

MONITULKINTAISTEN SANOJEN KÄÄNTÄMINEN  
KONEKÄÄNNÖSJÄRJESTELMILLÄ  
ENGLANNISTA SUOMEKSI

Erja Salminen

Tampereen yliopisto  
Informaatiotieteiden yksikkö  
Informaatiotutkimus ja  
interaktiivinen media  
Pro gradu -tutkielma  
Toukokuu 2012

Konekääntäminen on yksi kääntämismenetelmä kieltenvälisessä tiedonhaussa, ja monitulkintaisten sanojen kääntäminen on yksi kieltenvälisen tiedonhaun haasteista. Tässä tutkimuksessa testattiin, kuinka virheettömästi konekäännösjärjestelmät kääntävät englanninkielisiä monitulkintaisia sanoja suomeksi. Ensimmäisenä selvitettiin, onko virkeyhteysessä esiintyvien yksikäsitteisten ja monitulkintaisten sanojen käännösten laadussa eroja. Toisena tarkasteltiin, onko monitulkintaisten sanojen käännösten laadussa eroja eri konekäännösjärjestelmien välillä. Kolmantena selvitettiin, onko monitulkintaisten sanojen käännösten laadussa eroja erilaajuisten asiayhteyksien välillä. Neljäntenä tarkasteltiin, onko monitulkintaisten sanojen käännösten laadussa eroja sääntöperustaisen ja tilastollisten konekäännösjärjestelmien välillä. Lopuksi määriteltiin, minkä tyyppisiä virheitä konekäännösjärjestelmät tekevät virkeyhteysessä esiintyvien monitulkintaisten sanojen kääntämisessä.

Testattavat järjestelmät olivat TeemaPoint, Sunda, SDL Free Translation, Google Translate ja Bing<sup>TM</sup> Translator Beta, joista kolme ensimmäistä on sääntöperustaisia ja kaksi jälkimmäistä tilastollisia. Aineistona käytettiin 50 englanninkielisen monitulkintaisen sanan ympärille luotua 195 virkeparia, jotka koostuivat virkkeen lyhyestä ja pitkästä versiosta. Monitulkintaiset sanat käännettiin myös yksittäisinä sanoina ilman asiayhteyttä. Käännösten laadun evaluointimittarina käytettiin tarkkuutta.

Tulosten perusteella virkeyhteysessä esiintyvien monitulkintaisten sanojen käännösten laatu on huomattavasti heikompaa kuin yksikäsitteisten sanojen. Kaikkien erilaajuisissa asiayhteyksissä tehtyjen monitulkintaisten sanojen käännösten perusteella Sunda ja Google Translate ovat parhaat järjestelmät ja SDL Free Translation ja Bing<sup>TM</sup> Translator Beta huonoimmat. TeemaPoint jää parhaimpien ja huonoimpien järjestelmien välimaastoon. Järjestelmien keskinäinen paremmuusjärjestys vaihtelee käännettävän monitulkintaisen sanan asiayhteyden laajuudesta riippuen.

Yksittäisinä sanoina käännettyjen monitulkintaisten sanojen käännösten laatu on heikompaa kuin virkeyhteysessä käännettyjen sanojen. Käännösten laadussa ei kuitenkaan ole eroa erilaajuisten virkeyhteyksien välillä.

Kaikkien erilaajuisissa asiayhteyksissä tehtyjen monitulkintaisten sanojen käännösten perusteella sääntöperustaisen ja tilastollisen kääntämismenetelmän käännöslaadussa ei ole eroja. Tilastollinen menetelmä kääntää kuitenkin sääntöperustaista laadukkaammin yksin esiintyviä monitulkintaisia sanoja, kun taas sääntöperustainen menetelmä kääntää tilastollista laadukkaammin virkeyhteysessä esiintyviä monitulkintaisia sanoja.

Virkeyhteysessä esiintyvien monitulkintaisten sanojen käännökset voidaan luokitella kuuteen virhetyyppiin. Yleisin virhe on se, että monitulkintaisen sanan käännökseksi valitaan väärä vaihtoehto tarkasteltavan monitulkintaisen sanan käännösvaihtoehtoista. Toiseksi yleisin virhetyyppi on se, että tarkasteltava monitulkintainen sana jää kokonaan kääntämättä. Muita virhetyyppejä esiintyy selvästi vähemmän. Virhetyyppien osuuksien jakaumissa on eroja järjestelmien välillä.

Avainsanat: kieltenvälinen tiedonhaku, konekääntäminen, monitulkintaisuus, yksikäsitteistäminen

# Sisällysluettelo

1	JOHDANTO	5
2	KÄÄNTÄMINEN KIELTENVÄLISESSÄ TIEDONHAUSSA	8
2.1	Kääntämisstrategiat	9
2.1.1	Kyselyiden kääntäminen	9
2.1.2	Dokumenttien kääntäminen	10
2.1.3	Kieliriippumattoman väliesityksen käyttäminen	11
2.2	Kääntämismenetelmät	12
2.2.1	Sanakirjaperustainen menetelmä	12
2.2.2	Korpusperustainen menetelmä	13
3	KONEKÄÄNTÄMINEN	15
3.1	Yleistä	15
3.2	Käännettävän tekstin pituuden yhteys konekäännöksen laatuun	17
3.3	Konekääntämisen kääntämismenetelmät	19
3.3.1	Sääntöperustainen konekääntäminen	20
3.3.2	Tilastollinen konekääntäminen	23
3.3.3	Esimerkkiperustainen konekääntäminen	26
4	SANOJEN MONITULKINTAISUUS	28
4.1	Monitulkintaisuuden tyypit	28
4.2	Monitulkintaisten sanojen kääntäminen	29
4.3	Sanan merkityksen yksikäsitteistämismenetelmät	30
4.3.1	Sanaluokkien jäsentäminen	32
4.3.2	Rinnakkaiskorpusten hyödyntäminen	33
4.3.3	Sanojen yhteisesiintymisen todennäköisyyden hyödyntäminen	34
4.3.4	Kyselynlaajennus	35
4.3.5	Rakenteisten kyselyiden käyttäminen	35
4.3.6	Erikoissanakirjan käyttäminen	36
4.3.7	Kaksisuuntainen kääntämistekniikka	37
4.3.8	Koneoppimisen menetelmät	37
5	KONEKÄÄNNÖSJÄRJESTELMIEN EVALUOINTIMITTAREITA	38
5.1	BLEU	38
5.2	METEOR	40
5.3	Tarkkuus	41
6	TUTKIMUSASETELMA	42
6.1	Tutkimuskysymykset	43
6.2	Testatut konekäännösjärjestelmät	44
6.2.1	TeemaPoint	45
6.2.2	Sunda	46
6.2.3	SDL Free Translation	47
6.2.4	Google Translate	48
6.2.5	Bing™ Translator Beta	49
6.2.6	Yhteenvetoa testatuista konekäännösjärjestelmistä	49
6.3	Aineiston kerääminen	50
6.3.1	Käännettävien virkkeiden luominen	50
6.3.2	Sanojen ja virkkeiden kääntäminen konekäännösjärjestelmillä	54

6.4	Aineiston analysointi .....	54
6.4.1	Käännösten pisteyttäminen .....	55
6.4.2	Tarkkuusarvojen ja prosenttiosuuksien laskeminen .....	61
6.4.3	Tilastollisten merkitsevyydestien suorittaminen .....	64
7	TULOKSET .....	67
7.1	Käännöslaadun erot yksikäsitteisten ja monitulkintaisten sanojen välillä .....	67
7.2	Käännöslaadun erot konekäännösjärjestelmien välillä .....	69
7.3	Käännöslaadun erot erilaajusten asiayhteyksien välillä .....	69
7.4	Käännöslaadun erot kääntämismenetelmien välillä .....	73
7.5	Monitulkintaisten sanojen käännösten virhetyypit .....	75
8	POHDINTA .....	77
8.1	Menetelmien pohdinta .....	78
8.2	Tulosten pohdinta .....	80
	LÄHTEET .....	87
	LIITTEET	

# 1 JOHDANTO

Konekäännösjärjestelmän tuottaman käännöksen laatu on heikko, jos järjestelmä kääntää virkkeen ”*I play the flute.*” virkkeeksi ”*Pelaan huilua.*” tai ”*Esitän huilua.*”. Tässä esimerkissä konekäännösjärjestelmä ei osannut valita monitulkintaiselle sanalle *play* oikeaa käännöstä sanan monista käännösvaihtoehdoista, eli se ei yksikäsitteistänyt sanan merkitystä oikein. Konekääntäminen on yksi kieltenvälisessä tiedonhaussa (CLIR, cross-language information retrieval) käytettävistä kääntämismenetelmistä (Hutchins 2005b, 69), ja monitulkintaisten sanojen kääntäminen on yksi kieltenvälisen tiedonhaun suurimmista haasteista (Kishida 2005, 439). Kieltenvälisessä tiedonhaussa käyttäjän tekemä kysely on erikielinen kuin haun tuloksena saatavat dokumentit (Kishida 2005, 433). Lähdekielellä tarkoitetaan alkuperäisen kyselyn kieltä ja kohdekielellä kieltä, jolle käännetään (Hutchins & Somers 1992, 4).

Kieltenvälisen tiedonhaun kysymyksiä on käsitelty jo useita vuosikymmeniä (Kishida 2005, 433). Etenkin 1990-luvulla ilmestyneen World Wide Webin myötä käyttäjä on päässyt helposti erikielisille tiedonlähteille (Kettunen 2009, 103; Kishida 2005, 433). Tarjolla olevan monikielisen informaation määrän nopea kasvaminen on vain lisännyt tarvetta kieltenväliselle tiedonhauille (Braschler & Schäuble 2000, 273) sekä kieltenvälisen tiedonhaun tutkimukselle (Pirkola 1998a, 48). Kieltenvälinen tiedonhaku on tärkeää esimerkiksi internetin hakupalveluissa, koska suurin osa internetin sisällöstä on englanniksi, kun taas suurin osa hakupalveluiden käyttäjistä ei puhu äidinkielenään englantia (Wu, He, Ji & Grishman 2008, 71).

Oardin ja Dorrin (1996, 2) mukaan kieltenvälisestä tiedonhausta on hyötyä, jos tiedonhakijan kielitaito riittää haun tuloksena saatavien tietynkielisten dokumenttien ymmärtämiseen mutta ei kyselyn muodostamiseen kyseisellä kielellä. Myös Airio (2008, 760) päätyi tutkimuksessaan siihen, että kieltenvälisestä tiedonhausta on hyötyä erityisesti käyttäjälle, jolla on kohtalainen ja passiivinen kielitaito.

Konekääntäminen on automaattista tekstin tai puheen kääntämistä kieleltä toiselle. Konekäännösjärjestelmien kehittämisen taustalla on ollut haave, että eri kieliä puhuvat ihmiset voisivat kommunikoida keskenään. Konekääntäminen on kuitenkin haastavaa, ja konekäännösjärjestelmät tuottavat heikkolaatuisia käännöksiä, joita käyttäjän on jälkikäteen paranneltava. (Manning & Schütze 2000, 463.) Vaikka konekäännösjärjestelmi-

en käännöslaatu on vielä heikkoa, ainakin Gasparin (2006, 54) tutkimuksen perusteella käyttäjillä on myönteinen asenne ilmaisia konekäännösjärjestelmiä kohtaan, ja konekääntäminen koetaan hyödylliseksi aputyökaluksi.

