Konsta Kalenius

KONEKÄÄNTÄMISEN MENETELMÄT ENGLANTI–SUOMI-KONEKÄÄNNÖKSESSÄ

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2020

TIIVISTELMÄ

Kalenius, Konsta

Konekääntämisen menetelmät englanti–suomi-konekäännöksessä

Jyväskylä: Jyväskylän yliopisto, 2020, 40 s.

Tietojärjestelmätiede, kandidaatintutkielma

Ohjaaja: Clements, Kati

Tämä kirjallisuuskatsauksena toteutettu kandidaatintutkielma tarkastelee kolmea keskeistä konekääntämisen menetelmää sääntöpohjainen, tilastollinen ja neuroverkkokonekääntäminen ja niiden tarjoamien englanti–suomi-konekäännösten kehitystä ja ominaisuuksia. Neuroverkkokonekääntäminen vaikuttaa olevan kirjallisuuskatsauksen perusteella nopeasti kehittyvä menetelmä, joka tarjoaa yhä laadukkaampia englanti–suomi-kieliparin konekäännöksiä. Toisaalta Suomessa on vahvat perinteet kieliparin sääntöpohjaisen konekäännöksen kehittämisessä ja menetelmä on yhä relevantti englanti–suomi-konekääntämisessä. Tutkielman tarkastelemien tutkimusten valossa näyttääkin siltä, että neuroverkkokonekääntämisen ja sääntöpohjaisen konekääntämisen vahvuudet yhdistävä hybridikonekäännin olisi ihanteellinen menetelmä ratkaisemaan suomen kieliopin ja sanaston ja englanti–suomi-rinnakkaiskorpusten puutteiden aiheuttamia englanti–suomi-konekääntämisen ongelmia.

Asiasanat: sääntöpohjainen konekääntäminen, tilastollinen konekääntäminen, neuroverkkokonekääntäminen, englanti–suomi-konekäännös

ABSTRACT

Kalenius, Konsta

Machine translations from English to Finnish with three machine translation methods

Jyväskylä: University of Jyväskylä, 2020, 40 pp.

Information Systems, Bachelor's Thesis

Supervisor: Clements, Kati

This Bachelor's Thesis carried out as a literature review focuses on three essential machine translation methods rule-based, statistical and neural machine translation and machine translations from English to Finnish made with these methods and the chronological development and properties of the methods and translations. The literature review shows that neural machine translation is a fast-developing, state-of-the-art method, which offers ever-improving machine translations from English to Finnish. On the other hand, the long-term development of rule-based machine translation from English to Finnish (and vice versa) in Finland means that this older method is still relevant to the language pair. In fact, in the light of the studies reviewed in this thesis it seems that the ideal solution for machine translation from English to Finnish would be a hybrid method which combined the strengths of neural and rule-based machine translation in order to deal with the problems caused by the grammar and vocabulary of Finnish and the deficiencies in English–Finnish parallel corpora.

Keywords: rule-based machine translation, statistical machine translation, neural machine translation, machine translation from English to Finnish

TAULUKOT

SISÄLLYS

1 JOHDANTO

Luonnollisten kielten konekääntämisen tarve on kasvanut voimakkaasti 2000-luvulla kiihtyneen digitalisaation ja globalisaation seurauksena. Konekääntämistä tarvitsevat niin yksittäiset ihmiset, yritykset kuin julkisorganisaatiotkin. Tietokonetta käyttävien suomalaisten talouksien osuus kasvoi tasaisesti vuosina 2000–2009 vuosikymmenen alun 47 prosentista vuosikymmenen lopun peräti 81 prosenttiin. (Koskenniemi ym., 2012.) Suomalaisista 16–89-vuotiaista 90 prosenttia käytti Internetiä vuonna 2019 ja alle 65-vuotiaista nettiä käyttivät melkein kaikki. Samana vuonna 16–89-vuotiaista suomalaisista 79 prosenttia käytti Internetiä monta kertaa vuorokaudessa. (Tilastokeskus, 2019.) Älypuhelinten käytön yleistyminen on kasvattanut entisestään nettiä käyttävien osuutta ja verkon käyttämisen määrää. Koska englanti on Internetin ja yleensäkin kansainvälisen tiedonvälityksen valtakieli ja yhteistä äidinkieltä jakamattomien ihmisten yleisin lingua franca, voidaan vetää johtopäätös, että suomalaiset yksityishenkilöt, yritykset ja organisaatiot tarvitsevat konekääntämisessä eniten englannin ja suomen välisiä konekäännöksiä molempiin käännössuuntiin. Tämä pätee siitäkin huolimatta, että kansainvälisen kielikoulutukseen keskittyvän koulutusyhtiöiden ryhmän EF Education First mukaan suomalaisten englannin kielen taito on tutkituista maailman 100 maasta 7:nneksi paras ja Euroopan 33 maasta 5:nneksi paras (EF EPI, 2019). Edellä todetun englannin kielen valta-aseman vuoksi tässä tutkielmassa keskitytäänkin tarkastelemaan konekääntämisen tärkeimpiä menetelmiä ja niiden tuottamia englanti–suomi-konekäännöksiä. Kielipari on tärkeytensä lisäksi myös siinä mielessä kiinnostava tarkastelun kohde, että kielet ovat lingvistisesti hyvin erilaisia: englanti on germaaniseen kieliperheeseen kuuluva analyyttinen kieli, jossa esimerkiksi lauseen sanajärjestys on merkityksen luomisen kannalta tärkeä, kun taas suomi on suomalais-ugrilaiseen kieliperheeseen kuuluva voimakkaasti agglutinoiva eli taipuva kieli, jossa merkityksiä luodaan sanoihin kiinteästi liitettävillä johtimilla, päätteillä ja liitteillä sanajärjestyksen ollessa vapaampi. Tutkielmaan on valittu tarkasteltaviksi konekäännöksen menetelmiksi sääntöpohjainen, tilastollinen ja

neuroverkkokonekääntäminen, jotka ovat konekääntämisen tärkeimmät menetelmät.

Viestinnän näkökulmasta kielenkääntämisen tarve voidaan jakaa kolmeen käyttötarkoitukseen: kääntäminen tiedon hakemisen, tiedon jakamisen ja vuorovaikutuksen vuoksi. Kolmijako pätee riippumatta siitä, toimiiko kääntäjänä ihminen vai kone. (Nuutila, 2005.) Näihin moninaisiin tarkoituksiin kääntämistä tarvitessa käännöksen täydellinen tarkkuus ja oikeakielisyys ei ole välttämättä ratkaisevaa, vaan sen nopeus ja halpuus (Malmivaara, 2007). Digitalisaation ja globalisaation lisäksi konekääntämisen tarvetta lisääkin nimenomaan se, että konekäännöstä käytetään usein ns. raakakäännöksenä. Tavallinen käyttäjä hyödyntää Google Translaten ja Microsoft Translatorin kaltaista ilmaista konekäännintä raakakäännökseen, jonka avulla hän saa summittaisen käsityksen alkuperäistekstistä. Toinen tyypillinen konekääntämisen sovelluskohde on käyttää konekäännöstä raakakäännöksenä, jonka ammattilaiskääntäjä jälkieditoi julkaisukelpoiseksi käännökseksi. Suomessa erityisesti Maarit Koponen on tutkinut konekäännöksen jälkieditointia mm. sen vaatiman ajallisen ja kognitiivisen panostamisen näkökulmasta käsin. (Koponen, Aziz, Ramos & Specia, 2012; Koponen, 2016; Koponen & Salmi, 2017.) Konekäännöksen jälkieditointi on monelle ammattikääntäjälle nykyään arkipäivää ja se muodostuu tulevaisuudessa todennäköisesti yhä tärkeämmäksi osaksi kääntäjän toimenkuvaa. Tässä tutkielmassa sivutaan konekäännöksen jälkieditointia, joskin lähempi tarkastelu jätetään vähemmälle. Kolmas konekääntämisen sovelluskohteista on käyttää konekäännöstä lopullisen, julkaistavan käännöksen laatimiseen. Konekääntämisen voimakkaasta kehityksestä 2010-luvun puolivälistä lähtien huolimatta konekäännös soveltuu yhä vain harvoin ja rajallisesti lopullisesti julkaistavaksi, ihmisen jälkieditoimattomaksi käännökseksi. Varhainen esimerkki tällaisesta konekääntämisen menestyksekkäästä soveltamisesta on MÉTÉO-järjestelmän käyttö Kanadassa säätiedotteiden kääntämiseen englannista ranskaksi 1980-luvulla (Langlais, Leplus, Gandrabur & Lapalme, 2005.) Konekäännöstä voidaankin soveltaa parhaiten lopulliseksi, julkaistavaksi käännökseksi, jos lähdetekstin aihepiiri ja sanasto on tarkoin rajattua. Jos konekäännöstä käytetään julkaistavan käännöksen tekemiseksi ilman jälkieditointia, saattaa tulos olla tahattoman koominen varsinkin kaunokirjallisuutta käännettäessä. Suomessa nousi syksyllä 2016 otsikoihin suuressa kirjakauppaketjussa myyty William Shakespearen Othellon käännös, jonka suomennoksessa esiintyy mm. lause: ”'Kuningas on kauhea rage', sanoi gloucester.” (Määttänen, 2016.)

Tässä tutkielmassa siis keskitytään tarkastelemaan kieliteknologisista sovelluksista konekääntämistä. Muita kääntämiseen liittyviä sovelluksia ovat esimerkiksi ammattikääntäjien manuaalisen käännöstyönsä nopeuttamiseksi käyttämät käännösmuistit ja termipankit (Lopéz, 2010, s. 18). Konekääntämisen ja manuaalisen kääntämisen lisäksi kieliteknologia tarjoaa toki ratkaisuja laajemminkin, kuten mm. puheentunnistamisen, puhesynteesin ja tulkkaamisen sovelluksia sekä chatbotteja. Kuten jo aiemmin todettiin, konekääntämiseen keskittyminen on sen vuoksi perusteltua, että konekäännin on monen sekä

tavallisen käyttäjän arjessaan että organisaation perustoiminnassaan käyttämä sovellus. Konekääntäminen on myös kiinnostava automaattiseen tietojenkäsittelyyn liittyvä haaste luonnollisen kielen monimutkaisuuden vuoksi. Konekääntäminen on relevantti tutkimuskohde siinäkin mielessä, että se on verrattain vanha alan tutkimus- ja kehityskohde jo 1940-luvulta lähtien.

Tutkielma etenee siten, että kolmessa ensimmäisessä sisältöluvussa käsitellään konekääntämisen menetelmät niiden syntymisen järjestyksessä. Ensimmäinen luku käsittelee sääntöpohjaista konekääntämistä, joka on konekääntämisen vanhimpana menetelmänä eräänlaista klassista konekääntämistä. Sääntöpohjaisen konekääntämisen historian yhteydessä käsitellään hieman konekääntämisen historiaa yleisemminkin. Luvussa esitellään sääntöpohjaisen konekääntämisen alatyypit ja menetelmän vahvuuksia ja heikkouksia yleisellä, kieliriippumattomalla tasolla. Menetelmän englanti–suomi-käännösten vahvuuksia ja heikkouksia käsitellään neljännessä luvussa.

Toisessa sisältöluvussa käsitellään tilastollista konekääntämistä, sen alatyyppejä ja menetelmän yleisiä vahvuuksia ja heikkouksia. Tilastollisen konekääntämisen englanti–suomi-käännösten vahvuuksia ja heikkouksia käsitellään neljännessä luvussa.

Kolmannessa sisältöluvussa käsitellään uusinta konekääntämisen menetelmää neuroverkkokonekääntämistä. Koska neuroverkkoja käytetään konekääntämisen lisäksi yleisemminkin koneoppimisessa ja tekoälyssä, neuroverkkojen toimintaa tarkastellaan myös yleisestä informaatioteknologian näkökulmasta konekääntämisen näkökulman lisäksi.

Neljäs luku tekee kirjallisuuskatsauksen englanti–suomi-kieliparin konekääntämiseen. Tarkasteltavana on 2010-luvun tutkimuksia ja akateemisia esitelmiä, jotka käsittelevät englanti–suomi-konekäännöksiä eri näkökulmista, mm.: 1) käännösvirheiden, 2) kieliparin yleisen konekäännöksen tuen, 3) monimerkityksisten sanojen kääntämisen, 4) käännettävän tekstin ominaisuuksien, 5) käännöksen jälkieditoinnin ja 6) vuosittaisen World Machine Translation -konferenssin käännöskilpailun viime vuosien käännöstulosten kehityksen näkökulmasta. Lopuksi vedetään yhteen havainnot englanti–suomi-konekäännöksen nykytilanteesta ja ennakoidaan mahdollisia kehityskulkuja ja esitetään jatkotutkimuksen kohteita.

