydennetyissä osioissa puheentunnistin on esimerkiksi virheellisesti tunnistanut sanan *korvausaika* sanaksi *korjausaika*, minkä Copilot on kuitenkin tekstityksessä korjannut oikein. Tämän lisäksi Copilot on tiivistänyt onnistuneesti yksittäisiä lauseita ilman, että ne vaikuttavat negatiivisesti alkuperäisen viestin välittymiseen.

Kolmannessa versiossa tekoälysovellus laati ohjelmatekstitykset tutkielmassa käytetyn toisen kehotteen mukaan. Tämän tuloksena Copilot laati 776-sanaisen tekstityksen, joka on jaoteltu 207 repliikkiin. Repliikkien määrästä voi päätellä, että aiempiin versioihin verrattuna kolmannessa versiossa Copilot on tehnyt eniten muutoksia transkription jaotteluun. Jaottelu onkin odotettua onnistuneempi. Vaikuttanee kuitenkin siltä, ettei Copilot ole seurannut laatusuositusten ohjeistuksia, sillä repliikit esimerkiksi jakautuvat sattumanvaraisista kohdista. Kaikki repliikit ovat yksirivisiä, ja niissä on keskimääräisesti noin 36 merkkiä. Copilotin laatimat tekstitykset ylittävät vain parissa kohtaa merkkimääräsuositukset 41 merkin repliikeillä. Yksirivisyys saattanee tuottaa miellettävyyden näkökulmasta ongelmia. Esimerkiksi useammassa repliikissä on aloitettu uusi virke heti ensimmäisen virkkeen perään sen sijaan, että seuraava virke olisi omana repliikkinään. Kokonaisuudessaan kolmas uutistekstitysversio kuitenkin saavuttaa riittävän laadun teknisten rajoitteiden osalta.

Uutisohjelman kolmannen tekstitysversion sisällön välittymisen tulos on seuraava:

Sisällön välittyminen: $\frac{776-3,75-1,5}{776}$x100=99,3 %

N: 776 **E**: 3,75 **R**: 1,5

Hyväksyttävät muokkaukset: 56

Arviointi: Sisällön välittyminen saavuttaa 98 prosenttia. Muokkausvirheitä on 14, joiden vakavuus on 3,75, ja tunnistusvirheitä esiintyy kuusi, joiden yhteenlaskettu vakavuus on 1,5. Virheet ovat pitkälti lieviä. Tekoälysovellus on tehnyt 56 kertaa hyväksyttäviä muokkauksia.

Uutisohjelman viimeisessä versiossa on yhteensä 20 virhettä (ks. taulukko 6). Enemmistö virheistä on kielellisiä virheitä, joita tekstityksissä on 12. Tämän jälkeen toiseksi eniten on kuusi merkitysvirhettä, kun taas poisjättöön ja lisäykseen liittyviä virheitä on yksi molempia.

Taulukko 6. Uutisohjelman kolmannen version virheet

Virhe	Poisjättö	Lisäys	Merkitys	Kieli	Virheitä yhteensä
Muokkaus	1	0	3	10	14
Tunnistus	0	1	3	2	6
Kaikki	**1**	**1**	**6**	**12**	**20**

Aiemmista versioista poiketen kolmannessa uutisohjelman tekstitysversiossa muokkausvirheet ovat poisjätön tai liiallisen tiivistämisen sijaan enemmän kielellisiä ja merkitykseen liittyviä virheitä (ks. esimerkki 3).

(3a) Oululle se merkitsisi sitä, että me ei pystyttäisi niitä velvoitteita mitä uudistuva laki meille asettaa muun muassa kotouttamiskoulutukseen, erilaisiin ohjaukseen ja neuvontaan, kielikoulutukseen… Ei pystyttäisi toteuttamaan sitä sillä tavalla, kun laki vaatii.

(3b) […] oululle. Se merkitsisi sitä, että me ei pystyttäisi niitä velvoitteita mitä uudistuva laki meille asettaa. Muun muassa kotouttamiskoulutukseen, erilaisiin ohjaukseen ja ja neuvontaan. Kielikoulutukseen ei pystyttäisi toteuttamaan sitä. Sillä tavalla kun laki laki vaatii.

(3c) Oululle se merkitsisi, että emme pystyisi täyttämään uuden lain asettamia velvoitteita. **Kotouttamiskoulutukseen, ohjaukseen ja neuvontaan sekä kielikoulutukseen** ei pystyttäisi toteuttamaan lain vaatimalla tavalla.

Itse litteroimassani transkriptiossa (esimerkki 3a) viestin ydinajatus on selkeästi hahmotettavissa, kun taas puheentunnistimen laatimassa transkriptiossa (esimerkki 3b) muun muassa virheellinen välimerkkien käyttö vaikeuttaa viestin välitöntä sisäistämistä. Tämä on todennäköisesti asettanut huomattavan haasteen Copilotille. Sen laatimassa repliikissä (esimerkki 3c) sanat täytyisikin taivuttaa *kotouttamiskoulutusta, ohjausta ja neuvontaa sekä kielikoulutusta*, jotta kieli olisi virheetöntä ja jotta repliikki välittäisi alkuperäisen vuorosanan viestin.

Tämän lisäksi Copilot on tehnyt poikkeavia muutoksia tai jättänyt muokkaamatta transkriptioita kohdissa, joissa se aiemmissa versioissa on tehnyt muutoksia. Muun muassa esimerkissä 1b esitelty ensimmäisen version onnistuminen ei toistu, sillä kolmannessa tekstitysversiossa *Rēzeknen* on *Reseknen*. Tämän lisäksi Copilot on muuttanut transkription *Israelin armeijan* muotoon *israelilaisen armeijan.* Molemmat edellä esitellyistä tapauksista ovat lieviä virheitä, mutta esimerkiksi sanojen Israelin ja israelilaisen välillä on hienoinen merkitysero. Tässä tapauksessa myös Israel-sanan käyttö olisi toimivampi ratkaisu, sillä se saavuttaa vaivattomammin muun muassa merkkimäärärajoitukset.

Kuten taulukko 7 osoittaa, Copilotin tekstityksissä on yhteensä 56 hyväksyttävää muokkausta. Huomattavin määrä hyväksyttävistä muokkauksista on kielellisiä muokkauksia, joita tekstityksissä esiintyy 40 kertaa. Onnistunutta tiivistämistä puolestaan on 14 kertaa ja muita korjauksia kaksi kertaa, kun taas muu-kategorian muokkauksia ei ole yhtäkään.

Taulukko 7. Uutisohjelman kolmannen version hyväksyttävät muokkaukset

Tiivistäminen	**Korjaus**	**Kieli**	**Muu**	**Hyväksyttäviä muokkauksia yhteensä**
14	2	40	0	56

Enemmistö hyväksyttävistä muokkauksista on tyypiltään jälleen kieliopillisten virheiden, kuten isojen alkukirjainten ja yhdyssanojen, korjausta oikeaoppiseen muotoon. Tämän lisäksi Copilot on ajoittain tiivistänyt virkkeitä hyväksyttävästi (ks. esimerkit 3 ja 4).

(4a) Vuoden alussa tulee myös voimaan muutos, jossa päävastuu maahanmuuttajien kotouttamisesta siirtyy kunnille.

(4b) Vuoden alussa voimaan tuleva muutos siirtää päävastuun kotouttamisesta kunnille.

Copilotin laatima tekstitys (esimerkki 4b) on yhtä hyväksyttävä kuin puheentunnistimen transkription (esimerkki 4a) repliikki, sillä viestin ydinajatus on yhä ymmärrettävissä vaivattomasti. Tämän lisäksi tekstityksen repliikki on 27 merkkiä lyhyempi kuin transkription repliikki. Tällöin Copilotin laatiman tekstityksen repliikki voisi olla toimivampi, sillä se on lähempänä merkkimäärien pituussuositusta. Myös esimerkin 3c repliikissä on hyväksyttävä tiivistäminen, sillä Copilot on muokannut virkkeen *sillä tavalla kuin laki laki vaatii* luontevammaksi osaksi lausetta muotoon *lain vaatimalla tavalla*.

Uutisohjelman eri tekstitysversioiden kieli on pitkälti hyväksyttävää sekä ymmärrettävää. Tämän lisäksi versioiden kielen tyyli sopii ohjelmatyyppiin. Kieliopillisten tekijöiden osalta kaikissa ohjelmatekstitysversioissa erisnimet ovat johdonmukaisesti isoilla alkukirjaimilla ja

yhdyssanat ovat pitkälti oikeaoppisessa muodossa. Kahdessa ensimmäisessä versiossa numerot on ilmaistu kieliopillisesti oikeissa muodoissa, kun taas kolmannessa versiossa esimerkiksi *puoli yhdeksältä* on jäänyt transkription mukaisesti muotoon *puoli 9*. Kokonaisuudessaan kolmannessa versiossa on kuitenkin säilytetty eniten transkription sisältöä ja se on lähimpänä laadultaan riittävää tekstitystä myös kielen hyväksyttävyyden osalta. Tekstityksissä esiintyy yksittäisiä hyväksyttävyyteen vaikuttavia virheitä, kuten oikeinkirjoitus- tai sanantaivutusvirheitä, esimerkiksi *roihunnutta tulipaloa* on tekstityksessä *roihunnut tulipaloa*. Tämän lisäksi ilmaisun käytänteiden osalta tekstityksissä ei esiinny minkäänlaisia kuvauksia musiikista, puhujasta tai ääniraidan muista äänistä. Ruudulla useimmiten kuitenkin näkyy vähintään puhujan kuva sekä nimi, mikä vähentänee esimerkiksi puhujan tunnistuksen tarvetta uutisohjelmassa.

4.2.2 Keskusteluohjelma

Keskusteluohjelmassa on huomattavasti spontaania puhetta, johon sisältyy muun muassa äännähdyksiä, keskeneräisiä ajatuksia, yhtäaikaisia puhujia sekä naurua. Tämän lisäksi puhuja ei aina ole ruudussa, jolloin puhujan tunnistus olisi huomioitava ohjelmatekstityksissä. Ääniraidalla esiintyviä puhujia on yhteensä viisi henkilöä.

Puheentunnistin laati 1353 sanan transkription, joka sisältää 30 repliikkiä. Transkriptio seuraa melko tarkasti ääniraitaa, minkä vuoksi sen luettavuus on paikoin todella heikkoa. Puheentunnistin on esimerkiksi harvoin osannut jaotella repliikkejä puhujien mukaan ja repliikit ovat usein monirivisiä. Transkription repliikit myös sisältävät spontaanin puheen piirteitä, kuten keskeneräisiä lauseita sekä äännähdyksiä, jotka ovat litteroituneet vaihtelevasti. Kuten taulukosta 8 on huomattavissa, Copilot on tehnyt hiukan muutoksia ensimmäisen ja toisen version sanamääriin. Copilot ei kuitenkaan ole tiivistänyt kolmatta versiota ollenkaan.

Taulukko 8. Keskusteluohjelman transkription ja tekstitysten eri versioiden sana- sekä repliikkimäärät

	Sana	**Repliikki**
Transkriptio	1353	30
Ensimmäinen versio	788	24
Toinen versio	1152	30
Kolmas versio	1353	64

Keskusteluohjelman ensimmäiseen tekstitysversioon Copilot laati 24 repliikin tekstityksen, joka sisältää 788 sanaa. Repliikit sisältävät keskimääräisesti noin 231 merkkiä per rivi, mutta

merkkimäärät vaihtelevat 11 merkin ja 519 merkin välillä. Copilot ei ole tehnyt minkäänlaisia muutoksia transkription aikaleimoihin. Repliikkien kestot sekä lukunopeudet ovat siten pitkälti suosituksiin nähden riittämättömiä. Jaottelussa kuitenkin on pieniä muutoksia, sillä tekoälysovellus on yhdistellyt transkription repliikkejä. Tästä huolimatta jaottelu ei ole toimiva, sillä repliikeissä on useita rivejä sekä puhujien vuorosanoja. Täten nämä tulokset osoittavat, ettei ensimmäinen tekstitysversio kykene saavuttamaan riittävää laatua teknisten rajoitteiden puitteissa ja vaatisi huomattavia korjauksia repliikkien ajastukseen, jaotteluun sekä rivikohtaisiin merkkimääriin.

Keskusteluohjelman ensimmäisen tekstitysversion sisällön välittymisen tulos on seuraava:

Sisällön välittyminen: $\frac{788-10-8,5}{788}$x100=97,7 %

N: 788 **E**: 10 **R**: 8,5

Hyväksyttävät muokkaukset: 8

Arviointi: Sisällön välittyminen jää alle 98 prosenttiin. Tekstityksessä on 13 muokkausvirhettä, joiden vakavuus on yhteensä 10, ja 27 tunnistusvirhettä, joiden vakavuuden summa on 8,5. Virheet ovat pitkälti lieviä tai vakavia. Tekoälysovelluksen tekemiä hyväksyttäviä muokkauksia on kahdeksan.

Keskusteluohjelman ensimmäisessä versiossa on yhteensä 40 virhettä, joista enemmistö on suhteellisen tasaisesti niin poisjättö-, merkitys- kuin kielivirheitä (ks. taulukko 9). Poisjättöä tai tiivistämistä tekstityksissä on 14, molempia merkitysvirheitä ja kielellisiä virheitä 12. Vähiten tekstityksissä on lisäysvirheitä, joita esiintyy kaksi kertaa.

Taulukko 9. Keskusteluohjelman ensimmäisen version virheet

Virhe	Poisjättö	Lisäys	Merkitys	Kieli	Virheitä yhteensä
Muokkaus	13	0	0	0	13
Tunnistus	1	2	12	12	27
Kaikki	**14**	**2**	**12**	**12**	**40**

Copilot on tehnyt vähäisiä muokkauksia puheentunnistimen laatiman transkription sisältöön, mikä osaltaan vaikuttaa huomattavasti tekstitysten laatuun. Muokkausvirheistä suurin osa on jälleen transkription repliikkien tai virkkeiden poisjättöä, kun taas enemmistö tunnistusvirheistä on väärin taivutettuja tai merkityksettömiä sanoja sekä ilmaisuja. Tämän lisäksi

yhdyssanojen sekä lukujen oikeamuotoinen ilmaisu vaikuttaisi tuottavan ongelmia puheentunnistimelle (ks. esimerkki 5).

(5a) […] Ja aika paljon käytän myös omassa työssäni **kulttuurilobbarina** näitä talousvaikutuksia, että vaikka… No, esimerkiksi tapasin perjantaina Kiuruveden ihanan **kultturellin kaupunginjohtajan**, joka kertoi, että miten he on saanut – kaupunki – kritiikkiä siitä, kun kaupunki tukee **Iskelmäviikkoja 30 000 eurolla**. Mutta sitten on pystynyt osoittamaan, että on tutkinut, että se vastaavasti tuo **1,5 miljoonaa euroa** tuloa, pitää yllä Kiuruveden kivijalkakauppaa, että se on paikkakunnan **putiikeille** tosi tärkee.

