Elina I. López

LUONNOLLISTEN KIELTEN KÄÄNTÄMINEN JA KONEKÄÄNNÖS

Taustaa, teoriaa ja menetelmiä

Tietojärjestelmätieteen
kandidaatintutkielma
25.02.2010

Jyväskylän yliopisto
Tietojenkäsittelytieteiden laitos
Jyväskylä

TIIVISTELMÄ

López, Elina I.

Tietojärjestelmätieteen kandidaatintutkielma / Elina I. López
Jyväskylä: Jyväskylän yliopisto, 2010.
36 s.

Luonnollisten kielten kääntäminen on olennainen osa ihmisten elämää, erityisesti nykyisessä kansainvälisessä maailmassa. Ilman kääntämistä eivät esimerkiksi yritykset pysty toimimaan. Käännettävät tekstimassat kuitenkin kasvavat kasvamistaan ja käännöstyön nopeuttamiseksi on haettu apua tietokoneista. Konekääntämistä onkin tutkittu ensimmäisten tietokoneiden käyttöönotosta lähtien.

Tässä tutkielmassa käsitellään luonnollisia kieliä, perinteistä kääntämistä ja konekäännöksiä. Tutkielmassa käydään läpi luonnollisten kielten jaotteluita ja ominaisuuksia, jotka tekevät luonnollisen kielen automaattisen käsittelyn erityisen vaikeaksi. Sen jälkeen esitellään perinteisen kääntämisen teorioita ja tekniikoita, sekä luodaan katsaus tietokoneavusteiseen kääntämiseen. Lopuksi keskitytään konekääntämisen kehitysvaiheisiin ja erilaisiin aiemmin vallalla olleisiin ja nykyisiin konekäännösmenetelmiin 1950-luvulta 2000-luvun alkuun.

AVAINSANAT: CAT, konekäännös, kääntäminen, luonnollinen kieli, MT, NLP, tietokoneavusteinen kääntäminen

SISÄLLYSLUETTELO

1 JOHDANTO

Luonnollisten kielten kääntäminen on olennainen osa ihmisten elämää. Tänä päivänä lähes kaikki ovat tekemisissä erikielisten ihmisten tai ainakin tekstien kanssa jopa päivittäin. Yksi tekijä tässä on tietysti internet ja erityisesti World Wide Web. Käännöstyön helpottamiseen ja nopeuttamiseen onkin haettu apua tietokoneista heti niiden käyttöönotosta lähtien.

Ilman kääntämistä eivät esimerkiksi yritykset pysty toimimaan. Entistä kansainvälisemmässä yritysmaailmassa käännettävät tekstimassat vain kasvavat kasvamistaan, ja paine käännöstyön tehostamiseen on kova. Kääntäjiä ei kuitenkaan haluta palkata lisää, koska työvoima on kallista erityisesti Suomessa. Myös käännösten käyttötarpeet saattavat vaihdella, eikä aina olekaan välttämätöntä saada virheetöntä ja julkaisukelpoista käännöstä, vaan joskus niin sanottu raakakäännöskin on riittävä tekstin sisällön summittaiseen ymmärtämiseen. Tämä onkin yksi niistä käyttötarkoituksista, johon konekäännös jo nykyisellään soveltuu ja johon sitä myös työelämässä käytetään. Ilman konekäännöstä tällaiset käännökset saattaisivatkin jäädä kokonaan tekemättä (esim. Nuutila 2005:7). Toisaalta konekääntäjän tuottama raakakäännös voi toimia myös kääntäjän tekstipohjana, josta muokataan lopullinen, julkaistava käännösteksti.

Perinteistä kääntämistä koskevalla keskustelulla on pitkä historia, sillä sen juuret voidaan ulottaa jo Rooman valtakunnan aikaan, jolloin erityisesti Horatius ja Cicero käänsivät kreikankielisiä tekstejä latinaksi ja kirjasivat ylös ajatuksiaan kääntämisestä. Varsinaisena akateemisena tieteenalana käännöstiede on kuitenkin melko uusi.

Konekäännöstä on puolestaan lähestytty lähinnä teknisestä näkökulmasta ja sitä on tutkittu 1940-luvulta lähtien. Huolimatta konekäännöksenkin suhteellisen pitkästä kehityshistoriasta, sen laadullinen taso on ollut melko alhainen. Viime vuosina tutkijat ovat kuitenkin saaneet uusia lupaavia tuloksia. Konekäännök-

sen tason parantaminen on yksi IT-alan suurimmista haasteista myös tulevina vuosina.

Tässä tutkielmassa käsitellään luonnollisia kieliä, perinteistä kääntämistä ja konekäännöksiä. Tutkielma perustuu kirjallisuuteen ja tutkimusartikkeleihin eikä sisällä empiiristä tutkimusta. Ilmiöitä havainnollistetaan mahdollisuuksien mukaan esimerkein suomen ja englannin kielistä. Tarkoituksena on tuoda konekäännöksiin humanistista näkökulmaa teknisen puolen rinnalle eli pohtia perinteistä kääntämistä ja konekääntämistä suhteessa toisiinsa. Lisäksi pyritään selvittämään, miten konekäännös on kehittynyt alkuajoistaan sekä millaisia ratkaisuja konekääntämisprosessiin on esitetty.

Ensin esitellään luonnollisten kielten jaotteluita ja ominaisuuksia, jotka tekevät luonnollisen kielen automaattisen käsittelyn erityisen vaikeaksi. Sitten käydään lyhyesti läpi perinteisen kääntämisen teorioita ja tekniikoita, sekä esitellään tietokoneavusteista kääntämistä. Lopuksi keskitytään konekääntämisen kehitysvaiheisiin ja erilaisiin aiemmin vallalla olleisiin ja nykyisiin konekäännösmenetelmiin 1950-luvulta 2000-luvun alkuun.

Oma taustani on alun perin kielitieteessä, sillä olen suorittanut FM-tutkinnon vuonna 2003 pääaineenani englannin kieli. Valmistumiseni jälkeen olen toiminut suomi-englanti-suomi-kääntäjänä eräässä monikansallisessa suuryrityksessä. Työssäni käytän tietokonetta hyvinkin monipuolisesti, niin tekstinkäsittelyyn, tiedonhakuun internetistä kuin varsinaiseen käännöstyöhönkin. Käytössäni on CAT-työkaluista käännösmuisti ja termipankki. Työni ja nykyisten opintojeni myötä heräsi kiinnostukseni myös konekääntämisen hyödyntämismahdollisuuksiin. Olen kuitenkin omakohtaisesti huomannut myös sen, että erityisesti internetissä ilmaiseksi saatavilla olevia konekääntäjiä käytetään myös väärin, esimerkiksi sellaisiin kieliin, joita käyttäjä ei itse ollenkaan ymmärrä. Toistaiseksi kuitenkaan mikään ilmainen konekääntäjä ei tuota niin hyvää jäl-

keä, että niitä voisi käyttää kaupallisiin tarkoituksiin ilman minkäänlaista jälkieditointia. Tällaiseen ei pysty kovin moni maksullinenkaan tuote.

2 LUONNOLLISET KIELET

Luonnollinen kieli (natural language) on ihmisen luonnostaan käyttämä kieli. Se on sosiaalisuuden väline, mutta kuitenkin vain osittain yhteistä samankaan kulttuurin jäsenille. Se on jatkuvasti muuttuvaa, vaihtelevaa ja monitulkintaista, mikä tekee sen automaattisen käsittelyn tietokoneilla todella vaikeaksi. Esimerkiksi englannin kielessä moniselitteisyys on runsasta: iso osa juoksevan tekstin sanoista voitaisiin kontekstista irrallaan tulkita useammalla kuin yhdellä tavalla. Esimerkiksi ”hand” voi olla joko substantiivi (”my hand was hurt”) tai verbi (”please, hand me that paper”).

Luonnollisen kielen käsittely (natural language processing, NLP) tarkoittaa nimensä mukaisesti luonnollisten kielten prosessointia tietokoneilla niiden ymmärtämällä tavalla. Käsite viittaa sekä tekstin että puheen automaattiseen käsittelyyn, ja kattaa niin kielen tuottamisen kuin kielen tulkitsemisenkin. *Konekääntäminen (machine translation, MT)* on yksi NLP:n osa-alue, muita ovat esimerkiksi *tiedonhaku (information retrieval, IR)* ja *puheentunnistus (speech recognition)*.

Tässä luvussa esitellään luonnollisten kielten tyyppejä ja ominaisuuksia, sekä pohditaan esimerkkien avulla näiden vaikutusta kääntämiseen erityisesti suomen ja englannin välillä.

2.1 Kielten morfologiset päätyypit

Kielten morfologinen luokittelu jakaa maailman kielet ryhmiin niiden morfologian perusteella eli sen perusteella, millaisista irrallisista merkitysyksiköistä niiden sanat ja lauseet koostuvat. Morfologisia päätyyppejä tunnistetaan nykyään neljä: isoloivat, agglutinoivat, fuusioivat ja polysynteettiset kielet.