Sana on monitulkitainen, jos sillä on useampi kuin yksi merkitys. Kun monitulkitaista sanaa käytetään tekstissä tai puheessa, tavallisesti vain yksi sanan merkityksistä on oikea. Ihminen pystyy päättämään sanan oikean merkityksen tilanteen ja asiayhteyden perusteella. Tietokoneille oikean merkityksen tunnistaminen sen sijaan on haasteellinen tehtävä. (Kilgarriff & Palmer 2000, 1.) Koska oikean käännöksen tuottaminen edellyttää oikeaa yksikäsitteistämistä (Kilgarriff & Palmer 2000, 1), yksikäsitteistämällä on tärkeä rooli konekääntämisessä (Navigli 2009, 2).

Konekäännösjärjestelmien evaluointi on tärkeää, koska esimerkiksi Kettusen (2009, 114) tutkimuksen perusteella konekäännöksen laadusta voidaan ennustaa käännetyllä kyselyllä suoritettavan tiedonhaun tuloksellisuutta. McNameen ja Mayfieldin (2002, 159) mukaan käännösresurssien laatu on tärkein kieltenvälisen tiedonhaun tuloksellisuuden vaikuttava tekijä.

Tämän tutkimuksen tarkoituksena on evaluoida, kuinka virheettömästi konekäännösjärjestelmät kääntävät englanninkielisiä monitulkitaisia sanoja suomeksi. Testattavana on viisi verkossa toimivaa konekäännösjärjestelmää. Ensimmäisenä tutkimuksessa vertailaan, onko virkeyhteydessä esiintyvien yksikäsitteisten ja monitulkitaiden sanojen käännösten laadussa eroja. Toisena selvitetään, onko monitulkitaiden sanojen käännösten laadussa eroja testattavien konekäännösjärjestelmien välillä. Kolmantena vertailaan, onko monitulkitaiden sanojen käännösten laadussa eroja, kun tarkasteltava monitulkitainen sana esiintyy kolmessa erilaajuisessa asiayhteydessä. Useiden aiempien tutkimusten mukaan konekääntäminen on laadukkaampaa pitkien kuin lyhyiden kyselyiden kääntämisessä, koska konekääntäminen tarvitsee oikean käännöksen päättämiseksi tietoa sanojen asiayhteydestä (ks. esim. Chen & Gey 2004, 150; Kishida 2005, 437; Oard & Hackett, 1998, 1). Neljäntenä tarkastellaan, onko monitulkitaiden sanojen käännösten laadussa eroja kahden eri konekääntämisen kääntämismenetelmän välillä. Lopuksi selvitetään, minkä tyyppisiä virheitä konekäännösjärjestelmät tekevät virkeyhteydessä esiintyvien monitulkitaiden sanojen kääntämisessä.

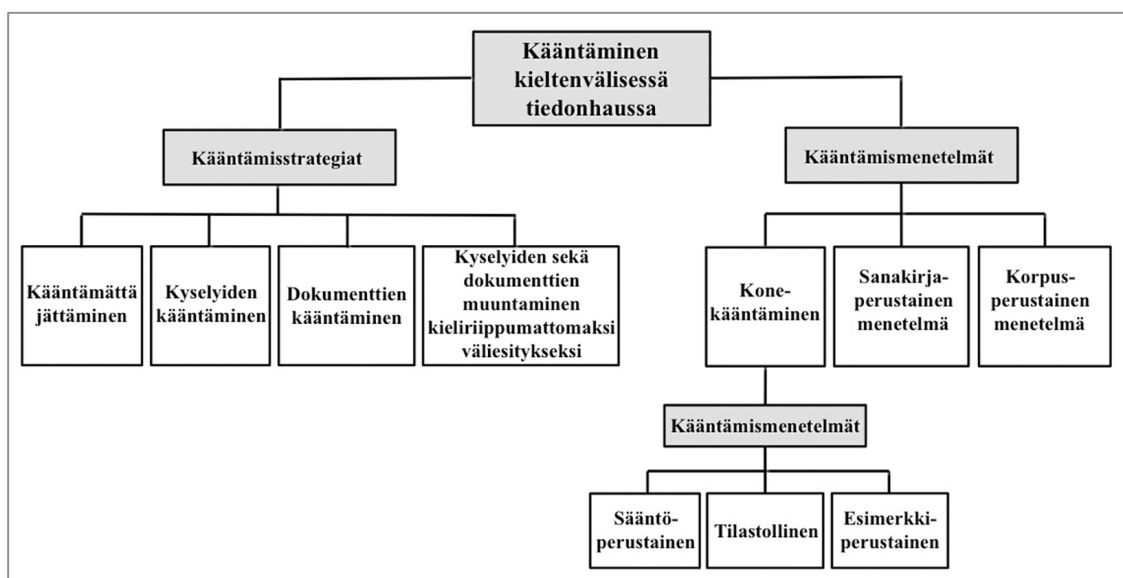
Täysin vastaavanlaista tutkimusta ei ole havaittu, joten tutkimuksen näkökulma ja rajaus ovat uusia. Tutkimusta puoltaa myös se, että konekäännösjärjestelmiä käsittelevissä

evaluointitutkimuksissa ei juurikaan ole tutkittu suomen kieltä. Suomen kielen vähäisen tutkimisen syynä voi olla se, että suomea ei ole vaihtoehtona kovinkaan monessa konekäännösjärjestelmässä eikä suomi ole muutenkaan yleinen kieli.

Tässä tutkielmassa tarkastellaan kääntämistä kieltenvälisessä tiedonhaussa pelkkää konekääntämistä laajemmin, jotta saadaan kokonaiskuva kieltenvälisen tiedonhaun eri osa-alueista ja kääntämisen vaihtoehdoista. Luvussa 2 keskitytään kääntämiseen kieltenvälisessä tiedonhaussa. Kyseisessä luvussa esitellään sekä kääntämisstrategioita että kääntämismenetelmiä. Koska konekääntäminen on tämän tutkielman keskeisin aihe, kyseistä kääntämismenetelmää käsitellään erikseen luvussa 3. Luvussa 4 keskitytään monitulkin-taisiin sanoihin, niiden kääntämiseen sekä sanan merkityksen yksikäsitteistämismene-telmiin. Luvussa 5 esitellään lyhyesti kolme konekääntämisen evaluoinnissa käytettävis-tä mittareista. Luvussa 6 puolestaan esitellään tutkimusasetelma, joka kattaa tutkimus-kysymykset, testattujen konekäännösjärjestelmien esittelyn sekä aineistonkeruun ja ana-lysoinnin kuvaamisen. Luvussa 7 raportoidaan tutkimustulokset. Luku 8 sisältää käytet-tyjen menetelmien ja tulosten arviointia sekä ideoita mahdollisille jatkotutkimuksille.

## 2 KÄÄNTÄMINEN KIELTENVÄLISESSÄ TIEDONHAUSSA

Zhun ja Wangin (2006, 593) tutkimuksen mukaan kääntäminen on tärkeässä asemassa kieltenvälisessä tiedonhaussa, koska tiedonhaun tuloksellisuus ja käännetyn kyselyn laatu korreloivat voimakkaasti. Myös Huang, Sun, Li, Pan ja Zhang (2007, 81) mainitsevat, että kysely on välttämätöntä kääntää oikein ja että käännökseen laatu vaikuttaa hakutuloksiin. Kettunen (2009, 114) kuitenkin toteaa, että on itsestäänselvää, että käännökseen laatu vaikuttaa tiedonhaun tuloksellisuuteen. Kieltenvälisen tiedonhaun tuloksellisuus antaa tietoa käytetyn konekäännösjärjestelmän laadusta, koska mitä paremmat hakutulokset ovat, sitä parempi käytetty käännosresurssi on (Kettunen 2009, 104). Yhteys voidaan nähdä myös niin päin, että konekäännökseen laadusta voidaan ennustaa käännetyllä kyselyllä tehtävän tiedonhaun tuloksellisuutta (Kettunen 2009, 114). Esimerkiksi McNamee ja Mayfield (2002, 165) osoittivat, että kieltenvälisen tiedonhaun tuloksellisuus heikkenee, kun kääntämisen sanastollisten resurssien kattavuus vähenee. Kyseisessä tutkimuksessa käytettyjä käännosresursseja huononnettiin ennen kyselyiden kääntämistä. Käännosresurssien laatu onkin tärkein kieltenvälisen tiedonhaun tuloksellisuuden vaikuttava tekijä. (McNamee & Mayfield 2002, 159.)



Kuva 1. Kääntämisen eri osa-alueita kieltenvälisessä tiedonhaussa (ks. Fujii & Ishikawa 2000, 23; Hutchins 2005b, 69; Kishida 2005, 436; Oard 1998, 473)

Kuvasta 1 ilmenevät kieltenvälisessä tiedonhaussa suoritettavan kääntämisen osa-alueet, joita luvuissa 2 ja 3 tarkastellaan. Kuvan luomisessa on käytetty usean eri lähteen tietoja



(ks. Fujii & Ishikawa 2000, 23; Hutchins 2005b, 69; Kishida 2005, 436; Oard 1998, 473). Luvussa 2.1 esitellään kääntämisstrategioita, jotka määräävät, mitä kieltenvälisessä tiedonhaussa käännetään. Kääntämismenetelmät puolestaan määräävät, mitä resursseja ja keinoja kääntämisessä käytetään. Kääntämismenetelmistä sanakirjaperustaista ja korpusperustaista menetelmää tarkastellaan luvussa 2.2 ja konekääntämistä luvussa 3. Kuten kuvasta 1 ilmenee, konekääntäminen voi kääntämismenetelmältään olla sääntöperustaista, tilastollista tai esimerkkiperustaista.

## **2.1 Kääntämisstrategiat**

Kieltenvälinen tiedonhaku perustuu kääntämiseen: käännetään joko kyselyt dokumenttien kielille tai dokumentit kyselyiden kielelle (Airio 2009, 28; Fujii & Ishikawa 2000, 23). Kolmantena mahdollisuutena on esittää sekä kyselyt että dokumentit kieliriippumattomassa muodossa (Fujii & Ishikawa 2000, 23). Kieliriippumattomassa väliesityksessä kaikki merkitykseltään samanlaiset virkkeet esitetään samalla tavalla riippumatta kielestä, jolla ne alun perin on ilmaistu (Jurafsky & Martin 2000, 812). Neljäs vaihtoehto on se, ettei käännetä mitään, vaan erikieliset sanat täsmäytetään sellaisenaan (Oard 1998, 473). Kääntämättä jätetään lähde- ja kohdekielellä samaa alkuperää olevat sanat, esimerkiksi erisnimet tai tekniset termit, joiden oletetaan täsmävän vastaavien erikielisten sanojen kanssa sellaisenaan (Kishida 2005, 434-435). Seuraavaksi tarkastellaan kyselyiden kääntämistä, dokumenttien kääntämistä sekä kieliriippumattoman väliesityksen käyttämistä kääntämisessä.

### **2.1.1 Kyselyiden kääntäminen**

Kyselyiden kääntäminen on helppoutensa takia yleisin kääntämisstrategia (Airio 2009, 28; Kishida 2005, 435). Kun käännetään kyselyt dokumenttien kielelle, hakujärjestelmän indeksejä ei tarvitse muuttaa (Kishida 2005, 435). Kieltenvälisen tiedonhaun tutkimukset ovatkin pääasiassa keskittyneet kyselyiden kääntämismenetelmiin, koska etenkin isoissa dokumenttikokoelmissa kyselyiden kääntäminen on kannattavampi ja käytännöllisempi vaihtoehto kuin dokumenttien kääntäminen (Ballesteros & Croft 1998, 64; Pirkola 1998a, 49).