2 SÄÄNTÖPOHJAINEN KONEKÄÄNTÄMINEN

2.1 Sääntöpohjaisen konekääntämisen historiaa

Sääntöpohjainen konekääntäminen on vanhin konekääntämisen tekniikka, eräänlainen klassinen lähestymistapa konekääntämiseen. (Tästä eteenpäin lyhenteellä RBMT (=rule-based machine translation) viitataan sekä sääntöpohjaiseen konekääntämiseen yleisesti menetelmänä että yksittäiseen sääntöpohjaiseen konekääntimeen eli käännösjärjestelmään.) Varsinaisia nykyaikaisia digitaalisia konekäännösjärjestelmiä edeltävistä automatisoiduista kääntämisjärjestelmistä mainittakoon lyhyesti neuvostoliittolaisen Petr Petrovitš Trojanskijin ja ranskalais-armenialaisen Georges Artsrounin järjestelmät. (Hutchins, 2005) Molemmat kehittäjät hankkivat tahoillaan patentin järjestelmilleen, joskaan niistä ei tullut erityisen suosittuja. Trojanskij esitteli vuonna 1933 koneensa ”sanojen valitsemiseen ja tulostamiseen kieleltä toiselle kääntämiseen”. Mekaaninen keksintö hyödynsi sanakortteja neljällä eri kielellä, filmikameraa, konekirjoitinta ja reikänauhaa. (Radwanski, 2017, s. 6–7) Artsrounin niin ikään vuonna 1933 patentoima kone ”mekaaninen aivo” ei ollut varsinaisesti tarkoitettu ainoastaan konekääntämiseen, vaan se oli yleisluontoinen tiedontallennus- ja tulostusväline, jonka yhtenä sovelluskohteena nähtiin karkeat sanasta sanaan -käännökset. (Radwanski, 2017, s. 6–7)

Varsinaisen konekääntämisen ja RBMT:n historian voidaan katsoa alkaneen heinäkuussa 1949 yhdysvaltalaisen matemaatikon Warren Weaverin julkaisemasta muistiosta *Translation* (Weaver, 1955, s. 15–23). Muistio määritteli tavoitteita ja metodeja, joilla kieltä voisi kääntää tietokoneen avulla. Weaverin ideat pohjasivat matematiikkaan, tietojenkäsittelytieteeseen, toisessa maailmansodassa saavutettuihin salakirjoitusmenetelmiin ja luonnollisten kielten universaaleihin ominaisuuksiin. Weaverin muistio oli vaikutusvaltainen ja synnytti kiinnostusta konekääntämisen kehittämiseen. (Hutchins, 2000, s. 20.)

7.1.1954 järjestettiin New Yorkissa Georgetownin yliopiston ja IBM:n näytöstilaisuus konekääntämisen mahdollisuuksista. Näytöksessä IBM 701 –tietokone käänsi automaattisesti 60 venäjänkielistä virkettä englanniksi. Koeasetelma oli varsin kontrolloitu, koska virkkeet olivat huolellisesti valikoituja siten, että ne eivät sisältäneet monimerkityksisyyttä. Vaikka Georgetownin yliopiston ja IBM:n näytös ei varsin tarkasti rajatun lähdekielisen tekstiaineksen perusteella täyttäkään nykyisiä vaatimuksia konekäännösjärjestelmän aidolle testaamiselle (ks. esim. (Malki ym. (toim.), 2012, s. 166–168)), näytös herätti suuren kiinnostuksen konekääntämisen mahdollisuuksia kohtaan ja kylmän sodan aikana sai ajan merkittävimmät valtiot suuntaamaan rahoitusta konekääntämisen kehittämiseen.

Jos Georgetownin ja IBM:n näytös vuonna 1954 synnytti suurta innostusta konekääntämisen kehittämistä kohtaan, niin kymmenen vuotta myöhemmin tapahtui täysin päinvastaista. Yhdysvaltojen hallitus perusti vuonna 1964 seitsemän tutkijan komitean arvioimaan yleisesti tietokonelingvistiikan ja erityisesti konekääntämisen edistystä. Kyseinen ALPAC (Automatic Language Processing Advisory Committee) –niminen komitea julkaisi vuonna 1966 raporttinsa, joka suhtautui kielteisesti siihenastiseen konekääntämisen tutkimukseen ja korosti tietokonelingvistiikan perustutkimuksen tarvetta konekääntämisen tutkimuksen sijasta. (Hutchins, 1996, s. 131–135) Raportti sai Yhdysvaltojen hallituksen vähentämään konekääntämisen tutkimukseen suunnattua rahoitusta merkittävästi. Kiinnostus konekääntämistä kohtaan ei kuitenkaan kuollut täysin, ja vuonna 1968 perustettiin konekääntämisen pioneereihin kuuluva SYSTRAN-yritys. Konekääntämisen tutkimus elpyi laajemmin Yhdysvalloissa 1980-luvulla, jolloin mm. tietokoneiden kasvanut laskentateho mahdollisti edistyneempien RBMT:iden kehittämisen. RBMT menetti 1980–90-luvun taitteessa asemansa vallitsevana (ja siihen asti käytännössä ainoana) konekääntämisen tekniikkana, kun tilastollinen konekääntäminen kehitettiin.

2.2 Sääntöpohjainen konekääntäminen menetelmänä

RBMT kuuluu ns. tietämyspohjaisiin konekääntämisen lähestymistapoihin (knowledge-based machine translation approach) erotuksena korpuspohjaisiin konekääntämisen lähestymistapoihin (corpus-driven approach), joita ovat tilastollinen, esimerkkipohjainen ja neuroverkkokonekääntäminen. (Okpor, 2014, s. 160) Tietämyspohjaisella lähestymistavalla tarkoitetaan sitä, että konekääntimen toimintaan tarvittava kielellinen tieto, kuten erilaiset kieliopit ja sanastot, on koodattu manuaalisesti järjestelmään. Korpuspohjaisen lähestymistavan mukaiseen konekääntimeen kielellinen tieto on johdettu korpuksista, jotka ovat laajoja, tyypillisesti miljoonien sanojen, järjestelmällisesti kerättyjä, usein kielikoodattuja, tekstikokoelmia (Tieteen termipankki, 2020).

RBMT:n kehittämiseen tarvitaan monenlaisia sanakirjoja: yksikielisiä sanakirjoja erikseen lähde- ja kohdekielelle, kaksikielisiä lähde- ja kohdekielen

välisiä sanakirjoja sekä mahdollisesti myös monikielisiä sanakirjoja. (Hutchins, 1994, s. 2) RBMT:n kehittämiseen tarvitaan myös monenlaisia kielioppikokoelmia. Ensinnäkin tarvitaan kielen morfologisen analyysin eli muoto-opillisen analyysin kuvaava kielioppi sekä kääntimen lähde- että kohdekielelle erikseen. (Hutchins, 1994, s. 2) Toisekseen tarvitaan kielen syntaksin eli lauserakenteen kuvaava kielioppi lähde- ja kohdekielelle erikseen. Sääntöpohjaisen konekääntimen kehittäminen vaatii siis huomattavaa lingvististä asiantuntemusta informaatioteknologisen osaamisen lisäksi.

RBMT:t voidaan teknisiltä toteutuksiltaan jakaa kolmeen alatyyppiin: 1) suora käännös / sanakirjapohjainen käännös (direct / dictionary-based translation), 2) siirtomenetelmä (transfer-method) ja 3) interlingvaalinen menetelmä (interlingual method). (Hutchins, 1994, s. 1) Seuraavaksi esitellään näitä teknisiä toteutuksia tarkemmin.

Suora käännös on vanhin ja tekniseltä toteutukseltaan yksinkertaisin RBMT:n tekniikka. Sen toimintaperiaate vastaa varhaisen käännösteorian sanasanaisen kääntämisen, suoran ekvivalenssin periaatetta (Vehmas-Lehto, 1999, s. 55). Suora käännös toimii yksinkertaistettuna seuraavasti. Ensimmäisessä vaiheessa järjestelmä analysoi lähdekielisen tekstisyötteen sanat. Sanoista poistetaan mahdolliset kieliopilliset morfeemit, esim. monikon tunnukset, jolloin saadaan sanojen sanakirjassa esiintyvä perusmuoto. Toisessa vaiheessa lähdekielisen kielisyötteen sanojen perusmuotojen kohdekieliset vastineet haetaan kohdekielisestä sanakirjasta. Lopuksi haettuihin kohdekielisiin vastinesanoihin lisätään mahdollisesti tarvittavat kieliopilliset morfeemit, kuten vaikkapa monikon tunnukset tai taivutuspäätteet ja mahdollisesti tehdään muita muokkausoperaatiota.

Vaikka suora käännös onkin nykystandardien valossa alkeellinen ja vanhentuneena pidetty konekääntämisen tekniikka, se soveltuu kääntämään yksinkertaistetulla kielellä kirjoitettuja tekstejä, joissa toistuvat samat sanat. Suoraa käännöstä onkin käytetty menestyksekkäästi esimerkiksi käyttöoppaiden kaltaisten toisteista ja kontrolloitua kieltä sisältävien tekstien kääntämiseen. (Hutchins, 1994, s. 13.)

Siirtomenetelmä (transfer-method) on suoraa käännöstä teknisesti edistyneempi. Siirtomenetelmässä teksti käännetään kolmessa vaiheessa. Ensin lähdekielinen tekstisyöte analysoidaan sen kieliopillisen rakenteen määrittämiseksi. Toiseksi kyseisessä analyysissa saatu lähdekielisen tekstin kieliopillinen rakenne siirretään kohdekielen mukaiseksi kieliopilliseksi rakenteeksi, joka sopii kohdekielisen tekstin tuottamiseen. Kolmannessa vaiheessa tuotetaan varsinainen kohdekielinen käännösteksti. (Hutchins, 1994, s. 2) Siirtomenetelmässä ja jäljempänä kuvattavassa interlingvaalisessa menetelmässä on yhteisiä piirteitä. Keskeinen ero on se, että siirtomenetelmällä toimiva käännin on kieliparikohtainen, eli tehty nimenomaan tietyn lähde- ja kohdekielen välisiin käännöksiin, mahdollisesti molempiin suuntiin. Interlingvaalinen menetelmä mahdollistaa kääntimen kääntää yhdellä teknisellä perustoteutuksella useasta kielestä useaan kieleen, eli sitä kautta monia kielipareja. (Hutchins, 1994, s. 2)

Interlingvaalinen menetelmä (interlingual method) on käsitteellisesti abstraktein ja teknisesti kunnianhimoisin sääntöpohjaisen konekääntämisen tekniikka. (Hutchins, 1994, s. 12) Interlingvaalista menetelmää käyttävä konekäännin kääntää tekstin siirtomenetelmää käyttävän kääntimen tavoin kolmessa vaiheessa. Ensiksi lähdekielinen tekstisyöte analysoidaan sen kieliopillisen rakenteen määrittämiseksi. Toiseksi saadusta kieliopillisesta rakenteesta laaditaan interlingua, kieliriippumaton abstrakti representaatio, joka sisältää tekstin semanttisen ja kieliopillisen merkityksen. Viimeiseksi interlinguasta laaditaan kohdekielen mukainen teksti. Interlingvaalinen menetelmä on teoriassa resursseja säästävä RBMT:n toteutus, koska se mahdollistaa uuden käännöskielen lisäämisen järjestelmään muita RBMT:n menetelmiä vaivattomammin. Käytännössä kuitenkin universaalin, kieliriippumattoman interlinguan määritteleminen on hyvin vaativaa, joitakin kieliä käytettäessä jopa mahdotonta, koska luonnolliset kielet voivat olla rakenteeltaan hyvin erilaisia. Usein siirtomenetelmä onkin interlingvaalista menetelmää mielekkäämpi RBMT:n toteutus kehityskustannuksiltaan. (Hutchins, 1994, s. 12)

2.3 Sääntöpohjaisen konekääntämisen yleisiä vahvuuksia ja heikkouksia

RBMT:n keskeinen vahvuus on, että käännöstulokset ovat verrattain johdonmukaisia (consistent). Toisin sanoen yleensä voidaan ennustaa, millaisessa käyttöyhteydessä, esimerkiksi tekstin aihepiirin ja ominaisuuksien puitteissa, käännös on hyvä. RBMT:ssä kieliopillisten poikkeustapausten käsitteleminen on jäljempänä esiteltävää tilastollista konekääntämistä helpompaa, koska järjestelmään voidaan manuaalisesti lisätä tarvittava lisäsääntö poikkeustapauksen selvittämiseksi. RBMT:n eräs vahvuus on myös sen potentiaali sanaliittojen kaltaisten yhteen käsitteeseen viittaavan monisanaisen ilmaisun, eli eri sanoiksi kirjoitettavan sanaparin tai -ryhmän, jonka merkitysseikat sitovat yhteen, hahmottaminen verrattuna tilastolliseen kääntämiseen. (Monti ym., 2013, s. 26–33)

RBMT:n heikkouksiin kuuluu se, että sen kehittäminen vaatii usein paljon resursseja, eli voi olla hidasta ja kallista. (Sreelekha, 2017, s. 5) RBMT:n kehittäminen vaatii tilastollisen konekääntimen kehittämistä enemmän lingvistien työtunteja, koska kielioppisäännöt koodataan järjestelmään manuaalisesti, kun taas korpuspohjaisten konekäänninten kehittämisessä käytetään pitkälle automatisoitua koneoppimista. Toinen tyypillinen RBMT:n heikkous on se, että sen käännöstulos ei ole yleensä yhtä sujuva kuin tilastollisen konekääntimen. (Sreelekha, 2017, s. 5) Käännös voi sinänsä olla kieliopillisesti oikeellinen, mutta se ei ole kovinkaan sujuvaa kohdekielen rakenteelle tyypillistä kieltä, jolloin käännöksessä näkyy "koneen jälki" esimerkiksi tekstin epäluontevana sanajärjestyksenä.