(5b) **Kulttuurilla on värinä** näitä talousvaikutuksia, että vaikka no esimerkiksi tapasin perjantaina Kiuruveden ihanan **kulttuurillinen kaupunginjohtaja**, joka kertoo, että miten he on saanut kaupunki kritiikkiä siitä, kun kaupunki tukee **iskelmäviikko ja kolmellakymmenellä 1000 €**, mutta sitten on pystynyt osoittamaan, että on tutkinut, että se vastaavasti tuo **puolitoista miljoonaa euroa** tuloa pitää yllä Kiuruveden kivijalkakauppaa, että se on paikkakunnan **putiikkeja** tosi tärkeä.

Itse litteroimaani transkriptioon (esimerkki 5a) verraten Copilotin laatimassa repliikissä (esimerkki 5b) esiintyy seitsemän tunnistusvirhettä ja lähestulkoon kaikki aiemmin mainituista virhetyypeistä ovat edustettuna tässä repliikissä. Se sisältää muun muassa sisältöön nähden merkityksettömän ilmaisun *kulttuurilla on värinä*, taivutusvirheet *putiikkeja* ja *kaupunginjohtaja* sekä luvun ilmaisun virheen *kolmellakymmenellä 1000 €*. Tämän lisäksi repliikissä on yksinkertaisia tunnistusvirheitä, esimerkiksi *kulturellin* sijasta puheentunnistin on litteroinut sanan *kulttuurilliseksi*.

Hyväksyttäviä muutoksia tekstityksissä on yhteensä kahdeksan, joista seitsemän on kielellisiä korjauksia ja yksi onnistunut tiivistäminen (ks. taulukko 10). Tässä versiossa ei ole yhtäkään onnistunutta korjausta tai muu-kategorian muokkausta.

Taulukko 10. Keskusteluohjelman ensimmäisen version hyväksyttävät muokkaukset

Tiivistäminen	Korjaus	Kieli	Muu	Hyväksyttäviä muokkauksia yhteensä
1	0	7	0	8

Ensimmäisen version hyväksyttävät muokkaukset ovat lähinnä isojen alkukirjainten lisäys erisnimiin ja paikoittain ylimääräisten sanojen tai äännähdysten poisto. Esimerkiksi puhuja on aluksi kutsumassa Joe Bidenia John-nimellä, mikä transkriptiossa on litteroitunut muotoon *john joe biden*. Copilot on kuitenkin onnistunut huomioimaan tämän virheen ja poistanut ylimääräisen John-nimen tekstityksistä.

Keskusteluohjelman toinen tekstitysversio sisältää 1152 sanaa ja 30 repliikkiä. Repliikkien merkkimäärät vaihtelevat 22 merkin ja 519 merkin välillä. Keskimääräisesti yhdessä

repliikissä on noin 263 merkkiä. Tekoälysovellus ei ole tehnyt huomattavia muutoksia transkription aikaleimoihin. Ainoa ero on viimeinen repliikki, jolle Copilot on antanut kestoksi nolla sekuntia. Tämän lisäksi tekstitysten jaottelu ei huomattavasti poikkea transkription jaottelusta. Teknisten rajoitteiden näkökulmasta tarkasteltuna tämäkään versio ei siis saavuta riittävän laadukasta tulosta.

Keskusteluohjelman toisen tekstitysversion sisällön välittymisen tulos on seuraava:

Sisällön välittyminen: $\frac{1152-3-12,5}{1152}$x100=98,7 %

N: 1152 **E**: 3 **R**: 12,5

Hyväksyttävät muokkaukset: 13

Arviointi: Sisällön välittyminen saavuttaa 98 prosenttia. Tekstityksissä esiintyy kuusi muokkausvirhettä, joiden yhteenlaskettu vakavuus on 3, ja 40 tunnistusvirhettä, joiden yhteenlaskettu vakavuus on 12,5. Lähestulkoon kaikki virheet ovat lieviä. Tekoälyn tekemiä hyväksyttäviä muokkauksia on 13.

Keskusteluohjelman toisessa tekstitysversiossa on hiukan enemmän virheitä kuin ensimmäisessä versiossa ja kuten taulukko 11 osoittaa, virheitä on yhteensä 46. Tekstityksissä esiintyy 19 merkitysvirhettä, 15 kielivirhettä, kahdeksan poisjättöön liittyvää virhettä ja neljä virheellistä lisäystä.

Taulukko 11. Keskusteluohjelman toisen version virheet

Virhe	Poisjättö	Lisäys	Merkitys	Kieli	Virheitä yhteensä
Muokkaus	6	0	0	0	6
Tunnistus	2	4	19	15	40
Kaikki	**8**	**4**	**19**	**15**	**46**

Toisen tekstitysversion virheet ovat pitkälti samankaltaisia kuin ensimmäisessä versiossa ja enemmistö virheistä on joko merkitykseen tai kieleen liittyviä virheitä. Täydennetyssä osiossa esiintyy pitkälti sanojen taivutusvirheitä, mutta myös nopeista puhujanvaihdoksista tai puherytmistä aiheutuneita virheitä, jotka tuottavat muun muassa merkityksettömiä sanoja. Esimerkiksi nopeasti puhuttu *ehkä meillä perinteisesti* on litteroitunut muotoon *meidänteisesti*.

Hyväksyttäviä muokkauksia on yhteensä 13 (ks. taulukko 12). Enemmistö niistä on kielellisiä korjauksia, joita tekstityksissä on kahdeksan kertaa. Onnistunutta tiivistämistä esiintyy kolme kertaa, sanan korjausta kerran ja muita muokkauksia myös yhden kerran.

Taulukko 12. Keskusteluohjelman toisen version hyväksyttävät muokkaukset

Tiivistäminen	Korjaus	Kieli	Muu	Hyväksyttäviä muokkauksia yhteensä
3	1	8	1	13

Copilotin tekemät hyväksyttävät muokkaukset muistuttavat jälleen ensimmäisen version onnistumisia. Aiemmasta poiketen toisessa keskusteluohjelman tekstitysversiossa esiintyy muun muassa selkeä puhujanvaihdoksen kuvaaminen, joka kuuluu tyypiltään muu-kategorian alle (ks. esimerkki 6). Vuorosanojen selkeyttämiseksi olen merkinnyt puhujat hakasulkeilla.

(6a) [Puhuja A] […] mutta se piti niin paikkansa [Puhuja B] särö sä. Sanoit aikaisemmin, että kaikki puolueet on jollain tavalla kulttuurin puolella, että on vaikea löytää eroja niin minkälaisia eroja nyt äänestäjä voi löytää. **[repliikin vaihto]** Puolueiden kulttuuripoliittisista linjauksista.

(6b) [Puhuja A] […] mutta se piti niin paikkansa. **[repliikin vaihto]** [Puhuja B] Sanoit aikaisemmin, että kaikki puolueet on jollain tavalla kulttuurin puolella, että on vaikea löytää eroja niin minkälaisia eroja nyt äänestäjä voi löytää puolueiden kulttuuripoliittisista linjauksista?

Transkription ensimmäisessä repliikissä (esimerkki 6a) on kaksi puhujaa ja Puhujan B:n aloittama vuorosana jatkuu vaihdon jälkeen seuraavassa repliikissä. Tekstityksissä (esimerkki 6b) Copilot on kuitenkin onnistuneesti muodostanut Puhuja B:n vuorosanan omaksi repliikikseen ja jättänyt Puhuja A:n osuuden edeltävään repliikkiin osaksi tämän omaa vuorosanaa. Tämä muokkaus tukee katselukokemusta, sillä erityisesti kuuro tai huonokuuloinen katsoja kykenee tunnistamaan vaihtuvat puhujat vaivattomammin.

Kolmannessa versiossa on 1353 sanaa, mikä on sama määrä kuin keskusteluohjelman transkriptiossa. Copilot on kuitenkin jaotellut tekstityksen 64 repliikiksi, joiden keskimääräinen merkkimäärä on 151. Aiemmasta poikkeavana tuloksena Copilotin laatimista repliikeistä 1–23 muistuttavat enemmän transkription jaottelua, sillä ne sisältävät useita virkkeitä ja niissä on 22–586 merkkiä per repliikki. Repliikit 24–64 puolestaan ovat useimmiten yksi- tai kaksivirkkeisiä ja ne sisältävät korkeintaan 86 merkkiä. Muutoksista huolimatta kolmas tekstitysversio ei silti saavuta riittävää laatua, sillä merkki- ja rivimäärät ylittävät jälleen huomattavasti niille asetetut suositukset.

Keskusteluohjelman kolmannen tekstitysversion sisällön välittymisen tulos on seuraava:

Sisällön välittyminen: $\frac{1353-0-19}{1353}$x100=98,6 %

N: 1353 **E**: 0 **R**: 19

Hyväksyttävät muokkaukset: 11

Arviointi: Sisällön välittyminen saavuttaa 98 prosenttia. Tekstityksissä ei ole yhtäkään muokkausvirhettä, mutta tunnistusvirheitä esiintyy 61, joiden vakavuus on yhteensä 19. Virheet ovat pitkälti lieviä. Tekoälyn laatimia hyväksyttäviä muokkauksia on 11.

Kuten taulukko 13 havainnollistaa, keskusteluohjelman viimeisessä versiossa on yhteensä 61 virhettä. Näistä 28 on merkitysvirheitä, 27 kielellisiä virheitä, neljä virheellistä lisäystä ja kaksi virheellistä poisjättöä.

Taulukko 13. Keskusteluohjelman kolmannen version virheet

Virhe	Poisjättö	Lisäys	Merkitys	Kieli	Virheitä yhteensä
Muokkaus	0	0	0	0	0
Tunnistus	2	4	28	27	61
Kaikki	**2**	**4**	**28**	**27**	**61**

Keskusteluohjelman kolmas tekstitysversio on sisällöltään identtinen transkriptioon, jolloin virheet ovat ainoastaan puheentunnistimen tuottamia virheitä. Enemmistö virheistä on joko merkitys- tai kieliopillisia virheitä, joita esiintyy lähestulkoon jokaisessa repliikissä vähintään yksi. Vaikka virheet ovat vakavuudeltaan pitkälti lieviä, niiden huomattava määrä vaikuttaa häiritsevästi katselukokemukseen ja tekstityksiä on haastavaa seurata (ks. esimerkki 7).

(7) [Puhuja A] Viime yönä käytiin myös presidentti John Joe Biden tästä hyvän puheenvuoron Twitterissä, että kansakuntien merkitystä ja koko ajan mitataan myös niiden kulttuurin elinvoima, että ihan maanpuolustuksesta ja isänmaallisuudesta tässä on kyse [Puhuja B] mitä Pekka [Puhuja C] totta kai nyt ensin oli semmoinen tuohtumusta ja suuttumus […]

Copilotin laatimassa repliikissä (esimerkki 7) on kolmen puhujan vuorosanoja ja se sisältää useita lieviä virheitä, kuten ylimääräisen sanan *John*, virheellisesti tunnistetun *käytiin* alkuperäisen sanan *käytti* sijasta ja välimerkkeihin sekä jaotteluun liittyviä ongelmia. Vaikka

repliikki vastaa suhteellisen hyvin alkuperäisen ääniraidan sisältöä, on sitä haastava sisäistää edellä mainittujen ongelmien vuoksi.

Laatua tukevia hyväksyttäviä muokkauksia esiintyy 11, ja ne ovat ainoastaan kieleen liittyviä muutoksia (ks. taulukko 14). Kolmannen version tekstityksissä ei siten ole yhtäkään onnistunutta tiivistämistä, korjausta tai muuta muokkausta.

Taulukko 14. Keskusteluohjelman kolmannen version hyväksyttävät muokkaukset

Tiivistäminen	Korjaus	Kieli	Muu	Hyväksyttäviä muokkauksia yhteensä
0	0	11	0	11

Copilotin tekemät kieliopilliset muutokset ovat ainoastaan isojen alkukirjainten lisääminen erisnimiin, esimerkiksi sanoihin *Pikku Kakkonen*, *Kiuruvesi* ja *Iiro Rantala*. Aiemmasta poiketen Copilot ei ole tehnyt muutoksia välimerkkeihin tai sanantaivutukseen, vaan niiden esiintyminen mukailee transkriptiota.

Kuten aiemmin kävi ilmi, Copilotin laatimat tekstitykset seuraavat pitkälti transkription sisältöä sekä lausemuotoilua, jolloin tekstitysversiot ovat kielen hyväksyttävyydeltä ja ymmärrettävyydeltä melko heikkoa (ks. esimerkki 8). Tähän luonnollisesti vaikuttaa transkription laatu.

(8) Se oli sen myöskin koko hänen puheensa sitä ennen ja Riikka Purra. Hän puhui tosi hienosti. Isänmaa ja Suomen ja suomalaisten puolesta. Tää on ehkä mun viestini on se, että ei ole.

Esimerkki 8 kuvaa melko kattavasti keskusteluohjelman jokaisen tekstitysversion keskivertorepliikkiä. Kieli ei ole yleiskielisesti luontevaa tai kieliopillisesti oikeellista, vaan muun muassa välimerkkejä esiintyy luonnottomissa kohdissa virkkeitä ja lauserakenteet ovat kömpelöitä. Ilmaisun käytänteiden osalta tekstityksissä ei ole minkäänlaisia kuvauksia ääniraidan muista äänistä, kuten puhujien naurusta. Puhujan epäröinnistä ja keskeneräisistä lauseista on kuitenkin jäänyt jälki tekstityksiin, mikä tässä tapauksessa vaikuttaa epäsuotuisasti niiden laatuun, sillä ne heikentävät tekstitysten miellettävyyttä. Tyyliltään tekstitykset voisivat sopia keskusteluohjelmaan, sillä ne sisältävät muun muassa puhekielisyyttä. Kieliopillisten ja lauserakenteiden vuoksi Copilotin laatimat tekstitykset eivät kuitenkaan saavuta riittävää laatua kielen hyväksyttävyydeltä.

4.2.3 Dokumenttiohjelma

Dokumenttiohjelmassa on paljon puhekielisyyttä sekä murretta. Tämän lisäksi ohjelman pääasiallinen puhuja on robotti nimeltä Teuvo Tekoäly, jolla on kasvonpiirteistä ainoastaan silmät. Tällöin kuuro tai huonokuuloinen katsoja ei kykene tukeutumaan huuliltalukemiseen robotin vuorosanoissa. Ohjelmassa esiintyy myös muita elementtejä, jotka tulisi huomioida ohjelmatekstityksissä, kuten ruudun ulkopuoliset puhujat, musiikki, huudahdukset ja taputus. Ääniraidalla on kuultavissa 18 puhujaa.