Kyselyiden kääntäminen ei välttämättä tapahdu suoraan lähdekieleltä kohdekielelle, vaan apuna saatetaan käyttää jotain kolmatta kieltä välikielenä. Välikieltä käytetään

korvaamaan puuttuvia kielellisiä resursseja. (Kishida 2005, 445; Lehtokangas, Airio & Järvelin 2004, 3.) Esimerkiksi käännettäessä japanista hollanniksi voidaan englantia käyttää välikielenä, koska käänösresursseja japanista englanniksi ja englannista hollanniksi on todennäköisemmin saatavilla kuin resursseja suoraan japanista hollanniksi. Ensin siis kyselyt käännetään japanista englanniksi ja tämän jälkeen englannista hollanniksi. (Kishida 2005, 445.) Lehtokankaan ja muiden (2004, 11) tutkimuksen mukaan välikieltä käyttävä kääntäminen on tiedonhaun tuloksellisuudella mitattuna lähes yhtä laadukasta kuin suoraan kieleltä toiselle kääntäminen ilman välikieltä.

### **2.1.2 Dokumenttien kääntäminen**

Dokumenttien kääntämisestä olisi hyötyä käyttäjälle, mutta laajan dokumenttikokoelman kääntäminen eri kielille on työlästä (Airio 2009, 30). Kieltenvälisessä tiedonhaussa dokumenttien kääntämistä hyödynnetäänkin harvoin (Kishida 2005, 435). Esimerkiksi Rosemblatin, Gemoetsin, Brownen ja Tsen (2003, 566) tutkimuksessa kyselyiden kääntäminen osoittautui paremmaksi kääntämisstrategiaksi kuin dokumenttien kääntäminen. Kyseisessä tutkimuksessa vertailtiin tiedonhaun tuloksellisuutta, kun käännettiin kyselyt tai kokoelman kaikki dokumentit konekäännösjärjestelmällä (Rosemblat ym. 2003, 566).

Oardin (1998, 474) tutkimustulos poikkeaa edellä mainitusta Rosemblatin ja muiden (2003, 566) tutkimustuloksesta: Oardin (1998, 474) mukaan dokumenttien kääntäminen on tiedonhaun tuloksellisuudella mitattuna parempi kääntämisstrategia kuin kyselyiden kääntäminen. Tutkimustulos perustuu näkemykseen, jonka mukaan dokumenteissa on kyselyitä enemmän tietoa sanojen asiayhteydestä, koska kyselyt ovat yleensä lyhyitä eivätkä useinkaan täydellisesti muodostettuja kokonaisia virkkeitä (Oard 1998, 474). Oardin (1998, 474) tutkimustulosten edellytyksenä olikin se, että konekäännösjärjestelmän oli kääntäessään pystyttävä hyödyntämään sanojen asiayhteyttä. Myös Oard ja Hackett (1998, 6) raportoivat, että dokumenttien kääntäminen kaupallisella konekäännösjärjestelmällä suoriutuu kyselyiden kääntämisestä paremmin. Tutkimuksessa evaluoitiin käännettyillä kyselyillä tehtyjen tiedonhakuisten tuloksellisuutta (Oard & Hackett 1998, 2).

Fujii ja Ishikawa (2000, 13-14) kehittivät menetelmän, jossa käännetään kysely ja osa haetuista dokumenteista. Ensin lähdekielinen kysely käännetään esimerkiksi konekäännösjärjestelmällä dokumenttien kielelle (Fujii & Ishikawa 2000, 14-15). Kohdekielille

käännetyllä kyselyllä haetaan kohdekielisiä dokumentteja. Haun jälkeen alkuperäisen kyselyn kielelle eli lähdekielelle käännetään konekääntämisellä ainoastaan relevanssilajiteltujen hakutulosten kärkidokumentit, esimerkiksi tulosten tuhat ensimmäistä dokumenttia. (Fujii & Ishikawa 2000, 14.) Lopuksi käännetyt kärkidokumentit lajitellaan vielä uudelleen näiden kärkidokumenttien ja alkuperäisen lähdekielisen kyselyn samankaltaisuuden perusteella (Fujii & Ishikawa 2000, 16).

Fujii ja Ishikawa (2000, 13) osoittivat tiedonhaun tuloksellisuudella mitattuna menetelmänsä paremmuuden pelkkien kyselyiden kääntämiseen verrattuna. Lähdekielelle käännettyjen kärkipään dokumenttien uudelleen järjestäminen vaikuttaa olevan oleellinen vaihe tässä menetelmässä. Menetelmällä vältetään kokoelman kaikkien dokumenttien kääntämisen työmäärältä (Fujii & Ishikawa 2000, 14). Etuna on myös se, että hakutuloksena saatavat dokumentit ovat samankielisiä kuin käyttäjän tekemä kysely.

### **2.1.3 Kieliriippumattoman väliesityksen käyttäminen**

Kääntämisessä voidaan käyttää myös kieliriippumatonta väliesitystä, johon sekä kysely että dokumentit muunnetaan (Kishida 2005, 436). Väliesitys on tiedonesitysmuoto, joka on riippumaton tavasta, jolla eri kielet ilmaisevat merkityksiä (Manning & Schütze 2000, 465). Kieliriippumattomassa väliesityksessä kaikki merkitykseltään samanlaiset virkkeet esitetään samalla tavalla riippumatta kielestä, jolla ne alun perin on ilmaistu (Jurafsky & Martin 2000, 812).

Joissain kielissä kieliriippumattomana väliesityksenä voidaan käyttää esimerkiksi WordNetiä, joka on koneen luettavissa oleva tesaurus (Kishida 2005, 436). WordNet käyttää synsettejä eli synonyymiryhmiä. Jokainen synonyymiryhmä edustaa yhtä käsitettä, ja ryhmät on yhdistetty toisiinsa semanttisten ja sanastollisten suhteiden perusteella. (WordNet 2011.) Koska tiettyä käsitettä edustava synonyymiryhmän numero vastaa tiettyä sanojen joukkoa kaikilla tesaurun tukemilla kielillä, lähdekielisen kyselyn sanat voidaan yhdistää niitä vastaaviin kohdekielisiin sanoihin ryhmän numeron perusteella (Kishida 2005, 436). WordNetiä ei kuitenkaan ole käytettävissä kaikille kielille.

## 2.2 Kääntämismenetelmät

Kieltenvälisessä tiedonhaussa käytetään kolmea kääntämismenetelmää: kääntämistä kaksikielisen, koneen luettavissa olevan sanakirjan avulla (machine-readable dictionary, MRD), korpuksiin perustuvaa menetelmää (corpora-based method) sekä konekääntämistä (machine translation, MT) (Kishida 2005, 436). Koska tämän tutkielman keskeisin aihe ja kääntämismenetelmä on konekääntäminen, tätä kolmatta kieltenvälisen tiedonhaun kääntämismenetelmää tarkastellaan erikseen luvussa 3. Kaksi muuta kääntämismenetelmää kuitenkin esitellään lyhyesti, jotta saadaan kokonaiskuva kääntämismenetelmien vaihtoehdoista kieltenvälisessä tiedonhaussa. Seuraavaksi ensin tarkastellaan sanakirjaperustaista menetelmää ja sen jälkeen esitellään korpusperustaista menetelmää.

### 2.2.1 Sanakirjaperustainen menetelmä

Sanakirjaperustainen menetelmä perustuu koneen luettavissa olevan sanakirjan käyttämiseen kääntämisessä (Ballesteros & Croft 1998, 64; Hedlund 2003, 26). Sanakirjaperustainen kääntäminen on kaikista yleisin kääntämismenetelmä kieltenvälisessä tiedonhaussa, koska käännessanakirjat ovat usein suhteellisen halpoja ja helppoja käyttää (Airio 2009, 30). Esimerkiksi kyselyiden kääntäminen on usein helpompaa toteuttaa sanakirjaperustaisella menetelmällä kuin korpusperustaisella menetelmällä, koska korpuksia ei ole valmiina saatavilla (Pirkola 1998b, 55). Myös Huang ja muut (2007, 82) mainitsivat, että sanakirjaperustainen kääntäminen on melko yksinkertaista.

Lähdekielisen kyselyn sanat kannattaa muuttaa perusmuotoon ennen kääntämistä, koska käännessanakirjat eivät sisällä sanojen taivutusmuotoja (Airio 2009, 30; Pirkola 1998a, 51). Vasta nämä perusmuotoon muutetut lähdekieliset sanat käännetään kohdekielelle (Airio 2009, 30). Sanakirjaperustaisessa menetelmässä lähdekielinen sana korvataan kaikilla sanakirjassa olevilla kohdekielisillä vastineilla, jotka kaikki sisällytetään varsinaiseen kyselyyn (Ballesteros & Croft 1998, 64; Hedlund 2003, 26; Pirkola 1998a, 50). Kääntämisen jälkeen, ennen hakemista, kyselyn sanat tarvittaessa muutetaan hakuindeksiä vastaavaan muotoon (Airio 2009, 30-31), esimerkiksi muuntamalla ne sanavartaloiksi.

Koska sanakirjaperustaisessa kääntämisessä kaikki käänösvaihtoehdot sisällytetään käännettyyn kyselyyn, oikeakin käänös on varmasti mukana, jos vain sanakirja on riit-

tävän kattava (Airio 2009, 30). Useiden käännosten sisällyttäminen kyselyyn saattaa toimia myös kyselynlaajenuksena, koska käännoissanakirjat sisältävät usein synonyymeja (Airio 2009, 30). Kyselynlaajennus on menetelmä, jossa käyttäjän syöttämään kyselyyn lisätään automaattisesti valittuja sanoja alkuperäisen kyselyn hakutulosten kärkidokumenteista. Kieltenvälisessä tiedonhaussa kyselyä voidaan laajentaa ennen kyselyn kääntämistä tai kääntämisen jälkeen. (Kishida 2005, 441.)

Kaikkien käännosvaihtoehtojen sisällyttäminen kyselyyn aiheuttaa myös ongelmia, koska useista käännosvaihtoehtoista johtuva monitulkintaisuus heikentää haun tuloksellisuutta (Ballesteros & Croft 1996, 800). Yhtenä sanakirjaperustaisen kääntämisen ongelmana onkin monitulkintaisuus ja asiaankuulumattoman informaation lisääminen kyselyyn (Ballesteros & Croft 1996, 800; Ballesteros & Croft 1997, 84). Myös Huang ja muut (2007, 82) mainitsevat sanojen monitulkintaisuuden aiheuttavan ongelmia sanakirjaperustaisessa menetelmässä. Monitulkintaisten sanojen kääntämistä ja käannosten yksikäsitteistämistä käsitellään lisää luvuissa 4.2 ja 4.3.

Sanakirjaperustaisessa menetelmässä on muitakin ongelmia kuin edellä mainittu sanojen monitulkintaisuus. Ongelmia ilmenee esimerkiksi sanakirjaan sisällyttämättömien sanojen kääntämisessä (Ballesteros & Croft 1997, 84; Huang ym. 2007, 82). Sanakirjan kattavuus onkin yksi menetelmän heikkous liittyen etenkin erikoisalojen termeihin, joita yleissanakirjoissa on vähän (Pirkola 1998a, 50). Sanakirjaperustaisella menetelmällä on vaikeuksia myös fraasien tunnistamisessa (Ballesteros & Croft 1997, 84; Pirkola 1998a, 50); fraasit kadotetaan, jos ne käännetään sana sanalta (Ballesteros & Croft 1996, 800). Pirkola, Hedlund, Keskustalo ja Järvelin (2001, 210) mainitsevat sanakirjaperustaisen menetelmän ongelmana myös sanojen taipumisen käsittelemisen.