Isoloivissa (isolaatio = eristäminen) kielissä sanat ovat muuttumattomia, selvästi erottuvia, irrallisia kokonaisuuksia. Niissä esimerkiksi substantiivin monikkoa

ja mennyttä aikaa ei ilmaista päätteillä, vaan merkitykset on pääteltävä puhetilanteen perusteella. Kiina ja vietnam ovat tällaisia kieliä. (Karlsson 2004:116)

Agglutinoivissa (agglutinaatio = kiinniliimautuminen) kielissä käytetään päätteitä, jotka normaalisti liittyvät muuttumattomaan sanavartaloon. Samaa merkitystä vastaa järjestelmällisesti samanasuinen pääte. Esimerkiksi turkki on agglutinoiva kieli. (Karlsson 2004:116) Suomenkin katsotaan kuuluvan tähän ryhmään, vaikka suomessa sanavartalo muuttuukin usein päätettä lisättäessä (esimerkiksi ”aurinko” + genetiivin pääte ”-n” -> ”auringon”).

Fuusioivissa (fuusio = sulautuminen) kielissä on sanavartalon sisäisiä, merkitykseen vaikuttavia äännevaihteluita. Erimerkityksisten morfeemien erottaminen sanavartalosta voi olla hankalaa tai jopa mahdotonta. (Karlsson 2004:117) Germaanisten kielten kuten englannin, ruotsin ja saksan epäsäännöllisten verbien imperfektimuodot ovat hyviä esimerkkejä tästä: esim. englannin juosta-verbi ”run”, jonka imperfektimuoto on ”ran”.

Polysynteettisissä (poly = monta, synteesi = yhdistyminen) kielissä on erityisen paljon sidonnaisia morfeemeja, joiden merkitys on myös usein runsaampi kuin tavanomaisen taivutuspäätteen tai johtimen. Tästä syystä polysynteettisissä kielissä sanan ja lauseen ero ei ole ollenkaan selvä. Yksi yhtenäinen ”sana” voi vastata kokonaista lausetta esimerkiksi suomeksi. Eskimokielet ja useat Pohjois-Amerikan intiaanikielet ovat tyypillisiä polysynteettisiä kieliä. (Karlsson 2004:117)

2.2 Englannin ja suomen rakenteen vertailua

Tässä esitellään pääpiirteittäin rakenneseikkoja, jotka erottavat englannin ja suomen kielet toisistaan. Tällä on tarkoitus havainnollistaa kääntämisen haasteita näiden kahden kielen välillä sekä tukea seuraavissa kappaleissa esitettyjä käännösesimerkkejä.

Englannin kieli kuuluu germaaniseen kieliryhmään ja on fuusioiva kieli. Kuitenkin englannin kielessä on vähän taivutusmuotoja verrattuna useimpiin muihin indoeurooppalaisiin kieliin verrattuna. Siinä ei ole esimerkiksi sijamuotoja, kieliopillista sukua eikä adjektiivien taivutusta. Sitä vastoin sen aikamuotojärjestelmä on monimutkainen, ja aikamuotoja ilmaistaankin liittosanoilla (”I have eaten” tai ”I would have been eating”). Epäsäännöllisiä verbimuotoja ja substantiivien monikkomuotoja on paljon.

Suomen kieli kuuluu suomalais-ugrilaiseen kieliryhmään ja on voimakkaasti taipuva, agglutinoiva kieli. Suomessa on sijamuototaivutus sekä yksikössä että monikossa. Lisäksi substantiiveihin voidaan liittää omistusliitteitä eli possessiivisuffikseja (”minun auto/-ni”, ”hänen kirja/-nsa”). Sanajärjestys on toissijainen. On siis sama sanotaanko ”Kissa söi hiiren.” vai ”Hiiren söi kissa”, koska lauseen tekijä (subjekti) ja tekemisen kohde (objekti) on ilmaistu sijapäätteillä. Suomen kielen verbit taipuvat persoonamuodoissa (kuusi persoonaa ja passiivi), aikamuodoissa (preesens, imperfekti, perfekti ja pluskvamperfekti) sekä tapaluokissa eli moduksissa (indikatiivi, imperatiivi, konditionaali ja potentiaali).

2.3 Luonnollisten kielten ominaisuuksia

Luonnollisilla kielillä on useita piirteitä, jotka tekevät niistä monitulkintaisia ja monipuolisia. Tässä kappaleessa nostetaan esiin joitakin sellaisia ominaisuuksia, jotka erityisesti vaikuttavat kääntämiseen. Näitä havainnollistetaan esimerkeillä suomen ja englannin kielistä.

Polysemiassa samalla sanalla on kaksi tai useampia merkityksiä (Karlsson 2004:213). Suomessa esimerkiksi sana ”kieli” on polyseeminen. Se taipuu samalla tavalla kaikissa merkityksissään. Englanniksi tämä sana kuitenkin kääntyy eri merkityksissään eri tavoilla, esimerkiksi ”tongue”, ”string” tai ”language”. Käännettäessä on otettava konteksti huomioon, jotta osataan tehdä oikea sanavalinta.

Homonymialla tarkoitetaan sitä, että kahden eri sanan äänneasut ovat identtiset (Karlsson 2004:213). Esimerkiksi sana ”kuusi” voi olla joko havupuulaji tai lukusana. Eri merkityksissään sana ”kuusi” taipuu eri tavoilla, esimerkiksi genetiivissä ”kuusen” ja ”kuuden”. Ainakin yksittäisenä perusmuotoisena sanana on mahdotonta tietää, pitäisikö ”kuusi” kääntää englanniksi ”fir” vai ”six”.

Synonymia tarkoittaa, että sama käsite voidaan ilmaista monella eri sanalla (Karlsson 2004:219). Eri kielissä on samalle asialle eri määrä synonyymejä ja niillä voi myös olla erilaisia vivahde-eroja. Jos esimerkiksi pitäisi kääntää englannin sana ”pine” suomeksi, se voi olla yhtä lailla ”mänty” kuin ”petäjä”. Tosin näistä yleiskielisempi eli ensisijainen valinta olisi kuitenkin ”mänty”. Koherenssin vuoksi yhdessä tekstissä on useimmiten järkevää käyttää samasta asiasta järjestelmällisesti yhtä termiä.

Kiertoilmaus eli parafraasi. Joskus jotakin asiaa ei haluta tai ei pystytä ilmaisemaan yksiselitteisellä käsitteellä tai ilmauksella, vaan se ilmaistaan monimutkaisemmin kiertoilmauksen avulla. Tuo merkitys voi olla upotettu diskurssiin myös lausumattomana: pätevä kuulija tai lukija ymmärtää sen kuitenkin. Jos puhutaan vaikkapa työpaikan kiusanhengestä, ei välttämättä haluta mainita hänen nimeään, vaan sanotaan esimerkiksi ”Taas se yksi teki tosi ärsyttävän tempun!”. Kuulijat, joille tämä viesti on tarkoitettu, myös ymmärtävät ketä ”se yksi” tarkoittaa. Jos tämä lausahdus pitäisi kääntää jollekin toiselle kielelle, on erityisesti koneelle vaikeaa tietää miten ”se yksi” pitäisi kääntää. Ihminen ymmärtää asiayhteydestä, että ”se yksi” voidaan korvata ainakin persoonapronominille (”he/she”), mutta kone tulkitsisi ilmaisun todennäköisesti kirjaimellisesti.

Anaforiset pronominit ovat pronomineja, jotka viittaavat tekstiyhteydessä aiemmin esitettyyn kohteeseen (Karlsson 2004:243). Lauseissa ”Kirsi pääsi opiskelemaan Jyväskylän yliopistoon. Hän aloittaa opintonsa ensi syksynä.” sana ”hän” on anaforinen ilmaisu, joka viittaa Kirsiin. Käännettäessä tätä anaforista ilmai-

sua englanniksi, pitää tietää, että Kirsi on naispuolinen henkilö, jotta osataan valita oikea persoonapronomini ”she” eikä ”he”.

Hyponymiassa käytetään abstraktimpaa käsitettä ilmaisemaan konkreettisempi käsite. Se on tekstin sisäinen sidoskeino (Karlsson 2004:244). Erona anaforiaan, hyponymialta edellytetään, että viittaava ilmaisu edustaa abstraktimpaa käsitettä. Esimerkiksi lauseissa "Erkin kissa piti lopettaa. Eläin oli jo niin sairas." sana "eläin" hyponyyminen. Tällainen rakenne on haastava kääntää esimerkiksi englanniksi, jossa on käytössä määräiset ja epämääräiset artikkelit. Pitää ymmärtää, että ”eläin” viittaa aiemmin mainittuun ”kissaan”, jotta osaa laittaa englanniksi oikean eli määräisen artikkelin ”the animal”.

Yhdyssanat ja sanaliitot. Yhdyssanoissa loppuosa ilmoittaa pääluokan ja alkuosa(t) alaluokan. Loppuosan tunnistaminen on tärkeää, koska sen mukaan esimerkiksi taivutetaan. Yhdyssanat ja sanaliitot voivat aiheuttaa hankaluutta kääntäjälle. Esimerkiksi englannissa on yleensä sanaliitto, kun suomessa on yhdyssana: sanaliitto ”apple tree” on suomeksi yhdyssana ”omenapuu”.

Kielikuvat. Kielikuvat ovat sanan tai lauseen kirjaimellisesta merkityksestä poikkeavalla tavalla käytettyjä sanoja tai lauseita. Kielikuvia ovat muun muassa *idiomit* ja *metaforat*. Näitä ei useinkaan pysty kääntämään suoraan kieleltä toiselle, vaan pitää tuntea kyseinen kielikuva, jotta sama tai edes samankaltainen merkitys pystytään välittämään lukijalle. Esimerkiksi suomeksi idiomi ”vetää nenästä” tarkoittaa ”vitsailla, huijata leikkimielisesti”. Sama asia ilmaistaan englanniksi idiomilla ”to pull someone’s leg” eli ”vetää säärestä”.