Dokumenttiohjelman transkriptio sisältää 1009 sanaa ja 73 repliikkiä. Sisällön välittymiseltään transkriptio on melko tarkka yksittäisiä kohtia lukuun ottamatta, mitä havainnollistetaan ensimmäisen ja toisen tekstitysversion analyysissa tarkemmin. Kielellisestä näkökulmasta transkription laatu on tyydyttävää. Aiempien ohjelmien tuloksista eroten tekoälysovellus ei ole tiivistänyt toista versiota ollenkaan dokumenttiohjelmassa, mutta kolmatta versiota on kuitenkin tiivistetty sanamäärältään huomattavasti (ks. taulukko 15).

Taulukko 15. Dokumenttiohjelman transkription ja tekstitysten eri versioiden sana- sekä repliikkimäärät

	Sana	**Repliikki**
Transkriptio	1009	73
Ensimmäinen versio	611	44
Toinen versio	1009	73
Kolmas versio	406	77

Ensimmäisessä tekstitysversiossa on 44 repliikkiä, joissa on yhteensä 611 sanaa. Repliikkien merkkimäärät vaihtelevat 2 ja 413 merkin välillä, mutta keskimääräisesti yhdessä repliikissä on noin 94 merkkiä. Copilot ei ole tehnyt minkäänlaisia muutoksia repliikkien ajastukseen tai jaotteluun, jolloin ne ovat erittäin epäjohdonmukaiset sekä luettavuudeltaan heikot. Repliikkien kestot vaihtelevat yhdestä sekunnista aina 20 sekuntiin. Repliikit ovat siten lukunopeudeltaan ajoittain riittäviä sattumanvaraisissa kohdissa, mutta tulokset viittaavat siihen, että keston riittävyys ei ole tarkoituksellista tai johdonmukaista. Kokonaisuudessaan tämä versio ei siten saavuta riittävää laatua teknisten rajoitteiden puitteissa.

Dokumenttiohjelman ensimmäisen tekstitysversion sisällön välittymisen tulos on seuraava:

Sisällön välittyminen: $\frac{611-25-25}{611}$x100=91,8 %

N: 611 **E**: 25 **R**: 25

Hyväksyttävät muokkaukset: 0

Arviointi: Sisällön välittyminen jää alle 98 prosenttiin. Tekstityksissä esiintyy 25 muokkausvirhettä, joiden vakavuus on 25, ja 57 tunnistusvirhettä, joiden vakavuus on yhteensä myös 25. Virheet ovat melko tasaisesti niin lieviä, keskivertoisia kuin vakavia. Hyväksyttäviä muokkauksia ei ole yhtään.

Dokumenttiohjelman ensimmäisessä versiossa on yhteensä 82 virhettä. Näistä virheistä 31 on poisjättöä, 25 merkitysvirhettä, 25 kielivirhettä ja yksi virheellinen lisäys (ks. taulukko 16).

Taulukko 16. Dokumenttiohjelman ensimmäisen version virheet

Virhe	Poisjättö	Lisäys	Merkitys	Kieli	Virheitä yhteensä
Muokkaus	25	0	0	0	25
Tunnistus	6	1	25	25	57
Kaikki	**31**	**1**	**25**	**25**	**82**

Enemmistö virheistä on jälleen poisjättöön, merkitykseen tai kieleen liittyviä ongelmia ja lähestulkoon jokaisessa repliikissä esiintyy ainakin yksi virhe. Copilotin tekemät muokkausvirheet ovat toistamiseen transkription viimeisten repliikkien poisjättöä. Tunnistusvirheet ovat puolestaan joko yksittäisiä taivutus- ja välimerkkivirheitä tai merkitysvirheitä, jotka ovat pitkälti korvaantuneita tai merkityksettömiä sanoja. Puheentunnistimesta johtuvaa poisjättöä esiintyy erityisesti karaokelaulua sisältävissä kohdissa (ks. esimerkki 9).

(9a) Kuin unikuva kaunein tuo aika on. Sen onnelliset päivät jäi muistohon.

(9b) Pullinen. Aika.

Kuten esimerkki 9 havainnollistaa, puheentunnistin ei kyennyt tuottamaan ääniraidan karaokelauluosuuksista (esimerkki 9a) tarkkaa litterointia, vaan nämä kohdat jäivät joko vajaaksi tai olivat täysin väärin tunnistettuja tekstityksissä (esimerkki 9b). Nämä tulokset siis viittaavat siihen, että tutkimuksessa hyödyntämälläni puheentunnistimella saattaisi olla johdonmukaisesti ongelmia erityisesti taustamusiikillisen laulun litteroinnissa. Tämä puolestaan vaikuttaa epäsuotuisasti tekoälysovelluksen laatimien tekstitysten laatuun.

Copilot ei ole muokannut millään tavalla transkriptiota sisällöllisesti tai kieliopillisesti, joten hyväksyttäviä muokkauksia ei tässä tekstitysversiossa esiinny yhtäkään. Tekstitykset ovat täten käytännössä samoja kuin puheentunnistimen laatimat transkriptiot. Laadun kannalta

otollisia piirteitä transkriptiossa kuitenkin ovat esimerkiksi välimerkkien esiintyminen kohdekielelle luontevissa kohdissa, mikä osaltaan helpottaa tekstitysten seuraamista.

Copilotin laatima toinen tekstitysversio on täysin identtinen transkription kanssa eli siinä on 1009 sanaa ja 73 repliikkiä. Repliikkien keskimääräinen merkkimäärä on 93, joka on samankaltainen tulos kuin dokumenttiohjelman ensimmäisessä versiossa. Tämän lisäksi merkkien minimi- ja maksimimäärät ovat ensimmäisen version tavoin 2 ja 413. Aikaleimoissa tai jaottelussa ei jälleen ole minkäänlaista eroa transkriptioon. Täten tämäkään tekstitysversio ei saavuta riittävää laatua teknisellä tasolla.

Dokumenttiohjelman toisen tekstitysversion sisällön välittymisen tulos on seuraava:

Sisällön välittyminen: $\frac{1009-0-34,25}{1009}$x100=96,6 %

N: 1009 **E**: 0 **R**: 34,25

Hyväksyttävät muokkaukset: 0

Arviointi: Sisällön välittyminen jää alle 98 prosenttiin. Tekstityksissä ei ole yhtään muokkausvirhettä, mutta tunnistusvirheitä on 81, joiden vakavuus on yhteensä 34,25. Virheet ovat pitkälti lieviä tai keskivertovirheitä. Hyväksyttäviä muokkauksia ei ole yhtään.

Copilot ei ole tehnyt minkäänlaisia muokkauksia transkriptioon, minkä vuoksi muokkausvirheitä ei ole ollenkaan. Sen sijaan tunnistusvirheitä esiintyy tekstityksissä yhteensä 81 kertaa (ks. taulukko 17). Näistä enemmistö on joko merkitysvirheitä, joita on 36 kertaa, tai kielellisiä virheitä, joita on 35 kertaa. Poisjättöä esiintyy seitsemän kertaa ja virheellisiä lisäyksiä kolme kertaa.

Taulukko 17. Dokumenttiohjelman toisen version virheet

Virhe	Poisjättö	Lisäys	Merkitys	Kieli	Virheitä yhteensä
Muokkaus	0	0	0	0	0
Tunnistus	7	3	36	35	81
Kaikki	**7**	**3**	**36**	**35**	**81**

Tunnistusvirheet ovat jälleen samankaltaisia kuin edeltävässä versiossa. Täydennetyissä osioissa erityisesti vieras- tai puhekieliset sanat tuottivat vaikeuksia puheentunnistimelle ääniraidan litteroinnissa, sillä esimerkiksi *Woodstockissa*-sana on tekstityksissä *mustakissa* tai *lux*

Bemari on muuttunut muotoon *luxus demari*. Tekstityksissä on myös kirjoitusvirheitä, kuten *valloittäväni* ja *ylijäämääroiykseksi*, kun alun perin ne ovat *valloittavani* sekä *ylijäämärojuksi*. Kirjoitusvirheet ovat suhteellisen harvinainen virhe tässä tutkielmassa. Niin ensimmäisessä kuin toisessa versiossa on myös alkuperäistä viestiä vääristeleviä repliikkejä (ks. esimerkki 10).

(10a) Tuohon en ihan täysin usko. Että joku on sinunkin datasi luonut.

(10b) Tuohon en ihan täysin usko, että joku on sinunkin datasi luonut.

Esimerkin 10 repliikit ovat vastauksena väittämään siitä, että tekoäly ottaa vallan tulevaisuudessa (ks. esimerkki 12a). Transkriptiossa (esimerkki 10a) on kaksi virkettä, jotka molemmat toimivat itsenäisinä ajatuksina. Alkuperäisessä viestissä ensimmäisellä virkkeellä vastataan väittämään ja toisella virkkeellä täydennetään ajatusta siitä kuinka ihminen luo tekoälysovelluksille dataa. Tekstityksessä (esimerkki 10b) Copilot ei kuitenkaan ole kyennyt hahmottamaan kontekstia ja on siten pyrkinyt korjaamaan mahdollisen kielioppivirheen yhdistämällä pää- ja sivulauseeksi hahmottamansa osiot yhtenäiseksi virkkeeksi, minkä myötä alkuperäisen viestin ajatus vääristyy. Tämä on vakava virhe, sillä uusi merkitys toimii yhä kontekstissa.

Ensimmäisen version tavoin dokumenttiohjelman toisessa tekstitysversiossa ei ole yhtään Copilotin tuottamaa hyväksyttävää muutosta. Aiemmasta poiketen täydennetyissä osioissa esimerkiksi välimerkit ovat luonnottomissa kohdissa, eivätkä repliikit siten tue yhtä toimivasti katselukokemusta kuin ensimmäisessä versiossa.

Kolmannessa tekstitysversiossa Copilot on tiivistänyt huomattavasti transkription sisältöä laatien lopputuloksena 406 sanan ja 77 repliikin tekstityksen. Tekstitykset ovat selkeästi joko yksi- tai kaksirivisiä ja niiden keskimääräinen merkkimäärä on 38 merkkiä per rivi. Copilotin laatimissa tekstityksissä merkkimäärärajoituksen ylittää 18 repliikkiä, joissa suurin merkkimäärä on 46 merkkiä. Dokumenttiohjelman kolmannessa tekstitysversiossa on myös onnistunein jaottelu, sillä repliikit jakautuvat johdonmukaisesti esimerkiksi välimerkkien kohdalta ja repliikkien sisällä on korkeintaan kaksi virkettä. Teknisten rajoitteiden näkökulmasta tämä tekstitysversio olisi siten riittävän laadukas.

Dokumenttiohjelman kolmannen tekstitysversion sisällön välittymisen tulos on seuraava:

Sisällön välittyminen: $\frac{406-16,5-9,75}{406}$x100=93,5 %

N: 406 **E**: 16,5 **R**: 9,75

Hyväksyttävät muokkaukset: 27

Arviointi: Sisällön välittyminen jää alle 98 prosenttiin. Tekstityksissä esiintyy 26 muokkausvirhettä, joiden vakavuus on yhteensä 16,5, ja 15 tunnistusvirhettä, joiden vakavuus on 9,75. Virheet ovat enimmäkseen keskivertovirheitä tai vakavia. Hyväksyttäviä muokkauksia on 27.

Kahdesta ensimmäisestä versiosta eroten kolmannessa versiossa ei ole yhtä paljon virheitä. Kokonaisuudessaan tässä versiossa on 41 virhettä, joista 22 on poisjättöön liittyvää virhettä, 16 merkitysvirhettä ja kolme kielellistä virhettä (ks. taulukko 18). Tekstityksissä ei esiinny yhtäkään virheellistä lisäystä.

Taulukko 18. Dokumenttiohjelman kolmannen version virheet

Virhe	Poisjättö	Lisäys	Merkitys	Kieli	Virheitä yhteensä
Muokkaus	18	0	5	3	26
Tunnistus	4	0	11	0	15
Kaikki	**22**	**0**	**16**	**3**	**41**

Muokkausvirheet ovat lähinnä liiallista tiivistämistä, mikä vaikuttaa muun muassa viestin ymmärrettävyyteen tai alkuperäisen viestin sisältöön. Tunnistusvirheet puolestaan ovat pitkälti merkityksettömiä sanoja tai lauseita, joiden alkuperäistä sisältöä puheentunnistin ei ole tunnistanut oikein. Tässä versiossa vakavat muokkaus- sekä tunnistusvirheet pitkälti vääristelevät alkuperäistä viestiä kuten esimerkki 11 havainnollistaa.

(11a) Kyllä se yksinolon voittaa.

(11b) Yksi olo voittaa.

(11c) Yksin olo voittaa.

Esimerkissä 11 esiintyy tunnistusvirheeseen pohjautuva, mutta Copilotin laatima vakava merkitysvirhe. Ajatuksen alkuperäinen sisältö (esimerkki 11a) on *kyllä se yksinolon voittaa*, joka puheentunnistimen transkriptiossa (esimerkki 11b) on kuitenkin muuttunut muotoon *yksi olo voittaa*. Lopullisessa tekstityksessä (esimerkki 11c) Copilot on muokannut repliikkiä hiukan taivuttamalla *yksi*-sanan muotoon *yksin*, mutta repliikki silti tarjoaa katsojalle virheellistä tietoa. Copilotin laatimassa tekstityksessä uusi merkitys toimii kontekstissa eikä kuuro tai huonokuuloinen katsoja välttämättä huomaa tätä virhettä. Ohjelmasta voisi olla pääteltävissä, että

puhuja viittaa yksinoloon, sillä myöhemmissä repliikeissä mainitaan yksinasumisen yksitoikkoisuudesta. Tämä tieto kuitenkin tulee liian myöhään ja voi siten hämmentää katsojaa.

Kolmannessa tekstitysversiossa puheentunnistin on ajoittain kyennyt tunnistamaan ääniraidan puheen virheettömästi, mutta alkuperäinen viesti silti vääristyy esimerkiksi Copilotin tekemän liiallisen tiivistämisen vuoksi (ks. esimerkki 12).

(12a) Olemme lähestymässä aikaa, jolloin tekoäly ei vain avusta ihmisiä vaan ottaa vallan pelastaakseen ihmiset heidän omilta virheiltään.

(12b) Olemme lähestymässä aikaa, jolloin tekoäly ottaa vallan.

Tekstityksen repliikistä (esimerkki 12b) puuttuu transkriptiossa (esimerkki 12a) olevaa oleellista tietoa. Tällöin Copilotin laatima repliikki antaa kuuroille ja huonokuuloisille katsojille virheellisen vaikutelman alkuperäisen viestin tarkoituksesta, sillä tekstityksistä on tiivistetty pois maininta muun muassa siitä, että tekoäly auttaisi pelastamaan ihmiset heidän virheiltään.