RBMT:lle on kuitenkin käyttötarkoituksensa, joihin se sopii nykystandardeinkin. RBMT sopii hyvin rajatun aihepiirin ja sanastoltaan ennakoitavien tekstien, kuten jo aiemmin mainittujen sääennusteiden ja käyttöoppaiden kääntämiseen. (Mäkinen (toim.), 2017, s. 39), (Hutchins, 1994, s. 13). RBMT voi olla pakkovalinta konekäännösjärjestelmän menetelmäksi pieniä kieliä käännettäessä. Tällaiset pienet kielet eivät välttämättä ole puhujamäärältään pieniä, vaan niille on tarjolla vähän rinnakkaiskorpusten kaltaisia digitaalisia resursseja, jolloin tilastollisesta konekääntimestä ei tulisi laadukasta. (Irvine, 2013, s. 54)

3 TILASTOLLINEN KONEKÄÄNTÄMINEN

3.1 Tilastollisen konekääntämisen historiaa

Tilastollinen konekääntäminen kuuluu korpuspohjaisiin konekääntämisen tekniikoihin. (Tästä eteenpäin lyhenteellä SMT (=statistical machine translation) viitataan sekä tilastolliseen konekääntämiseen yleisesti menetelmänä että yksittäiseen tilastolliseen konekääntimeen eli käännösjärjestelmään.) Muita korpuspohjaisia konekääntämisen tekniikoita ovat esimerkkipohjainen konekääntäminen ja neuroverkkokonekääntäminen, joista ensin mainittu jätetään tutkielman tarkastelun ulkopuolelle, koska se ei ole suomen kielen konekääntämisessä relevantti.

Ensimmäiset hahmotelmat tilastollisten menetelmien hyödyntämisestä konekääntämisessä esitteli jo edellisessä luvussa mainittu Warren Weaver vuoden 1949 muistiossaan. (Weaver, 1955, s. 15–23) Tuon ajan tietokoneiden laskentateho ei kuitenkaan olisi riittänyt tilastollisten menetelmien soveltamiseen tietokonelingvistiikassa, joten SMT:n varsinainen synty siirtyi. 1980-luvulla kiinnostus konekääntämisen kehittämiseen voimistui 1960–luvun jälkipuoliskon ja 1970-luvun hiljaiselon jälkeen ja koneiden laskentateho kasvoi. IBM:n tutkijat tarttuivat 1980-luvun lopulla Weaverin ideoihin. IBM TJ Watson -tutkimuskeskus käynnisti Candide-projektin, jonka tarkoituksena oli kehittää kokeellinen tilastollisia menetelmiä käyttävä konekäännösjärjestelmä. (Berger ym., 1994) Candide-projektin perustavana ajatuksena oli se, että SMT:n vaatima jopa vuosikymmeniä kestävä kielioppien, sanakirjojen ja käännössääntöjen kirjoittaminen oli kestämätön lähtökohta konekääntämiselle. Koettiin, että koska ammattilaiskääntäjän ”kääntämisalgoritmia” ei voida perusteellisesti mallintaa, tarvitaan uusi lähestymistapa konekääntämiseen. (Berger, 1998) Candide-projektin johtoajatus olikin, että konekäännösjärjestelmä oppisi itse, kuinka käännetään. Candidessa luodussa konekäännösjärjestelmässä koneoppimiseen käytettiin Kanadan parlamentin istuntojen englannin- ja

ranskankielisiä pöytäkirjoja eli ns. Hansard-papereita. Huomionarvoista on, että suuri osa Candide-projektin tutkijoista ei puhunut ollenkaan tai paljonkaan ranskaa, joten koneoppimiseen luotettiin toden teolla. (Berger, 1998)

SMT oli noin 25 vuoden ajan vallitseva konekääntämisen menetelmä 1980–90-luvun vaihteesta 2010-luvun puoliväliin asti. Menetelmä saavutti sellaisen kypsyysasteen, että pohdittiin jopa SMT:n sopivuutta kaunokirjallisuuden kääntämiseen. (Toral ym., 2014) Vuodesta 2016 alkaen neuroverkkokonekääntäminen alkoi syrjäyttää SMT:tä, kun Google alkoi käyttämään käännöskoneessaan suurimpien kielten kääntämiseen neuroverkkopohjaista teknologiaa käytettyään SMT:tä noin 10 vuotta.

3.2 Tilastollinen konekääntäminen menetelmänä

SMT on korpuspohjainen konekäännösjärjestelmä, mikä tarkoittaa sitä, että sen koneoppimisvaiheen kouluttamisessa ja toiminnassa käytetään rinnakkaiskorpuksia. Edellisessä pääluvussa todettiin, että korpus on laaja, yleensä miljoonien sanojen, tekstikokoelma, joka sisältää jossakin aidossa käyttöyhteydessä esiintyviä tekstejä. Rinnakkaiskorpus tarkoittaa saman korpuksen erikielistä versiota. Käytännössä rinnakkaiskorpukset ovat siis keskenään saman tekstikokoelman alkukielinen ja ihmisen kääntämä muunkielinen versio. SMT:iden kehittämiseen yleisesti käytettyjä korpuksia ovat mm. Euroopan unionin istuntojen pöytäkirjoja sisältävä Europarl (Koehn, 2005) ja Yhdistyneiden kansakuntien istuntojen tekstejä sisältävä UNPC (Ziemski ym., 2016). SMT:ihin sopivia rinnakkaiskorpuksia on tarjolla yleisesti myös avoimen lähdekoodin projekteissa, esimerkiksi OPUS-kokoelmassa. OPUS sisältää mm. Suomen valtion lakikokoelman sisältävän Finlex-korpuksen, jota voi hyödyntää suomi–ruotsi-SMT:n ja jossakin määrin myös suomi–englanti-SMT:n kehittämisessä (Tiedemann, 2004).

SMT:n toiminta perustuu kahteen tilastolliseen lähtökohtaan. Ensimmäisen lähtökohdan mukaan mikä tahansa kohdekielinen käännös on jollakin todennäköisyydellä lähdekielisen tekstin sopiva käännös. Tilastollisessa englanti–suomi-konekääntämisessä todennäköisyys sille, että SMT tuottaa sopivan suomenkielisen käännöksen s englanninkieliselle lähdetekstille e voidaan merkitä kaavalla $P(s \mid e)$. Todennäköisyys on väliltä [0, 1] eli 0–100 %, ja SMT valitsee käännösvaihtoehdoista sen, joka on suurimmalla todennäköisyydellä sopiva. Toisessa lähtökohdassa mallia tarkennetaan vaihtamalla käännettävien kielten rooleja ja tarkastellaan, millä todennäköisyydellä englanninkielinen alkuperäisteksti e voisikin olla suomenkielisen käännöstekstin s sopiva käännös eli $P(e \mid s)$. Tästä päästään soveltamaan Bayesin teoreemaa muodossa $P(s \mid e) = P(e \mid s) * P(s) / P(e)$. Kaavassa $P(s)$ viittaa jonkin suomenkielisen sanajonon esiintymisen todennäköisyyteen kaksikielisessä korpuksessa ja P(e) englanninkielisen sanajonon esiintymisen todennäköisyyteen.

SMT:n toiminnassa on keskeistä kaksi alajärjestelmää: käännösmalli ja kohdekielen kielimalli. Käännösmalli mallintaa lähde- ja kohdekielen väliset kielelliset vastaavuudet, jotka johdetaan rinnakkaiskorpuksista. Kohdekielen kielimalli pyrkii tekemään käännöksestä mahdollisimman sujuvan ja kohdekielen rakenteen mukaisen. Kohdekielen kielimalli johdetaan yksikielisestä kohdekielisestä korpuksesta. SMT:n käännösprosessissa on olennaista rinnakkaiskorpusten sisältämien sanojen kohdistaminen toisiinsa (word alignment) (Och ym., 2003), koska käännöstekstissä on saatettu käyttää hyvinkin erilaista sanajärjestystä kuin alkuperäistekstissä, etenkin jos kielet noudattavat erilaista perussanajärjestystä (esimerkiksi SVO = subjekti–verbi–objekti vs. SOV = subjekti–objekti–predikaatti) (Genzel, 2010). SMT:n kehittäminen on nopeampaa kuin RBMT:n, koska sen kehittämisessä käytetään koneoppimista ja se vaatii vähemmän kieliasiantuntijoiden manuaalista työtä. (Sreelekha, 2017)

SMT:itä voidaan luokitella sen perusteella, minkä pituisia kielellisiä yksiköitä ne prosessoivat ja kääntävät kerrallaan. SMT voi olla sana kerrallaan kääntävä sanapohjainen käännin (word-based statistical machine translator) (Koehn, 1999, s. 6), usean sanan kerrallaan kääntävä fraasipohjainen käännin (phrase-based statistical machine translator) (Koehn, 1999, s. 8) tai kokonaisia virkkeitä kerrallaan kääntävä syntaksiperusteinen käännin (syntax-based statistical machine translator). (Yamada ym., 2001)

3.3 Tilastollisen konekääntämisen yleisiä vahvuuksia ja heikkouksia

Aloitetaan tilastollisen konekääntämisen yleisten vahvuuksien ja heikkouksien tarkastelu vertaamalla keskenään SMT:iden alatyyppejä.

Sanapohjaista SMT:tä ei suosita nykyään, koska se ei yleensä tuota hyviä käännöstuloksia. Tekniikan keskeinen ongelma on, että monimerkityksistä sanaa kääntäessään käännin ei pysty ottamaan huomioon ympäröiviä sanoja, jotka vaikuttavat monimerkityksisen sanan disambiguaatioon eli merkityksen valintaan. (Carpuat ym., 2005) Toisin sanoen sanapohjainen SMT ei pysty kunnolla ottamaan huomioon kielellistä kontekstia, joka viime kädessä ratkaisee merkityksen. Tekniikassa saattaa myös esiintyä ongelmia sanaliittoja käännettäessä, koska käännin hahmottaa sanaliiton osasanat semanttisesti erillisinä sanoina sanaliiton tarkoittaman yhteismerkityksen sijaan.

Fraasipohjainen SMT kääntää siis fraasin, eli useampia sanoja kerrallaan. Fraasilla ei tässä tarkoiteta kieliopillista lauseketta kuten nominilauseketta, vaan useamman sanan muodostamaa loogista kokonaisuutta, joka on esimerkiksi muotoa prepositio + substantiivi. Fraasipohjainen SMT pystyy käännösprosessissaan ottamaan paremmin huomioon kontekstin, mikä mahdollistaa paremmat käännöstulokset monimerkityksisiä sanoja käännettäessä (Zens ym., 2002).