Kirjoitusvirheet. Kirjoitusvirheitä ei koskaan pystytä kokonaan poistamaan, huolimatta tekstinkäsittelyohjelmiin sisältyvistä oikolukuohjelmista. Kirjoitusvirheet vaikeuttavat tekstin ymmärtämistä. Erityisesti tämä vaikeuttaa konekäännöstä, koska kone tulkitsee merkkijonoja. Jos kyseistä merkkijonoa ei löydy sen sanastosta, kyseinen sana jää kääntämättä. Jos taas väärinkirjoitettu sana sattu-

malta saakin saman muodon kuin joku oikea sana, kone tulkitsee sen saamiensa ohjeiden mukaan kyllä oikein, mutta tekstin sisältöön nähden väärin.

Tekstin sisäiset viittaukset. Joskus voi olla vaikeaa, jopa mahdotonta, tulkita tekstin sisäiset viittaukset varmasti oikein. Esimerkiksi lauseessa ”poista kumiset tuet ja levyt” ei voida asiaa täsmällisesti tuntematta tietää, onko myös levyt tehty kumista. Tämä on merkityksellinen seikka, jos kyseisessä laitteessa on useammasta materiaalista tehtyjä levyjä.

3 KÄÄNTÄMINEN

Kääntämisellä tarkoitetaan tekstin merkityksen siirtämistä yhdeltä luonnolliselta kieleltä toiselle. Kirjallista kääntämistä on ollut niin kauan kuin kirjoitustaitoakin, suullista kääntämistä eli tulkkausta sitäkin kauemmin (Vehmas-Lehto 2002:21).

Luonnollisten kielten kääntäminen kieleltä toiselle on vaikea tehtävä. Puhtaasti kaksikielinenkään henkilö ei välttämättä osaa kääntää, sillä kääntäminen vaatii erityistaitoja, kuten kielen rakenteiden ymmärtämistä, kieliopin ja oikeinkirjoitussääntöjen tuntemusta, eri tekstilajien tuntemusta ja sujuvaa kirjoitustaitoa. Kääntäminen vaatii myös erityisvalmiuksia, joilla ei ole mitään tekemistä kielitaidon kanssa, kuten kulttuurien, eri elämänalojen ja eri aihepiirien tavanomaista syvempää tuntemusta (Tommola 2004:9).

3.1 Kääntäminen tieteenalana

Ensimmäiset kääntämistä koskevat pohdinnat on johdettavissa muinaisiin roomalaisiin, lähinnä Horatiukseen ja Ciceroon, jotka elivät ensimmäisellä vuosisadalla ennen ajanlaskumme alkua. He käänsivät kreikankielisiä tekstejä latinaksi ja heidän tallensivat näkemyksiään kääntäjän työstä omiin kirjoituksiinsa (Bassnett 1995:61). Kristinuskon leviämisen myötä raamatun kääntäminen eri kielille lisäsi keskustelua kääntämisen säännöistä, koska nyt kääntämisestä tuli

jumalan sanan levittämisen väline ja samalla politiikan väline (Bassnett 1995:63-65). Kirjapainotaidon kehittymisen myötä 1400-luvulla kääntämisen rooli kuitenkin muuttui merkittävästi lähinnä käännöstyön määrän lisääntyessä (Bassnett 1995:71). Juuri tuolta ajalta ovatkin peräisin varhaisimmat varsinaiset kääntämistä koskevat teoriat. Yliopistoissa kääntämistä on tutkittu ja tutkitaan edelleenkin niin kielitieteen kuin kirjallisuustieteenkin piirissä. Tiittulan (2008) mukaan käännöstiede itsenäisenä tieteenalan onkin varsin uusi ja sen itsenäistyminen on yhteydessä kääntäjänkoulutuksen kehittymiseen yliopistotasoiseksi koulutukseksi. Sen piiriin kuuluu sekä kääntämisen että tulkkauksen tutkimus. Keskeisiä tutkimusalueita ovat muun muassa kääntämisen ja tulkkauksen prosessit ja tuotokset, käännösteknologia, kääntämisen historia ja sosiologia sekä kääntäjän ja tulkin asiantuntijuus ja koulutus.

Käännösteoreettista keskustelua on siis käyty vuosisatojen ajan, ja käsitykset kääntämisestä ja erityisesti niin sanotusta käännösihanteesta ovat vaihdelleet aikakaudesta ja koulukunnasta toiseen. Käsittelen seuraavassa lyhyesti käännösteoreettisen keskustelun eri piirteitä.

3.2 Kääntämisen teoriaa

Käännösteoreettisessa keskustelussa on ollut kaksi perustavanlaatuista kysymystä, joita on pohdittu (Vehmas-Lehto 1999:23). Ensimmäinen on kysymys siitä, onko kääntäminen ylipäätään mahdollista. Toinen puolestaan liittyy käännösihanteeseen, eli siihen miten pitäisi kääntää.

Kysymys kääntämisen mahdottomuudesta on sinällään turha, koska kääntämistä kuitenkin koko ajan harrastetaan. Siinä on kuitenkin Vehmas-Lehdonkin (1999:24) mukaan totuuden siemen, koska käännettäessä kieleltä toiselle joudutaan aina jossain määrin tekemään kompromisseja sekä sisällön että rakenteen suhteen. Tätä on Vehmas-Lehdon (1999:24-25) mukaan perusteltu seuraavilla seikoilla (kuvaukset ja esimerkit omiani):

1) Todellisuutta jaotellaan eri kielissä eri tavoin. On esimerkiksi kieliä, joissa ei ole lukusanoja, on vain ”yksi” ja ”monta”.

2) Kielissä on ns. vastineetonta sanastoa. Yksi usein esitetty esimerkki on se, että eskimokielissä (ja suomessakin!) on monia eri sanoja, jotka kuvailevat lunta, riippuen lumityypistä. Suomeksi meillä on esimerkiksi sanat ”räntä”, ”loska”, ”suojalumi” ja ”pakkaslumi”. Näille ei löydy kaikilla kielillä vastaavia käsitteitä, vaan niille pitää antaa kuvailevia käännöksiä.

3) Konnotaatiot eli sivumerkitykset vaihtelevat kielestä toiseen. Esimerkiksi suomalaiset liittävät ”hämäläisyyteen” hitauden. Espanjassa taas hitaiksi ja tyhmiksi käsitetään Galician maakunnan gallegot. Tekstiä käännettäessä tällainen asia pitää tietää, jotta tekstin merkitys säilyy edes suunnilleen saman.

4) Sanaleikkejä tai vitsejä voi olla mahdotonta kääntää. Ne voivat olla joko kielisidonnaisia tai kulttuurisidonnaisia. Esimerkiksi suomalaisten kertomat ruotsalaisvitsit eivät naurata muita kuin suomalaisia. Englantilaiset kertovat vastaavia vitsejä irlantilaisista.

5) Murteelliset ilmaukset voivat myös aiheuttaa ylitsepääsemättömiä ongelmia. Murteella puhumiseen liittyy myös aina tietty konnotaatio puhujasta itsestään, jota ei käännöksessä pysty välittämään.

Käännösihannetta koskeva keskustelu on käynnissä jatkuvasti. Vehmas-Lehto (1999:26-30) erottaa kolme päätapaa kääntää: mielivaltainen käännös, muodollisesti vastaava käännös ja vapaa käännös.

Mielivaltaisen kääntämisen kannattajat ovat sitä mieltä, että käännettävän tekstin kanssa voi tehdä mitä tahansa. Tekstiä voi esimerkiksi lyhentää tai pidentää mielensä mukaan. Tämä käännösihanne oli vallalla 1800-luvulle saakka. Nykyäänkin tämä on noussut esille, sillä katsotaan, että joskus tekstin adaptointi on

sallittua ja jopa välttämätöntä. Se ei kuitenkaan saa olla mielivaltaista, vaan sen pitää olla hyvin perusteltua, eikä lukijaa saa johtaa harhaan tekstin alkuperästä.

Muodollisesti vastaavan käännöksen ihanteen mukaan käännetyn tekstin tulee noudattaa mahdollisimman sanatarkasti alkuperäistä tekstiä. Myös lauserakenteita on tarkoitus imitoida sanajärjestystä myöten. Nykyisessä käännösteoreettisessa ajattelussa muodollista vastaavuutta vastustetaan. Esimerkiksi kääntäjien koulutuksessa pyritään vähentämään sanatarkkuutta ja erityisesti sanaluokkauskollisuutta. Tämän on nimittäin todettu huonontavan käännöksen ymmärrettävyyttä ja nautittavuutta, kun lukijan huomio kiinnittyy sisällön sijaan kummallisiin kielirakenteisiin.

Vapaan käännöksen ihanteen mukaan käännöksen tarkoituksena on välittää tarkasti lähtötekstin asiasisältö kohdekielellä huomioiden, että kohdekielinen teksti on luontevaa. Sana ”vapaa” tarkoittaa siis tästä vapautta muodollisuudesta, sisällön suhteen ei kovin suuria vapauksia kuitenkaan ole tarkoitus ottaa. Useimmissa tekstityypeissä asiasisältö on tekstin tärkein anti.