### **2.2.2 Korpusperustainen menetelmä**

Korpusperustainen menetelmä hyödyntää rinnakkais- tai vastinkorpuksia (Pirkola 1998a, 49), joista johdetaan kääntämistietoa (Talvensaari 2008a, 29). Rinnakkaiskorpuksissa kaikki dokumentit on käännetty korpusten kaikille kielille (Ballesteros & Croft 1998, 65; Pirkola 1998a, 49). Vastinkorpuksissa dokumenteista ei sen sijaan ole suoria käännoksiä, vaan erikielisten korpusten dokumentit vain käsittelevät samaa aihealuetta (Ballesteros & Croft 1998, 64; Pirkola 1998a, 49). Esimerkkinä vastinkorpuksesta Talvensaari (2008a, 29) mainitsee suomen- ja ruotsinkieliset sanomalehtiartikkelikokoel-

mat samalta ajanjaksolta. Tässä esimerkkitapauksessa vastinkorpus luodaan etsimällä suomenkielisen kokoelman artikkeleille samoja aiheita tai tapahtumia käsittelevät artikkelit ruotsinkielisestä kokoelmasta. Jokaiselle lähdekieliselle dokumentille ei välttämättä löydy paria kohdekielisistä dokumenteista, koska kaikkia samoja suomenkielisten sanomalehtien tapahtumia ja aiheita ei välttämättä esiinny ruotsinkielisissä lehdissä. (Talvensaari 2008a, 29.) Vastinkorpuksen dokumenttien lukumäärä voikin olla pienempi kuin lähdekielisten (Talvensaari 2008a, 29) tai kohdekielisten dokumenttien lukumäärä.

Eri tekijät vaikuttavat rinnakkais- tai vastinkorpusten laatuun kieltenvälisessä tiedonhaussa. Ensinnäkin korpusten aihealueiden pitää sopia kyselyiden aihealueisiin. (Talvensaari 2008b, 114.) Toiseksi korpusten on oltava riittävän laajat, ja niiden on sisällettävä myös harvinaisia sanoja (Talvensaari 2008b, 124). Kolmantena vaikuttavana tekijänä on rinnastuksen laatu (Talvensaari 2008b, 114). Talvensaaren (2008b, 114) tutkimuksen perusteella tärkein näistä korpusten laatuun vaikuttavista tekijöistä on korpusten ja kyselyiden aihealueiden läheisyys.

Rinnakkaiskorpuksia pidetään vastinkorpuksia parempina, koska rinnakkaiskorpuksiset sisältävät vastinkorpuksia täsmällisempää kääntämistietämystä (Talvensaari 2008a, 29). Talvensaari, Pirkola, Järvelin, Juhola ja Laurikkala (2008, 19) kuitenkin osoittivat, että rinnakkaiskorpus yleisellä sanastolla tarjoaa heikompa kääntämistietämystä kuin vastinkorpus tietyn aihealueen erityissanastolla. Rinnakkaiskorpusten puuttuessa kieltenvälisessä tiedonhaussa käytetään usein vastinkorpuksia (Talvensaari 2008a, 29). Koska käsin tehdyt korpuksiset ovat tavallisesti kooltaan pieniä, kalliita luoda eivätkä riittävän kattavia, viime aikoina on kehitetty automaattisia menetelmiä korpusten luomiseksi (Talvensaari 2008b, 114).

## 3 KONEKÄÄNTÄMINEN

Konekääntäminen on sanakirjaperustaisen ja korpusperustaisen menetelmän ohella kolmas kääntämismenetelmä kieltenvälisessä tiedonhaussa (Zhu & Wang 2006, 593). Sanakirja- ja korpusperustainen menetelmä esiteltiin luvussa 2, joten tässä luvussa keskitytään ainoastaan konekääntämiseen. Ensin määritellään konekääntämiseen liittyviä peruskäsitteitä sekä tarkastellaan konekääntämistä yleisellä tasolla ja suhteessa esimerkiksi sanakirjaperustaiseen menetelmään. Yleisen esittelyn jälkeen tarkastellaan käännettävän tekstin pituuden yhteyttä konekäännöksen laatuun. Lopuksi keskitytään konekääntämisen kääntämismenetelmiin.

### 3.1 Yleistä

Konekääntäminen on automaattista kääntämistä kieleltä toiselle tietokoneita käyttäen (Lopez 2008, 1). Joissain kielissä konekääntämisestä käytetään käsitettä automaattinen kääntäminen (Hutchins 2005a, 1). Konekäännösjärjestelmät ovat oiva työkalu kieltenvälisessä tiedonhaussa, koska hyvä konekäännösjärjestelmä helpottaa tiedonhakua (Kishida 2005, 437). Konekäännösjärjestelmien käyttäminen on yleistynyt etenkin kyselyiden kääntämisessä, koska konekäännösjärjestelmiä on viime vuosina ollut aiempaa helpommin tarjolla monilla kielillä ja niiden käännösten laatu on parantunut (Kettunen 2009, 103). McNameen ja Mayfieldin (2002, 159) mukaan konekääntäminen on helpoin lähestymistapa kyselyn kääntämiseen, mutta dokumenttien kääntäminen saattaa konekääntämiselläkin olla liian työlästä. Fujii ja Ishikawa (2000, 13) esittävän eriävän näkemyksen mainitessaan, että konekääntäminen on tehokas tapa dokumenttien kääntämiseen. Airion (2009, 29) mukaan konekääntäminen on ainoa menetelmä, jota voidaan soveltaa dokumenttien kääntämisessä.

Konekäännösjärjestelmä voi olla kaksi- tai monikielinen (Hutchins & Somers 1992, 69). Kaksikielisessä konekäännösjärjestelmässä voidaan kääntää vain kahden kielen välillä, kun taas monikielinen järjestelmä kääntää useamman kuin kahden kielen välillä (Hutchins & Somers 1992, 69-70). Konekäännösjärjestelmä voi olla yksi- tai kaksisuuntainen (Hutchins & Somers 1992, 70). Yksisuuntainen järjestelmä kääntää kieleltä toiselle vain yhteen suuntaan (Hutchins & Somers 1992, 70), esimerkiksi vain suomesta englanniksi mutta ei englannista suomeksi. Kaksisuuntainen järjestelmä puolestaan kääntää

kummaltakin kieliparin kieleltä toiselle (Hutchins & Somers 1992, 70). Järjestelmästä, käyttötarkoituksesta ja tavoitteista riippuu, kuinka monen kielen välillä järjestelmä kääntää ja kääntääkö se yksi- vai kaksisuuntaisesti (Hutchins & Somers 1992, 70).

Konekäännösjärjestelmä valitsee tavallisesti vain yhden käännöksen monista lähdekielisellemme sanalle olevista kohdekielisistä vaihtoehdoista (Kishida 2005, 437; McNamee & Mayfield 2002, 159; Wu ym. 2008, 74). Kaikki käännösvaihtoehdot kyselyyn sisällyttävään sanakirjaperustaiseen menetelmään verrattuna konekääntämisessä onkin vaarana, että oikea käännös kadotetaan (Airio 2009, 30). Näin ollen alkuperäisen kyselyntietoa katoaa (McNamee & Mayfield 2002, 159). Lisäksi kun valinta rajoitetaan vain yhteen käännökseen, estetään tiedonhakujärjestelmää laajentamasta alkuperäistä kyselyä synonyymeilla (Nie, Simard, Isabelle & Durand 1999, 74). Toisaalta Oardin (1998, 481-482) tutkimuksen mukaan satunnaisesti kaksikielisestä sanakirjasta valittu yksittäinen käännös on tiedonhaun tuloksellisuuden perusteella yhtä laadukas kuin kaikkien mahdollisten käännösten sisällyttäminen kyselyyn.

Rosemblatin ja muiden (2003, 566) mukaan sanakirjaperustaiseen kääntämiseen verrattuna konekääntäminen onnistuu käsittelemään sekä sanakirjan ulkopuolisista sanoista että sanojen monitulkintaisuudesta johtuvia ongelmia. Esimerkiksi Huangin ja muiden (2007, 81) sekä Oardin (1998, 474) tutkimukset osoittivat, että käännettyillä kyselyillä tehtyjen tiedonhakujen tuloksellisuuden perusteella konekääntäminen suoriutuu kyselyiden kääntämisestä sanakirjaperustaista menetelmää paremmin. Myös Wu ja muut (2008, 76) päätyivät samaan tulokseen. Wu ja muiden (2008, 72) tutkimuksessa kyselyiden kääntämisessä vertailtavina käännösresursseina olivat sanakirjaperustainen menetelmä sekä Google Translate -konekäännösjärjestelmä. Toisaalta taas Oard (1998, 481) on eripituisilla kyselyversioilla osoittanut sanakirjaperustaisen menetelmän konekääntämistä paremmaksi. Tulosten ristiriitaisuus ja sanakirjaperustaisen menetelmän parimmuus saattavat johtua siitä, että kyselyt ovat usein lyhyitä sanojen joukkoja eivätkä ne tarjoa konekääntämistä varten riittävän laajaa asiayhteyttä (Kishida 2005, 437). Luvussa 3.2 käsitellään käännettävän tekstin pituuden yhteyttä konekäännöksen laatuun.

Konekäännösjärjestelmien tutkimisen ongelmana on se, ettei tutkija voi kontrolloida järjestelmiä tai niiden asetuksia (Rosemblat ym. 2003, 566). Tästä on poikkeuksena tutkijan itse kehittämä konekäännösjärjestelmä. Wu ja muut (2008, 72) kuitenkin toteavat,



että tilanne vastaa käyttäjän toimintaa, koska käyttäjäkään ei muuta järjestelmien asetuksia.

Verkossa toimivia konekäännösjärjestelmiä ovat esimerkiksi Babelfish, Systran, Google Translate, SDL Free Translation, WorldLingo, ProMt (Kit & Wong 2008, 311), Bing<sup>TM</sup> Translator (Toral, Gaspari, Naskar & Way 2011, 15), TeemaPoint, Sunda ja Kielikoneen MOT Translation. Kit ja Wong (2008, 299) vertailivat kuuden verkossa toimivan ilmaisen konekäännösjärjestelmän käännöslaatua lakitekstien kääntämisessä eri kieliltä englanniksi. Testattavat konekäännösjärjestelmät olivat Babelfish, Google Translate, ProMt, SDL Free Translation, Systran ja WorldLingo (Kit & Wong 2008, 311). Käännettävät tekstit olivat muun muassa EU:n lakitekstejä ja YK:n dokumentteja (Kit & Wong 2008, 310-311). Lähdekieliset tekstit käännettiin konekäännösjärjestelmillä virke kerrallaan (Kit & Wong 2008, 313). Käännösten laatua evaluoitiin kahdella automaattisella konekäännösten evaluointimittarilla (Kit & Wong 2008, 299).

Kitin ja Wongin (2008, 319) tutkimuksessa ei havaittu minkään tietyn konekäännösjärjestelmän kääntävän toisia laadukkaammin kaikilla kielipareilla, vaan eri järjestelmien käännösten laatu vaihteli eri kielipareilla. Babelfish, Google Translate ja Systran olivat kuitenkin käännöslaadultaan muita testattuja järjestelmiä parempia (Kit & Wong 2008, 319). Kitin ja Wongin (2008, 314-315) tutkimuksen mukaan useissa eri konekäännösjärjestelmissä esiintyvät suosituimmat kieliyhdistelmät tuottavat keskimäärin laadukkaampia käännöksiä kuin kieliyhdistelmät, joita ei ole käytettävissä monessa eri järjestelmässä. Kit ja Wong (2008, 319) mainitsevat, että heidän tutkimustuloksiaan vertailukelpoisemmat tulokset saisi useammalla tekstityypillä ja useammilla kielipareilla kuin käyttämällä pelkästään lakitekstejä ja rajoittamalla kielet muutamaan suosituimpaan.