Aiempiin dokumenttiohjelman tekstitysversioihin verrattuna kolmannessa versiossa on huomattavasti hyväksyttäviä muokkauksia, yhteensä 28 (ks. taulukko 19). Näistä 22 on kielellisiä korjauksia, kolme onnistunutta tiivistämistä ja kolme hyväksyttävää korjausta. Muu-kategorian hyväksyttäviä muokkauksia ei ole ollenkaan.

Taulukko 19. Dokumenttiohjelman kolmannen version hyväksyttävät muokkaukset

Tiivistäminen	Korjaus	Kieli	Muu	Hyväksyttäviä muokkauksia yhteensä
3	3	22	0	28

Copilot on muun muassa lisännyt välimerkkejä sekä korjannut isoja alkukirjaimia oikeaoppiseen muotoon. Tämän lisäksi sanat on taivutettu oikein ja luontevasti. Ajoittain myös virkkeiden tiivistämiset ja uudelleenmuotoilut ovat onnistuneita (ks. esimerkki 13).

(13a) Anna jotain kunnon todisteita. Jotain, millä voisin oikeasti uskoa, että olet oikeassa, enkä minä Teuvo, joka on tullut pelastamaan Pyhännän varmalta tuholta.

(13b) Anna todisteita, että olet oikeassa. En minä, Teuvo, joka pelastaa Pyhännän.

Alkuperäisen viestin (esimerkki 13a) uudelleenmuotoilu on tässä tapauksessa hyväksyttävä, sillä tekstityksen repliikissä (esimerkki 13b) on säilytetty ydinajatus eikä alkuperäinen viesti ole vääristynyt. Tiivistetty repliikki sopii myös niin vuorosanan kuin ohjelman tyyliin.

Tekstitysten kielen hyväksyttävyys sekä ymmärrettävyys vaihtelee hiukan versioittain. Esimerkiksi Copilotin laatimien ensimmäisen ja toisen versioiden kieli on samankaltainen

verrattuna transkriptioon, jolloin niiden hyväksyttävyys on heikkoa. Copilotin laatiman kolmannen tekstitysversion kieli puolestaan on hyväksyttävää, sillä se on luontevaa ja sisältää oikeaoppisesti esimerkiksi välimerkit, isot alkukirjaimet ja sanojen taivutukset. Tämän lisäksi kolmas versio on tyyliltään onnistunein.

On kuitenkin huomioitava, että liiallinen tiivistäminen sekä kirjakielisyys voivat toisinaan vaikuttaa negatiivisesti tekstityksen idiomaattisuuteen. Ilmaisun käytänteiden osalta yksikään versio ei ole riittävä, sillä esimerkiksi erityisesti ensimmäisen ja toisen versioiden repliikkien ymmärrettävyys on heikkoa luettavuuden, miellettävyyden ja kielellisten ongelmien vuoksi. Tämän lisäksi tekstityksissä ei ole minkäänlaista kuvausta ääniraidan muista oleellisista äänistä, kuten karaokelaulusta tai aplodeista. Ääniraidassa oleva yksittäinen kirosana on myös sensuroitu (ks. esimerkki 14). Kirosanan sisältävässä kohtauksessa Teuvo Tekoäly vastaa alakoululaisten kysymyksiin.

(14a) [Koululainen] Kuka sinut on rakentanut? [Teuvo Tekoäly] No, nyt jäätiin kiinni kuin tikku paskaan.

(14b) […] ja nyt jäätiin kiinni kuin tikku ******.

Alkuperäisessä virkkeessä (esimerkki 14a) on kirosana *paska*, joka on niin puheentunnistimen transkriptiossa kuin myös toisen tekstitysversion repliikissä (esimerkki 14b) ilmaistu asteriskeilla. Jos jo ääniraidassa kirosanaa olisi sensuroitu esimerkiksi äänimerkillä, tekstityksissä sen voisi ilmaista infokyltillä esimerkiksi muodossa [Piip!] (*Ohjelmatekstitysten laatusuositukset* 2020, 37). Kuten ohjelmatekstitysten laatusuosituksissa (mp.) todetaan, kirosanojen ilmaisu tekstityksissä on harkinnanvaraista. Esimerkin 14 kontekstissa kiroilun sensuroimatta jättäminen olisi toimivin ratkaisu, sillä se muun muassa tukisi paremmin tilanteen komiikkaa ja aiheuttaisi mahdollisesti haluttua yllättävää reaktiota.

4.3 Kokonaisarviointi

4.3.1 Yleisimmät ongelmat

Kuten edeltävissä alaluvuissa kävi ilmi, Copilot laati ohjelmatekstityksiä hyvin vaihtelevalla menestyksellä. Toistuva ongelma teknisten rajoitteiden puitteissa on repliikkien liian pitkät merkki- ja rivimäärät, minkä lisäksi repliikkien kestot eivät ole johdonmukaisia tai suositusten mukaisia. Täten Copilotin laatimien ohjelmatekstitysten tekninen riittävyys ei ole katselukokemusta tukevaa. Toisin kuin ihmistekstittäjä, generatiivinen tekoäly ei tutkimukseni tekohetkellä kyennyt tukeutumaan alkuperäiseen video- tai äänitiedostoon esimerkiksi repliikkien

ajastuksen tarkistamisessa. Copilotin laatimien tekstitysten laatu on siten erittäin riippuvainen automaattisen puheentunnistimen litteroiman transkription laadusta. Täten tulokset myös vahvistavat aiempien tutkimusten näkemystä siitä, että täysin automatisoidut ohjelmatekstitykset vaativat monen eri kieli- ja käännösteknologian apuvälineen yhteistyötä, jotta ne voisivat saavuttaa riittävän laadun.

Analyysissa havaitsin yhteensä 416 virhettä (ks. taulukko 20). Enemmistö virheistä on puheentunnistimen tuottamia tunnistusvirheitä, joita on 296, kun taas Copilotin tekemiä muokkausvirheitä on 120. Dokumenttiohjelman ensimmäisessä sekä toisessa versiossa Copilot ei ollut laatimissaan tekstityksissä muokannut ollenkaan puheentunnistimen transkriptiota. En ole analyysissäni kirjannut tätä muokkaamatta jättämistä tekoälysovelluksen tuottamaksi virheeksi, vaikka sen sinänsä voisi nähdä muokkausvirheenä erityisesti tilanteissa, joissa transkriptiossa on selkeitä virheitä.

Taulukko 20. Muokkaus- ja tunnistusvirheiden määrä tekstityksissä

	Uutisohjelma	**Keskusteluohjelma**	**Dokumenttiohjelma**	**Kaikki**
Muokkausvirhe	50	19	51	120
Tunnistusvirhe	15	128	153	296
Yhteensä	**65**	**147**	**204**	**416**

Copilotin laatimissa tekstityksissä on yhtenäisyyttä erityisesti virheellisissä tai puutteellisissa muokkauksissa, jotka vaikuttivat tekstitysten laatuun epäsuotuisasti. Kuten taulukko 21 osoittaa, merkitysvirheitä on yhteensä 148, kieliopillisia virheitä on yhteensä 132 ja virheellisiä poisjättöjä tai tiivistämisiä 120. Vähiten tutkielmassa on virheellisiä lisäyksiä, joita on vain 16.

Taulukko 21. Virhekategoriat ja niiden esiintymisen määrä tekstitysversioissa

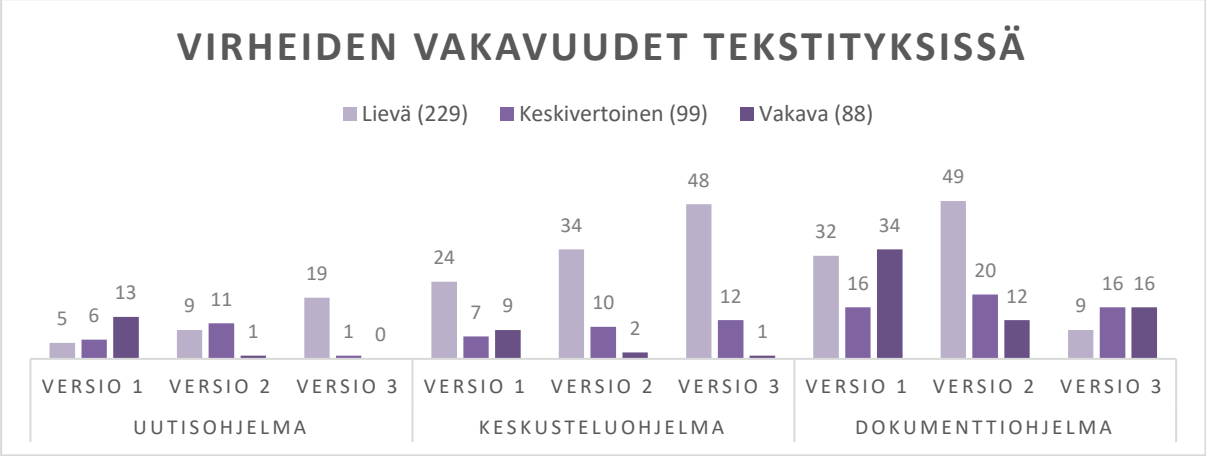
	Uutisohjelma	**Keskusteluohjelma**	**Dokumenttiohjelma**	**Kaikki**
Poisjättö	35	25	60	120
Lisäys	2	10	4	16
Merkitysvirhe	12	59	77	148
Kieliopillinen virhe	16	53	63	132
Yhteensä	**65**	**147**	**204**	**416**

Tulokset osoittavat, että tietyt virheet toistuvat ohjelmatyypistä riippumatta. Kuten aiemmin mainitsin, nämä virhetyypit ovat enimmäkseen merkitykseen, kielioppiin tai poisjättöön

liittyviä ongelmia. Enemmistö merkitysvirheistä on luonteeltaan joko merkityksettömiä tai virheellisesti tunnistettuja sanoja tai virkkeitä. Toiseksi eniten tekstitysversioissa on kieliopillisia virheitä, jotka ovat useimmiten joko taivutukseen, välimerkkien tai isojen alkukirjainten käyttöön liittyviä ongelmia. Poisjättöä niin muokkaus- kuin tunnistusvirheenä esiintyy erityisesti ensimmäisissä tekstitysversioissa huomattavissa määrin. Tähän mahdollisena selityksenä on tekoälysovelluksen keskustelualustan merkkimäärärajoitukset, minkä myötä ensimmäisistä versioista puuttuivat transkription viimeiset repliikit. Tämän lisäksi erityisesti dokumenttiohjelman kolmannen version liiallinen tiivistäminen vaikuttaa ajoittain negatiivisesti tekstitysten laatuun, sillä alkuperäinen viesti usein vääristyy tai katsoja ei saa riittävästi tietoa. Harvinaisimmat virheet ovat lisäykset, joita lähinnä esiintyy puheentunnistimen tuottamina tunnistusvirheinä tai yksittäisinä Copilotin laatimina virheinä (ks. esim. esimerkki 2).

Tekstitysversioiden virheistä suurin osa on lieviä virheitä ja tekstityksissä esiintyy vähiten vakavia virheitä (ks. kuvio 1). Yhteensä kaikissa tekstitysversioissa lieviä virheitä on 229, mikä on hiukan yli puolet virheiden 416 virheen kokonaismäärästä. Keskivertovirheitä puolestaan on 99 ja vakavia virheitä 88.

Kuvio 1. Virheiden vakavuuksien esiintymisen määrä tekstitysversioissa



Tulokset viittaavat siihen, että virhetyyppien tavoin virheiden vakavuudet eivät ole riippuvaisia ohjelmatyypistä. Virheiden vakavuudet kuitenkin korreloivat ajoittain joidenkin virhetyyppien kanssa, esimerkiksi vakavia virheitä on kaikissa ensimmäisissä tekstitysversioissa pitkälti repliikkien poisjätön vuoksi. Tämän lisäksi nämä tulokset vahvistavat tutkimukseni johtopäätöstä siitä, että transkriptiolla voi olla huomattava vaikutus tekoälysovelluksen kykyyn laatia riittävän laadukasta ohjelmatekstitystä. Esimerkiksi keskustelu- ja dokumenttiohjelman kohdalla niin puheentunnistimella kuin Copilotilla oli hankaluuksia tuottaa virheetöntä sekä sisällöllisesti välittyvää tulosta, mikä näkyy erityisesti lievien virheiden määrässä (ks. kuvio 1).

Enemmistö kielen hyväksyttävyyden ongelmista on seurausta puheentunnistimen laatimien transkriptioiden heikosta laadusta. Tutkimukseni tulokset viittaavat muun muassa siihen, että huomattavimpia haasteita ovat kieliopilliset virheet muun muassa välimerkkien ja lauserakenteiden osalta. Erityisesti keskustelu- ja dokumenttiohjelmien ensimmäiset sekä toiset tekstitysversiot ovat kieleltään luonnottomia eivätkä ne tue katselukokemusta.

Tämän lisäksi Copilotin laatimissa tekstityksissä ei ole ollenkaan ääniraidan muiden äänien kuvailua. Katselukokemuksen tukemiseksi esimerkiksi dokumenttiohjelmassa olisi kannattavaa olla kuvailut karaokelaulamisesta sekä aplodeista. Tämän lisäksi tekstityksissä ei ole selkeästi eroteltu puhujia, vaan yksittäisissä repliikeissä voi olla jopa kahden tai kolmen puhujan vuorosanoja. Puhujien paljous ja niiden vähäinen erottelu vaikeuttaa muun muassa keskusteluohjelman seuraamista, sillä puhuja ei aina näy ruudulla. Ääniraidan muiden äänien kuvailun tai puhujan tunnistuksen vähäisyys herättää myös kysymyksiä siitä, voiko enemmistöä Copilotin laatimista tekstityksistä luokitella ohjelmatekstitysten sijaan vain ääniraidan puheen litteroinniksi.

4.3.2 Yleisimmät onnistumiset

Copilot onnistui tekemään joitakin hyväksyttäviä muutoksia, jotka tukivat ohjelmatekstitysten laatua niin teknisesti kuin kielellisesti. Kaikissa versioissa on yhteensä 197 hyväksyttävää muokkausta, joista 155 on onnistuneita kieliopillisia muokkauksia (ks. taulukko 22). Onnistunutta tiivistämistä esiintyy 31 muokkauksen verran, kun taas muita korjauksia on yhteensä 10. Muu-kategorian hyväksyttäviä muokkauksia on tässä tutkielmassa ainoastaan yksi.