3.3 Kääntämisen tasot

Luonnolliset kielet ovat monitasoisia ja erittäin monimutkaisia järjestelmiä. Kielitiede jakaa ne viiteen eri osajärjestelmään (Karlsson 2004:15). Fonologia tutkii äännerakennetta. Sanastoa puolestaan tutkii leksikologia. Morfologialla tarkoitetaan sanojen sisäistä osajärjestelmää. Syntaksi viittaa lauserakenteeseen. Kielen merkityksen tutkimusta kutsutaan semantiikaksi. Kun kahden eri kielen rakenteita lähdetään vertailemaan, tulee vertailua tehdä kolmella eri tasolla: morfologisella, syntaktisella ja semanttisella tasolla.

Morfologinen analyysi tutkii yksittäisten sanojen sisäistä rakennetta. Esimerkiksi suomenkielisessä sanassa ”talossani” on useita morfeemeja eli merkityksellisiä yksiköitä: perusmuoto ”talo”, sijapääte ”-ssa” ja omistusliite ”-ni”. Jos tämä

sama sana halutaan kääntää englanniksi, sitä vastaakin useampi sana ”in my house”.

Syntaktinen analyysi tarkastelee lauserakennetta. Otetaan esimerkiksi lause ”Kadulla on kuusi autoa”. Tämä lauseen lauseenjäsenet ovat subjekti (”kuusi autoa”), predikaatti (”on”) ja adverbiaali (”kadulla”). Käännettäessä tätä lausetta englanniksi (”There are six cars on the street.”) vastaavat lauseen jäsenet ovat muodollinen subjekti (”there”), predikaatti (”are”), varsinainen subjekti (”six cars”) ja adverbiaali (”on the street”).

Semanttinen analyysi keskittyy sanojen, lauseiden ja tekstien merkityksiin. Merkityksen muodostumiseen vaikuttaa suuresti konteksti eli ympäristö, jossa kyseinen sana, lause tai teksti esiintyy. Sama sana, lause tai kokonainen tekstikin voi saada eri merkityksen riippuen kontekstista, jossa se esiintyy. Näin ollen myös sen käännökseksi jollekin toiselle kielelle voi olla useita vaihtoehtoja. Otetaan esimerkiksi vaikkapa lause ”Sammuta se!”. Tämän lauseen englanninkielinen käännös olisi ”Switch it off!” tai ”Turn it off!”, jos kyse on esimerkiksi television sammuttamisesta. Jos taas kyse onkin tulipalon sammuttamisesta, tarvitaan erilainen käännös, kuten ”Put it out!”

Eugene A. Nida (1964:146) tarkastelee käännösprosessia seuraavanlaisen mallin avulla. Mallissa lähdetään liikkeelle lähtökielisestä tekstistä, jonka analysoidaan toiseen muotoon lähtökielellä. Sitten tehdään sisällön siirto kohdekielelle, jonka jälkeen muotoillaan lopullinen kohdekielinen käännös. Tässä voidaan ajatella, että analyysivaihe kattaa kaikki kolme edellä esitettyä analyysia, eli morfologisen, syntaktisen ja semanttisen analyysin. Siirtovaiheessa haetaan kohdekieliset vastaavat sanat ja rakennesäännöt, joiden perusteella sitten muotoiluvaiheessa rakennetaan kohdekielinen käännös. Tällaista prosessimallia on sovellettu myös konekääntämiseen.

3.4 Kääntämisen tekniikoita

Kääntäjällä on käytössään useita tekniikoita käännöksen sujuvuuden parantamiseksi. Näitä ovat esimerkiksi transpositio, deverbalisaatio, ikonisuus ja relevanssi (Chesterman 1996). Käännösesimerkit ovat omiani.

Transpositiolla tarkoitetaan sanaluokan vaihtamista lähtökielisen ja kohdekielisen lauseenjäsenen välillä. *Deverbalisaatiolla* puolestaan tarkoitetaan sitä, että nostetaan virkkeen tai tekstin merkitys pintarakenteen alta ja ilmaistaan se sitten kohdekielellä. Tarkoitus on siis irrottautua lähtökielisen tekstin rakenteesta, jotta kohdekielisestä tekstistä tulee mahdollisimman sujuva. Seuraava esimerkki havainnollistaa näitä molempia tekniikoita. Jos käännettävänä on lause ”Suomessa on paljon puusta tehtyjä taloja.”, se voitaisiin kääntää englanniksi säilyttäen sen rakenne ja käyttäen samojen sanaluokkien sanoja, eli muotoon ”In Finland there are many houses made of wood.” Huomattavasti sujuvampi käännös olisi kuitenkin ”There are many wooden houses in Finland.”

Ikonisuudella tarkoitetaan sitä, että teksti on selkeämpi, jos sen rakenne vastaa sen merkitystä. Esimerkiksi kehotus ”Ennen kuin menet saunaan, käy suihkussa.” olisi jo suomeksikin selkeämpi, jos se sanottaisiin näin: ”Käy suihkussa, ennen kuin menet saunaan.” Jälkimmäisessä virkkeessä ilmaistaan tekeminen sen todellisessa järjestyksessä. Kääntäjä voi tehdä tämän parannuksen käännökseensä, eli jos käännettävänä on virke ”Ennen kuin menet saunaan, käy suihkussa.” hän voi kääntää sen muotoon ”Take a shower before entering the sauna.”

Relevanssi tarkoittaa kääntämisessä sitä, että kääntäjä pitää käännöksen lukijan mielessään käännöstä tehdessään. Tekstiin voi esimerkiksi lisätä jotain selvennykseksi tai jotain lukijan näkökulmasta epäolennaista voi jättää pois. Varmasti vaikkapa Amazonin viidakoista löytyy eläinlajeja, jotka on nimetty joillakin intiaanikielillä, mutta ei suomeksi. Ja vaikka kyseisellä eläimellä olisikin suomenkielinen nimi, se ei välttämättä kerro suomalaiselle lukijalle mitään, ellei tekstin

kohdeyleisönä satu olemaan kyseistä elinpiiriä tutkiva suomalainen tiedeyhteisö. Näin ollen kääntäjän pitäisi pystyä selittämään asia sillä tavalla, että lukija voi edes jossain määrin käsittää millaisesta eläimestä on kyse. Kääntäjä voi esimerkiksi lisätä suomenkieliseen tekstiin selostuksen eläimen ulkonäöstä tai muista erityispiirteistä. Toisaalta taas jotain sellaista, jota käännöksen lukija ei kulttuurieroista johtuen todennäköisesti ymmärrä, voidaan jättää poiskin. Konekääntäjä ei tällaisia ratkaisuja pysty tekemään.

3.5 Tietokoneavusteinen kääntäminen

Tietokoneavusteinen kääntäminen (computer-assisted/computer aided translation, CAT) tarkoittaa tietokoneen ja tietokoneohjelmien käyttöä käännöstyössä. Se kattaa niin tekstinkäsittelyohjelmat, elektroniset sanakirjat ja tietosanakirjat, verkkosanakirjat, oikoluku- ja kielentarkistusohjelmat, *termipankit (term bank), käännösmuistit (translation memory, TM)* kuin konekääntämisenkin. Nykyisin ammattikääntäjillä on yleensä käytössään nämä kaikki apuvälineet. Näistä termipankit ja käännösmuistit ovat yleensä vain ammattikääntäjien käytössä, muut apuvälineet puolestaan ovat toki kaikkien tietokoneen ja internetin käyttäjien saatavilla. *Termipankkeja* käytetään erikoisalojen termistön hallintaan pitämään dokumenteissa käytetty termistö yhtenäisenä kautta linjan. Termipankki toimii yleensä integroituna käännösmuistiohjelmaan. Käytännössä tämä tarkoittaa sitä, että aina kun tekstissä tulee vastaan termipankkiin syötetty termi, ohjelma tarjoaa sille tiettyä kielivastinetta, jonka kääntäjä voi sitten siirtää tekstiinsä. *Käännösmuistiohjelmat* puolestaan perustuvat virkepareihin, joita on tallennettu tietokantaan. Tietokannassa on siis sama virke vaikkapa suomeksi ja englanniksi. Käännettäessä ohjelma tutkii virkkeitä merkkijonoina, ja jos kyseinen virke löytyy tietokannasta, se tarjoaa sen käännöstä kääntäjän käyttöön. Myös alle sataprosenttisia osumia tuodaan tarjolle, jolloin kääntäjä muokkaa käännöstä vastaamaan hänen tekstissään olevaa lausetta. Esimerkiksi, jos käännettävänä on virke ”Pojalla on *punainen* paita.” ja tietokannasta löytyy virkepari ”Pojalla on *sininen* paita. The boy is wearing a *blue* shirt.” ohjelma kuitenkin tar-

joaa virkettä kääntäjälle. Kääntäjän täytyy vain sitten muuttaa virkkeen käännöstä yhden sanan verran, eli muotoon ”The boy is wearing a *red* shirt.” Samalla tämä uusi virkepari tallentuu tietokantaan jatkokäyttöä varten. Käännösmuistiohjelmista on hyötyä lähinnä toistuvien tekstimassojen kääntämisessä, kuten esimerkiksi teknisissä käyttö- ja huolto-ohjeissa.