### **3.2 Käännettävän tekstin pituuden yhteys konekäännöksen laatuun**

Konekäännösjärjestelmät pyrkivät hyödyntämään sanan asiayhteyteen liittyviä vihjeitä (Oard & Hackett 1998, 1), ja ne on tavallisesti optimoitu kääntämään virkkeitä eikä niinkään yksittäisiä sanoja (Zhu & Wang 2006, 599). Pyrkimyksenä on sana sanalta käännösten sijaan tuottaa kieliopillisesti oikeita virkkeitä (Airio 2009, 29; Zhu & Wang 2006, 594). Näin ollen konekääntämiseen perustuva kyselyn kääntäminen ei voi toimia parhaalla mahdollisella tavalla, jos kysely on lyhyt, epätäydellinen virke (Chen & Gey 2004, 150; Lehtokangas ym. 2004, 3).

Wun ja muiden (2008, 72) tutkimuksen yhtenä tutkimuskysymyksenä oli, vaikuttaako kyselyn pituus käännöksen laatuun. Tiedonhaun tuloksellisuuden perusteella konekääntäminen suoriutuu jopa lyhyiden kyselyiden kääntämisestä, mutta se suoriutuu paremmin pitkien kuin lyhyiden kyselyiden kääntämisestä (Wu ym. 2008, 75). Myös Oardin ja Hackettin (1998, 1) mukaan kyselyiden kääntäminen onnistuu konekäännösjärjestelmiltä, mutta riittävän asiayhteyden puuttuminen vain muutaman sanan kyselyssä saattaa rajoittaa järjestelmien kykyä valita sopivin käännös. Esimerkiksi Oardin (1998, 481) tutkimuksessa pitkien kyselyiden kääntämisessä konekääntäminen oli tiedonhaun tuloksellisuudella mitattuna sanakirjaperustaista menetelmää parempi. Sen sijaan lyhyiden, enintään kolmen sanan pituisten kyselyiden kääntämisessä sanakirjaperustainen menetelmä oli tiedonhaun tuloksellisuudella mitattuna parempi kääntämismenetelmä kuin konekääntäminen (Oard 1998, 481). Oardin (1998, 481) mukaan tulos ei ollut yllättävä, koska tutkimuksessa käytetty konekäännösjärjestelmä on suunniteltu saavuttamaan paras tuloksellisuus oikein muodostettujen virkkeiden kääntämisessä. On kuitenkin otettava huomioon, että pitkä kysely ei välttämättä tarkoita samaa kuin oikein muodostettu virke; kysely voi sisältää monta sanaa, mutta se ei välttämättä muodosta virkettä.

Kettusen (2009, 112) tutkimuksen perusteella yleinen uskomus siitä, että lyhyet kyselyt olisivat konekäännösjärjestelmille pitkiä kyselyitä selvästi ongelmallisempia kääntää, ei pidä paikkansa. Kyseisessä tutkimuksessa konekäännösjärjestelmien tuottamien käännösten laadussa ei ollut suuriakaan eroja käännettäessä kyselyitä erilaajuisina. Vertailtavina käännöksinä käytettiin tietyn testikokoelman hakuaiheiden otsikoita sekä otsikoiden ja kuvailutekstien alkujen yhdistelmiä. Erilaajuisien kyselyiden käännösten vertailu toteutettiin ohjelmalla, joka vertailee käännöksiä merkkitasolla. Käännöslaatuun perustuvien tulosten mukaan kahdeksan järjestelmää kahdestatoista testatusta järjestelmästä käänsi hakuaiheiden otsikot samalla tavalla kuin otsikoiden ja kuvailutekstien alkujen yhdistelmien alut erilaajuisista asiayhteyksistä huolimatta. (Kettunen 2009, 112.) Mainittakoon, että hakuaiheet ovat yksi tiedonhaketutkimuksissa käytettävien testikokoelmien osa. Hakuaiheet sisältävät tavallisesti otsikon sekä aiheen lyhyen ja pitkän kuvailun. Otsikkoa ja lyhyttä kuvailua voidaan käyttää kyselynä, kun taas pitkä kuvailu määrittelee, millaiset dokumentit ovat relevantteja kyseiselle hakuaiheelle. (Croft, Metzler & Strohmman 2010, 305-306.)

Käännettävän tekstin pituuteen liittyy myös Wun ja muiden (2008, 75) sanakirjaperustaisen menetelmän ja konekääntämisen vertailun yhteydessä tekemä havainto. Tämän

havainnon mukaan konekäännösjärjestelmän ylivoima sanakirjaperustaiseen kääntämiseen verrattuna pienenee, kun kyselyt pitenevät. Tulos on poikkeava, koska yleisesti ajatellaan, että konekäännösjärjestelmän tuloksellisuus on parempi pitkien kuin lyhyiden kyselyiden kääntämisessä. (Wu ym. 2008, 75.)

Kyselynlaajennus voidaan nähdä asiayhteyden lisäämisenä eli kyselyn pituuden kasvattamisena. Wun ja muiden (2008, 74) tutkimuksessa kyselyitä toteutettiin erilaisilla kyselynlaajennusversioilla tai ilman laajennusta. Kyselynlaajennuksen vaihtoehtoina olivat kyselyn laajentaminen ennen kyselyn kääntämistä, kääntämisen jälkeen tai sekä ennen kyselyn kääntämistä että kääntämisen jälkeen (Wu ym. 2008, 74). Kyselynlaajennuksesta oli eniten hyötyä lyhyille kyselyille; pitkissä kyselyissä haun tuloksellisuus saattoi jopa huonontua käytetystä kyselynlaajennustavasta riippuen. Näin ollen kyselynlaajennuksen tarjoama hyöty pienenee, kun kysely pitenee. (Wu ym. 2008, 75.)

### **3.3 Konekääntämisen kääntämismenetelmät**

Konekääntämisen menetelmät voidaan jakaa sääntö- ja korpusperustaiseen menetelmään, joista korpusperustainen menetelmä voidaan jakaa edelleen tilastolliseen ja esimerkkiperustaiseen menetelmään (Hutchins 2005b, 63). Bennett ja Gerber (2003, 178) käyttävät tilastollisesta ja esimerkkiperustaisesta menetelmästä käsitettä dataohjatut menetelmät, koska niissä käytetään koneoppimisen menetelmiä käännösmallien automaattisessa oppimisessa. Konekääntämisen menetelmät voidaan kuitenkin jakaa myös suoraan kolmeen vaihtoehtoon, jolloin konekääntäminen voi olla sääntöperustaista (rule-based machine translation, RBMT), tilastollista (statistical machine translation, SMT) tai esimerkkiperustaista (example-based machine translation, EBMT) (Hutchins 2005b, 69).

1980-luvun loppuun saakka konekäännösjärjestelmät olivat sääntöperustaisia, mutta 1980-luvun lopussa menetelmän valta-asema heikkeni uusien korpusperustaisten menetelmien kehittymisen myötä (Hutchins 2005a, 11). Kahdessa vuosikymmenessä tilastollinen konekääntäminen on tullut hallitsevaksi (Hutchins 2005a, 12; Lopez 2008, 2). Ei kuitenkaan ole olemassa yhtä menetelmää, jolla saavutettaisiin laadukas konekäännös, vaan kunkin menetelmän parhaita ominaisuuksia kannattaa yhdistää (Hutchins 2005a, 14). Esimerkiksi esimerkkiperustainen menetelmä käyttää tekniikoita sekä tilastollisesta että sääntöperustaisesta menetelmästä (Hutchins 2005b, 69). Tilastollinen menetelmä

puolestaan hyödyntää esimerkkiperustaisen menetelmän tapaan aiempaa enemmän fraasiperustaista rinnastamista (Hutchins 2005a, 12).

Seuraavaksi tarkastellaan näitä kolmea konekääntämisen kääntämismenetelmää. Ensin esitellään sääntöperustainen menetelmä, minkä jälkeen keskitytään tilastolliseen konekääntämiseen. Lopuksi käsitellään esimerkkiperustaista konekääntämistä. Tilastollisen ja esimerkkiperustaisen konekääntämisen yhteydessä menetelmiä myös vertaillaan, mikä selkeyttää etenkin esimerkkiperustaisen konekääntämisen määrittelemistä.

### **3.3.1 Sääntöperustainen konekääntäminen**

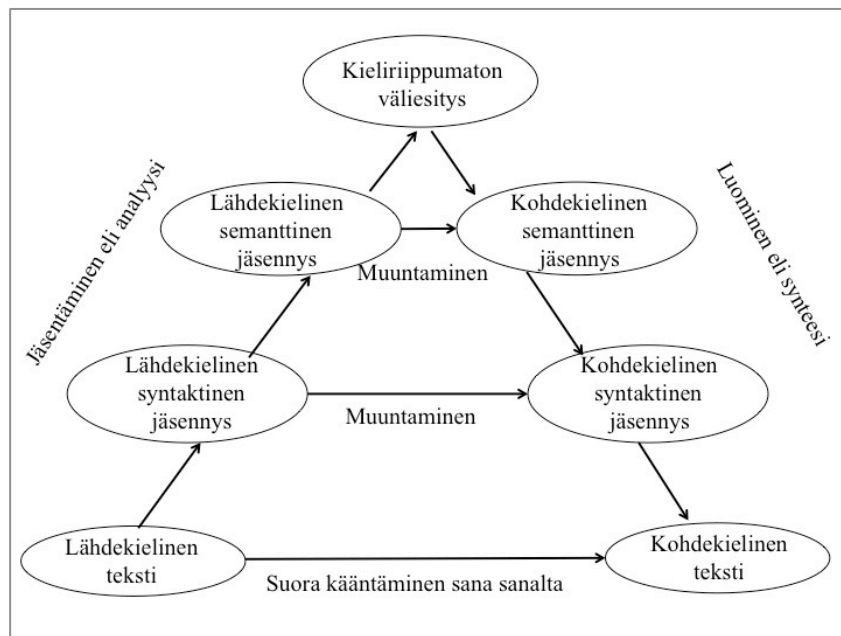
Sääntöperustaisessa konekääntämisessä käytetään erilaisia kielenanalysointisääntöjä (Hutchins 2005a, 11; Zhu & Wang 2006, 594), esimerkiksi morfologiselle tai syntaktiselle analyysille (Hutchins 2005a, 11). Sääntöperustaisen konekääntämisen ytimen muodostavatkin kaksikielinen sanasto ja erinäiset säännöt (Hutchins 2005b, 66). Näin ollen sääntöperustainen konekääntäminen hyödyntää kieltenvälisen tiedonhaun sanakirjaperustaista kääntämismenetelmää. Sääntöperustaisen konekääntämisen kahdesta resurssista Zhun ja Wangin (2006, 599-600) tutkimuksen perusteella järjestelmien kehittämisessä olisi oleellisempaa keskittyä sanastojen laajentamiseen kuin sääntöjen lisäämiseen, koska sanakirjan koolla on suurempi vaikutus tiedonhaun tuloksellisuuteen kuin sääntöjen määrällä. Zhu ja Wang (2006, 593) testasivat tutkimuksessaan erikseen sääntöjen ja sanaston vähentämisen vaikutusta sekä käännöksen laatuun että käännettyillä kyselyillä tehtyjen tiedonhakujen tuloksellisuuteen.

Sääntöjen luomisessa käytetään apuna kielitieteilijöiden tietämystä ja säännöt kirjoitetaan käsin (Bennett & Gerber 2003, 176). Sääntöjen luomisessa tutustutaan teksteihin ja yritetään kehittää niissä havaituista yksittäisistä ilmiöistä yleisiä sääntöjä. Sääntöjä on luotava mahdollisimman monipuolisista ja aidoista teksteistä, koska konekäännösjärjestelmien syötteinä saamat käännettävät tekstitkin ovat aitoja, käyttäjän luonnollisella kielellä ilmaistuja eivätkä yksinkertaistettuja kielioppiesimerkkejä. (Bennett & Gerber 2003, 178.)