Syntaksipohjaisessa SMT:ssä kohdekielen syntaksi- eli lauseopillinen rakenne on täysin mallinnettu ja järjestelmä kääntää fraaseja pidempiä yksiköitä, tyypillisesti virkkeitä, kerrallaan. Syntaksipohjainen SMT voi saavuttaa fraasipohjaista SMT:tä parempia käännöstuloksia, mutta toisaalta se toimii hitaammin. Näin ollen fraasipohjainen SMT tarjoaa parhaimman käännöstarkkuuden ja –nopeuden kompromissin ja on siten ollut pitkään suosituin SMT:n alatyyppi.

SMT sopii erityisen hyvin runsasresurssisten kielten käänninten kehittämiseen. Runsasresurssisella kielellä tarkoitetaan tässä kieltä, jolle on saatavilla paljon digitaalista aineistoa. Tällaisella kielellä on tyypillisesti paljon puhujia ja vahva kansainvälinen asema, kuten esimerkiksi englannilla, espanjalla ja ranskalla. Pieniresurssisella kielellä tarkoitetaan sitä vastoin kieltä, jolla on tyypillisesti vain vähän äidinkielisiä puhujia eikä vahvaa kansainvälistä asemaa. Valtaosa maailman kielistä onkin edellä mainituin kriteerein pieniresurssisia. SMT, jonka lähde- tai kohdekieli on pieniresurssinen kieli, saattaa käyttää kääntämisessään välikielenä englantia tai muuta suuriresurssista kieltä. Pieniresurssisillakin kielillä on usein saatavilla rinnakkaiskorpuksia, joiden toinen kieli on englannin kaltainen valtakieli. Jos käännösprosessissa käytetään kolmatta kieltä välikielenä, todennäköisyys käännösvirheisiin kasvaa, koska SMT:ssä, kuten missä tahansa konekääntämisessä, saattaa suuriresurssisten sukulaiskieltenkin välillä käännettäessä tapahtua käännösvirheitä. RBMT:ssähän kyseistä ongelmaa ei ole, koska käännös tapahtuu lähde- ja kohdekielten välillä ilman kolmatta luonnollista kieltä. (Mahdollista interlinguaa ei lasketa varsinaiseksi kieleksi).

SMT:ssä saavutetaan usein parhaat käännöstulokset keskipitkiä virkkeitä käännettäessä. Lyhyet virkkeet eivät välttämättä rajaa tarpeeksi kontekstia monimerkityksisyyden selvittämiseksi. Hyvin pitkien virkkeiden kääntämisessä esiintyy usein sanajärjestysongelmia, kieliopillisten fraasien sekoittumisia yms. pitkistä virkkeensisäisistä etäisyyksistä aiheutuvia ongelmia. (Vilar ym., 2006)

SMT:n keskeisimpiä ongelmia RBMT:hen nähden on, että käännöslaatu ei ole yhtä hyvin ennustettavissa. Vaikka SMT:iden opettamiseen käytetyt rinnakkaiskorpukset ovat yleensä valtavia tekstimassoja, voi olla vaikeaa ennakoida, mitkä sanat eivät esiinny korpuksissa ja sen seurauksena aiheuttavat käännöstekstissä ns. domeenin eli aihepiirin ulkopuolisina sanoina (out-of-domain words) käännösvaikeuksia. Jos käännettävässä lähtötekstissä käytetään ns. kontrolloitua kieltä (Koehn, 2009, s. 21), joka muistuttaa sanastoltaan ja syntaktiselta rakenteeltaan järjestelmän opettamiseen käytettyjä korpuksia, todennäköisyys hyvään käännöstulokseen on suurempi.

4 NEUROVERKKOKONEKÄÄNTÄMINEN

4.1 Neuroverkkokonekääntämisen historiaa

Neuroverkkokonekääntäminen on uusin konekääntämisen menetelmä. (Tästä eteenpäin lyhenteellä NMT (=neural machine translation) viitataan sekä neuroverkkokonekääntämiseen yleisesti menetelmänä että yksittäiseen neuroverkkokonekääntimeen eli käännösjärjestelmään.) Ensimmäiset NMT:n ratkaisuja esittäneet tieteelliset artikkelit julkaistiin jo 1980–90-luvulla, mutta NMT alkoi kehittyä varsinaisesti vasta 2010-luvun puolivälistä alkaen (Koehn, 2017). Google Translate on käyttänyt konekäännösjärjestelmässään NMT:tä syksystä 2016 lähtien suurimpien kielten kääntämisessä, aluksi rinnakkain fraasipohjaisen SMT:n kanssa, sitten yksipuolisemmin NMT:hen siirtyen (Wu ym., 2016). Suomen kielen kääntämisessä Google Translate on käyttänyt NMT:tä keväästä 2017 lähtien. Neuroverkkoja on käytetty kieliteknologiassa ennen konekääntämistä mm. puheentunnistamiseen. (Sutskever ym., 2014)

4.2 Neuroverkkokonekääntäminen menetelmänä

NMT perustuu neuroverkkoihin, joita käytetään tekoälyssä laajemminkin, esimerkiksi kieliteknologian ulkopuolella visuaaliseen tunnistamiseen (Sutskever ym., 2014). Neuroverkot perustuvat nimestään huolimatta vain löyhästi neuroneihin eli ihmisten hermosoluihin tai ylipäänsä ihmisaivojen toimintaan. Neuroverkot koostuvat tuhansista keinotekoisista yksiköistä, neuroneista. Neuroverkon neuronin toiminta muistuttaa biologista neuronia kuitenkin siinä mielessä, että sen tuottama tuloste tai aktivaatio riippuu ärsykkeestä, jonka se vastaanottaa toisilta neuroneilta ja niiden yhteyksien voimakkuudesta, joita pitkin nämä ärsykkeet kulkevat. Seuraavaksi tarkastellaan NMT:n toimintaa sen eri tasoilla.

NMT käyttää toiminnassaan ns. enkoodaus–dekoodaus-mallia (Cho ym., 2014). Malli sisältää neuroverkkoja kerroksittain: sisääntulokerrokset, piilokerrokset ja ulostulokerrokset. Enkoodaus–dekoodaus-malli toimii perusperiaatteeltaan siten, että sisääntulokerrokset ottavat vastaan ja käsittelevät lähdekielistä kieliainesta, joka piilokerrosten lisäkäsittelyn jälkeen kulkee ulostulokerroksiin, jotka tuottavat kohdekielisen käännöksen. Yksinkertaistaen enkoodari ensin käsittelee lähdekielisen tekstisyötteen, jonka dekoodari sitten kääntää kohdekieliseksi tekstitulosteeksi. NMT:ssä sanoja tai sanoja lyhyempiä jaksoja kuten yksittäisiä merkkejä ja niiden muodostamia jaksoja käsitellään rinnakkaisesti ja hajautetusti. Suurissa neuronisarjoissa käytetään kunkin neuronin aktivaatiotilaa muodostamaan sanojen ja niiden kontekstien hajautettuja representaatioita. (Cho ym., 2014.) Representaatio tarkoittaa tässä neuronien aktivaatiotiloja useamman neuronin muodostamassa erityisessä ryhmässä, kerroksessa. Representaatio koostuu kiinteän kokoisesta sarjasta eli vektorista, joka sisältää neuronien lukuarvoja, esimerkiksi (+0,80; –0,10; +0,23; –0,05; +0,21; …). Käännös luodaan käyttäen näitä representaatioita. Vektorit saadaan representoimaan käännettävän kieliaineksen kaltaista monimutkaista tietoa tekemällä niistä hyvin moniulotteisia, selvästi yli arkiajattelulla vielä helposti hahmotettavan yli kolmen ulottuvuuden. On huomionarvoista, että edellä kuvatut representaatiot ovat tavallisesti syviä, eli ne rakennetaan vaiheittain muista pinnallisemmista representaatioista ja kerroksista. Yksi kerros sisältää tavallisesti satoja neuroneja. Kerroksen neuroneille annetut painoarvot kytkevät ne seuraavan kerroksen kaikkiin neuroneihin tietyllä painolla, jolloin neuronien väliset yhteydet mitataan tuhansissa.

Tarkastellaan NMT:n kehittämisestä järjestelmän opetusvaihetta. NMT:n opetusvaiheessa järjestelmä halutaan opettaa lukemaan lähdekielisiä virkkeitä ja muodostamaan niistä neuroneihin hajautettuja representaatioita (eli neuronijoukkojen aktivaatioita), joista muodostetut kohdekieliset käännetyt virkkeet vastaavat mahdollisimman hyvin opetusdatan mallikäännöksiä. Järjestelmän opettamisessa onkin siis keskeistä määritellä sellaiset neuronien välisten yhteyksien voimakkuudet, joilla saavutetaan haluttu käännöstulos. (Neubig, 2017.) Opettamiseen tarvitaan hyvin suuria harjoituskorpuksia; ihannetapauksessa voidaan käyttää vähintään yhtä suuria korpuksia kuin SMT:n opettamiseen. Opettaminen tapahtuu kierroksittain siten, että jokaisen opetuskierroksen jälkeen neuronien painoa muutetaan siten, että järjestelmän tuottama käännös poikkeaa mahdollisimman vähän harjoituskorpuksen mallikäännöksestä. Tätä poikkeamaa kutsutaan virhefunktioksi tai katofunktioksi (Shen ym., 2015). Opetusalgoritmit toistavat edellä kuvattua opetusprosessia niin monta kierrosta, kunnes virhefunktio on minimaalinen tai riittävän pieni.

NMT:n toimintaa voidaan tehostaa erilaisin menetelmin. Enkoodaus–dekoodaus-arkkitehtuurissa käytetään usein lisänä huomiomekanismia (Luong ym., 2015). Huomiomekanismi koostuu omista lisäneuronikerroksista ja -yhteyksistä. Huomiomekanismi toimii siten, että enkooderin tuottaman viimeisimmän representaation lisäksi (esimerkiksi kokonainen virke ”Let's

translate this English sentence.”) se kiinnittää huomiota aiempiin representaatioihin (”Let's”, ”Let's translate” jne.) kumuloituvasti. Huomiomekanismin avulla NMT voi tarpeen vaatiessa kiinnittää erityistä huomiota virkkeen vaikeasti käännettäviin kohtiin.

4.3 Neuroverkkokonekääntämisen yleisiä vahvuuksia ja heikkouksia

NMT:tä on luontevaa verrata ensisijaisesti SMT:hen, koska molemmat ovat korpuspohjaisia konekääntämisen tekniikoita. Miten kyseiset konekääntämisen tekniikat sitten vertautuvat toisiinsa? NMT:n opettaminen vaatii moninkertaisesti aikaa SMT:n opettamiseen nähden, vuorokausista jopa kuukausiin (Dam ym. (toim.), 2018). Toisaalta NMT vaatii vähemmän dataa opetusvaiheessa, koska sen opettamiseen riittää pienempi rinnakkaiskorpus. NMT yleensä kääntää SMT:tä hitaammin, koska enkoodaus–dekoodaus-mallin dekoodausvaihe on tyypillisesti SMT:n suoraviivaisempaa toimintaa hitaampi. Nykyään neuroverkkopohjaisen Google Translaten nopeus saavutetaan mm. moninkertaisella rinnakkaislaskennalla. Tavallisesti NMT kuitenkin vaatii SMT:tä enemmän laskentatehoa. Toisaalta paikalliset NMT:t vaativat vähemmän levytilaa mm. pienemmän rinnakkaiskorpuksen johdosta.
NMT:n käännöksissä on vähemmän morfologisia, syntaktisia ja kongruenssivirheitä (yhden lauseenjäsenen taivutus mukautuu toisen lauseenjäsenen taivutuksen mukaan). NMT sopii paremmin monikielisen konekäännösjärjestelmän menetelmäksi, koska dekoodaus-vaiheen soveltaminen helpottaa useamman kohdekielen käyttöä. NMT suoriutuu Luongin ym. (2014) mukaan SMT:tä huonommin harvinaisista sanoista jättäen ne usein kokonaan kääntämättä.

Kaiken kaikkiaan NMT:tä pidetään nykyään yleisesti edistyneimpänä konekääntämisen menetelmänä, ja se tarjoaa keskimäärin parhaat käännöstulokset (Bentivogli ym., 2016). NMT:n merkittävänä etuna nähdään se, että SMT:n kanssa samoihin käännöstuloksiin päästäkseen NMT ei vaadi yhtä suuria korpuksia kuin SMT. Tämä on suuri vahvuus, kun käännöskielenä on suomen kaltainen pieniresurssinen kieli.