Termipankeista ja käännösmuisteista kumpikaan ei kuitenkaan ole automaattinen työkalu, vaan molemmat toimivat vuorovaikutuksessa kääntäjän kanssa. Kuitenkin, erityisesti käännösmuistit ovat vaikuttaneet kääntäjän työhön siten, että erilaiset käännösteoriat ja käännösihanteet ovat ehkä jääneet tai ainakin jäämässä taka-alalle. Kiire ja tehokkuusvaatimukset koskevat myös kääntäjiä, ja käännösmuistit mahdollistavat lähes automaattisen kääntämisen, erityisesti jos muistin koko on suuri, eli sieltä löytyy runsaasti käännösvastineita. Tällöin esimerkiksi tekstin koherenssi saattaa kärsiä, koska kääntäjä ei ehdi käsitellä tekstiä kokonaisuutena, vaan käy sitä läpi yksittäisinä virkepareina. Erilaiset tekstinsisäiset viittaukset saattavat jäädä puuttumaan tai mennä väärinkin, erityisesti jos kääntäjä ei ehdi ollenkaan lukea käännöstään läpi.

4 KONEKÄÄNNÖKSEN HISTORIAA JA MENETELMIÄ

Konekäännöksellä (machine translation, MT) tarkoitetaan tietokoneen käyttöä tekstien kääntämiseen eri luonnollisten kielten välillä. Tämä onkin yksi niistä käyttötarkoituksista, joihin tietotekniikalla on pyritty sen käyttöönotosta lähtien. Tänä päivänä on jo olemassa useita erilaisia järjestelmiä, jotka ainakin jollain tasolla pystyvät vastaamaan tähän haasteeseen.

4.1 Konekäännöksen kehitysvaiheet

Heti kun ensimmäiset tietokoneet, elektroniset laskukoneet, otettiin käyttöön, heräsi kiinnostus myös tietokoneiden käyttämiseen kielenkäännöksiin. Ensimmäisenä idean toivat esiin Andrew Booth ja Warren Weaver vuosien 1946 ja 1947 aikana (Hutchins 2006:375-376). Muutamassa vuodessa konekäännösten tutkimus alkoi yhdysvaltalaisissa yliopistoissa (Hutchins 2006:376). Tutkimusta rahoitti USA:n hallitus, ja kiinnostus johtui lähinnä sotilaallisista tarkoituksista.

1960-luvun alkupuolella saatettiin alulle kolme erilaista lähestymistapaa konekääntämiseen: suora kääntäminen (direct translation model), välikielen kautta kääntäminen (interlingua model) ja muunnosmalli (transfer model). *Suoran kääntämisen mallia* käytetään kääntämään tietyltä lähtökieleltä tietylle kohdekielelle tekemällä vain hyvin vähäistä syntaktista analyysia. Homonyymien ja moniselitteisyyden ongelmat ratkaistiin antamalla sanastoissa vain yksi, mahdollisimman kattava vastine jokaiselle sanalle. 1950-luvun puolivälistä 1960-luvun puoliväliin tutkimus keskittyikin lähinnä kaksikielisiin sanakirjoihin, joiden avulla pystyttiin tuottamaan sanasta sanaan -käännöksiä. Huomiota ei niinkään kiinnitetty syntaksiin tai semantiikkaan, joten konekääntäjien tuottamat tekstit olivat melko epäselviä. *Välikielen kautta kääntäminen* puolestaan tarkoittaa sitä, että lähtökielinen teksti analysoidaan ja siitä luodaan jonkinlainen kielestä riippumaton esitys, josta muokataan kohdekielinen käännös. Vielä 1960-luvulla välikieleen keskittyvä tutkimus oli kuitenkin lähes pelkästään teoreettista, eli käy-

tännön sovelluksia ei juurikaan saatu aikaan. *Muunnosmallin* mukaiset konekääntäjät taas analysoivat ja muuntavat lähtökielisen tekstin yksinkertaistettuun, yksiselitteiseen muotoon. Sitten luodaan vastaava kohdekielinen käännös, josta muokataan lopullinen kohdekielinen teksti. 1960-luvulla pyrittiin kehittämään syntaktisia analysaattoreita, mutta tulokset jäivät kuitenkin laihoiksi. (Hutchins 2006:376-377)

1960-luvulla usko konekäännöksiin alkoi hiipua, kun kielelliset ongelmat tulivat entistä ilmeisimmiksi (Slocum 1985:1). Parempaan laatuun haluttiin kuitenkin pyrkiä, mutta tarpeeksi nopeaa kehitystä ei tapahtunut. Vuonna 1966 ALPAC (Automatic Language Processing Advisory Committee) tuomitsi konekääntäjät kalliiksi ja epäluotettaviksi. Komitean raportti vaikutti tutkimuksen rajuun vähenemiseen erityisesti Yhdysvalloissa. (Hutchins 2006:376-378)

1980-luvun alussa tutkimusta alkoi olla yhä useammissa maissa. Kilpailuun tulivat mukaan muiden muassa Logos, Ariane ja Eurotra. Nyt pyrittiin kehittämään järjestelmiä, joiden pääperiaatteena oli muunnosmallin mukainen kääntäminen. (Hutchins 2006:378-379). 1980-luvun puolivälistä alkaen myös välikielimalli alkoi saada uudestaan huomiota (Hutchins 2006:379).

1990-luvun alusta lähtien vallalla ovat olleet niin sanotut *korpuspohjaiset konekäännösjärjestelmät*. 1990-luvun alkupuolella IBM julkaisi ensimmäiset tulokset Candide-järjestelmästään, joka perustui puhtaasti *tilastollisiin (statistical) menetelmiin*. Japanissa alettiin käyttää *esimerkkipohjaisia (example-based) menetelmiä*. (Hutchins 2006:380) Kummatkin eroavat aikaisemmin käytetyistä *sääntöpohjaisista (rule-based) menetelmistä*. Esittelen näitä tarkemmin seuraavissa luvuissa.

1990-luvun lopussa ja 2000-luvun alussa ovat jatkuneet 1990-luvun alussa alkaneet trendit. Tutkimusta on tehty erityisesti tilastollisista ja esimerkkiperustaisista järjestelmistä. 1990-luvun puolivälistä alkaen internetillä on ollut suuri vaikutus. Internet ja erityisesti World Wide Web ovat lisänneet räjähdysmäisesti käännettävien tekstien lukumäärää, joka puolestaan on edelleen lisännyt

kiinnostusta konekääntämiseen. Internetissä onkin nykyään saatavilla monia eri työkaluja kaikille käyttäjille, muun muassa www-sivujen sekä sähköpostiviestien kääntämistä varten (Hutchins 2006:382). Lisäksi internetin myötä myös konekääntäjien pohjaksi käytettävien olemassa olevien korpusten määrä on lisääntynyt huimasti, eikä korpuksia tarvitse varsinaisesti tuottaa vain konekääntäjiä varten.

4.2 Sääntöpohjainen konekäännös

Sääntöpohjainen konekäännös (rule-based machine translation, RBMT) oli vallalla oleva konekäännösmenetelmä 1980-luvun loppupuolelle saakka (Hutchins 2005:1). Se perustuu erilaisiin koneelle syötettyihin kieltä koskeviin sääntöihin. Säännöt koskevat esimerkiksi lähtökielisen tekstin syntaktista analyysia, leksikaalista eli sanaston siirtoa, kohdekielen syntaksin muodostamista sekä morfologiaa. Sääntöpohjaisen konekäännöksen ongelmakohtana ovat loputtomat poikkeukset sääntöihin (Krikke 2006:4), mikä lienee pääsyy sille, miksi sen ei yksinään enää katsota soveltuvan konekääntämiseen. Toiseksi sääntöpohjainen konekäännös soveltuu erityisen huonosti myös toisistaan rakenteellisesti eroavien kielten väliseen kääntämiseen, esimerkiksi kääntämiseen englannin ja japanin välillä (Hutchins 2005:1).

4.3 Esimerkkipohjainen konekäännös

Esimerkkipohjainen konekäännös (example-based machine translation, EBMT) perustuu erikielisiin tekstikorpuksiin. Korpusten on oltava sisällöltään toisiaan vastaavia ja niiden pitää olla toisiaan vastaan *linjattu (aligned)* siten, että kone tietää mikä segmentti (tyypillisesti virke) vastaa mitäkin segmenttiä toisella kielellä (Somers 1999:118). Korpusten käsittely voi olla suhteellisen yksinkertaista, jos kyseessä on kaksi toisiaan typologisesti vastaavaa kieltä (Somers 1999:118), kuten esimerkiksi espanja ja italia. Toisaalta se voi olla hyvinkin vaikeaa, jos kyseessä on kaksi toisistaan suuresti poikkeavaa kieltä (Somers 1999:118), kuten

vaikka jo aiemminkin mainittu kielipari englanti-japani, jotka eivät eroa ainoastaan rakenteellisesti, vaan myös merkkijärjestelmiltään.

Esimerkkipohjainen konekäännösjärjestelmä linjaa tekstejä: se vertaa syötettyä virkettä korpuksesta löytyviin virkkeisiin (eli ”esimerkkeihin”), valitsee sopivat kohdekieliset sanat ja lausekkeet ja yhdistelee sitten löytämänsä sanat ja lausekkeet hyväksyttäväksi kohdekieliseksi virkkeeksi (Hutchins 2005:1).