Sääntöperustainen konekääntäminen voi kääntämisstrategialtaan olla suoraa tai välivaiheita käyttävää. Välivaiheita käyttävä strategia voidaan jakaa edelleen väliesitys- ja muuntamisstrategiaan. (Hutchins & Somers 1992, 71-72.) Tässä yhteydessä kääntämis-

strategia ei tarkoita täysin samaa kuin luvussa 2.1 esiteltyt kieltenvälisessä tiedonhaussa käytettävät käänntämisstrategiat, jotka määrittelevät sen, mitä käännetään.

Sääntöperustaisen konekääntämisen käänntämisstrategiat eroavat toisistaan lähdekielisen tekstin jäsentämisen laajuudelta (Dorr, Hovy & Levin 2006, 383). Kuvassa 2 on kuvattu nämä kolme sääntöperustaisen konekääntämisen käänntämisstrategiaa pyramidina, joka esimerkiksi Arnoldin (2003, 122) mukaan on alun perin Vauquoisin (1976) tapa esittää käänntämisstrategiat. Pyramidin vasenta puolta ylöspäin edettäessä ilmenevät eri strategioiden jäsentämisvaiheet ja oikeaa puolta alaspäin tultaessa luomisvaiheet. Pyramidin poikki kulkevat puolestaan muuntamisvaiheet. Pyramidissa ylöspäin eteneminen lisää tekstin jäsentämisen työmäärää (Dorr ym. 2006, 383). Samalla kuitenkin vähenee työmäärää, joka vaaditaan käsittelemään kieltenväliset erot muuntamisvaiheessa (Dorr ym. 2006, 383).



Kuva 2. Kolmen käänntämisstrategian pyramidi (ks. Arnold 2003, 123; Dorr ym. 2006, 383; Jurafsky & Martin 2000, 812; Knight 1997, 82; Manning & Schütze 2000, 464)

Alimpana kuvan 2 pyramidissa on suora käänntämisstrategia, jossa lähdekielinen teksti käännetään sana sanalta kohdekieliseksi tekstiksi (Dorr ym. 2006, 383). Lähdekielistä tekstiä ei siis juurikaan jäsennetä, eikä käänntämisessä ole välivaiheita, vaan lähdekielistä tekstistä päädytään melko suoraan kohdekieliseen käänntökseen. (Arnold 2003, 122.) Ennen käänntämistä saatetaan ainoastaan muuttaa taivutusmuotoiset sanat perusmuotoon (Hutchins & Somers 1992, 72). Näin ollen suorassa käänntämisessä on pelkkä muuntamisvaihe (Arnold 2003, 123). Suoraan käänntämisstrategiaan perustuva käänntä-

minen on tarkoitettu yleensä yhdelle kieliparille (Jurafsky & Martin 2000, 816), eli se ei sovellu monikielisiin konekäännösjärjestelmiin.

Suoraa kääntämisstrategiaa käyttävien sääntöperustaisten konekäännösjärjestelmien käännöslaatu on melko heikkoa. Järjestelmät tekevät virheitä sekä sanastollisella että syntaktisella tasolla. (Hutchins & Somers 1992, 72.) Virheitä ilmenee etenkin monitul-kintaisten sanojen kääntämisessä, koska sana sanalta kääntämisessä on vaikea tietää monitul-kintaisen sanan oikeaa merkitystä (Knight 1997, 82; Manning & Schütze 2000, 463). Koska lähdekielistä tekstiä ei jäsennetä, kohdekielisen käännöksen syntaksiraken- teet kuvastavat usein lähdekielen rakennetta (Hutchins & Somers 1992, 72-73).

Suoran kääntämisen jälkeen kuvan 2 pyramidia ylöspäin edettäessä ilmenee syntakti- seen jäsentämiseen perustuva muuntamisstrategia ja tämän jälkeen myös semanttiseen jäsentämiseen perustuva muuntamisstrategia (Dorr ym. 2006, 383). Muuntamisstrategi- assa kääntäminen koostuu kolmesta vaiheesta: jäsentämisestä eli analyysistä, muunta- misesta sekä luomisesta eli synteesisestä (Arnold 2003, 122; Hutchins 2005a, 3; Jurafsky & Martin 2000, 807). Jäsentämisvaiheessa lähdekielinen teksti jäsennetään ja siitä tuo- tetaan lähdekielen ominaisuuksia sisältävä esitys (Arnold 2003, 122; Manning & Schüt- ze 2000, 464), esimerkiksi jäsennyyspuu, josta ilmenee adjektiivin ja substantiivin keski- näinen järjestys (Jurafsky & Martin 2000, 808). Toisessa vaiheessa eli muuntamisvai- heessa lähdekielinen esitys muunnetaan eli korvataan vastaavalla kohdekielisel- lällä (Arnold 2003, 122). Vasta tämän muuntamisvaiheen jälkeen lähdekieliset sanat korvataan kohdekielisel- lillä sanoilla (Jurafsky & Martin 2000, 809). Viimeisessä vaihees- sa, yhdistämis- eli luomisvaiheessa, tekstiä voidaan vielä käsitellä esimerkiksi taivutta- malla sanoja tai lisäämällä tietyille kielelle tyypillisiä tiettyyn sanaluokkaan kuuluvia apusanoja (Jurafsky & Martin 2000, 809). Jurafskyn ja Martinin (2000, 821) mukaan muuntamisstrategia vaatii vain syntaktista jäsentämistä. Manning ja Schütze (2000, 464- 465) sekä Dorr ja muut (2006, 383) puolestaan mainitsevat myös semanttisen jäsentä- misen, jota voidaan tehdä syntaktisen jäsentämisen jälkeen. Semanttista jäsentämistä voidaan hyödyntää esimerkiksi sanan merkityksen yksikäsitteistämässä (Manning & Schütze 2000, 464-465).

Kuvassa 2 aivan pyramidin huipulla on kaikista eniten jäsentämistä vaativa kieliriippu- matonta väliesitystä käyttävä strategia (Dorr ym. 2006, 383). Väliesitysstrategiassa käännetään lähdekieleltä kohdekielelle kieliriippumattoman tiedonesitysmuodon kautta

(Manning & Schütze 2000, 464). Kieliriippumattoman esityksen käyttämistä kääntämissä käsiteltiin jo luvussa 2.1, jossa tarkasteltiin yleisesti kääntämisen strategioita kielenvälisessä tiedonhaussa. Väliesitys on siis tiedonesitysmuoto, joka on riippumaton vasta, jolla tietyt kielet ilmaisevat merkityksiä (Manning & Schütze 2000, 465). Se edustaa samanaikaisesti sekä lähde- että kohdekielistä tekstiä (Dorr ym. 2006, 383). Karkeasti sanottuna väliesitysstrategiassa kääntäminen on kaksivaiheinen: ensin käännetään lähdekieleltä väliesitykseksi ja tämän jälkeen väliesityksestä kohdekielelle (Hutchins 2005a, 3).

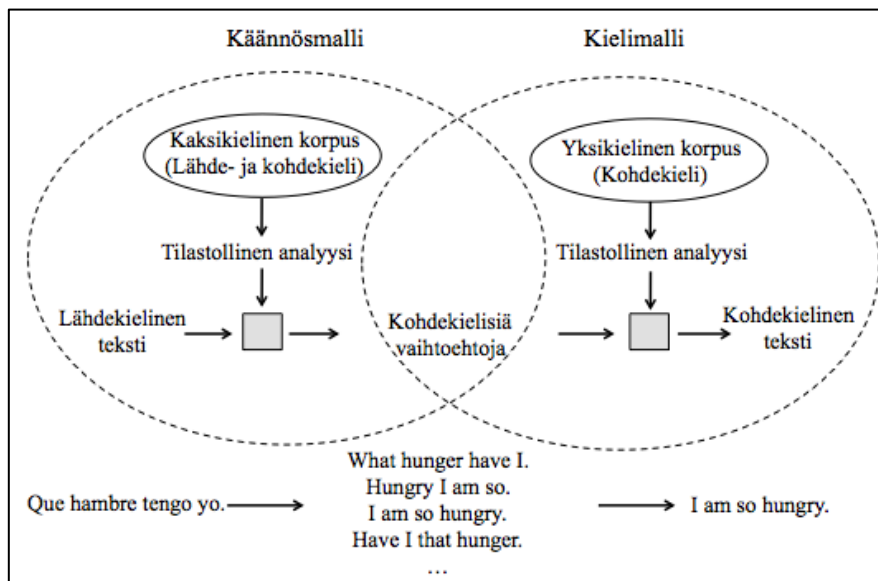
Muuntamisstrategiassa jäsenetyt esitykset riippuvat sekä lähde- että kohdekielestä (Hutchins & Somers 1992, 75). Muuntamisstrategian yksi ongelma onkin se, että strategia vaatii erillisiä kääntämissääntöjä jokaiselle kieliparille, mikä ei ole optimaalista monikielisellä järjestelmällä käännettäessä (Jurafsky & Martin 2000, 811-812). Väliesitystä käyttävän strategian etuna onkin se, että strategia ratkaisee kääntämisongelman suurelle kielimäärälle (Manning & Schütze 2000, 465), jos vain kielistä voidaan luoda kieliriippumattomat esitykset. Kieliriippumattomien esitysten luomisen vaikeus on kuitenkin yksi syy, miksi muuntamisstrategiaa pidetään väliesitysstrategiaa parempana (Hutchins & Somers 1992, 76).

### **3.3.2 Tilastollinen konekääntäminen**

Tilastollinen konekäännösjärjestelmä oppii olemassa olevista käännöksistä sanojen tai fraasien väliset vastaavuudet koneoppimisen menetelmillä kieliopista ja kielellisestä jäsentämisestä välittämättä (Bennett & Gerber 2003, 179), joten menetelmä poikkeaa kielellisiä sääntöjä käyttävästä sääntöperustaisesta menetelmästä. Tilastollinen menetelmä eroaa sääntöperustaisesta menetelmästä myös siinä, että sääntöperustaisen menetelmän sanakirja on tilastollisessa menetelmässä korvattu kaksikielisillä rinnastetuilla korpukilla. Rinnastettujen korpusten avulla rinnastetaan lähdekieliset sanat ja virkkeet kohdekielisiin sanoihin ja virkkeisiin. (Hutchins 2005b, 66.) Tilastollinen konekäännösjärjestelmä tutkii suurta määrää rinnakkaiskorpuksissa olevia ihmisten tekemiä käännöksiä ja oppii automaattisesti, miten pitää kääntää (Lopez 2008, 1-2). Järjestelmä tallentaa tietoja käännösten vastaavuuksista ja niiden esiintymisten todennäköisyyksistä (Bennett & Gerber 2003, 179). Oppimisen tuloksena tilastollinen konekäännösjärjestelmä pystyy kääntämään tekstejä, joita se ei ole aiemmin nähnyt (Lopez 2008, 2). Koneoppimisen hyödyntäminen erottaa tilastollisen ja sääntöperustaisen kääntämismenetelmän toisis-

taan; sääntöperustainen järjestelmä ei yleensä opi, eli se ei tallenna tietoa käänöksistä myöhempää käyttöä varten (Carl, Pease, Iomdin & Streiter 2000, 225).