5 KIRJALLISUUSKATSAUS ENGLANTI–SUOMI-KONEKÄÄNNÖKSEN TILANTEESEEN

Tässä luvussa luodaan kirjallisuuskatsaus Suomessa tehtyyn englanti–suomi-konekäännöksen tutkimukseen. Englanti–suomi-konekääntämistä on tutkittu ja kehitetty Suomen ulkopuolellakin mm. sen vuoksi, että suomi on ollut joinakin vuosina mukana käännettävänä kielenä World Machine Translation -konferenssin konekäännöskilpailutehtävissä. Toisaalta suomen kielen konekääntämistä koskeva ulkomainen tutkimus on verrattain fragmentaarista em. soveltaviin tuloksiin tähtäämisen vuoksi, kun taas Suomessa on tehty ja tehdään enemmän aihepiirin perustutkimusta. Sen vuoksi tarkastelu rajataan tässä enimmäkseen kotimaiseen tutkimukseen.

Kirjallisuuskatsauksen tutkimukset ovat 2010-luvulta ulottuen vuosikymmenen alusta sen loppuun, joten katsaus antaa osittaisen kuvan englanti–suomi-konekäännöksen ajallisesta kehittymisestä. Tarkastellun ajanjakson vuoksi vanhimmat tutkimukset rajoittuvat RBMT:hen ja SMT:hen, kun taas uusimmat tutkimukset käsittelevät joko kaikkia kolmea konekääntämisen menetelmää tai yksinomaan NMT:tä. Koska konekääntäminen on monipuolinen teknologinen, lingvistinen ja kielipoliittinenkin ilmiö, on tarkasteluun valittu varsin erilaisiin fokuksiin keskittyviä tutkimuksia. Tutkimukset käsittelevät seuraavia aiheita: englanti–suomi-konekäännösten käännösvirheiden määrä ja tyypit, suomen kielen yleinen digitaalinen tuki ja erikseen konekäännöksen tuki, monitulkintaisten sanojen suomentaminen, käännöstekstien ominaisuuksien vaikutus konekäännöksen laatuun ja WMT-konferenssin konekäännöskilpailun suomen kielen käännöstulokset. Siten tässä luvussa päästään konkreettisten esimerkkien kautta sivuamaan myös sellaisia konekääntämisen teemoja, joita ei kandidaatintutkielman rajallisessa mitassa voida perusteellisemmin käsitellä. Luvun lopussa listataan taulukkoihin tutkimuksissa ilmenneitä englanti–suomi-käännösten vahvuuksia ja heikkouksia konekäännösmenetelmittäin. Listauksissa on havaintoja myös sellaisista tutkimuksista, joita ei esitellä tarkemmin.

5.1 Maarit Koposen tutkimus (2010)

Maarit Koposen tutkimusartikkeli vuodelta 2010 tarkastelee englanti–suomi-konekäännösten laatua käännösten sisältämien virheiden tyypin ja määrän näkökulmasta. (Koponen, 2010.) Koponen käyttää tutkimuksensa konekääntiminä Sunda Systems Oy:n RBMT:n ilmaista verkossa toimivaa demoversiota ja Google Translatea, joka oli tuolloin SMT. Konekäännöksiä verrataan manuaalisesti ihmisen tekemiin referenssikäännöksiin. Käännettävinä tekstiaineistoina toimivat Euroopan yhteisöjen komission Vihreä paperi vuodelta 2009, National Geographic -lehden artikkeli vuodelta 2008 ja Symantecin ohjelmiston käyttöohje vuodelta 2006. Kaikki tekstit ovat siis asiatekstejä. Asiakirja ja lehtiartikkeli ovat kielenkäytöltään monimutkaisempia mm. sisältäen pitempiä virkkeitä (keskimäärin 33 ja 26 sanaa virkkeessä) kuin käyttöohje, joka on kielenkäytöltään yksinkertaisempi sisältäen enimmäkseen käskylauseita ja lyhyempiä virkkeitä (keskimäärin 11 sanaa virkkeessä). Tekstit ovat kokonaispituudeltaan suunnilleen yhtä pitkiä. Koponen jakaa käännösvirheet käsitteitä koskeviin ja käsitteiden välisiä suhteita koskeviin virheisiin. Käsitteitä koskevat virheet jaetaan käännöksessä poistettuihin käsitteisiin, lisättyihin käsitteisiin, kääntämättömiin käsitteisiin, väärin käännettyihin käsitteisiin ja vaihdettuihin käsitteisiin (käännöksen käsite ei ole suora leksikaalinen vastine alkuperäistekstin käsitteelle, mutta voidaan ajatella kontekstissa toimivana korvikkeena). Käsitteiden välisiä suhteita koskevat virheet jaetaan edellisen kaltaisiin tyyppeihin siten, että kukin virhetyyppi jaetaan vielä alatyyppeihin sen mukaan, koskeeko virhe relaation osapuolia vai itse relaatiota.

Koposen evaluaatio on siis varsin analyyttinen, mitä pidetään yleisesti ihmisen tekemän konekääntämisen evaluoinnin etuna automaattiseen evaluointiin nähden (Koehn ym., 2006). Tutkimuksen tulos oli, että Sunda Systems Oy:n RBMT teki yhteensä 289 virhettä, joista 42 % koski käsitteitä ja 58 % käsitteiden välisiä suhteita. Googlen SMT teki lähes kaksinkertaisen määrän virheitä: 516, joista 32 % koski käsitteitä ja 68 % käsitteiden välisiä suhteita. Sundan RBMT:n huomattavasti pienempää virheiden määrää selittänee osaltaan se, että Sunda Systems Oy:n RBMT pohjautuu Kielikone Oy:n jo 1980-luvulta alkaen kehittämään teknologiaan, jota on kehitetty manuaalisesti lingvistien asiantuntemusta hyödyntäen, nimenomaan englanti–suomi-kieliparille optimoiden. Google Translate on puolestaan julkaistu vuonna 2006, eli sitä ei ollut tutkimuksen hetkellä kehitetty vielä yhtä pitkään, ja sille suomi on vain yksi monista käännettävistä kielistä eikä pienenä kielenä erityisen huomion ja optimoinnin kohde. Suomen kieli ei myöskään tarjoa SMT:n kehittämiseen optimaalista määrää rinnakkaiskorpusten kaltaisia kieliresursseja, kuten englannin tai ranskan kaltaiset maailmankielet. Toisaalta konekäännöksen laadun evaluoinnissa virheiden määrä ei sinänsä kerro koko totuutta konekäännösten laadusta. Sujuvat mutta semanttisesti harhaanjohtavat, liian suuressa määrin väärän merkityksen tuottavat, konekäännökset ovat usein käyttäjälle ongelmallisempia kuin vähemmän

sujuvat ja enemmän kielioppivirheitä sisältävät mutta tekstin alkuperäismerkityksen riittävästi välittävät käännökset (Gimenez ym., 2007). Joka tapauksessa Koposen tutkimus tarjoaa virheiden monipuolisella luokittelullaan analyyttisen metodiikan konekäännöksen laadun manuaaliseen eli ihmisen tekemään arviointiin. Yleensä suositaan automaattisia arviointimenetelmiä manuaalisten sijasta, koska automaattiset ovat nopeampia ja halvempia toteuttaa, mutta manuaalinen arviointi antaa yleensä tarkemman tuloksen konekäännöksen kvantitatiivisesta ja kvalitatiivisesta laadusta.

5.2 META-NET-tutkimusryhmän raportti (2012)

META-NET-tutkimusryhmän raportti vuodelta 2012 tarjoaa katsauksen Euroopan unionin virallisten kielten ja siten myös suomen kielen kieliteknologian tukeen. (Koskenniemi ym., 2012.) Raportti tarkastelee kieliteknologian monia sovelluksia: puheentunnistusta, puhesynteesiä, kieliopillista analyysia, semanttista analyysia, tekstin tuottamista ja konekäännöstä. META-NET on Euroopan komission rahoittama huippuosaamisen verkosto, joka vuonna 2012 muodostui 54 tutkimuskeskuksesta 33 Euroopan maassa. META-NET määrittelee tavoitteekseen rakentaa monikielisen tietoyhteiskunnan teknologista perustaa. Raportin suomen kieltä koskevan osuuden on laatinut ryhmä, joka koostuu Helsingin yliopiston ja Kotimaisten kielten tutkimuskeskus Kotuksen tutkijoista.

Raportissaan tutkijaryhmä luokitteli suomen kielen konekäännöksen tuen tilanteen vuonna 2012 heikoimpaan kategoriaan ”heikko tai olematon tuki”. Vertailun vuoksi puheenkäsittelyn suomen tuki luokiteltiin kahta korkeampaan kategoriaan ”kohtuullinen tuki”, tekstianalyysi yhtä korkeampaan kategoriaan ”osittainen tuki” ja puhe- ja tekstiaineistot kategoriaan ”osittainen tuki”. Konekäännöksen tuen luokittelussa otettiin huomioon kaikki silloiset 22 Euroopan unionin virallista kieltä, eli suomi lähdekielenä konekäännettäessä 21 muulle EU-kielelle ja suomi kohdekielenä konekäännettäessä 21 muusta EU-kielestä. Konekäännöksen tuen / laadun arviointimittarina käytettiin automaattista BLEU-mittaria, joka vertaa konekäännöksen ja ihmisen tekemän referenssikäännöksen yhdenmukaisuutta (Papineni ym., 2002). Kaikkien EU-kielten mukaan laskeminen heikensi suomen kielen konekäännöksen tuen luokittelua aina heikoimpaan kategoriaan asti, koska mukana oli harvoin käännettyjä kielipareja, esimerkiksi malta–suomi. Tämän tutkielman fokuksessa oleva kielipari englanti–suomi oli kuitenkin vahvin kielipareista, joissa suomi on konekäännöksen kohdekielenä. Kieliparin konekäännöksen BLEU-pisteytys oli 38,6. Pistemäärä 80 vastaa ammattilaiskääntäjän tekemää käännöstä. Vertailun vuoksi konekäännös sukulaiskieli virosta suomeen sai 37,7 BLEU-pistettä ja konekäännös muista 20 EU-kielestä suomeen jäi pistevälille 25,8–32,4. Konekäännöksen laatu suomesta muihin EU-kieliin käännettäessä oli parempi: suomi–englanti-kielipari sai kohtalaiset 49,3 pistettä ja muut kieliparit suomi

lähdekielenä saivat 19,4–40,6 pistettä. Tutkimusryhmä arvioi RBMT:n olleen SMT:tä laadukkaampaa käännettäessä sekä suomesta että suomeen. Yhteiskunnallis-poliittiseksi syyksi suomen kielen verrattain heikkoon konekäännöksen tukeen arvioitiin Suomen valtion vähentynyt konekäännösteknologian tutkimuksen ja kehittämisen rahoittaminen 1980- ja 90-luvun jälkeen. Lingvistisiksi syiksi arvioitiin suomen kielen suhteellisen vapaa sanajärjestys ja morfologinen monimutkaisuus.

5.3 Erja Salmisen tutkimus (2012)

Erja Salminen tutkii pro gradu -tutkielmassaan vuodelta 2012 monitulkintaisten sanojen konekääntämistä englannista suomeksi. (Salminen, 2012.) Tutkimuksen fokus on relevantti, koska kielen monitulkintaisuus on konekääntämisen vaikeimpia haasteita (Vickrey ym., 2005). Tutkimuksensa koeasetelmassa Salminen käyttää kolmea RBMT:tä SDL Free Translation, Sunda ja TeemaPoint ja kahta SMT:tä Bing Translator Beta ja Google Translate. Käännettävä tekstiaineisto koostuu 50 englanninkielisen monitulkintaisen sanan ympärille luodusta 195 virkeparista, jotka koostuvat virkkeen lyhyestä ja pitkästä muodosta. Virkeparin pitkä virke alkaa samoilla sanoilla kuin lyhyt virke. Käännösten laadun arviointimittarina käytetään tarkkuutta muiden ihmisarvioinnin mittarien kuten sujuvuuden sijasta (Snover ym., 2009). Tarkkuudessa arvioidaan vain tutkittujen monitulkintaisten sanojen käännös ja vertailun vuoksi jokaisesta virkkeestä satunnaisesti poimittu yksitulkintainen sana puolikkaan painoarvolla monitulkintaiseen sanaan nähden. Virkkeen muiden sanojen käännöstä ei arvioida.