Korpuksen koko johtaa sekä mahdollisuuksiin että ongelmiin esimerkkipohjaisessa konekäännöksessä. Somers (1999:119) esittää, että on olemassa joku raja esimerkkien määrälle, jonka jälkeen määrän lisääminen ei enää paranna käännöstulosta. Suuressa korpuksessa voi olla päällekkäisiä esimerkkejä. Hyvä tilanne on se, että kaksi esimerkkiä esittää saman asian samalla tavalla eli ne vahvistavat toisiaan. Mutta voi käydä myös niin, että samalle asialle onkin kaksi erilaista käännöstä, johtuen ainoastaan epäjohdonmukaisuudesta. (Somers 1999:121)

4.4 Tilastollinen konekäännös

Tilastollinen konekäännös (statistical machine translation, SMT) perustuu myös tekstikorpuksiin. Näiden korpuksien perusteella kone opetetaan algoritmin avulla kääntämään. Algoritmilla selvitetään eri sanojen esiintymistiheyksiä ja todennäköisyyksiä, joiden perusteella valitaan tilastollisesti parhaimmat vaihtoehdot, joista kohdekielinen virke rakennetaan. Voidaan sanoa, että tulokset ovat sitä parempia, mitä suurempi korpus järjestelmän pohjana on.

Toistaiseksi tilastollisessa konekäännöksessä käytetyt tekstikorpukset ovat olleet *rinnakkaisia (parallel)*, eli ne ovat koostuneet ihmisten kääntämistä, toisiaan vastaavista erikielisistä teksteistä, kuten esimerkiksi Euroopan parlamentin pöytäkirjoista (Koehn 2005). Nämä korpukset ovat siis olleet olemassa jo valmiiksi, niitä ei ole varta vasten tuotettu konekäännösten pohjaksi. Nyt ollaan kuitenkin kehittämässä keinoja, joiden avulla myös sellaisia erikielisiä korpuk-

sia, jotka eivät sisällöllisesti ole toisiaan vastaavia, voitaisiin hyödyntää konekäännösjärjestelmissä (esim. Munteanu & Marcu 2005).

Tilastollinen konekäännös toimii erityisen hyvin sen alan teksteille, joiden perusteella se on oppinut, mutta ei niinkään jonkun toisen alan teksteille (Munteanu & Marcu 2005:477). Tämä ei liity pelkästään sanastoon ja lausetyyppeihin, vaan myös tyyliseikat pitää ottaa huomioon. Ajatellaan vaikkapa jotakin teknistä laitetta koskevaa tekstiä. On ihan eri asia kertoa laitteen teknisistä ominaisuuksista sen käyttöohjeessa, kuin jos ominaisuudet esitellään mainoksessa.

2000-luvulla tilastollinen konekäännös on mennyt laadullisesti eteenpäin isoin harppauksin. Och & Ney (2004:418) toteavat tämän johtuvan erityisesti siitä, että tilastollisten konekääntäjien opettamiseen on tarjolla entistä suurempia tekstimassoja. Toiseksi syyksi he mainitsevan sen, että mallinnus, opettaminen ja hakutekniikat ovat myös kehittyneet hurjasti siitä, kun tilastollista konekääntämistä alettiin alun perin tutkia 1980-luvun lopulla.

4.5 Uusimmat virtaukset

Viimeisimmät ehdotukset uusista konekäännösjärjestelmistä ovat hybridijärjestelmiä, jotka yhdistävät tilastollisia menetelmiä esimerkkipohjaisiin tai sääntöpohjaisiin menetelmiin (Krikke 2006:4, Groves & Way 2005). On myös kehitelty algoritmeja, jotka hoitavat sekä järjestelmän opettamisen että varsinaisen kääntämisen. Esimerkiksi Language Weaver toimittaa myös kustomoitavia järjestelmiä yrityksille (Language Weaver 2010). Konekääntäjä voidaan opettaa kääntämään nimenomaan perustuen kyseisen yrityksen olemassa oleviin teksteihin.

5 KONEKÄÄNTÄMINEN JA SUOMEN KIELI

5.1 Tärkeimmät suomen kieltä koskevat projektit

Suomessa on myös tehty omaa suomen kieltä koskevaa konekäännöstutkimusta. Arppen (2007) mukaan kieliteknologisella liiketoiminnalla on Suomessa vahvat akateemiset juuret ja suurin osa suomalaisista kieliteknologisista ohjelmistoyrityksistä voidaankin johtaa yksittäisiin tutkijoihin ja tutkijaryhmiin suomalaisissa yliopistoissa ja korkeakouluissa. Helsingin yliopistossa tutkimus aloitettiin 1980-luvulla professori Lauri Carlsonin johdolla. Ensimmäisenä varsinaisena konekäännösprojektina voidaan pitää Mentor/F-nimistä englanti-suomi-konekäännösjärjestelmää, jota Carlson kehitti yhdessä Krister Lindénin ja Seppo Koskenniemen kanssa vuosina 1988 - 1992. Kyseistä järjestelmää ei kuitenkaan koskaan tuotteistettu. (Arppe 2007) Lauri Carlson on kuitenkin tänä päivänäkin tärkein suomalainen akateeminen konekääntämisen ja kieliteknologian tutkija.

Tähän mennessä tärkein suomalainen konekäännösprojekti on kuitenkin ollut alun perin Suomen itsenäisyyden juhlarahaston (SITRA) rahoittama Kielikone-projekti, joka käynnistettiin vuonna 1982. Jäppisen, Hartosen, Kulikovin, Nykäsen & Ylä-Rotialan mukaan (1993:173) projekti lähti liikkeelle tavoitteenaan mallintaa suomen kieltä sillä tavalla, että näitä malleja voitaisiin sitten myöhemmässä vaiheessa käyttää kieliteknologisten sovellusten, kuten konekäännösjärjestelmän, pohjana. Mallien pohjalta syntyikin nopeassa tahdissa esimerkiksi morfologinen analysaattori, kielentarkastusohjelma ja tavutusalgoritmi. Konekääntämiseen liittyen ensimmäisenä kehitettiin elektroninen suomi-englanti-sanakirja (Jäppinen ym. 1993:173).

Jäppisen ym. (1993:173-174) mukaan Kielikone-projekti keskittyi vuodesta 1986 alkaen ainoastaan konekäännösjärjestelmän kehittämiseen. Silloinen Telenokia Oy oli pilottiasiakas suomi-englanti-järjestelmälle ja Finnair Oy englanti-suomi-

järjestelmälle. Kieliteknologisten tuotteidensa kaupallistamista varten Kielikone-projektista tuli Kielikone Oy -niminen yhtiö vuonna 1987. (Arppe 2007)

Kielikone Oy:n tänä päivänäkin markkinoima konekääntäjä oli alun alkaen nimeltään TranSmart, nykyään se on MOT Translation. Se on saatavilla sekä suomi-englanti- että englanti-suomi-käännöksiin. Kielikoneen omien www-sivujen (2009) mukaan se on tarkoitettu ensimmäisen käännösversion tuottamiseen, jonka pohjalta käyttäjä voi sitten itse muokata lopullisen tekstin. Kielikone Oy:n tuotteet, mukaan lukien konekääntäjä, ovat myös räätälöitävissä asiakkaan kielelliset tarpeet huomioiden, esimerkiksi termistön suhteen. Tästä on esimerkkinä Rautaruukki Oyj:n ja Kielikone Oy:n yhteistyö (Laitinen 2008).

5.2 Muu suomen kieltä koskeva konekäännöstutkimus

Koehn (2005) on tutkinut ja testannut omaa tilastollista konekäännösjärjestelmäänsä, joka pohjautuu Euroopan parlamentin pöytäkirjoista kerättyyn korpukseen. Kyseinen korpus käsittää samansisältöiset tekstit 11 Euroopan unionin virallisella kielellä. Koehn käytti käännöstuloksen ja laadun arviointiin niin sanottua BLEU-asteikkoa, jota käytetään yleisesti konekäännösjärjestelmien arvioinnissa. Koehnin (2005:83) tulokset osoittivat, että tutkittujen yhdentoista kielen joukossa suomen kielen kääntäminen oli kaikkein vaikeinta, niin suomeen päin kuin suomesta johonkin toiseen kieleen. Kaikkein huonoimman tuloksen sai kääntäminen hollannista suomeen. Koehn (2005:84) toteaakin, että tämä johtuu suomen rikkaasta morfologiasta eli sen moninaisista taivutusmuodoista. Samassa yhteydessä hän mainitsee, että suomenkielisessä korpuksessa olikin tästä johtuen noin viisi kertaa suurempi sanasto kuin englanninkielisessä korpuksessa.

Runsaasti taivutusmuotoja sisältävien kielten osalta on 2000-luvulla tehty jonkin verran tutkimusta, koskien lähinnä tilastollisia konekäännösjärjestelmiä. Tutkimuksella on pyritty löytämään menetelmiä, jotka parantaisivat sellaisten sanamuotojen käännöstulosta, joita ei ole esiintynyt järjestelmän pohjana olleen-

sa korpuksessa. Esimerkiksi, jos korpuksessa on ollut vaikkapa sanat ”koiralle” ja ”kissa”, ja käännettävässä tekstissä tulee vastaan sana ”kissalle”, järjestelmä osaisi tulkita oikein sijapäätteen ”-lle” ja yhdistää sen sanaan ”kissa” ja näiden kahden asian perusteella sitten kääntää kyseisen sanan oikein kohdekielelle. Kyseessä on siis morfologinen jäsentäminen. Tätä ovat tutkineet esimerkiksi Virpioja, Väyrynen, Creutz & Sadeniemi (2007) sekä El-Kahlout & Oflazer (2006). Näiden tutkimusten tulokset ovat kuitenkin ainakin toistaiseksi jääneet laihoiksi, sillä kummassakaan ei saatu merkittävästi aikaisempaa parempia BLEU-tuloksia. Tutkimus jatkuu edelleen.