Tilastollisessa konekäännösjärjestelmässä tarvitaan jokin mekanismi valitsemaan käännettävälle sanalle käännös monien mahdollisten vaihtoehtojen joukosta. Tarvitaan myös mekanismi, joka muuttaa kohdekielisen käännöksen sanajärjestyksen oikeaksi, koska toisiaan vastaavat erikieliset sanat eivät välttämättä esiinny samassa järjestyksessä lähde- ja kohdekielisessä tekstissä. (Lopez 2008, 4.) Näihin ongelmiin ovat ratkaisuna käännös- ja kielimalli, joilla tilastollisessa menetelmässä sääntöperustaisen menetelmän säännöt on korvattu (Hutchins 2005b, 66). Käännös malli sisältää tiedon lähde- ja kohdekielisten sanojen välisistä vastaavuuksista ja kielimalli tiedon kohdekielisten sanojen yhteisesiintymisistä kohdekielisissä teksteissä (Hutchins 2005b, 66). Tilastollinen konekääntäminen perustuukin sanojen esiintymistiheyksiin ja esiintymisten todennäköisyyksiin. Kääntämisessä valitaan kaikista todennäköisimmät kohdekieliset vastineet lähdekielisille sanoille ja juuri valituille kohdekielisille sanoille valitaan kaikista todennäköisin kohdekielinen kokonaisuus. (Hutchins 2005b, 63.)



Kuva 3. Tilastollisen kääntämismenetelmän käännös- ja kielimalli (ks. Knight & Koehn 2003, 3)

Kuvasta 3 ilmenee käännös- ja kielimallin muodostama kokonaisuus ja tilastollisen menetelmän kääntämisen periaate. Kuva on hieman muunnettu versio Knightin ja Koehnin (2003, 3) esittämästä kuvasta. Kuvan esimerkissä käännös mallin kaksikielinen korpus sisältää englannin- ja espanjankielisiä tekstejä ja kielimallin yksikielinen korpus englanninkielisiä tekstejä. Lähdekielenä on espanja ja kohdekielenä englanti. (Knight & Koehn 2003, 3.) Käännös malli ottaa syötteenä lähdekielisiä sanoja (Hutchins 2005b,



66). Esimerkissä syötteenä on espanjankielinen virke ”*Que hambre tengo yo.*” (Knight & Koehn 2003, 3). Syötteenä olevaa lähdekielistä tekstiä täsmäytetään käännoismallia vastaan ja tuotetaan kohdekieliset sanat (Hutchins 2005b, 66). Esimerkissä käännoismallin tilastolliset menetelmät tuottavat kaksikielisestä korpuksesta monta englanninkielistä käännoisvaihtoehtoa, esimerkiksi ”*I am so hungry.*” ja ”*Have I that hunger.*” (ks. Knight & Koehn 2003, 3). Käännoismallin tuottamat tekstit toimivat kielimallin syötteenä. Englanninkielisen korpuksen ja tilastollisen analyysin perusteella lopulliseksi käännoiseksi valitaan ”*I am so hungry.*” (Knight & Koehn 2003, 3), joka on todennäköisin käännoismallin tarjoamista englanninkielisistä vaihtoehtoista. Kielimalli siis ikään kuin tarkastaa käännoismallin ehdottamien kohdekielisten virkkeiden oikeellisuuden ja auttaa välttämään syntaksiltaan virheellisiä käännoiksi (Zens, Och & Ney 2002, 19).

Tilastollisessa konekääntämisessä perusyksikkönä on sana, mutta viime aikoina yksittäisiä sanoja pidempiäkin sanajonoja on otettu huomioon (Hutchins 2005b, 63). Yksittäisten sanojen täsmäyttämässä on se heikkous, että kun todennäköisyydet perustuvat ainoastaan yksittäisiin sanoihin, sanojen asiayhteyttä ei huomioida, vaikka monien sanojen käännoiset riippuvat sanoja ympäröivistä sanoista. Yksi tapa ottaa asiayhteys huomioon on opettaa kääntämistä yksittäisten sanojen sijaan fraaseilla. (Zens & Ney 2004.) Tässä yhteydessä fraasi tarkoittaa peräkkäisiä sanoja (Lopez 2008, 8; Zens & Ney 2004), eikä kyseessä siis välttämättä ole kieliopillinen fraasi (Knight & Koehn 2003, 3; Lopez 2008, 8). Perusideana fraasiperustaisessa kääntämisessä on jakaa annettu lähdekielinen virke fraaseiksi, kääntää jokainen fraasi ja lopulta yhdistää käännoiset fraasit tietyssä järjestyksessä kohdekieliseksi virkkeeksi (Lopez 2008, 8). Zensin ja muiden (2002, 30) tutkimuksen mukaan kaksikielisten fraasien käyttäminen käännoismallissa yksittäisten sanojen sijaan parantaa huomattavasti käännoisten laatua. Fraasiperustainen kääntäminen onnistuu yleensä esimerkiksi tuottamaan sanaperustaista kääntämistä paremmin oikean kohdekielisen sanajärjestyksen (Lopez 2008, 9). Fraasiperustaista kääntämistä käytetäänkin laajalti (Lopez 2008, 9).

Toralin ja muiden (2011, 19) tutkimuksessa käännoisten laadussa ilmeni eroja tilastollisten ja sääntöperustaisten konekääntöjärjestelmien välillä. Kyseisessä tutkimuksessa vertailtiin CoSyne-konekääntöjärjestelmän käännoislaatua neljän ilmaisen verkossa toimivan konekääntöjärjestelmän käännoislaatuun (Toral ym. 2011, 17). CoSyne on Amsterdamin yliopistossa kehitetty tilastollinen konekääntöjärjestelmä monikielisten wikisivujen kääntämistä varten (Toral ym. 2011, 13-14). CoSynen lisäksi muut testatta-

vat järjestelmät Toralin ja muiden (2011, 17) tutkimuksessa olivat tilastolliset Google Translate ja Bing<sup>TM</sup> Translator sekä sääntöperustaiset Systran ja SDL Free Translation. Aineistona käytettiin uutisaiheisia tekstejä (Toral ym. 2011, 15). Käännökset tehtiin saksasta, italiasta ja hollannista englanniksi (Toral ym. 2011, 13). Käännöslaadun evaluoinnissa käytettiin useaa automaattista konekäännösten evaluointimittaria (Toral ym. 2011, 14).

Tämän tutkimuksen kannalta ei ole oleellista, miten Toralin ja muiden (2011) tutkimuksessa testatut konekäännösjärjestelmät suoriutuivat suhteessa CoSyne-järjestelmään, vaan miten ne suoriutuivat suhteessa toisiinsa. Tässäkin tutkimuksessa kun osa testattavista konekäännösjärjestelmistä on tilastollisia ja osa sääntöperustaisia. Oleellista on myös se yhtäläisyys, että kahden muun konekäännösjärjestelmän lisäksi tässäkin tutkimuksessa testataan Google Translatea, Bing<sup>TM</sup> Translator Betaa sekä SDL Free Translationia.

Toralin ja muiden (2011, 19) tutkimuksessa kaikki kolme tilastollista konekäännösjärjestelmää olivat kaikilla kieliyhdistelmillä ja kaikilla evaluointimittareilla mitattuina parempia kuin kumpikaan sääntöperustaisista järjestelmistä. Näin ollen Toralin ja muiden (2011, 19) tutkimuksen perusteella uutisaiheisella aineistolla tilastolliset konekäännösjärjestelmät ovat käännöslaadultaan parempia kuin sääntöperustaiset. Eri kieliyhdistelmien välillä oli pieniä eroja tilastollisten järjestelmien keskinäisestä paremmuudesta, mutta Google Translate oli keskimäärin paras järjestelmä (Toral ym. 2011, 19).

### **3.3.3 Esimerkkiperustainen konekääntäminen**

Tilastollisen konekääntämisen ohella toinen korpusperustainen konekääntämismenetelmä on esimerkkiperustainen konekääntäminen, josta on käytetty myös käsitettä muisti-perustainen konekääntäminen. Kyseisessä menetelmässä kääntäminen on käännettävää tekstiä vastaavien esimerkkien löytämistä; taustalla on tieto siitä, miten tietty ilmaisu on aiemmin käännetty. (Hutchins 2005a, 12.) Esimerkkiperustaisessa menetelmässä oleellisinta onkin se, että tietämys tulee esimerkeistä (Somers 1999, 148) ja menetelmän ydinprosessina on valita ja poimia lähdekielisiä katkelmia vastaavat kohdekieliset katkelmat (Hutchins 2005b, 66).

Sääntöperustaiseen menetelmään verrattuna esimerkkiperustaisessa menetelmässä sanakirja on korvattu rinnastetuilla kaksikielisillä korpuksilla eli esimerkkien joukolla ja

säännöt on korvattu korpuksen kohdekielisillä esimerkeillä (Hutchins 2005b, 66). Perusyksikkönä kääntämisessä on virke tai sanojen sarja eikä yksittäinen sana (Hutchins 2005b, 63). Koska esimerkkiperustainen menetelmä asettuu sääntöperustaisen ja tilastollisen menetelmän väliin, se hyödyntää sekä tilastollisia ominaisuuksia että kielellisiä keinoja. Kyseinen menetelmä onkin kolmesta konekääntämisen menetelmästä kaikista vaikein määritellä. (Hutchins 2005b, 69.)

Kaksikieliset korpuksot ovat osa esimerkkiperustaisen menetelmän määritelmää, mutta ne eivät ole riittävä määritelmä, koska monet konekäännösjärjestelmät hyödyntävät jollain tavalla korpuksia (Somers 1999, 148). Sääntöperustaisissa järjestelmissä esimerkivirkkeitä voidaan käyttää sääntöjen tietolähteenä ja toisaalta tilastollisten järjestelmien tiedot johdetaan esimerkkitekstien korpuksesta. Pelkkiä kaksikielisiä korpuksia rajoittavampi määritelmä esimerkkiperustaiselle menetelmälle onkin se, että esimerkkejä käytetään kääntämisen aikana. (Hutchins 2005b, 65.) Somersin (1999, 148-149) mukaan tämä määritelmä erottaa tilastollisen menetelmän esimerkkiperustaisesta, koska tilastollisessa menetelmässä käytettävä data johdetaan etukäteen ennen kääntämisprosessia. Menetelmien erona on myös se, että kun esimerkkiperustainen konekääntäminen perustuu kielellisiin esimerkkiteksteihin, niin tilastolliset menetelmät käyttävät tilastotietoja sanojen ja fraasien esiintymisistä, ja sääntöperustaiset menetelmät toimivat sanojen merkitysten perusteella. (Hutchins 2005b, 69.) Lisäksi tilastollisessa menetelmässä syöte hajotetaan yksittäisiksi sanoiksi, kun taas esimerkkiperustaisessa menetelmässä syöte hajotetaan useamman kuin yhden sanan katkelmiksi (Hutchins 2005b, 69). Fraasiperustainen tilastollinen konekääntäminen on kuitenkin hämärtänyt tilastollisen ja esimerkkiperustaisen menetelmän eroja (Hutchins 2005b, 69).

Koska esimerkkiperustainen menetelmä perustuu todellisiin teksteihin, käännosten pitäisi olla luettavampia ja paremmin asiayhteyden huomioivia kuin sääntöperustaisella menetelmällä tehtyjen käännosten. Esimerkkiperustaista menetelmää puoltaa mahdollisuus parantaa käännetyn tekstin laatua. Esimerkkiperustainen menetelmä voi myös käsitellä sääntöperustaista menetelmää paremmin tapauksia, joissa lähde- ja kohdekielen välillä on suuria rakenteellisia eroja. Esimerkkiperustaisia konekäännösjärjestelmiä voidaan parantaa lisäämällä esimerkkejä kaksikieliseen korpukseen, kun taas sääntöperustaisen järjestelmien kehittäminen vaatii sääntöjen ja sanaston lisäämistä ja muuntelua. (Hutchins 2005b, 64.) Esimerkkiperustaisesta kääntämismenetelmästä onkin tulossa yhä suosituampi (Hutchins 2005b, 63).