Tutkimuksen keskeinen tulos oli, että kaikkien viiden konekäännösjärjestelmän kaikkien käännösten keskimääräinen tarkkuus monitulkintaisten sanojen kääntämisessä oli vain 0,47. Tutkimuksen mukaan siis oli todennäköisempää, että konekäännin käänsi monitulkintaisen sanan väärin kuin oikein. RBMT:iden ja SMT:iden suoriutumisen välillä ei ollut merkittäviä eroja: sääntöpohjainen Sunda oli paras ja tilastollinen Google Translate toiseksi paras. Sääntöpohjainen SDL Free Translation ja tilastollinen Bing Translator Beta olivat kaksi selvästi huonoiten suoriutunutta käännintä. TeemaPoint oli lähellä Google Translaten tasoa. Se, että Bing Translator Betasta käytettiin betaversiota ja muista kääntimistä varsinaista versiota, vaikutti todennäköisesti Bing-kääntimen suoriutumiseen. Käännöstuloksissa ei ollut merkittävää eroa lyhyitä ja pitkiä virkkeitä käännettäessä, mutta pitkiä virkkeitä käännettäessä saavutettiin odotetusti paremmat tulokset pitkien virkkeiden tarjoaman tarkemman kontekstin vuoksi.

Monitulkintaisten sanojen konekääntämistä koskeva tutkimus olisi hyödyllistä toistaa nyt 2020-luvulla, jotta voisi tarkastella NMT:n suoriutumista asiassa. Uudessa tutkimuksessa voisi käyttää enemmän ja vaihtelevampia virkkeitä mm. pituuden, sanaston ja tekstilajin puitteissa, jolloin saataisiin monipuolisempia tutkimustuloksia.

5.4 Ari Gröhnin tutkimus (2019)

Ari Gröhnin pro gradu -tutkielma vuodelta 2019 tarkastelee erilaisten tekstien soveltuvuutta englanti–suomi-NMT:ssä (Gröhn, 2019). Tutkielma on hyödyllinen avaus aihepiiriin, koska NMT:n laatua on tutkittu vain vähän tarkastellen laaja-alaisesti käännettävän tekstin ominaisuuksia. Tutkimuksen NMT:inä toimivat Google Cloud Translation ja suomalaisen Lingsoftin käännin. Gröhn käyttää tutkimuksessaan kolmea muodostamaansa tekstikorpusta eri genreistä: fiktiosta, virallisista kirjeistä ja virallisista dokumenteista. Fiktiokorpus sisältää Sir Arthur Conan Doylen Baskervillen koira -teoksen vuodelta 1902 ja sen Yrjö Veilinin suomennoksen vuodelta 1904. Virallisten kirjeiden korpus sisältää EUR-Lex-korpuksesta valitut 27 kirjeenvaihtoa ja niiden suomenkieliset käännökset. Virallisten dokumenttien korpus sisältää EUR-Lex-korpuksesta valitut 20 tekstiä ja niiden suomenkielistä käännöstä: 10 päätöstä ja 10 raporttia. Jokainen tutkimukseen muodostettu korpus sisältää alkuperäisen englanninkielisen lähdetekstin, ihmiskääntäjän tekemän suomenkielisen referenssikäännöksen ja kahden NMT:n tekemän käännöksen. Korpukset arvioidaan automaattisesti käyttämällä LeBLEUta, joka on laajennos yleisesti käytettyyn BLEU-arviointimenetelmään (Virpioja ym., 2015). Jokaisesta korpuksesta otetaan lisäksi otos, jonka virheet luokitellaan tarkemmin manuaalisesti. Virheluokittelussa tarkastellaan jokaisen korpuksen virhejakaumaa ja verrataan sitä toisiin korpuksiin ja molempien kääntimien välillä. Lisäksi tarkastellaan, onko korpusten keskimääräisten lausepituuksien ja niiden saamien arviointitulosten välillä korrelaatiota.

Molemmat NMT:t suoriutuivat virallisia kirjeitä ja virallisia dokumentteja sisältäneiden korpusten kääntämisestä paremmin kuin fiktiokorpuksen kääntämisestä. Googlen kääntäjä oli kaikkien korpusten kääntämisessä parempi kuin Lingsoftin. Toisaalta tutkimusasetelma on siinä mielessä ongelmallinen, että NMT:iden opettamiseen käytetään yleisesti juuri EU-tekstejä, joita kaksi asiatekstikorpusta sisälsivät. LeBLEUn käyttäminen automaattisena evaluaatiomittarina lienee ollut ongelmallista fiktiokorpuksen konekäännöstä arvioitaessa, koska mittari vertaa konekäännöstä ihmisen tekemään käännökseen ja laskee LeBLEU-arvon käännösten väliseen poikkeamaan perustuen. Ihmiskääntäjä ottaa yleensä fiktion kääntämisessä enemmän vapauksia kuin asiatekstien kääntämisessä: sanajärjestys on vapaampi ja alkuperäistekstin virkkeitä yhdistellään tai jaetaan. Konekääntimen tekemä käännös on yleensä alkuperäistekstin rakenteelle uskollisempi, joten poikkeama ihmisen tekemään käännökseen voi olla suurikin. Tutkimuksessa ei voinut luotettavasti ennustaa käännöstuloksia lauseiden pituuden perusteella. Keskimäärin lyhyempiä lauseita sisältävä fiktiokorpus oli konekääntimille vaikeampi kääntää kuin asiatekstikorpukset. Toisaalta käännöksen laatu keskimäärin parani lauseiden lyhentyessä ainakin kuuteen sanaan asti. Alle kolmen sanan lauseiden kääntäminen osoittautui tutkimuksessa vaikeaksi. Konekääntämisessä on tyypillistä, että mitä vähemmän lauseessa on sanoja, sitä todennäköisemmin käännin kääntää joko paljon oikein tai paljon väärin.

Lyhyen lauseen yksinkertaisempi rakenne voi johtaa hyvään tulokseen mutta toisaalta epäselvempi konteksti huonoon tulokseen.

5.5 Tommi Niemisen esitelmä (2019)

Tommi Niemisen esitelmä konekäännöksestä suomalaisen kääntäjän näkökulmasta ei ole varsinaisesti tieteellinen tutkimus vaan akateeminen esitelmä. Esitelmä kuitenkin otetaan tarkasteluun, koska se tarjoaa näkymän englanti–suomi-konekäännösten kehittymiseen viime vuosina. (Nieminen, 2019) Nieminen on käännöstieteiden tutkija, jonka julkaisuihin on viitattu 342 kertaa tutkielman kirjoitushetkellä Google Scholarin mukaan, joten esitelmä luokiteltakoon validiksi osaksi tätä kirjallisuuskatsausta, vaikka esitelmä ei olekaan formaatiltaan aivan tyypillinen kirjallisuuskatsauksen aineisto. Nieminen tarkastelee osana esitelmäänsä kansainvälisen World Machine Translation -konferenssin konekäännöskilpailun englanti–suomi-kieliparin kilpailutehtävien tuloksia vuosilta 2017–19. Vuoden 2017 tuloksissa voidaan verrata RBMT:llä, SMT:llä ja NMT:llä kilpailussa saavutettuja käännöstuloksia. Vuoden 2018 ja 2019 kilpailun käännökset ovat SMT:llä ja NMT:llä tuotettuja. Seuraavaksi tarkastellaan käännösesimerkkejä tarkemmin. Arvioni ovat omiani.

Esimerkki 1: alkuperäinen käännettävä lause

> ”This isn't a good experience for people and we plan to address the issue,” a Facebook spokesperson said in an e-mailed statement.”

Esimerkkivirkkeen pituuden ei sinänsä pitäisi aiheuttaa pitkille käännettäville tekstijaksoille tyypillistä merkityksen katoamista. Virkkeen lausejärjestys, jossa johtolause tulee ennen lainausta, saattaa olla kääntimelle vaikeampi kuin päinvastainen lausejärjestys, koska esimerkin mukainen lausejärjestys tarjoaa kääntimelle vaikeamman käännöskontekstin. Esimerkkivirke ei sisällä erityisen harvinaisia sanoja. Virke sisältää nominilausekkeita, joiden attribuutit voivat vaikeuksia: ”a **Facebook** spokesperson”, ”an **e-mailed** statement”. (Koehn ym., 2003)

Esimerkki 1: sääntöpohjainen konekäännös

> ”Tämä ei ole hyvä kokemus ihmisille ja me suunnittelemme esittämään kysymyksen, Facebook puhehenkilö sanoi e-mailed julkilausumassa.

RBMT:n käännös ei ole kovinkaan hyvä. Jaksot ”address the issue” ja ”a Facebook spokeperson” on käännetty väärin siten, että niiden merkitys on muuttunut. Käännin ei ole kääntänyt attribuuttina käytettyä verbin partisiippimuotoa ”e-mailed” lainkaan todennäköisesti siksi, että sen käyttö attribuuttina on harvinaisempaa kuin predikaattina.

Esimerkki 1: tilastollinen konekäännös

> ”Ei hyvä kokemus ihmisille ja aiomme puuttua asiaan”, Facebookin tiedottaja sanoi – julkilausumaan.

SMT:n käännös ei ole kovinkaan hyvä. Aloitus on epäselvä ja yleiskielen normien vastainen. Sana ”e-mailed” on pudonnut käännöksestä pois siten, että sen tilalla on vain ajatusviiva. Sana ”julkilausumaan” on väärin taivutettu käännösvirkkeen kontekstissa.

Esimerkki 1: neuroverkkokonekäännös

> ”Tämä ei ole hyvä kokemus ihmisille ja me aiomme käsitellä asiaa”, Facebookin tiedottaja sanoi sähköisessä lausunnossa.

NMT:n käännös on edellisiin käännöksiin nähden kohtalaisen hyvä. NMT on ainoa käännin, joka on edes jotenkin kääntänyt sanan ”e-mailed”, tosin väärin, mutta virheellinen käännös ei muuta tässä alkuperäistä merkitystä ainakaan harhaanjohtavasti.

NMT:n tuottamat englanti–suomi-käännökset ovat kehittyneet huomattavasti vuoden 2017 jälkeen, mitä osoittaa seuraava vuoden 2019 WMT-kilpailun uutissarjan konekäännös. Valitettavasti lähdekielinen teksti puuttuu, mutta käännöksen sujuvuuden voidaan olettaa korreloivan käännöksen tarkkuuden ja yleisen laadukkuuden kanssa.

> Huolimatta siitä, että Katalonian itsenäisyyttä kannattavat puolueet saavuttivat elintärkeän, joskin niukan voiton viime joulukuussa järjestetyissä aluevaaleissa, ne ovat ponnistelleet pitääkseen vauhtia yllä tänä vuonna monien tunnetuimpien johtajiensa ~~ollessa joko itse pakotettuja~~ [pitäisi olla: jouduttua lähtemään] maanpakoon tai [pitäisi olla: ollessa] pidätettyinä odottamassa oikeudenkäyntiä roolistaan kansanäänestyksen järjestämisessä ja sitä seuranneessa itsenäisyysjulistuksessa.

Esimerkin käännettävä virke on erittäin pitkä ja sen polveileva lauseenvastikkeiden ja sivulauseiden käyttö voisi helposti aiheuttaa konekääntimelle vaikeuksia, mutta NMT on suoriutunut käännöksestä erittäin hyvin. Ainoa varsinainen käännösvirhe ”ollessa joko itse pakotettuja” on sekin sellainen, että alkuperäistekstin merkitys välittyy, vaikkakin huonolla englannin kielen rakenteen mukaisella ilmaisulla.

Edellä mainitun esimerkin virke kääntyi pituudestaan ja lausevastikkeiden ja sivulauseiden käytöstään huolimatta hyvin mm. sen vuoksi, että se oli tekstilajiltaan uutistekstinä sanastoltaan verrattain standardoitua kieltä ja rakenteeltaan kohesiivinen. Kieltä kuitenkin käytetään monissa tekstilajeissa vapaammin tajunnanvirran omaisesti, jolloin teksti ei ole niin koherenttia. Vähemmän koherentti teksti tekee virkkeen eri osien väliset suhteet epäselvemmiksi, mikä on omiaan aiheuttamaan vaikeuksia

konekääntimelle (Smith ym., 2016). Seuraava WMT 2019 -kilpailun lähdeteksti oli NMT:llekin erityisen vaikea.

Esimerkki 3: alkuperäisteksti

> And yet, it goes beyond being an informed citizen when you find yourself on hour six of watching a panel of experts debate Bob Woodward's use of "deep background"sourcing for his book "Fear,"Paul Manafort's $15,000 ostrich-leather bomber jacket ("a garment thick with hubris,"The Washington Post said) and the implications of Stormy Daniels's lurid descriptions of Mr. Trump's, um, anatomy.