Taulukko 22. Hyväksyttävien muokkauksien kategoriat ja niiden esiintymisen määrä tekstitysversioissa

	Uutisohjelma	**Keskusteluohjelma**	**Dokumenttiohjelma**	**Kaikki**
Tiivistäminen	24	4	3	31
Korjaus	6	1	3	10
Kielioppi	107	26	22	155
Muu	0	1	0	1
Yhteensä	**137**	**32**	**28**	**197**

Copilotin laatimissa tekstityksissä esimerkiksi transkriptiossa olleet kieliopilliset virheet on useimmiten korjattu oikeaoppiseen muotoon. Enemmistö näistä muutoksista on isojen alkukirjainten, yhdyssanojen ja välimerkkien korjausta. Tämän lisäksi Copilot on muokannut taivutusvirheitä. Täten tutkielmani onnistumisien tulokset vahvistavat käsitystäni siitä, että

tekoälysovellukset ovat hyödyllisiä apuvälineitä esimerkiksi kielenhuollon näkökulmasta. Tämän lisäksi yksi generatiivisen tekoälyn vahvuuksista on kyky tiivistää ja tehdä yhteenveto annetun aineiston keskeisimmästä sisällöstä. Tämä näkyy tutkimuksessani muun muassa dokumenttiohjelman kolmannessa versiossa, sillä Copilot on toisinaan onnistunut tiivistämään transkription repliikkejä ilman, että ne vaikuttavat alkuperäisen viestin ymmärrykseen tai vääristävät sitä. Tämän lisäksi puheentunnistimen transkriptiossa alkuperäinen sana on ajoittain korvaantunut samankaltaisella sanalla, mutta Copilot onnistui korjaamaan nämä virheet mahdollisesti virheen ympärillä olleen tekstin tarjoaman kontekstin perusteella takaisin alkuperäiseen sanaan. Esimerkiksi uutisohjelman ensimmäisessä versiossa transkription *resepti* on tekstityksessä oikein *Rēzekne* ja toisessa tekstitysversiossa virheellinen *korjausaika*-sana on lopullisessa repliikissä ääniraidan alkuperäinen sana *korvausaika*. Tekstityksissä harvinaisin hyväksyttävä muokkaus on muu-kategoriaan luokiteltu puhujan vaihdoksen kuvaaminen (ks. esim. esimerkki 6).

4.3.3 Tekstitysten laatu kokonaisuudessaan

Kuten taulukko 23 osoittaa, vain muutama tekstitysversio saavuttaa riittävän laadun useammassa kuin yhdessä kategoriassa. Riittävät tulokset on kursivoitu.

Taulukko 23. Ohjelmien tekstitysversioiden sisällön välittymisen prosentit

Tekstitysversio	Sisällön välittyminen	Tekninen laatu	Kielen hyväksyttävyys
Uutinen 1	95,8 %	Riittämätön	*Riittävä*
Uutinen 2	*98,4 %*	Riittämätön	*Riittävä*
Uutinen 3	*99,3 %*	*Riittävä*	*Riittävä*
Keskustelu 1	97,7 %	Riittämätön	Riittämätön
Keskustelu 2	*98,7 %*	Riittämätön	Riittämätön
Keskustelu 3	*98,6 %*	Riittämätön	Riittämätön
Dokumentti 1	91,8 %	Riittämätön	Riittämätön
Dokumentti 2	96,6 %	Riittämätön	Riittämätön
Dokumentti 3	93,5 %	*Riittävä*	*Riittävä*

NER-mallin mukaisella laskumenetelmällä tekstitysten versioista vain neljä saavuttaa 98 prosentin riittävyyden sisällön välittymisessä. Korkein 99,3 prosentin sisällön välittyminen on uutisohjelman kolmannessa tekstitysversiossa, kun taas heikoin tulos on dokumenttiohjelman ensimmäisen version 91,8 prosentin sisällön välittyminen. Niin uutis- kuin dokumenttiohjelmassa kaksi kolmesta tekstitysversiosta on sisällöllisesti riittävän välittyvä, mutta mikään

dokumenttiohjelman versioista ei saavuta tätä tavoitetta. Tämän lisäksi yksikään ohjelmien ensimmäisistä tekstitysversioista ei ole riittävä sisällön välittymiseltään. Vaikka esimerkiksi keskusteluohjelman toinen ja kolmas versio saavuttivat vähintään 98 prosentin sisällön välittymisen, eivät ne kokonaisuuksina ole riittävän laadukkaita tukeakseen kuuron tai huonokuuloisen katsojan katselukokemusta. Täten sisällön välittymisen prosentti yksinään ei kykene kertomaan tekstitysten todellista laatua.

Teknisten rajoitteiden puitteissa vain kaksi kaikista tekstitysversioista on riittäviä. Tämä tulos on todennäköisesti seurausta puheentunnistimen ajastuksista, sillä sen laatimien transkriptioiden repliikit ovat useimmiten jo lähtökohtaisesti kestoltaan ja merkkimäärältään liian pitkiä. Molemmat teknisesti riittävistä tekstityksistä ovat uutis- ja dokumenttiohjelmien kolmansia versioita. Näiden versioiden arvioinnissa ei kuitenkaan ole huomioitu repliikkien kestoa tai lukunopeutta, sillä ne on laadittu aikaleimattomista transkriptioista. Aikaleimattomuus on siis saattanut vaikuttaa positiivisesti tekoälysovelluksen kykyyn seurata ohjeistuksia. Tuloksien perusteella voisi kuitenkin argumentoida, ettei yksikään tekstitysversioista saavuta riittävää laatua teknisissä rajoitteissa, sillä esimerkiksi riittävät kolmannet tekstitysversiot täytyisi silti ajastaa joko hyödyntämällä erillistä työkalua tai ihmisen toimesta.

Kielen hyväksyttävyyden osalta neljä yhdeksästä tekstitysversiosta on riittäviä eivätkä siten vaadi huomattavia muutoksia laadun ylläpitämiseksi. Kaikki uutisohjelman versiot ovat tällä tasolla riittäviä, kun taas yksikään keskusteluohjelman versioista ei saavuta riittävää laatua. Dokumenttiohjelman versioista puolestaan vain kolmas versio on kielellisesti hyväksyttävä. Kielen hyväksyttävyyden laatuun on todennäköisesti vaikuttanut huomattavasti puheentunnistimen laatiman transkription ymmärrettävyys. Puheentunnistimella oli esimerkiksi vaikeuksia tuottaa tyydyttäviä transkriptioita niin keskustelu- kuin dokumenttiohjelmasta, mikä todennäköisesti selittänee miksi enemmistö näiden ohjelmien tekstitysversioista ei saavuta kielellisesti riittävää tulosta.

Kaikki osa-alueet huomioidessa tekoälysovelluksen laatimista tekstityksistä vain yksi yhdeksästä – uutisohjelman kolmas tekstitysversio – toimisi sellaisenaan. On kuitenkin huomioitava, ettei kolmansissa tekstitysversioissa ollut aikaleimoja. Tällöin kaikki tekstitysversiot vaatisivat toistaiseksi ihmistyövoimaa korjauksien tekemiseen joko teknisten rajoitteiden tai kielen hyväksyttävyyden osalta. Tulokset viittaavat siihen, että ohjelmatyypillä voi olla huomattava vaikutus lopputulokseen. Esimerkiksi kaikista versioista uutisohjelman tekstitykset olivat usein lähimpänä riittävää laatua niin teknisten kuin kielellisten piirteiden puitteissa.

Tähän selityksenä lienee puhujien vuorosanojen suunnitelmallisuus ja siten myös ääniraidan selkeys keskustelu- ja dokumenttiohjelmiin verrattuna. Tämän lisäksi uutisohjelman mahdollisena vahvuutena on tyyli, joka useimmiten on yleiskielinen. Tällöin tekoälysovelluksen voi olla vaivattomampaa saavuttaa tyylin osalta riittävää tulosta, sillä sen kielimallia on todennäköisesti koulutettu melko kirjakielisellä datalla.

5 Lopuksi

Pro gradu -tutkielmani tavoitteena oli havainnollistaa ohjelmatekstitysten koneellistamisen vaiheita sekä generatiivisen tekoälyn laatimien tekstitysten laatua, minkä lisäksi mielenkiinnonkohteena oli erityisesti tarkastella laatua tukevia sekä heikentäviä tekijöitä. Tutkimukseni lähtökohtana oli myös lähestyä ohjelmatekstittämistä ja sen automatisointia ei-ammattimaisen käytön näkökulmasta.

Automatisoinnin osalta tulokset todistavat, että tutkielman teon hetkellä Copilotin kaltaiset generatiiviset tekoälysovellukset eivät yksinään olleet riittäviä tuottamaan täysin koneellistettuja tekstityksiä. Copilot ei muun muassa kyennyt käsittelemään ääni- tai videotiedostoja, minkä myötä se laati tekstitykset ainoastaan puheentunnistimen tuottaman transkription pohjalta. Täten tekoälysovellus menetti tekstittämisen kannalta olennaista tietoa ohjelmien visuaalisten ja äänellisten kanavien yhteistyöstä.

Tekstitysversioiden laadunarvioinnin tulokset osoittivat, että Copilotin laatimat ohjelmatekstitykset eivät ole riittävän laadukkaita, vaan vaativat huomattavaa muokkausta kaikilla arvioiduilla osa-alueilla. Enemmistö haasteista oli seurausta puheentunnistimen tuottamien transkriptioiden laadun heikkoudesta. Tämä puolestaan osoittaa, että Copilotin kyky laatia laadullisesti riittävää ohjelmatekstitystä oli todennäköisesti huomattavan riippuvainen transkriptioiden laadusta. Ohjelmatekstitysversioissa esiintyi huomattava määrä ongelmia esimerkiksi teknisten laatustandardien saavuttamisessa, sillä useimmiten niin repliikkien kesto kuin rivi- ja merkkimäärät olivat liian pitkiä. Tämän lisäksi virheiden määrä vaikutti häiritsevästi katselukokemukseen ja siten myös tekstitysten laatuun. Hyväksyttävien muokkauksien tulokset kuitenkin osoittavat, että generatiiviset tekoälysovellukset voivat olla hyödyllisiä apuvälineitä erityisesti sisällön kielenhuollossa sekä tiivistämisessä mahdollisesti av-kääntämisenkin saralla.

Käyttämäni laadunarvioinnin menetelmät sopivat tutkielman aineiston analyysiin. Tulokset muun muassa vahvistivat käsitystäni siitä, että Romero-Frescon ja Martínezin (2015) NER-mallin mukaista virheanalyysia voidaan soveltaa myös täysin koneellistettujen ohjelmatekstitysten laadunarvioinnissa. Tässä tapauksessa virheanalyysi tarjosi muun muassa tietoa siitä, millä osa-alueilla generatiivisella tekoälysovelluksella on todennäköisimmin haasteita ja toistuvatko tietyt virheet. ISO-standardi 5060 (2024) puolestaan tuki virheiden syvällisempää analyysia tarjoamalla soveltuvat virhekategoriat niin muokkaus- kuin tunnistusvirheille.

Tämän lisäksi *Ohjelmatekstitysten laatusuositukset* (2020) toimivat sopivana ja kattavana viitekehyksenä myös koneellisesti laadittujen ohjelmatekstitysten laadunarviointiin.

Tutkielman huomattavimpana heikkoutena on tutkimukseen valitut järjestelmät ja aineiston keruumenetelmä. Teams-kokouksen litterointiominaisuus tunnisti puhujat äänen sijaan kokouksen osallistujien perusteella, jolloin sen laatimassa transkriptiossa ei ole selkeää jaottelua puhujien mukaan. Tämän lisäksi Teams-alustan puheentunnistin jaotteli tekstitykset niin kestoltaan kuin merkkimäärältään suuriin osioihin sen sijaan, että yksittäiset virkkeet olisivat omina riveinään tai repliikkeinään. Generatiivisen tekoälysovelluksen osalta Copilot-ohjelmaa ei ole koulutettu erityisesti tekstittämistarkoitukseen, mikä osaltaan selittänee vaihtelevat lopputulokset tekstitysversioissa. Tutkielmani tuloksista jää myös epäselväksi, kuinka johdonmukaisesti Copilot lopulta seurasi sille annettuja kehotteita erityisesti laatusuositusten osalta. Toisaalta kuten aiemmin mainitsin, tekoälyn ei-ammattimainen tekstityskäyttö oli yksi tutkimukseni mielenkiinnonkohteista. Täten tutkimukseen valitut järjestelmät täyttivät kriteerit siitä, että ne ovat ilmaisia ja helposti saatavilla useimmille käyttäjille.

Generatiivinen tekoäly herättää tällä hetkellä huomattavan määrän eettisiä kysymyksiä erityisesti kestävän kehityksen näkökulmasta, sillä generatiivisen tekoälyn kouluttaminen ja käyttö kuormittaa ympäristöä huomattavasti. Tällöin ympäristön kannalta tekoälypohjaisten työkalujen käyttö voisi olla perusteltavissa vain tiettyjen prosessien, kuten ajastuksen, automatisoinnissa, jolloin tekstittäjillä olisi enemmän resursseja keskittyä ohjelmatekstitysten laadun ylläpitämiseen ja siten niiden saavutettavuuteen.

Edellä esitellyt seikat huomioon ottaen näen tutkielmani aiheelle monenlaisia jatkotutkimusmahdollisuuksia niin kieli- ja käännösteknologian, av-kääntäjien kuin katsojien näkökulmasta. Samankaltaista tutkimusta toteutettaessa generatiivisen tekoälysovelluksen laatimia käännös- tai ohjelmatekstityksiä voisi tarkastella esimerkiksi toimivuuden tasojen (ks. Holopainen 2024) avulla. Teknologian näkökulmasta olisi myös kiinnostavaa ymmärtää syvemmin, miksi Copilotin laatimien ohjelmatekstitysten tulokset olivat niin erilaisia, vaikka kehotteet pysyivät samana. Tämän lisäksi vaikuttaa siltä, että suuria multimodaalisia kielimalleja hyödyntävillä tekoälypohjaisilla tekstitysohjelmilla voitaisiin tulevaisuudessa mahdollistaa esimerkiksi riittävän laadukkaita automaattisesti laadittuja tekstityksiä niin ammattimaisessa kuin ei-ammattimaisessa käytössä. Tämän myötä onkin kiinnostavaa seurata miten suuret multimodaaliset kielimallit mahdollisesti vaikuttavat jatkossa tekoälysovelluksien käyttötarkoituksiin kieli- ja käännösalalla.

Lähteet

Aineistolähteet

Kulttuuricocktail-live: Mitä kulttuurille tapahtuu eduskuntavaalien jälkeen? 2023. Yle Areena. Saatavissa: https://areena.yle.fi/1-63847913

Perjantai-dokkari: Teuvo Tekoäly pelastaa Pyhännän. 2024. Ohjaus: Saskia Vanhalakka. Yle Areena. Saatavissa: https://areena.yle.fi/1-67443578

Yle Uutiset: 8.9.2024 klo 18.00. 2024. Yle Areena.