6 POHDINTA

Kuten todettiin, luonnolliset kielet ovat monimutkaisia järjestelmiä. Kääntäminen eri kielien välillä on vaikeaa ihmisellekin, vaikka henkilöllä olisi kääntämiseen vaadittavat taustatiedot ja koulutuskin. Luonnollisten kielten automaattinen käsittely tietokoneilla ja erityisesti konekääntäminen on näistä syistä osoittautunut erityisen haasteelliseksi.

Kirjallinen kääntäminen on verbaalista viestintää, jossa kielellisiä, kulttuurisia ja kontekstiin liittyviä seikkoja joudutaan analysoimaan normaalia, yksikielistä viestintätilannetta tarkemmin. Jos verrataan kirjallista kääntämistä tulkkaukseen, voidaan sanoa, että kirjallinen kääntäminen on toisaalta yksinkertaisempaa ja toisaalta monimutkaisempaa kuin tulkkaaminen. Yksinkertaisempaa sen voi ajatella olevan siitä syystä, että kirjallisiin teksteihin ei liity nonverbaalia viestintää, eli kääntäjän ei tarvitse välittää esimerkiksi eleiden tai ilmeiden merkityksistä. Lisäksi kirjallisen käännöksen tekijällä on enemmän aikaa pohtia käännöstään kuin tulkilla. Samasta syystä kirjallista käännöksen tekemisen voi sanoa olevan monimutkaisempaa kuin tulkkaamisen. Kaikki pienetkin merkityserot olisi saatava irti pelkästään lähtötekstiä lukemalla. Ja usein käy niin, että mitä tarkemmin lähtötekstiä lukee, sitä enemmän sieltä nousee esiin vaihtoehtoisia tulkintoja. Ihmiskääntäjällä kuitenkin on useimmiten (tai ainakin pitäisi olla) mahdollisuus konsultoida lähtötekstin kirjoittajaa eli saada lisäselvitystä epäselviin kohtiin. Konekääntäjä kääntää tekstin ainoastaan sen pohjalta, mitä tekstiin on kirjoitettu.

Kieleltä toiselle kääntämisessä ei useinkaan ole kyseessä merkityksen siirtäminen vain kieleltä toiselle, vaan myös kulttuurista toiseen. Käännösviestintä on siis samalla myös kulttuurienvälistä viestintää. Siihen ei riitä pelkkä sanojen ja lauserakenteiden suora kääntäminen. Kaikkea ei edes välttämättä pysty kääntämään.

Toisaalta kääntämistä voi pohtia myös suhteessa perinteisiin viestinnän teorioihin. Lähdetään liikkeelle siitä, että joku ihminen kirjoittaa tekstin. Hän haluaa välittää tekstillään jonkun viestin jollekin toiselle ihmiselle. Viesti voi olla tarkoitettu vain yhdelle ihmiselle, jollekin tietylle ryhmälle tai suurelle, epämääräiselle joukolle ihmisiä. Jos kirjoittaja osaa asiansa, hän ottaa tekstiä kirjoittaessaan huomioon kohdeyleisönsä niin tekstin rakenteen kuin sanavalintojenkin suhteen. Mutta entäpä jos kyseinen teksti on tarkoitus kääntää jollekin toiselle kielelle ja jonkun toisen kulttuurin edustajalle? Tätäkin kirjoittajalla on mahdollisuus ennakoida, jos hän tuntee kohdekieltä ja -kulttuuria edes jossain määrin. Toisaalta, jos kirjoittaja tietää, että tekstin tulee kääntämään joku asiantunteva kääntäjä, hän voi luottaa siihen, että kääntäjä osaa välittää hänen viestinsä sopivalla tavalla viestin kohteelle. Tällaisessa tapauksessa kirjoittajan itsensä ei tarvitse pohtia sitä kirjoittaessaan. Joskus tämä voidaan viedä niin pitkälle, että alkuperäisen tekstin kirjoittaja kirjoittaa ainoastaan niin sanotun raakatekstin, josta lopullinen kohdekielinen teksti laaditaan (esimerkiksi tekninen kirjoittaminen voi olla tällaista). Pienemmässäkin mittakaavassa tätä ajatusta voi soveltaa esimerkiksi tapaukseen, jossa suomalaisen yrityksen edustaja lähettää joulun alla sähköpostia aasialaiselle asiakkaalle. Hän saattaa suomenkieliseen viestiinsä laittaa loppuun toivotuksen ”Hyvää joulua!”, jota asiantunteva kääntäjä ei käännä suoraan muotoon ”Merry Christmas!” vaan ”Season’s Greetings!”, koska aasialaisissa kulttuureissa joulua ei juhlita samalla tavalla ja samasta syystä kuin meillä.

Toisaalta on olemassa paljon tekstejä, joita ei ole lainkaan tarkoitettu käännettäviksi toisille kielille (tai edes kenenkään ulkopuolisen luettaviksi), mutta joita kuitenkin käännetään. Tällaisia ovat esimerkiksi joidenkin julkisuuden henkilöiden kirjeet ja päiväkirjat, joita julkaistaan kirjoina. Tällaisia tekstejä käännetään ihmiskääntäjänkin toimesta, mutta erityisesti konekäännös on mahdollistanut sen, että mitä tahansa kirjallista materiaalia voidaan ainakin yrittää kääntää kieleltä toiselle, ilman että se vaatii kovin paljoa aikaa tai resursseja. Interne-

tistä löytyvillä ilmaisilla konekäännösohjelmilla kuka tahansa internetin käyttäjä voi yrittää saada selkoa mistä tahansa vieraskielisestä tekstistä (edellyttäen tietysti, että joku konekääntäjä tukee kyseisiä kieliä).

Käännösprosessi on monivaiheinen, erityisesti jos sitä pohtii viestinnällisestä näkökulmasta. Ensimmäisenä tekijänä on lähtötekstin kirjoittaja ja hänen kykynsä välittää viestinsä selkeästi kirjoittamalla. Tähän vaikuttaa hänen kielellinen ja kirjallinen lahjakkuutensa. Tämän jälkeen joissakin tapauksissa (esimerkiksi yritysmaailmassa) joku toinen henkilö vielä editoi alkuperäistä lähtökielistä tekstiä. Sitten kääntäjä tekee oman tulkintansa tekstin rakenteesta, sisällöstä ja tarkoituksesta joko yksin tai alkuperäisen kirjoittajan avustuksella. Sitten hän tuottaa kohdekielisen tekstin pyrkien välittämään alkuperäisen tekstin viestin siten, että kohdekielisen kulttuurin edustaja ymmärtäisi sen niin kuin alkuperäisen tekstin kirjoittaja on sen tarkoittanut. Joskus varsinainen kääntäminen voi vielä jakautua osaprosesseihin. Jos osaavaa kääntäjää ei ole tarjolla, voidaan käännös tehdä myös välikielen, usein englannin, kautta, esimerkiksi suomi - englanti - hindi. Tällainen toimintatapa luonnollisesti lisää suuresti virhetulkintojen tekemisen mahdollisuutta, koska prosessiin tulee kolmannen kielen kääntäminen mukaan. Viimeisessä vaiheessa käännetyn tekstin lukija tekee oman tulkintansa lukemastaan. Lukijalla ei useimmissa tapauksissa ole mitään yhteyttä alkuperäisen tekstin kirjoittajaan, eli vastavuoroisuutta ei käännösviestinnässä yleensä ole.

Edellä esittämieni pohdintojen perusteella voisi ajatella, että konekääntämiseen pitäisi suhtautua hyvinkin skeptisesti. Toisaalta näin onkin, sillä kuten aiemmin esitin, konekääntäjä kääntää tekstiä ainoastaan sen pohjalta, mitä tekstiin on kirjoitettu. Konekääntäjä ei siis pysty tulkitsemaan tekstiä kovinkaan syvällisesti, eikä huomioimaan esimerkiksi kirjoitusvirheitä tai huomioimaan käännetyn tekstin tulevaa lukijaa tai kohderyhmää. Kuitenkin konekääntämisessä on runsaasti mahdollisuuksia, jos konekääntäjän tuottamaa tekstiä käytetään tarkoituksenmukaisella tavalla. Automaattisesti käännetty teksti voi esimerkiksi toi-

mia niin sanottuna raakakäännöksenä, josta ammattikääntäjä muokkaa lopullisen, julkaistavan tekstin. Toisaalta se voi auttaa ymmärtämään tekstiä summittaisesti ja määrittämään esimerkiksi sen, olisiko kyseinen teksti syytä käännättää julkaisukelpoiseen muotoon halutulle kielelle. Joskus konekääntäjän tuotos voi puolestaan olla riittävä sellaisenaankin, esimerkiksi sotilaallisissa tarkoituksissa tai vaikkapa markkinoilla olevia mahdollisia kilpailijoita selvitettäessä.