Esimerkin 3: neuroverkkokonekäännös

> Ja silti, se menee pidemmälle kuin on perillä kansalainen, kun löydät itsesi tunnin kuusi katsomassa paneeli asiantuntijoiden keskustelua Bob Woodward käyttö "syvä tausta"hankinta kirjaansa "Pelko", Paul Manafort n 15000 dollaria strutsi-nahka pommikone takki ("vaate paksu hubris", Washington Post sanoi) ja vaikutukset Stormy Daniels lurid kuvaukset Mr. Trump, um, anatomia.

Konekäännös ei välitä tekstin merkitystä juurikaan. Huomattavaa on se, että käännös sisältää paljon kielioppivirheitä, vaikka NMT yleensä tuottaa kieliopillisesti verrattain oikeellisen tekstin. Toisaalta esimerkin käännös ei runsaine kielioppivirheine luo sitä potentiaalisesti vaarallistakin skenaariota, että käännös sisältäisi sinänsä sujuvaa ja kieliopillisesti oikeellista mutta merkitykseltään aivan päinvastaista tekstiä, mitä alkuperäisteksti ilmaisee.

Niemisen mukaan suomea on pidetty pitkään vaikeana kielenä konekäännettäväksi, koska suomen kielen taivutus on monimutkaista ja suomen kielen sanajärjestys on joustava. Niemisen mukaan WMT-konferenssin konekäännöskilpailun tulosten perusteella suomen taivuttaminen ja sanajärjestys eivät aiheuta vaikeuksia NMT:lle. Nieminen lisää, että suomen kielessä merkityksiä ilmaistaan usein erilaisin rakentein kuin sen yleisimmissä lähde- ja kohdekielissä, mutta erot eivät vaikuta olennaisesti konekäännösten laatuun. Edellä mainitusta Nieminen vetää johtopäätöksen, että suomi ei ole konekääntämiseen erityisen vaikea kieli, vaan sille voidaan tehdä toimivia konekääntimiä helposti kieliriippumattomin menetelmin, kunhan aineistoa on saatavilla riittävästi.

Niemisen havainto WMT-konferenssin käännöskilpailuissa NMT:illä saavutetuista hyvistä ja vuosien mittaan parantuneista englanti–suomi-käännöstuloksista on sinänsä perusteltu. Toisaalta WMT:n kilpailussa käytetään tapahtumaan varta vasten kehitettyjä konekäännösjärjestelmiä, joita ei useinkaan lanseerata kilpailun jälkeen yleiseen kuluttajakäyttöön. Kilpailun tuloksista ei siis voida vetää suoraan johtopäätöstä, että tavallisen käyttäjän ulottuvilla olisi välttämättä edellä esitellyn tasoiset englanti–suomi-konekäännökset. Joka tapauksessa NMT on voimakkaasti kehittyvä teknologia, joka parantaa huomattavasti suomenkin kaltaisen vähäresurssisen kielen konekäännöksen laatua.

TAULUKKO 1 Sääntöpohjaisen konekääntämisen (RBMT) vahvuudet ja heikkoudet englanti–suomi-konekäännöksessä

Vahvuudet	Heikkoudet
– RBMT tekee parhaimmillaan vain yli puolet käännösvirheitä SMT:hen nähden. (Koponen, 2010) – RBMT tekee SMT:tä vähemmän käsitteiden välisiä suhteita koskevia virheitä. (Koponen, 2010) – Paras monitulkintaisten virkkeiden kääntämisestä suoriutunut käännin oli RBMT, joskaan erot RBMT:iden ja SMT:iden välillä eivät olleet suuria. (Salminen, 2012) – RBMT sopii hyvin kieliopillisesti standardien tekstien kuten uutistekstien kääntämiseen. (Hurskainen & Tiedemann, 2017) – RBMT:n toimintaan tarvittavien sääntöjen määrää voidaan pienentää käyttämällä oletuskäännöstä (default interpretation). (Hurskainen, 2018) – RBMT:hen voidaan tarvittaessa lisätä sääntöjä, joilla suoriudutaan paremmin englanti–suomi-konekäännöksen erityishaasteista, esimerkiksi agentin sisältävien passiivilauseiden kääntämisestä. (Hurskainen, 2017) – RBMT:n käännöslaatu on johdonmukaista ja ennustettavaa. (Ashraf & Ahmad, 2015)	– RBMT ei yleensä käännä hyvin kieliopillisesti vähemmän standardinmukaisia tekstejä. (Hurskainen & Tiedemann, 2017) – RBMT ei yleensä suoriudu käännösprosessissaan hyvin erisnimistä ja akronyymeistä. (Hurskainen, 2018) – RBMT:n käännökset ovat yleensä vähemmän sujuvia mm. käännöksessä esiintyvien tarpeettomien pronominien vuoksi. (Koponen, 2019) – RBMT:n toiminnan vaatimien sanakirjojen kehittäminen ja muutosten tekeminen RBMT-järjestelmään on usein kallista. (Okpor, 2014) – RBMT:n kehittäminen englanti–suomi-kääntämiseen vaatii paljon resursseja, koska suomi on morfologisesti monimutkainen ja hyvin erilainen kieli englantiin verrattuna. (Koskenniemi ym., 2012)

TAULUKKO 2 Tilastollisen konekääntämisen (SMT) vahvuudet ja heikkoudet englanti–suomi-konekäännöksessä

Vahvuudet	Heikkoudet
– SMT kääntää NMT:tä paremmin hyvin pitkiä (yli 40-sanaisia) virkkeitä. (Toral & Sánchez-Cartagena, 2017) – SMT:n tyypillisesti heikompaa	– SMT:t jättävät usein sanoja kääntämättä. (Koponen, 2010) – SMT:llä on käännösprosessissaan usein sanankohdistamisongelmia

kääntämistä suomen kaltaisille voimakkaasti taipuville ja ns. matalaresurssisille kielille voidaan parantaa tekniikalla, jossa SMT:n käännöksen perusyksikkö on sanan sijasta morfeemi, kuitenkin samalla kiinnittäen huomiota sananrajoihin. (Luong ym., 2019) – SMT tuottaa RBMT:tä sujuvampia, luonnollisemmalta vaikuttavia käännöksiä, koska se käyttää kohdekielelle räätälöityä erillistä kielimallia. (Koehn, 2009, s. 181–216)	(word alignment), mikä johtaa sanajärjestysongelmiin. – SMT ei yleensä käännä hyvin sanoja, jotka eivät sisälly sen koulutukseen käytettyihin rinnakkaiskorpuksiin. (Daumé III & Jagarlamudi, 2011) – SMT:n käännöksissä esiintyviä ongelmia voi olla vaikea ennustaa järjestelmän rinnakkaiskorpusten mahdollisen epätasaisen kattavuuden vuoksi.

TAULUKKO 3 Neuroverkkokonekääntämisen (NMT) vahvuudet ja heikkoudet englanti–suomi-konekäännöksessä

Vahvuudet	Heikkoudet
– NMT kääntää virallisten kirjeiden ja virallisten dokumenttien kaltaisia asiatekstejä hyvin. (Gröhn, 2019) – NMT käänsi vuoden 2017 WMT-konferenssin englanti–suomi-käännöstehtävän uutistekstejä paremmin kuin osittain vastaavia tekniikoita hyödyntävät SMT:t. (Nieminen, 2019) – NMT voi käyttää toimintansa tehostamiseksi dataa lisääviä (data augmentation) tekniikoita kuten takaisinkäännöstä (back translation). (Östling ym., 2017) – NMT kääntää muita kääntimiä paremmin lauserakenteeltaan monimutkaisia virkkeitä. (Nieminen, 2019) – NMT tekee SMT:tä vähemmän taivutukseen, sanajärjestykseen ja leksikaalisiin valintoihin (puuttuva sana, ylimääräinen sana, väärä sana) liittyviä virheitä. (Toral & Sánchez-Cartagena, 2017) – NMT kääntää SMT:tä paremmin tavanomaisen pituisia (alle 40-	– NMT kääntää lyhyitä lauseita epätasaisesti: joko monta sanaa oikein tai monta sanaa väärin. (Gröhn, 2019) – NMT saattaa tuottaa sujuvuudeltaan uskottavia mutta virheellisiä käännöksiä. (Koehn & Knowles, 2017) – NMT:t vaativat yleensä enemmän laskentatehoa sekä järjestelmän koulutus- että käyttövaiheessa. (Wu ym., 2016)

sanaisia) virkkeitä. (Toral & Sánchez-Cartagena, 2017) – NMT:n käännöksen jälkieditointi vaatii vähemmän teknistä ja ajallista vaivannäköä. (Koponen, 2019) – NMT:n käännöksen jälkieditoinnissa tarvitsee muokata pienempi määrä lauseenosia kuin muilla kääntimillä. (Koponen, 2019)	

6 YHTEENVETO JA JATKOTUTKIMUSAIHEET

Tässä tutkielmassa on tarkasteltu konekääntämisen kolmea keskeisintä menetelmää sääntöpohjainen, tilastollinen ja neuroverkkokääntäminen. Tarkastelussa on ollut menetelmien historia, toteutus, yleiset vahvuudet ja heikkoudet sekä erikseen menetelmien englanti–suomi–käännösten vahvuudet ja heikkoudet. NMT vaikuttaa olevan tässäkin kieliparissa vahvin menetelmä nyt ja tulevaisuudessa, mutta toisaalta suomen kielen morfologinen monimutkaisuus ja englanti–suomi-RBMT:n kehittämisen vahvat perinteet maassamme pitävät RBMT:n varteenotettavana vaihtoehtona. NMT:n ja RBMT:n parhaat puolet yhdistävä hybridikonekäännin olisi ihanteellinen ratkaisu kattamaan suomen kieliopin ja sanaston aiheuttamat englanti–suomi-konekäännöksen erityishaasteet ja toisaalta paikkaamaan suomenkielisten korpusten puutteita. Toisaalta yhä digitalisoituvassa maailmassa tekstien määrä ja sen myötä suomenkielisen korpusaineiston määrä kasvaa, joten korpuspohjaisten menetelmien laatua voidaan kehittää tältä osin.

Tutkielmassa on tarkasteltu englanti–suomi-kieliparin konekääntämisen tilannetta ja kehitystä niillä keinoin, kuin se on käsiteltävissä kirjallisuuskatsauksen keinoin. Aihepiiri olisi omiaan myös itse laaditulle empiiriselle tutkimusasetelmalle. Kirjallisuuskatsaus ei mahdollista kaikkien konekääntämisen kysymysten kovin yksityiskohtaista tarkastelua mutta yleistyksiä, oletuksia ja johtopäätöksiä voidaan tehdä. Konekääntämisen tematiikkaa onkin tarkasteltu eri näkökulmista jonkin rajatumman fokuksen perusteellisemman tarkastelun sijasta. Konekääntämisen akateemista tutkimusta ei useinkaan harjoiteta tietyn kieliparin näkökulmasta käsin, vaan tutkimuksen keskiössä on usein jokin konekäännösjärjestelmän tekninen ratkaisu tai konekääntämistä koskeva yleislingvistinen aihe. Mielestäni kuitenkin englanti–suomi-kieliparin konekääntämisen erityiskysymysten käsittelylle on tilausta konekääntämisen käytön voimakkaan kasvun ja englannin kielen hegemonisen aseman vuoksi. Kysyntää on sekä aihepiirin akateemiselle perustutkimukselle että parempaan konekääntämisen teknologiaan johtavalle soveltavalle tutkimukselle. Teeman nykyisen käsitykseni valossa esitän jatkotutkimuksen aiheeksi ainakin tarkastella

lähdetekstin ominaisuuksien vaikutusta konekääntämisen laatuun. Jos konekääntämisen laatua pystytään luotettavasti ennakoimaan, voidaan ennustaa paremmin mm. käännöksen luotettavuutta sellaisenaan ja julkaistavaksi tarkoitetun konekäännöksen mahdollista jälkieditoinnin tarvetta: aikaa, vaivaa ja kustannuksia. Huolimatta konekääntämisen laadun huomattavasta kasvusta viimeisen noin viiden vuoden aikana uskon, että ihmiskääntäjän ei tarvitse menettää yöuniaan työnsä konekääntimelle menettämisen pelossa aikoihin, vaan ihmiskääntäjän työtehtävät todennäköisesti sisältävät tulevaisuudessa enemmän (raaka)konekäännösten edellyttämää jälkieditointia.