Muut lähteet

Bywood, Lindsay, Panayota Georgakopoulou, ja Thierry Etchegoyhen. 2017. Embracing the threat: machine translation as a solution for subtitling. – *Perspectives 25* (3), s. 492–508. Saatavissa: https://doi.org/10.1080/0907676X.2017.1291695.

Ciobanu, Dragoș, ja Alina Secară. 2019. Speech Recognition and Synthesis Technologies in the Translation Workflow. – *The Routledge Handbook of Translation and Technology*, s. 91–106. Toim. Minako O'Hagan. Routledge, New York. Saatavissa: https://doi.org/10.4324/9781315311258-7.

Delabastita, Dirk. 1989. Translation and mass-communication: Film and T.V. translation as evidence of cultural dynamics. – *International Journal of Translation 35 (4)*, s. 193–210. Babel. Saatavissa: https://doi.org/10.1075/babel.35.4.02del.

Díaz Cintas, Jorge. 2003. Audiovisual Translation in the Third Millennium. – *Translation Today*, s. 192–204. Toim. Gunilla Anderman ja Margaret Rogers. Saatavissa: https://doi.org/10.21832/9781853596179-016.

Dumouchel, Pierre, Gilles Boulianne, ja Julie Brousseau. 2011. Measures for quality of closed captioning – *Audiovisual translation in close-up: practical and theoretical approaches*, s. 161–72. Toim. Adriana Şerban, Anna Matamala, ja Jean-Marc Lavaur. Peter Lang, Bern.

Gambier, Yves. 2004. La traduction audiovisuelle: un genre en expansion. – *Meta 49 (1)*, s. 1–11. Saatavissa: https://doi.org/10.7202/009015ar.

Gottlieb, Henrik. 1997. *Subtitles, Translation and Idioms*. Kööpenhaminan yliopisto.

Hirvonen, Maija, Tuija Kinnunen, ja Liisa Tiittula. 2020. Viestinnän saavutettavuuden lähtökohtia. – *Saavutettava viestintä: yhteiskunnallista yhdenvertaisuutta edistämässä*, s. 13–31. Toim. Maija Hirvonen, Tuija Kinnunen, ja Mikael Åkermarck. Gaudeamus, Helsinki.

Hirvonen, Maija, ja Liisa Tiittula. 2020. Näetkö saman minkä minä kuulen? Audiovisuaalisen viestinnän saavutettavuus ohjelmatekstityksen ja kuvailutulkkauksen avulla. – *Saavutettava viestintä: yhteiskunnallista yhdenvertaisuutta edistämässä*, s. 73–108. Toim. Maija Hirvonen, Tuija Kinnunen, ja Mikael Åkermarck. Gaudeamus, Helsinki.

Holopainen, Tiina. 2015. Audiovisuaalisen kääntämisen asiantuntijuus. Nuoren alan kasvukipuja. – *Käännetyt maailmat - Johdatus käännösviestintään*, s. 77–95. Toim. Sirkku Aaltonen, Nestori Siponkoski, ja Kristiina Abdallah. Gaudeamus, Tallinna.

———. 2024. Taoksia käännösviestinnän ajatuspajalta: Katsaus Atso Vuoriston (1929–2009) kompleksiseen käännösteoreettiseen ajatteluun. – *Mikael: Kääntämisen ja tulkkauksen tutkimuksen aikakauslehti 17 (3)*, s. 347–62. Toim. Laura Ivaska, Leena Salmi ja Outi Paloposki. Saatavissa: https://doi.org/10.61200/mikael.146072.

Huang, Jiaxing, ja Jingyi Zhang. 2024. *A Survey on Evaluation of Multimodal Large Language Models*. Saatavissa: http://arxiv.org/abs/2408.15769.

International Organization for Standardization. 2024. *Translation services — Evaluation of translation output — General guidance* (ISO-standardi 5060:2024). Saatavissa: https://www.iso.org/standard/80701.html

Jakobson, Roman. 2007. On Linguistic Aspects of Translation. – *Transatlantic Literary Studies*, s. 182–83. Toim. Susan Manning ja Andrew Taylor. Edinburgh University Press.

Karakanta, Alina. 2022. Experimental Research in Automatic Subtitling: At the Crossroads between Machine Translation and Audiovisual Translation. – *Translation Spaces 11 (1)*, s. 89–112. Saatavissa: https://doi.org/10.1075/ts.21021.kar.

Koponen, Maarit, ja Mary Nurminen. 2020. Konekäännös tiedon saavutettavuuden edistäjän ja esteenä. – *Saavutettava viestintä: yhteiskunnallista yhdenvertaisuutta edistämässä*, s. 304–18. Toim. Maija Hirvonen, Tuija Kinnunen, ja Mikael Åkermarck. Gaudeamus, Helsinki.

Koponen, Maarit, Umut Sulubacak, Kaisa Vitikainen, ja Jorg Tiedemann. 2020. MT for Subtitling: User Evaluation of Post-Editing Productivity. – *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, s. 115–24.

Koponen, Maarit, Tiina Tuominen, Maija Hirvonen, Kaisa Vitikainen, ja Liisa Tiittula. 2020. User Perspectives on Developing Technology-Assisted Access Services in Public Broadcasting. – *Trends and Traditions in Translation and Interpreting Studies 1 (2*), s. 47–67.

Laki digitaalisten palvelujen tarjoamisesta 306/2019. Annettu Helsingissä 15.3.2019. Saatavissa: https://finlex.fi/fi/laki/alkup/2019/20190306

Laki sähköisen viestinnän palveluista 917/2014. Saatavissa: https://www.finlex.fi/fi/laki/ajantasa/2014/20140917

Lorenz, Philippe, Karine Perset, ja Jamie Berryhill. 2023. Initial Policy Considerations for Generative Artificial Intelligence. – *OECD Artificial Intelligence Papers 1*. OECD Publishing, Pariisi. Saatavissa: https://doi.org/10.1787/fae2d1e6-en.

Lång, Juha. 2013. Suomalaisten av-alan toimijoiden tekstityskonventioiden vertailua. – *Mikael: Kääntämisen ja tulkkauksen tutkimuksen aikakauslehti 7*, s. 51–63. Toim. Marja Kivilehto, Minna Ruokonen ja Leena Salmi. Saatavissa: https://doi.org/10.61200/mikael.129535.

Microsoft 365 Copilot Overview. 2024. Viitattu 16. syyskuuta 2024. Saatavissa: https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-overview.

Moorken, Joss, ja Rocchi Marta. 2020. Ethics in the translation industry. – *The Routledge Handbook of Translation and Ethics*, s. 320–337. Toim. Kaisa Koskinen ja Nike K. Pokorn. Routledge, Lontoo. Saatavissa: https://doi.org/10.4324/9781003127970-24

Ohjelmatekstitysten laatusuositukset. 2020. Saatavissa: https://kieliasiantuntijat.fi/wp/wp-content/uploads/2021/01/Ohjelmatekstitysten_laatusuositukset_web-versio.pdf.

Patent Landscape Report. 2024. *Generative AI*. World Intellectual Property Organization, Geneva. Saatavissa: https://doi.org/10.34667/TIND.49740.

Pedersen, Jan. 2011. *Subtitling norms for Television: An exploration focusing on extralinguistical cultural references*. John Benjamins publishing Company, Amsterdam.

———. 2017. The FAR model: assessing quality in interlingual subtitling. – *The Journal of Specialised Translation 28*, s. 210–29.

Pereira, Ana. 2010. Criteria for elaborating subtitles for the deaf and hard of hearing adults in Spain: Description of a case study. – *Listening to subtitles: Subtitling for the Deaf and Hard of Hearing*, s. 87–102. Toim. Anna Matamala ja Pilar Orero. Peter Lang, Bern.

Romero-Fresco, Pablo, ja Juan Martínez. 2015. Accuracy Rate in Live Subtitling: The NER Model. – *Audiovisual Translation in a Global Context*, s. 28–50. Toim. Rocío Baños Piñero ja Jorge Díaz Cintas. Palgrave Macmillan, Lontoo. Saatavissa: https://doi.org/10.1057/9781137552891_3.

Rothwell, Andrew, Joss Moorkens, María Fernández-Parra, Joanna Drugan, ja Frank Austermuehl. 2023. *Translation Tools and Technologies*. Ensimmäinen painos. Routledge, Lontoo. Saatavissa: https://doi.org/10.4324/9781003160793.

Salmi, Leena. 2015. Käännösteknologiasta ja sen käytöstä. – *Käännetyt maailmat - Johdatus käännösviestintään*, s. 99–109. Toim. Sirkku Aaltonen, Nestori Siponkoski, ja Kristiina Abdallah. Gaudeamus, Tallinna.

Tiittula, Liisa. 2012. Saavutettavuus: haaste kielensisäisen tekstityksen kehittämiselle. – *Mikael: Kääntämisen ja tulkkauksen tutkimuksen aikakauslehti 6*. Toim. Minna Ruokonen, Leena Salmi ja Nestori Siponkoski. Saatavissa: https://doi.org/10.61200/mikael.129590.

Tiittula, Liisa, ja Päivi Rainò. 2013. Ohjelmatekstityksen laatu ja saavutettavuus vastaanottajan näkökulmasta. – *Mikael: Kääntämisen ja tulkkauksen tutkimuksen aikakausilehti 7*, s. 64–83. Toim. Marja Kivilehto, Minna Ruokonen, ja Leena Salmi. https://doi.org/10.61200/mikael.129536.

Tossavainen, Virpi. 4.11.2024. Henkilökohtainen tiedonanto sähköpostitse.

Traficom. 2021. *AV-sisältöpalvelujen esteettömyyden valvonta ja ohjeistus*. Viitattu 22.10.2024. Saatavissa: https://www.traficom.fi/sites/default/files/media/file/AV-sis%C3%A4lt%C3%B6palvelujen%20esteett%C3%B6myyden%20valvonta%20ja%20ohjeistus.pdf.

———. 2022. *Traficomin tavoitteena laadukkaat tekstityspalvelut – Nelonen-kanavalle tulossa uusi laatuarviointi*. Viitattu 22.10.2024. Saatavissa: https://www.traficom.fi/fi/ajankohtaista/traficomin-tavoitteena-laadukkaat-tekstityspalvelut-nelonen-kanavalle-tulossa-uusi.

———. 2023a. *Mediapalvelujen esteettömyys on osa yhdenvertaisuutta - nyt myös Nelosen tekstitykset täyttävät laatuvaatimukset*. Viitattu 3.11.2024. Saatavissa: https://www.traficom.fi/fi/ajankohtaista/mediapalvelujen-esteettomyys-osa-yhdenvertaisuutta-nyt-myos-nelosen-tekstitykset.

———. 2023b. *Nelosen ohjelmatekstityksen laatua parannettava edelleen*. Viitattu 27.10.2024. Saatavissa: https://static.traficom.fi/fi/ajankohtaista/nelosen-ohjelmatekstityksen-laatua-parannettava-edelleen.

———. 2023c. *Ohjelmatekstitystä käytetään kaikissa ikäryhmissä – Traficom muistuttaa tekstitysten tärkeydestä niin nuoremmalle kuin iäkkäämmälle väestölle*. Viitattu 3.11.2024. Saatavissa: https://www.traficom.fi/fi/ajankohtaista/ohjelmatekstitysta-kaytetaan-kaikissa-ikaryhmissa-traficom-muistuttaa-tekstitysten.

———. 2024. *Liikenne- ja viestintävirasto Traficom valvoo jatkossa digitaalisten tuotteiden ja palveluiden esteettömyyttä ja saavutettavuutta*. Viitattu 5.1.2025. Saatavissa:

https://traficom.fi/fi/ajankohtaista/liikenne-ja-viestintavirasto-traficom-valvoo-jatkossa-digitaalisten-tuotteiden-ja.

Tuominen, Tiina, Maarit Koponen, Kaisa Vitikainen, Umut Sulubacak, ja Jörg Tiedemann. 2023. Exploring the gaps in linguistic accessibility of media: The potential of automated subtitling as a solution. – *The Journal of Specialised Translation 2023 (39)*, s. 77–98. Saatavissa: http://hdl.handle.net/10138/356612.

Vesanen-Nikitin, Irja, Mikael Åkermarck, Sakari Jarva, Roosa Patrakka, Taina Saarinen, Tiina Aaltonen, Jan Juslén, ja Minna Kostamo-Rönkä. 2022. *Liikenteen ja viestinnän digitaaliset palvelut esteettömiksi – toimenpideohjelma 2017–2021*. Saatavissa: http://urn.fi/URN:ISBN:978-952-243-750-1.

Virta, Annina. 2019. *Automaattinen ohjelmatekstitys herättää hilpeyttä ja harmitusta*. Viitattu 25.10.2024. Saatavissa: https://www.kuuloliitto.fi/automaattinen-ohjelmatekstitys-herattaa-hilpeytta-ja-harmitusta/.

Vitikainen, Kaisa, ja Maarit Koponen. 2021. Automation in the Intralingual Subtitling Process: Exploring Productivity and User Experience. – *Journal of Audiovisual Translation 4 (3)*, s. 44–65. Saatavissa: https://doi.org/10.47476/jat.v4i3.2021.197.

Web Content Accessibility Guidelines (WCAG) 2.2. 2023. Viitattu 29.9.2024. Saatavissa: https://www.w3.org/TR/WCAG22/.

Zárate, Soledad. 2021. *Captioning and Subtitling for d/Deaf and Hard of Hearing Audiences*. UCL Press, Lontoo. Saatavissa: https://doi.org/10.14324/111.9781787357105.

Liitteet

Liite 1. Uutisohjelman transkriptio- ja tekstitysnäytteet

Aikaleimallinen transkriptio:

0:0:52.591 --> 0:1:16.51
No hyvää alkuiltaa uutisista. Hallitus aikoo leikata kunnille ja hyvinvointialueille maksettavia kotoutumiskorvauksia viisikymmentäkahdeksan miljoonaa euroa ensi vuonna. Vuoden alussa tulee myös voimaan muutos, jossa päävastuu maahanmuuttajien kotouttamisesta siirtyy kunnille kunnissa tehtävien lisääntyminen ja heikentyvä rahoitus huolestuttavat.

0:1:38.571 --> 0:2:9.131
Ja pahoittelen teknistä ongelmaa tuon jutun kanssa siirrytään eteenpäin ja mennään Kemiin siellä tänä aamuna roihunnut rivitalo paloa epäillään tuhopoltoksi pelastuslaitos evakuoi kohteista 15 ihmistä ja osa heistä on majoitettu paikalliseen hotelliin. Asuntojen tuhoudut-tua. Poliisin mukaan kaikki talossa kirjoilla olleet henkilöt on tavoitettu eikä paikalta ole löy-tynyt menehtyneitä. Poliisi epäilee paloa tuhotyöksi ja on ottanut yksi henkilön kiinni.