Suomen kielen näkökulmasta konekääntäjät ovat kuitenkin vielä aika lailla lapsenkengissään. Suomen kielen automaattinen kääntäminen onnistuu kunnolla ainoastaan suomi-englanti-suomi-kieliparille eikä tällaisiakaan varteenotettavia järjestelmiä ole oikeastaan muita kuin Kielikone Oy:n MOT Translation. Kokemukseni perusteella internetissä saatavilla olevat ilmaisohjelmat tuottavat erittäin heikkoa laatua erityisesti suomeen päin käännettäessä. Tosin niidenkin tuotokset saattavat soveltua sen selvittämiseen, mistä joku teksti ylipäätään kertoo. Tosin ei ole olemassa kovinkaan montaa konekäännösohjelmaa, joka ylipäätään tukee suomea. Jos tuki löytyy, se ei kuitenkaan useinkaan kata kovin montaa kieltä englannin ja suomen välisten käännösten lisäksi.

Tutkielman pohjalta uskallan kuitenkin todeta, että konekääntäjästä olisi varmasti hyötyä omassakin työssäni. Toistuvia tekstimassoja, kuten käyttö- ja huolto-ohjeita, voisi hyvin kääntää automaattisesti, ehkä jopa ilman jälkieditointia. Kyseiset tekstit ovat kuitenkin melko standardoituja niin sisällöltään kuin muodoltaankin. Kaikkiin minunkaan työssäni esiintyviin tekstityyppeihin konekääntäminen ei kuitenkaan sovellu ainakaan ilman etukäteis- ja jälkieditointia. Toisaalta näen, että jos editointia tarvitaan, vastaa kattava käännösmuistikin konekääntäjää. Käännösmuistiakin voi käyttää puoliautomaattisesti siten, että käännösmuistiohjelma pysähtyy ainoastaan sellaisten virkkeiden kohdalle, joita se ei löydä täsmälleen samassa muodossa. Tällöin voi kuitenkin luottaa siihen, että käännösmuistista löytyneet virkkeet tulevat käännettyä oikein (koska ne ovat varmennettuja käännöksiä) ja muut virkkeet tulee itse käännettyä tekstin edetessä käännösmuistiohjelman editorissa.

7 YHTEENVETO

Tässä tutkielmassa on pyritty käännösteoreettiseen keskusteluun pohjautuen ja esimerkkien kautta tuomaan esille luonnollisen kielen automaattisen käsittelyn, erityisesti kääntämisen, haasteita. Luonnolliset kielet ovat ominaisuuksiltaan sellaisia, että ainoastaan ihmisellä on luontainen kyky käyttää ja ymmärtää niitä. Ne ovat monitulkintaisia, kontekstisidonnaisia ja jatkuvasti muuttuvia. Näistä syistä luonnollisten kielten kääntäminen on haasteellinen tehtävä ihmisellekin. Konekääntämistä koskevassa tutkimuksessa ei useinkaan oteta kovinkaan syvällisesti kantaa kielitieteellisiin tai käännösteoreettisiin kysymyksiin, vaan pääpaino on teknisissä ratkaisuissa. Kuitenkaan näitä kysymyksiä ei pitäisi ohittaa, erityisesti kun ajatellaan konekäännösjärjestelmien käyttäjiä. Käyttäjällä pitäisi olla jonkinlainen käsitys luonnollisten kielten ja kääntämisen luonteesta yleensä, jotta hän osaisi käyttää myös konekääntäjää tarkoituksenmukaisesti.

Uusimmat konekääntäjät, erityisesti kaupalliset järjestelmät, ovat jo käyttökelpoisia vähintään raakakäännöksen tuottamiseen. Jotkut erityiseen kieliparii tai johonkin erikoisalaan keskittyvät järjestelmät saattavat pystyä hyvinkin sujuvaan lopputulokseen, mutta ne vaativat toisaalta lähtötekstiltä tiettyä rajattua terminologiaa ja tietynlaisia rakenteita. Toisaalta paljon dokumentteja tuottavilla yrityksillä on usein joka tapauksessa tietyt mallit ja säännöt siitä, miten dokumentteja kirjoitetaan, joten niitä ei välttämättä tarvitse erityisen paljon enää muokata konekäännöstä varten. Erityisen haasteellista konekääntäminen on kuitenkin siitä syystä, että luonnollinen kieli on jatkuvassa muutoksessa. Uusia sanoja ja rakenteitakin tulee koko ajan lisää, kun ihmiset kieltä käyttävät.

Tässä tutkielmassa ei pystytty kovinkaan syvällisesti paneutumaan konekääntäjien teknisiin toteutuksiin. Nykytilanteessa erityisen mielenkiintoista olisi perehtyä tarkemmin tilastolliseen konekäännökseen ja morfologisiin jäsentimiin. Tämä olisi ajankohtaista erityisesti suomen kielen konekääntämisen kannalta. Lisäksi olisi mielenkiintoista tehdä empiirinen, vertaileva tutkimus saatavilla

olevista suomi-englanti-suomi-konekääntäjistä kokeilemalla niitä käytännössä ja analysoimalla niiden tuottamia käännöksiä. Käännöslaatua voisi tutkia miettimällä tässä tutkielmassa käsiteltyjä kääntämisen eri tasoja, eli morfologiaa, syntaksia ja semantiikkaa. Saatujen tulosten perusteella saattaisi olla mahdollista esittää myös parannusehdotuksia konekääntäjien toteutukseen.

LÄHDELUETTELO

Arppe, A. 2007. Ei yhtä ainoaa polkua - Suomalaisia kokemuksia matkalla kieliteknologisesta tutkimuksesta liiketoimintaan [viitattu 22.2.2010]. Saatavilla www-muodossa http://www.ling.helsinki.fi/~aarppe/Publications/arppe-fi.shtml

Bassnett, S. 1995. Teoksesta toiseen. Jyväskylä: Gummerus Kirjapaino Oy.

Chesterman, A. 1996. Psst! Theory Can Be Useful! Kääntäjä-Översättaren 10/1996.

El-Kahlout, I. D. & Oflazer K. 2006. Initial Explorations in English to Turkish Statistical Machine Translation. Teoksessa Proceedings of the Workshop in Statistical Machine Translation, New York City, June 2006, 7-14.

Groves, D. & Way, A. 2005. Hybrid Data-Driven Models of Machine Translation. Machine Translation 19, 301-323.

Hutchins, J. 2005. Towards a Definition of Example-Based Machine Translation [viitattu 5.11.2009]. Saatavilla www-muodossa http://www.hutchinsweb.me.uk/MTS-2005.pdf.

Hutchins, J. 2006. Machine Translation: History. Teoksessa Brown, K. (toim.) Encyclopedia of Language & Linguistics, Vol. 7, 375-383.

Jäppinen, H., K. Hartonen, L. Kulikov, A. Nykänen & A. Ylä-Rotiala 1993. Kielikone Machine Translation Workstation. Teoksessa Nirenburg, S. (toim.) Progress in Machine Translation, 173-184. Oxford: IOS Press.

Karlsson, F. 2004. Yleinen kielitiede. Helsinki: Yliopistopaino.

Kielikone Oy:n kotisivut 2009 [viitattu 22.2.2010]. Saatavilla www-muodossa http://www.kielikone.fi/fi/tuotesivu/?intProductID=1379.

Koehn, P. 2005, Europarl: A Parallel Corpus for Statistical Machine Translation. Teoksessa Proceedings of the 10th Machine Translation Summit, Phuket, Thailand, 79-86.

Krikke, J. 2006. Machine Translation Inching Towards Human Quality. IEEE Intelligent Systems, March/April 2006, 4-6.

Laitinen, H-R. 2008. Euromap HLT Case Study: TranSmart® - konekäännösjärjestelmä Rautaruukissa [viitattu 22.2.2010]. Saatavilla www-muodossa https://kitwiki.csc.fi/twiki/bin/view/FiLT/RautaruukkiFi

Language Weaverin kotisivut 2010 [viitattu 22.2.2010]. Saatavilla www-muodossa http://www.languageweaver.com.

Munteanu, D. S. & Marcu, D. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Computational Linguistics 31(4), 477-504.

Nida, E.A. 1964. Toward a Science of Translating. Leiden: Brill.

Nuutila, P. 2005. Rough Machine Translation in the Communication Process. Licenciate Thesis. University of Tampere. School of Modern Languages and Translation Studies.

Och, F. J & Ney, H. 2004. The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics 30(4), 417-449.

Slocum, J., 1985. A Survey of Machine Translation: Its History, Current Status, and Future Prospects. Computational Linguistics, 11(1), 1-17.

Somers, H. 1999. Review Article: Example-Based Machine Translation. Machine Translation 14, 113-157.

Tiittula, L. 2008. Käännöstiede [viitattu 22.2.2010]. Saatavilla www-muodossa http://www.edu.fi/page.asp?path=498,1329,1393,86675,67374,67400.

Tommola, J. (toim.) 2004. Kieli, teksti ja kääntäminen. Language, text and Translation. Turku: Painosalama Oy.

Vehmas-Lehto, I. 2002. Kopiointia vai kommunikointia. Johdatus käännösteoriaan. Helsinki: Hakapaino Oy.

Virpioja, S., Väyrynen, J.J., Creutz, M. & Sadeniemi, M. 2007. Morphology-Aware Statistical Machine Translation Based on Morphs Induced in an Unsupervised Manner. Teoksessa Proceedings of the Machine Translation Summit XI, September 2007, Copenhagen, 491-498.