LÄHTEET

Ashraf, N., & Ahmad, M. (2015). Machine translation techniques and their comparative study. *International Journal of Computer Applications*, 125(7), 28. Haettu osoitteesta http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.7439&rep=rep1&type=pdf

Barreiro, A., Redol, R. A., Monti, J., Orliac, B., & Batista, F. (2013). When Multiwords Go Bad in Machine Translation. Haettu osoitteesta http://www.mt-archive.info/10/MTS-2013-W4-Barreiro.pdf

Bentivogli, L., Bisazza, A., Cettolo, M. & Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. arXiv preprint arXiv:1608.04631. Haettu osoitteesta https://www.aclweb.org/anthology/D16-1025.pdf

Berger, A. L., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., ... & Ureš, L. (1994). The Candide system for machine translation. Haettu osoitteesta https://www.researchgate.net/publication/2454978_The_Candide_System_for_Machine_Translation

Berger, A. (1998) Statistical Machine Translation. Haettu 23.03.2020 osoitteesta http://www.cs.cmu.edu/~aberger/mt.html

Carpuat, M. & Wu, D. (2005). Word sense disambiguation vs. statistical machine translation. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics,* 387–394. Haettu osoitteesta https://dl.acm.org/doi/10.3115/1219840.1219888

Chéragui, M. A. (2012). Theoretical overview of machine translation. *Proceedings ICWIT*, 160. Haettu osoitteesta http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.416.1463&rep=rep1&type=pdf#page=176

Cho, K., Van Merriënboer, B., Bahdanau, D. & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259. Haettu osoitteesta https://arxiv.org/pdf/1409.1259.pdf

Dam, H. V., Brøgger, M. N. & Zethsen, K. K. (toim.). (2018). *Moving boundaries in translation studies.* Routledge.

Daumé III, H., & Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* 407-412. Haettu osoitteesta https://www.aclweb.org/anthology/P11-2071.pdf

Education First (EF) (2019). EF English Proficiency Index. Haettu 22.03.2020 osoitteesta https://www.ef.fi/epi/regions/europe/finland/

Genzel, D. (2010). Automatically learning source-side reordering rules for large scale machine translation. *Proceedings of the 23rd International Conference on Computational Linguistics,* 376–384. Haettu osoitteesta https://www.aclweb.org/anthology/C10-1043.pdf

Giménez, J. & Màrquez, L. (2007). Linguistic features for automatic evaluation of heterogenous MT systems. *Proceedings of the Second Workshop on Statistical Machine Translation.* 256–264. Haettu osoitteesta https://dl.acm.org/doi/10.5555/1626355.1626393

Gröhn, A. (2019). *Suitability of Neural Machine Translation for Different Types of Texts* (pro gradu -tutkielma, Helsingin yliopisto). Haettu osoitteesta https://pdfs.semanticscholar.org/e3a9/f7c2f9905446ea27ecb8725872f97fd8bed2.pdf

Hutchins, J. (1996). ALPAC: the (in) famous report. *Readings in machine translation*, *14*, 131-135. Haettu osoitteesta http://hutchinsweb.me.uk/ALPAC-1996.pdf

Hutchins, J. (2005). The history of machine translation in a nutshell. Haettu osoitteesta http://hutchinsweb.me.uk/Nutshell-2005.pdf

Hutchins, J. (2000). *Warren Weaver and the launching of MT. Early Years in Machine Translation.* Amsterdam: John Benjamins.

Irvine, A. (2013). Statistical machine translation in low resource settings. *Proceedings of the 2013 NAACL HLT Student Research Workshop.* (s. 54-61). Haettu osoitteesta https://www.aclweb.org/anthology/N13-2008.pdf

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT summit* (5) 79–86. Haettu osoitteesta http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.459.5497&rep=rep1&type=pdf

Koehn, P. (2017). Neural machine translation. arXiv preprint arXiv:1709.07809. Haettu osoitteesta https://arxiv.org/pdf/1709.07809.pdf

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press

Koehn, P. & Knight, K. (2003). Feature-rich statistical translation of noun phrases. *Proceedings of the 41st Annual Meeting of the association for Computational Linguistics.* 311–318. Haettu osoitteesta https://www.aclweb.org/anthology/P03-1040.pdf

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*. Haettu osoitteesta https://arxiv.org/pdf/1706.03872.pdf

Koehn, P. & Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. *Proceedings on the Workshop on Statistical Machine Translation.* 102-121. Haettu osoitteesta http://homepages.inf.ed.ac.uk/pkoehn/publications/shared-task-wmt2006.pdf

Koponen, M. (2010). Assessing machine translation quality with error analysis. *Electronic proceeding of the KaTu symposium on translation and interpreting studies*. Haettu osoitteesta https://www.sktl.fi/@Bin/40701/Koponen_MikaEL2010.pdf

Koponen, M. (2016). *Machine Translation: Post-editing and Effort Empirical Studies on the Post-editing Process* (väitöskirja, Helsingin yliopisto). Haettu osoitteesta https://helda.helsinki.fi/bitstream/handle/10138/160256/machinet.pdf

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. *Proceedings of WPTP*, 11–20. Haettu osoitteesta http://www.mt-archive.info/AMTA-2012-Koponen.pdf

Koponen, M., & Salmi, L. (2017). Post-editing quality: Analysing the correctness and necessity of post-editor corrections. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, *16*. Haettu osoitteesta https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/439/394

Koskenniemi, K., Lindén, K., Carlson, L., Vainio, M., Arppe, A., Lennes, M. Westerlund, H., Hyvärinen, M., Bartis, I., Nuolijärvi, P. & Piehl, A. (2012). Suomen kieli digitaalisella aikakaudella.

Langlais, P., Leplus, T., Gandrabur, S., & Lapalme, G. (2005). From the RealWorld to Real Words: the METEO case. *EAMT-05, Budapest.* Haettu osoitteesta http://mt-archive.info/EAMT-2005-Langlais.pdf

López, E. I. (2010). *Luonnollisten kielten kääntäminen ja konekäännös – Taustaa, teoriaa ja menetelmiä* (pro gradu -tutkielma, Jyväskylän yliopisto). Haettu osoitteesta https://jyx.jyu.fi/bitstream/handle/123456789/24655/Elina.Lopez.pdf?sequence=1&isAllowed=y

Luong, M. T., Pham, H. & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025. Haettu osoitteesta https://arxiv.org/pdf/1508.04025.pdf

Luong, M. T., Sutskever, I., Le, Q. V., Vinyals, O. & Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. arXiv preprint arXiv:1410.8206. Haettu osoitteesta https://www.aclweb.org/anthology/P15-1002.pdf

Malmivaara, A. (2007). "*Sokea idiootti*" – *Konekääntämisen mahdollisuuksia* (pro gradu -tutkielma, Tampereen yliopisto). Haettu osoitteesta https://trepo.tuni.fi/bitstream/handle/10024/77995/gradu01858.pdf?sequence=1

Määttänen, J. (25.8.2016). Suomalaisen kirjakaupan konekäännetty Shakespeare hämmästyttää netissä: ”Kuningas on kauhea rage”. *Helsingin sanomat*.

Neubig, G. (2017). Neural machine translation and sequence-to-sequence models: A tutorial. arXiv preprint arXiv:1703.01619. Haettu osoitteesta https://arxiv.org/pdf/1703.01619.pdf

Nieminen, T. (31.01.2019). Konekäännös suomalaisen kääntäjän näkökulmasta. Haettu 23.03.2020 osoitteesta https://puolukka.uta.fi/~textmine/events/tew-tampere/TNieminenTre.pdf

Nirenburg, S. (1989). Knowledge-based machine translation. *Machine Translation*, 4(1), 5-24. Haettu osoitteesta https://www.jstor.org/stable/40008396?seq=1

Nuutila, P. (2005). *Rough Machine Translation in the Communication Process* (lisensiaatintyö, Tampereen yliopisto). Haettu osoitteesta https://trepo.tuni.fi/bitstream/handle/10024/76408/lisuri00034.pdf?sequence=1&isAllowed=y

Och, F. J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1), 19-51. Haettu osoitteesta https://www.mitpressjournals.org/doi/pdfplus/10.1162/089120103321337421

Okpor, M. D. (2014). Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, 11(5), 159-165. Haettu osoitteesta https://search.proquest.com/docview/1617937023?pq-origsite=gscholar&fromopenview=true

Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics.* 311–318. Haettu osoitteesta https://dl.acm.org/doi/10.3115/1073083.1073135

Radwanski, M. C. (2017). *Maschinelles Dolmetschen* (maisterintutkielma, Wienin yliopisto). Haettu osoitteesta http://othes.univie.ac.at/49650/1/52088.pdf

Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., & Liu, Y. (2015). Minimum risk training for neural machine translation. arXiv preprint arXiv:1512.02433. Haettu osoitteesta https://www.aclweb.org/anthology/P16-1159.pdf

Smith, K. S., Aziz, W. & Specia, L. (2016). The trouble with machine translation coherence. *Proceedings of the 19th Annual Conference of the European Association for Machine Translation.* 178–189. Haettu osoitteesta https://www.aclweb.org/anthology/W16-3407.pdf

Snover, M., Madnani, N., Dorr, B. J. & Schwartz, R. (2009). Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. *Proceedings of the Fourth Workshop on Statistical Machine Translation*. 259-268. Haettu osoitteesta https://dl.acm.org/doi/10.5555/1626431.1626480

Sreelekha, S. (2017). Statistical Vs (sic) Rule Based Machine Translation; A Case Study on Indian Language Perspective. *arXiv preprint arXiv*, *1708*. Haettu osoitteesta https://arxiv.org/ftp/arxiv/papers/1708/1708.04559.pdf

Sutskever, I., Vinyals, O. & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 3104–3112). Haettu osoitteesta http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

Tiedemann, J., & Nygaard, L. (2004, May). The OPUS Corpus-Parallel and Free: http://logos. uio. no/opus. Haettu osoitteesta http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.9191&rep=rep1&type=pdf

Tieteen termipankki (2020). Haettu 22.03.2020 osoitteesta https://tieteentermipankki.fi/wiki/Nimitys:korpus

Tilastokeskus, Suomen virallinen tilasto (SVT) (2019). Väestön tieto- ja viestintätekniikan käyttö. Haettu 22.03.2020 osoitteesta http://www.stat.fi/til/sutivi/2019/sutivi_2019_2019-11-07_kat_001_fi.html

Toral, A. & Way, A. (2014). Is Machine Translation Ready for Literature?. *Proceedings of Translating and the Computer*, 36, 174-176. Haettu osoitteesta http://www.mt-archive.info/10/Asling-2014-Toral.pdf

Tuohisaari, A. (2019). *Clasificación de errores y comparación de las traducciones español–finés de los traductores automáticos MT@ EC y eTranslation de la Comisión Europea* (pro gradu -tutkielma, Turun yliopisto). Haettu osoitteesta https://www.utupub.fi/bitstream/handle/10024/147083/Tuohisaari_Annika_opinnayte.pdf?sequence=1

Vehmas-Lehto, I. (1999). *Kopiointia vai kommunikointia? Johdatus käännösteoriaan.* Helsinki: Yliopistopaino.

Vickrey, D., Biewald, L., Teyssier, M. & Koller, D. (2005). Word-sense disambiguation for machine translation. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 771-778. Haettu osoitteesta http://ai.stanford.edu/~dvickrey/wordtrans_final.pdf

Vilar, D., Xu, J., Luis Fernando, D. H. & Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. *LREC,* 697–702. Haettu osoitteesta http://www.mt-archive.info/LREC-2006-Vilar.pdf

Virpioja, S., Grönroos, S. A. (2015). LeBLEU: N-gram-based Translation Evaluation Score for Morphologically Complex Languages. *Proc. of The Tenth Workshop on Statistical Machine Translation* (WMT15), 411-416. Haettu osoitteesta https://www.aclweb.org/anthology/W15-3052.pdf

Weaver, W. (1955). Translation. *Machine translation of languages*, 14, 15–23. Haettu osoitteesta http://htl.linguist.univ-paris-diderot.fr/_media/biennale/et09/supportscours/leon/leon.pdf

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. Haettu osoitteesta https://arxiv.org/pdf/1609.08144.pdf

Yamada, K. & Knight, K. (2001). A syntax-based statistical translation model. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics,* 523–530. Haettu osoitteesta https://www.aclweb.org/anthology/P01-1067.pdf

Zens, R., Och, F. J. & Ney, H. (2002). Phrase-based statistical machine translation. *Annual Conference on Artificial Intelligence,* 18-32. Haettu

osoitteesta https://link.springer.com/content/pdf/10.1007%2F3-540-45751-8.pdf

Ziemski, M., Junczys-Dowmunt, M. & Pouliquen, B. (2016). The United nations parallel corpus v1. 0. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 3530–3534. Haettu osoitteesta https://www.aclweb.org/anthology/L16-1561.pdf