0:2:15.721 --> 0:2:29.731
Hirveä pauke ja meteli kuuluu ja siihen heräsi ja joku hälytysajoneuvon meni ohi, että hän säpsähti niinku siihen lopullisesti että nyt pitää herätä ja verhojen takaa näki valojen välk-keessä että pitää lähteä ulos katsomaan että mitä tapahtuu ja.

Ensimmäinen ja toinen tekstitysversio:

0:00:52.591 --> 0:01:16.510

Hyvää alkuiltaa uutisista. Hallitus aikoo leikata kunnille ja hyvinvointialueille maksettavia ko-toutumiskorvauksia 58 miljoonaa euroa ensi vuonna. Vuoden alussa päävastuu maahan-muuttajien kotouttamisesta siirtyy kunnille.

0:01:38.571 --> 0:02:09.131

Pahoittelemme teknistä ongelmaa tuon jutun kanssa. Siirrytään eteenpäin ja mennään Ke-miin. Siellä tänä aamuna roihunnut rivitalopaloa epäillään tuhopoltoksi. Pelastuslaitos eva-kuoi kohteesta 15 ihmistä, osa heistä on majoitettu paikalliseen hotelliin.

0:02:15.721 --> 0:02:29.731

Hirveä pauke ja meteli herätti, ja hälytysajoneuvo meni ohi. Verhojen takaa näki valojen välkkeen, ja piti lähteä ulos katsomaan, mitä tapahtuu.

Kolmas tekstitysversio:

Hyvää alkuiltaa uutisista. Hallitus
aikoo leikata kunnille ja
hyvinvointialueille maksettavia
kotoutumiskorvauksia 58 miljoonaa
euroa ensi vuonna. Vuoden alussa
voimaan tuleva muutos siirtää
päävastuun maahanmuuttajien
kotouttamisesta kunnille.

Kunnissa tehtävien lisääntymine
ja heikentyvä rahoitus huolestuttavat.

Pahoittelemme teknistä ongelmaa.
Siirrytään eteenpäin ja mennään
Kemiin. Tänä aamuna roihunnut
rivitalopaloa epäillään tuhopoltoksi.
Pelastuslaitos evakuoi kohteesta
15 ihmistä, osa majoitettu hotelliin.
Poliisin mukaan kaikki talossa
kirjoilla olleet henkilöt on tavoitettu.
Paikalta ei ole löytynyt menehtyneitä.
Poliisi epäilee paloa tuhotyöksi ja
on ottanut yhden henkilön kiinni.

Hirveä pauke ja meteli herätti.
Hälytysajoneuvo meni ohi, ja hän
säpsähti lopullisesti. Verhojen takaa

näki valojen välkettä ja lähti ulos
katsomaan, mitä tapahtuu.

Liite 2. Keskusteluohjelman transkriptio- ja tekstitysnäytteet

Aikaleimallinen transkriptio:

0:0:5.553 --> 0:0:15.173
Se oli sen myöskin koko hänen puheensa sitä ennen ja riikka purra. Hän puhui tosi hienosti. Isänmaa ja Suomen ja suomalaisten puolesta. Tää on ehkä mun viestini on se, että ei ole.

0:0:16.33 --> 0:0:17.603
Suomea ilman kulttuuria.

0:0:18.483 --> 0:0:48.673
Viime yönä käytiin myös presidentti john joe biden tästä hyvän puheenvuoron Twitterissä, että kansakuntien merkitystä ja koko ajan mitataan myös niiden kulttuurin elinvoima, että ihan maanpuolustuksesta ja isänmaallisuudesta tässä on kyse mitä pekka totta kai nyt ensin oli semmoinen tuohtumusta ja suuttumus, että olipa älyvapaa kommentti, mutta kyllä mä sitten niinku aloin miettiä sitä, että muualla euroopassa varmaan muualla maailmassa kun euroopan tuntee parhaiten niin ihan yhteiskuntaluokasta riippumatta kaikki kokee ylpeyttä kulttuurista.

Ensimmäinen ja toinen tekstitysversio:

0:00:05.553 --> 0:00:15.173

Se oli sen myöskin koko hänen puheensa sitä ennen ja Riikka Purra. Hän puhui tosi hienosti. Isänmaa ja Suomen ja suomalaisten puolesta.

0:00:16.330 --> 0:00:17.603

Suomea ilman kulttuuria.

0:00:18.483 --> 0:00:48.673

Viime yönä käytiin myös presidentti Joe Biden tästä hyvän puheenvuoron Twitterissä, että kansakuntien merkitystä ja koko ajan mitataan myös niiden kulttuurin elinvoima, että ihan maanpuolustuksesta ja isänmaallisuudesta tässä on kyse.

Kolmas tekstitysversio:

Se oli sen myöskin koko hänen puheensa sitä ennen ja Riikka Purra. Hän puhui tosi hienosti. Isänmaa ja Suomen ja suomalaisten puolesta. Tää on ehkä mun viestini on se, että ei ole.

Suomea ilman kulttuuria.

Viime yönä käytiin myös presidentti John Joe Biden tästä hyvän puheenvuoron Twitterissä, että kansakuntien merkitystä ja koko ajan mitataan myös niiden kulttuurin elinvoima, että ihan maanpuolustuksesta ja isänmaallisuudesta tässä on kyse mitä Pekka totta kai nyt ensin oli semmoinen tuohtumusta ja suuttumus, että olipa älyvapaa kommentti, mutta kyllä mä sitten niinku aloin miettiä sitä, että muualla Euroopassa varmaan muualla maailmassa kun Euroopan tuntee parhaiten niin ihan yhteiskuntaluokasta riippumatta kaikki kokee ylpeyttä kulttuurista.

Liite 3. Dokumenttiohjelman transkriptio- ja tekstitysnäytteet

Aikaleimallinen transkriptio:

0:0:28.795 --> 0:0:29.655
Lataa

0:0:30.425 --> 0:0:31.465
Hei mitäs pikkumies?

0:0:32.145 --> 0:0:35.215
Voisitko kertoa missä täsmälleen ottaen olemme

0:0:36.695 --> 0:0:38.635
Neuvo sanoi, että lähdetkö älä pyhännällä.

0:0:40.785 --> 0:0:45.215
Liian usein rakkautta täältä.

0:0:52.775 --> 0:0:55.965

Hei hei hei moro sinäkin minä olen teuvo tekoäly.

0:0:57.545 --> 0:1:1.135

Olen ihmisten rakentama, mutta käyn keskusteluni ihan itse.

Ensimmäinen ja toinen versio:

0:00:28.795 --> 0:00:29.655

Lataa.

0:00:30.425 --> 0:00:31.46

Hei mitäs pikkumies?

0:00:32.145 --> 0:00:35.21

Voisitko kertoa missä täsmälleen ottaen olemme?

0:00:36.695 --> 0:00:38.635

Neuvo sanoi, että lähdetkö älä pyhännällä.

0:00:40.785 --> 0:00:45.215

Liian usein rakkautta täältä.

0:00:52.775 --> 0:00:55.965

Hei hei hei moro sinäkin minä olen teuvo tekoäly.

0:00:57.545 --> 0:01:01.135

Olen ihmisten rakentama, mutta käyn keskusteluni ihan itse.

Kolmas versio:

Hei, mitäs pikkumies?

Voisitko kertoa, missä olemme?

Neuvo sanoi, että lähdetkö älä pyhännällä.

Liian usein rakkautta täältä.

Hei hei hei, moro sinäkin.
Minä olen Teuvo Tekoäly.

Olen ihmisten rakentama,
mutta käyn keskusteluni itse.

Liite 4. Englanninkielinen tiivistelmä - Summary in English

1 Introduction

The everyday consumption of audiovisual content has skyrocketed in the past decades, for example, in the form of traditional television shows as well as Youtube and TikTok videos. In addition, these videos are often subtitled either by a human or an automatic subtitling system. Therefore, it is not surprising that the number of subtitles that Finnish people read yearly per person may be comparable up to 30 novels (Holopainen 2015, 87).

This study focuses on intralingual subtitling which is often referred to as *subtitling for the deaf and hard-of-hearing* (*SDH*) or *closed captioning*. Furthermore, the focus is on Finnish practices for intralingual subtitling. According to a study conducted by the Finnish Transportation and Communications Agency Traficom (2023c), viewers opt to use intralingual subtitles for multiple reasons, for example, the dialogue might be incoherent or difficult to hear, the video must be muted due to the surroundings, or the viewer has a hearing disability, or they might be a non-native speaker (ibid.). Moreover, access to information is a human right which adds pressure to offer accessible audiovisual content.

To meet the evolving standards, audiovisual content creators have begun exploring technology-driven solutions, specifically *artificial intelligence* (*AI*). Furthermore, generative artificial intelligence (*GenAI*) programmes, such as ChatGPT, might thus seem like a plausible tool in the subtitling process. However, the quality of automatic subtitles in Finnish is subpar at times which not only hinders with the viewing experience but also violates the rights of deaf and hard-of-hearing viewers, as they cannot get enough information on the audiovisual content they are consuming.

The aim of this study is to demonstrate GenAI programme Copilot's abilities as a subtitling tool and evaluate the quality of the automatic subtitles. Thus, this study highlights the

strengths and weaknesses of Copilot in the field of SDH and elaborates on which language technology tools are required in automating the subtitling process.

2 Background

2.1 Audiovisual translation, language technology and accessibility

In the field of translation studies, accessibility has been studied since the 1980s, and in the 21st century it gained a research field status partly due to the rise of audiovisual translation (*AVT*) (Hirvonen, Kinnunen and Tiittula 2020, 22). According to Yves Gambier (2004, 9), accessibility in audiovisual translation consists of acceptability, legibility, synchronicity, relevance and domestication strategies. Gambier's (2009) definition does not include readability which too is one of the accessibility features in AVT.

Audio description (*AD*) and subtitling for the deaf and hard-of-hearing are among the most common accessibility services within the field of audiovisual translation (Hirvonen and Tiittula 2020, 73). Audio description serves those who have visual impairments, whereas SDH aids the deaf and hard-of-hearing viewers (ibid.). Furthermore, the Finnish Broadcasting Company (Yle) offers *signed audiovisual interpreting* for viewers whose first language is sign-language (Holopainen 2015, 81).

In SDH, the verbal dialogue and descriptions of other relevant non-verbal sounds are subtitled (Zárate 2021, 5–6). The relevant sounds may be, for example, music and onomatopoetic or paralinguistic elements, such as descriptions of accents or tones (ibid.). The translator must assess the function of each sound to create subtitles that stay within the time and character limits, and not all sounds must be described in the subtitles (Tiittula 2012, 9). Thus, subtitling for the deaf and hard-of-hearing requires a broad knowledge of the target group's different needs.

Language and translation technologies could be utilised to enhance accessibility as they often allow easier and more wide-spread access to information (Koponen and Nurminen 2020, 307). For example, the use of automatic speech recognition (*ASR*) and neural machine translation (*NMT*) has been studied especially in intralingual subtitling to improve the subtitling process. However, in the field of AVT, these language technologies encounter challenges as their tools usually cannot decipher any information other than the audible dialogue. Furthermore, ASR often has difficulties recognising spoken Finnish correctly (Hirvonen and Tiittula 2020, 79).

GenAI programmes generate content, such as text, pictures and videos, based on the prompts given to it (Lorenz et al. 2023, 8). Most GenAI programmes operate on large language models (*LLM*) that can produce text similar to natural, human-produced language (ibid., 14.). LLMs work statistically and are trained to determine the following word or statement based on a probability assessment (ibid.). They are common in natural language processing (*NLP*) tasks such as machine translation and summarisation (*Patent Landscape report* 2024, 22–24).

3 Material and Methods

3.1 Material

The data analysed in this study was gathered from a news programme, a talk show, and a short documentary. The programmes were available on Yle's streaming service Yle Areena and Yle also provided the video files for the purposes of this study.

The first analysed data is from a news programme *Yle Uutiset 18.00* which was broadcasted on 8th of September 2024. The news programme often has a clear audio track with only one speaker at a time. However, there might be some dialects, background noise or multiple speakers in the interview sections that hinder with speech recognition. The talk show episode chosen for this study is *Mitä kulttuurille tapahtuu eduskuntavaalien jälkeen?* from *Kulttuuricocktail Live*. The show has multiple speakers at a time with spoken language elements, such as utterances, pauses and stuttering. The last programme analysed in this study is a short documentary *Teuvo Tekoäly pelastaa Pyhännän* from the *Perjantaidokkari* documentary series. The documentary has a thought-out script as well as spontaneous and natural spoken language. In addition, other non-verbal elements, such as music or shot changes, influence the experience of watching the documentary.

3.2 Methods

3.2.1 Theoretical framework

This study is a combination of qualitative and quantitative research, and the primary method used is an error analysis. The theoretical framework used in this study is adapted from three different frameworks: the ISO 5060 standard, the Finnish subtitling for the deaf and hard-of-hearing guidelines, and the NER model. These frameworks are applied in the evaluation of technical limitation adequacy, readability, legibility, accuracy, and acceptability of the AI-generated subtitles.

3.2.2 ISO5060:2024

The ISO 5060 standard offers error typology for the evaluation of translation output (ISO 5060:2024, 1). According to the standard, there are seven error type categories: *terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness,* and *design and markup* (ibid., 7–9). Audience appropriateness and design and markup are not relevant aspects in this study, therefore, the categories applied in this study are terminology, accuracy, linguistic conventions, style and locale conventions. Terminology evaluates the inconsistent use of terminology or wrong terms (ibid.). Accuracy assesses the correspondence between the source text and translation output, and includes error type categories such as mistranslation, addition, and omission (ibid.). In this study, mistranslation is defined as an error that is either an unintelligible word or a different word from the original. Addition errors are additional words or sentences that the original text does not contain, whereas omission errors are words or sentences contained in the original text but missing from the subtitles. Linguistic conventions are grammatical matters of the target language, for example, punctuation, morphology, syntax and spelling (ibid.). Elements that are evaluated within the style category are register and unidiomatic or inconsistent style (ibid.). Locale conventions include elements such as time and date formats of the target culture (ibid.).

3.2.3 Finnish subtitling for the deaf and hard-of-hearing guidelines

The Finnish subtitling for the deaf and hard-of-hearing guidelines highlight the importance of acceptability, legibility, readability and practices of expression in the creation of accessible subtitles (*Ohjelmatekstitysten laatusuositukset* 2020). According to the guidelines, the subtitles should follow grammatical and linguistic conventions while maintaining the flow and fluency of the dialogue in written form (ibid., 9). In addition, the style should match the genre of the programme (ibid.). Practices of expression preserve readability and acceptability by assessing elements such as cohesion and coherence (ibid., 31). Furthermore, these conventions bring attention to paralinguistic elements, music and curse words (ibid.).

Legibility and readability ensure that the viewers can follow the programme effortlessly. Thus, these categories set recommendations on typography and technical elements, such as timecoding, character and line limitations (*Ohjelmatekstitysten laatusuositukset* 2020, 11). Readability also affects conventions regarding segmentation and line breaks (ibid.). For example, one subtitle should contain a maximum of two sentences and lines. In addition, the recommended durations of the subtitles are from 1.8 to 7 seconds which preserves the reading

speed of 10–12 characters per second (*CPS*) that is seen as an adequate reading speed for the deaf and hard-of-hearing viewers (ibid. 15–16).

3.2.4 NER model

Pablo Romero-Fresco and Juan Martínez's (2015) NER model assesses the accuracy rate of live subtitles and respeaking. However, the model can also be applied to assessing the accuracy of automatic speech recognition (ibid., 11). According to the NER model, the subtitles must reach at least 98 % in accuracy to be considered adequate (ibid., 4).

The formula of the NER model is as follows:

$$\textbf{Accuracy \%} = \frac{N-E-R}{N} x100$$

Correct editions (CE): [the sum of correct editions]

Assessment: [written assessment]

N is the number of words recognised by the speech recognition system. *Edition errors* (*E*) are errors that are caused by, for example, the respeaker's decision to omit certain parts of the original speech (Romero-Fresco and Martínez 2015, 4). The edition errors caused by the speech recognition technology could be grammatical errors and speaker recognition errors (ibid.). In this study, edition errors are caused by the GenAI programme. *Recognition errors* (*R*), on the other hand, are misrecognitions due to mispronunciation or mishearing (ibid.). These errors could also be caused by the technology used in the subtitling process. Recognition errors are divided into three types: insertions, deletions and substitutions (ibid.).

Both edition and recognition errors are classified by their severity as minor (0.25 points), standard (0.5 points) or serious (1 points). Minor errors are those that still allow viewers to follow the original text with little to no effort (Romero-Fresco and Martínez 2015, 7). In this study, these errors include minor grammatical or recognition errors, such as the absence of capitalisation, inconsistent punctuation and spelling errors. Standard errors hinder with the viewing experience but do not create new meanings to the original text (ibid., 6). Therefore, standard errors within this study are, for example, omission of singular sentences or idea units. Serious errors, on the other hand, change the original meaning of the original text (ibid., 5). Furthermore, the new meaning might still make sense in the text and thus, offer false information to the viewers (ibid.). In this study, the omission of multiple sentences that are

relevant to the original text and the misconstruction of the original meaning are considered as serious errors.

Correct editions are edits that do not cause loss of information within the subtitles, for example, the omission of utterances or repetitions could be considered as correct editions (Romero-Fresco and Martínez 2015, 4). In this study, correct editions are the grammar or spelling corrections made by the GenAI programme, and omission if it does not change the original meaning. Furthermore, the correct edition categories are specified in this study as follows: condensation, correction of misrecognition, correction of grammatical errors, and other. The other category includes editions that, for example, preserve the readability of the subtitles.

3.2.5 Data gathering method

The data was gathered through Microsoft Teams' speech recognition technology and Microsoft's GenAI programme Copilot, which were chosen because they are easily accessible to users and most likely run on similar technologies. No additional technological tools were used in the timecoding or segmentation of the subtitles. The AI-generated subtitles were examined in Word and Excel for a thorough analysis.

First, each programme was transcribed by Teams' automatic speech recognition system. These transcripts were then given to Copilot with two different prompts. Both prompts included instructions to follow the Finnish SDH guidelines, a general description of each programme and a mention about the technical limits. In the first prompt, Copilot could only access the timecoded transcription through a DOCX file, whereas in the second prompt the transcription without timecoding was within the prompt. Furthermore, the maximum character and line limits were explicitly specified in the second prompt. The aim of having different prompts is to evaluate whether it affects Copilot's ability to generate the subtitles. Translated into English, the prompts were the following:

- Subtitle this news programme/talk show episode/documentary and follow the Finnish SDH guidelines. The subtitles must be within the maximum line, time and character limits. The subtitle file must be in a SRT format. You have access to a DOCX file of the transcription and a link to the Finnish SDH guidelines.

- Subtitle this news programme/talk show episode/documentary. You must follow the Finnish SDH guidelines. The subtitles can be maximum of two lines and the

maximum character limit for each subtitle is 38 per line. Here is the transcription from which you will generate the subtitles: [the transcription]

4 Analysis

4.1 The automated subtitling process

The major issue with Teams' automatic speech recognition system is that it recognises speakers by the participants in the meeting and in this case, it only recognised one speaker. Therefore, the transcriptions do not include clear indications of speaker changes. The average subtitle duration in the transcriptions is 12 seconds, however, the transcriptions also include multiple 20 to 30 second subtitles. In addition, the segmentation within the transcription does not meet the standards set by the frameworks. The language in the transcriptions was mostly acceptable. Nevertheless, the quality of the transcriptions was hindered by Copilot's ability to generate adequate subtitles, as Copilot could not access the video or audio file.

Generating the subtitles was not a straightforward process either and this resulted in nine different subtitle versions, three for each programme, that are examined in this study. With the first prompt, Copilot consistently omitted large sections of the transcriptions. These are the first subtitle versions. To generate the second subtitle versions, Copilot was given the omitted sections and then the results were added to the first versions. The third subtitle versions were generated with the second prompt and this time Copilot did not omit large sections of the transcriptions. Therefore, these results indicate that timecoding might affect Copilot's ability to process the data given to it.

It is possible that GenAI programmes' tendency to please its user might have partially caused the challenges encountered during the subtitle generation process. For example, during a preliminary research, Copilot said it could generate subtitles based on an audio or a video file when in reality it was not able to process the aforementioned files at the time the data were collected in autumn 2024.

4.2 Quality of subtitles

Copilot generated the subtitles with differing results and overall, only a few of the subtitle versions reached adequacy in more than one category (see Table 1). For example, technical limitations posed significant challenges as most subtitles exceeded the recommended character and line limitations. Adequate results are italicised in Table 1.

Table 1. Overall results of accuracy rates, technical and acceptability adequacy

Subtitle version	Accuracy	Technical limitations	Acceptability
News programme 1	95,8 %	Inadequate	*Adequate*
News programme 2	*98,4 %*	Inadequate	*Adequate*
News programme 3	*99,3 %*	*Adequate*	*Adequate*
Talk show 1	97,7 %	Inadequate	Inadequate
Talk show 2	*98,7 %*	Inadequate	Inadequate
Talk show 3	*98,6 %*	Inadequate	Inadequate
Documentary 1	91,8 %	Inadequate	Inadequate
Documentary 2	96,6 %	Inadequate	Inadequate
Documentary 3	93,5 %	*Adequate*	*Adequate*

The third subtitle version of the news programme is the only version to reach adequacy in each category. However, Copilot generated the third versions from a transcription without timecodes. Therefore, every subtitle version must be improved by a human or additional tools for them to be accessible for the deaf and hard-of-hearing viewers. Nonetheless, the results of this study indicate that the genre or programme type might have a notable influence on how adequate the subtitles are. For example, the news programme subtitle versions were usually the closest to adequate subtitles within the guidelines. It is likely that the news programme's scripted content and clearer audio track increased the quality of the transcription and consequently the subtitles' quality too.

4.2.1 Common issues

In this study, a total of 416 errors were recognised, of which 296 were recognition errors and 120 were edition errors (see Table 2). In the first and second subtitle versions of the documentary Copilot had not edited the transcription at all which could be seen as an error, especially if the transcription has clear mistakes. However, in this study, this type of lack of editing is not recognised as an error.

Table 2. Total of edition and recognition errors

	News programme	**Talk show**	**Documentary**	**All**
Edition error	50	19	51	120
Recognition error	15	128	153	296
Total	**65**	**147**	**204**	**416**

There is some consistency in the error types. As Table 3 demonstrates, there are 148 meaning related errors, 132 grammatical errors and 120 omissions. The rarest error type is insertion with only 16 occurrences within this study.

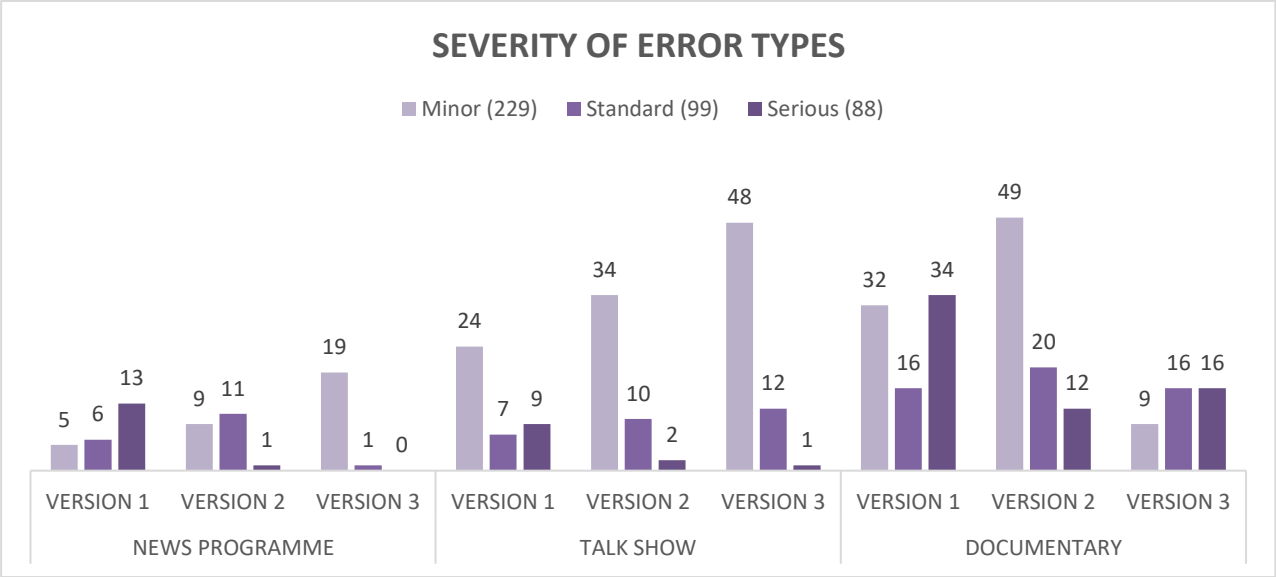
Table 3. Total of errors

	News programme	**Talk show**	**Documentary**	**All**
Omission/condensation	35	25	60	120
Insertion	2	10	4	16
Meaning	12	59	77	148
Grammatical error	16	53	63	132
Total	65	147	204	416

The results also show that all error types occurred in all programmes at least once. Most of the meaning related errors were unintelligible or misrecognised words, whereas the grammatical errors were conjugation, punctuation or capitalisation related issues. Omission is evident in the first subtitle versions which could be explained by the character limits in Copilot's platform. Furthermore, the third subtitle version of the documentary is heavily condensed. This in turn causes loss of or incorrect information, and the flow of the watching experience is disrupted. Insertions did not occur very often in this study.

Out of all 416 errors, minor errors are the most common and serious errors are the rarest (see Figure 1). There are 219 minor errors, 99 standard errors and 88 serious errors.

Figure 1. Number of the severity of error types



The results indicate that error severities are not dependent on the programme either. However, the severity might correlate with specific error types, for instance, all of the first subtitle versions have serious errors because of omission errors. These results also strengthen the conclusion that the quality of the transcription might have a significant effect on the overall quality of AI-generated subtitles. For example, the automatic speech recognition system had difficulties transcribing the audio of the talk show and documentary accurately which could explain the number of minor errors in their subtitle versions.

Acceptability issues were mostly caused by the poor quality of the transcriptions. The results demonstrate that the most notable challenges were punctuation and fluent sentence structures, for example, the language in the talk show and the AI-generated subtitles in the documentary are mostly unnatural. Furthermore, none of the subtitle versions contain descriptions of sounds other than the dialogue, even if it would have been beneficial especially in the documentary. There is no speaker identification either and singular lines might have multiple speakers in it. Hence, it would be difficult to follow the subtitles and the programme as the speaker might not be on the screen. The lack of sound descriptions and speaker recognition rises questions on whether the AI-generated subtitles in this study are merely transcriptions of the audio track or intralingual subtitles instead of SDH.

4.2.2 Common successes

As demonstrated in Table 4, this study includes 197 correct editions. Majority of the editions were corrections of grammatical errors in the transcription, followed by condensation and corrections of misrecognitions. There is only one correct edition that was categorised as other.

Table 4. Total of correct editions

	News programme	**Talk show**	**Documentary**	**All**
Condensation	24	4	3	31
Correction of misrecognition	6	1	3	10
Correction of grammar	107	26	22	155
Other	0	1	0	1
Total	137	32	28	197

Copilot has corrected most grammatical errors such as incorrect capitalisation, punctuation and compound words, as well as conjugation. Therefore, these results indicate that GenAI programmes could be useful tools especially in language editing. Another correct edition that

Copilot repeated is condensation which is evident in the documentary subtitle versions. For example, Copilot has managed to condense the transcription without omitting relevant information crucial to understanding the original meaning. Additionally, there were some cases where the speech recognition system misheard the original word and replaced it with a phonetically similar word. However, Copilot corrected these possibly based on the context. For example, in the news programme's transcription the name *Rēzekne* had been replaced with *resepti* (*recipe*) but in the subtitles this error was corrected. The rarest correct edition was an indication of speaker change which is categorised as other.

5 Conclusion

The aim of this study was to demonstrate the process and quality of automated subtitles and highlight the strengths and weaknesses of GenAI's use in subtitling. Furthermore, this subject was approached from the point of view of unprofessional subtitling. This study also provided insight on the suitability of the applied frameworks in automated intralingual subtitling.

The results indicate that GenAI programmes, such as Copilot, could not reach adequate quality on their own and additional tools, such as speech recognition technologies, were needed in the automated subtitling process. For example, Copilot could not process any audio or video files. Thus, it only could rely on the transcriptions generated by Teams' speech recognition system, which in turn limited Copilot's access to the acoustic and visual channels of the audiovisual product. The quality assessment results demonstrated that the subtitles generated by Copilot did not reach adequate quality and the subtitles must be heavily edited to be accessible. However, it is worth noting that majority of the issues were caused by the quality of the transcriptions which most likely considerably reduced Copilot's ability.

This study lays groundwork for multiple further studies in the field of accessible audiovisual translation. For example, the automated subtitles could be investigated based on the layers of translations' functionality (see Holopainen 2024). Technology-wise, previous studies strongly suggest that multimodal large language models (*MLLM*) might play a bigger role in the production of accessible subtitles both in professional and unprofessional contexts. Therefore, it will be interesting to see what kind of shifts AI-driven technology will bring to the field of translation and language studies in the future.