## 4 SANOJEN MONITULKINTAISUUS

Jos sanalla on useita merkityksiä, sana on monitulkintainen. Sanoista on tullut monitulkintaisia kielen kehittymisen myötä, ja enemmistö kielten sanoista on monitulkintaisia. (Turdakov 2010, 310.) Seuraavaksi ensin esitellään sanojen monitulkintaisuuden tyyppejä. Monitulkintaisuuden tyyppien esittelyn jälkeen tarkastellaan monitulkintaisten sanojen kääntämistä ja kääntämiseen liittyviä ongelmia. Lopuksi esitellään monitulkintaisten sanojen merkityksen yksikäsitteistämistä sekä yksikäsitteistämismenetelmiä.

### 4.1 Monitulkintaisuuden tyypit

Monitulkintaisuus voidaan jakaa morfologiseen, leksikaaliseen, semanttiseen, syntaktiseen ja pragmaattiseen monitulkintaisuuteen (Turdakov 2010, 310). Tässä tutkielmassa monitulkintaisuudella tarkoitetaan pääasiassa morfologista tai leksikaalista monitulkintaisuutta mutta osin myös semanttista monitulkintaisuutta.

Morfologinen monitulkintaisuus tarkoittaa sitä, että sanan merkitys voi viitata eri sanaluokkiin. Esimerkiksi englanninkielinen sana *look* voi olla verbi tai substantiivi. Tämän tyyppinen monitulkintaisuus on tyypillistä etenkin englannin kielelle. (Turdakov 2010, 310.) Hutchins ja Somers (1992, 85) käyttävät käsitettä kategorinen monitulkintaisuus, jos sana edustaa useampaa kuin yhtä kieliopillista tai syntaktista kategoriata. Sanaluokka on esimerkki syntaktisesta kategoriasta (Hutchins & Somers 1992, 85).

Leksikaalisessa monitulkintaisuudessa sanan merkitykset viittaavat tiettyyn sanaluokkaan mutta eroavat toisistaan (Turdakov 2010, 310). Leksikaalinen monitulkintaisuus kattaa homonyymit ja polyseemit (Pirkola ym. 2001, 216), joiden välinen ero on vaikea määrittellä (Hedlund 2003, 19). Homonyymejä ovat sanat, joilla on sama muoto mutta toisiinsa liittymättömät merkitykset (Kilgarrieff & Palmer 2000, 2; Lyons 1982, 146-147). Esimerkiksi *bank* tarkoittaa rahoituslaitosta tai rantaa. Jos sanalla on useampi kuin yksi merkitys ja nämä merkitykset liittyvät toisiinsa, kyseessä on polyseemi. Esimerkiksi *neck* voi olla kehon tai pullon osa. (Lyons 1982, 146-147.) Leksikaalinen monitulkintaisuus voidaan morfologiaan perustuen jakaa perus- ja taivutusmuotoiseen monitulkintaisuuteen. Perusmuotoisessa monitulkintaisuudessa yhdellä sanalla on vähintään kaksi merkitystä. Taivutusmuotoisessa monitulkintaisuudessa sen sijaan vähintään kahdella

sanalla on vähintään yksi yhteinen taivutusmuoto. (Pirkola ym. 2001, 216.) Esimerkki taivutusmuotoisesta monitulkinntaisuudesta on sana *hauissa*, jonka perusmuoto on joko *hauki* tai *haku*. Tässä tutkielmassa tarkastellaan ainoastaan perusmuotoista monitulkinntaisuutta.

Semanttisella monitulkinntaisuudella tarkoitetaan sanan käyttämistä kuvainnollisessa merkityksessä. Esimerkiksi sana *fox* voi ketun lisäksi tarkoittaa viekasta henkilöä. Syntaktista monitulkinntaisuutta puolestaan esiintyy, jos lause voi käsittää useamman kuin yhden syntaktisen rakenteen. Esimerkiksi lauseesta *Flying planes can be dangerous* ei voida tietää, ovatko lentävät koneet vaarallisia vai onko vaarallista lentää lentokoneella. Pragmaattista monitulkinntaisuutta esiintyy puolestaan pronomien yhteydessä. Esimerkiksi virkkeestä ”*She has dropped a pencil on a table and has broken it.*” ei voida yksiselitteisesti tietää, menikö rikki kynä vai pöytä. (Turdakov 2010, 310.)

## 4.2 Monitulkinntaisten sanojen kääntäminen

Käännösten monitulkinntaisuus on yleistä kieltenvälisessä tiedonhaussa (Hedlund 2003, 19). Käännösten epäselvyysoongelmat johtuvat siitä, että monilla sanoilla ei ole vain yhtä yksikäsitteistä käännöstä ja vaihtoehtoisten käännösten merkitykset saattavat poiketa toisistaan huomattavasti (Zhang & Vines 2005, 1). Yksi suurimmista haasteista kieltenvälisen tiedonhaun tuloksellisuuden parantamisessa onkin monitulkinntaisuuden vähentäminen kääntämisessä (Ballesteros & Croft 1998, 69). Kääntämisen monitulkinntaisuuteen liittyvä ongelma on otettava huomioon riippumatta siitä, mitä kääntämismenetelmää käytetään (Ballesteros & Croft 1998, 64).

Asiayhteydestä riippuen tietynkielinen sana voidaan kääntää toiselle kielelle monella eri tavalla (Turdakov 2010, 309). Tietyissä asiayhteydessä tiettyä sanaluokkaa edustava sana tiettyjen sanojen ympäröimänä käännetään eri tavalla kuin tämän esiintyessä jossain toisessa asiayhteydessä (Brunning, de Gispert & Byrne 2009, 110). Oleellista tietoa sanan asiayhteydestä ovatkin esimerkiksi kyseisen sanan sanaluokka, edeltävät ja seuraavat sanat, edeltävien ja seuraavien sanojen sanaluokka sekä jopa laajemman tekstialueen asiayhteys (Brunning ym. 2009, 111). Monet tilastolliset konekäännösjärjestelmät hyödyntävät vain vähän tietoa asiayhteydestä käännösvaihtoehdon valitsemisessa (Carpuat & Wu 2007b, 73), vaikka tutkimukset ovat osoittaneet asiayhteystiedon tärkeyden monitulkinntaisten virkkeiden kääntämisessä (Aurélien, Rafik & Langlais 2008, 114).

Aurélienin ja muiden (2008, 119) tutkimuksen perusteella asiayhteyttä hyödyntävä järjestelmä pystyy tuottamaan parempia käännöksiä kuin asiayhteyttä hyödyntämätön järjestelmä myös käännettäessä heikosti taipuvalta lähdekieleltä vahvasti taipuvalle kohdekielelle. Tämän tutkielman tutkimuksessa käännetään heikosti taipuvalta englannin kieleltä vahvasti taipuvalle suomen kielelle. Taivutusmuotoja ei kuitenkaan tarkastella.

Kuten luvussa 2.2.1 mainittiin, sanakirjaperustaisen menetelmän ongelmana on se, kun käännökseen sisällytetään käännettävän sanan kaikki mahdolliset käännösvaihtoehdot. Näin ollen varsinaisessa kyselyssä on myös asiaankuulumattomia sanoja (Ballesteros & Croft 1997, 84). Jos käännettyssä kyselyssä on alkuperäiselle kyselylle merkityksettömiä sanoja, haun tuloksellisuus heikkenee (Kishida 2005, 439). Zhangin ja Vinesin (2005, 1) mukaan myös käännettävän tekstin lyhyys aiheuttaa ongelmia monitulkintaisten sanojen kääntämisessä. Monitulkintaisten sanojen kääntäminen on erityisen haastavaa, jos käännettävä teksti on vain kaksi tai kolme sanaa (Zhang & Vines 2005, 1).

Monitulkintaisuus on ongelma etenkin välikieltä hyödyntävässä kääntämisessä. Monitulkintaisuutta kun saattaa esiintyä välikieltä hyödyntävän kääntämisprosessin kaikissa vaiheissa; käännettävät sanat saattavat olla monitulkintaisia niin lähde-, väli- kuin kohdekielellä. Epärelevanttien sanojen määrä kyselyssä todennäköisesti kasvaa joka kerta, kun kääntäminen suoritetaan. (Lehtokangas ym. 2004, 4.) Jos esimerkiksi japaninkielinen kysely koostuu neljästä sanasta ja jokaiselle sanalle on neljä englanninkielistä vastinetta ja edelleen kaikille englanninkielisille sanoille on neljä hollanninkielistä vastinetta, lopullisessa hollanninkielisessä kyselyssä on huomattavasti enemmän sanoja kuin alkuperäisessä japaninkielisessä kyselyssä. Joukossa on myös asiaankuulumattomia sanoja. (Kishida 2005, 445.) Kaikki sanayhdistelmät eivät kuitenkaan ole mielekkäitä.

### **4.3 Sanan merkityksen yksikäsitteistämismenetelmät**

Sanan merkityksen automaattista tunnistamista asiayhteyden perusteella sanotaan sanan merkityksen disambigoinniksi (Word Sense Disambiguation, WSD) eli yksikäsitteistämiseksi (Navigli 2009, 1; Turdakov 2010, 309). Yksikäsitteistäminen auttaa lähdekielisen sanan oikean merkityksen tunnistamisessa monien mahdollisten kohdekielisten vaihtoehtojen joukosta (Carpuat & Wu 2007a, 43; Jurafsky & Martin 2000, 597). Sanan merkityksen yksikäsitteistämisiongelma sai alkunsa 1950-luvulla konekääntämisen alitehtävänä. Siitä lähtien lukuisia menetelmiä on kehitetty yksikäsitteistämisen ratkaise-

miseksi. (Turdakov 2010, 309.) Yksikäsitteistäminen onkin yksi kieltenvälisen tiedonhaun tutkimuksen haastavimmista ongelmista (Kishida 2005, 439), ja sillä on tärkeä rooli konekääntämisessä (Navigli 2009, 2).

Nykyisin yksikäsitteistämisjärjestelmille on tarjolla standardoituja evaluointikehyksiä (Jurafsky & Martin 2000, 639). Jurafskyn ja Martinin (2000, 639) mukaan erityisesti Senseval (Kilgarriff & Palmer 2000) tarjoaa toimivat puitteet evaluoinnille. Senseval oli ensimmäinen avoin kokoontuminen yksikäsitteistämisjärjestelmien evaluomiselle. Ensimmäinen Senseval-kokoontuminen järjestettiin vuonna 1998 (Kilgarriff & Palmer 2000, 1). Senseval terävöitti yksikäsitteistämistutkimusta mahdollistamalla yksikäsitteistämisjärjestelmien suoran vertailun ja pakottamalla tietyntasoiseen yhteisymmärryseen yksikäsitteistämistehtävien määrittelemisessä (Kilgarriff & Palmer 2000, 11).

Yksikäsitteistämisessä käytetään tietoa monitulkintaisen sanan asiayhteydestä. Asiayhteys nähdään kahdella tavalla: asiayhteys on tietynkokoinen sanojen muodostama ikkuna kohdesanan ympärillä ilman sanojen suhteiden huomioimista tai asiayhteydessä huomioidaan kohdesanaan suhteessa olevat sanat. (Turdakov 2010, 312.) Kohdesanalla tarkoitetaan käännettävää monitulkintaista sanaa. Näitä kohdesanan ja asiayhteyden sanojen välisiä suhteita ovat esimerkiksi syntaktiset suhteet (Turdakov 2010, 312). Ikkunalähestymistavassa kohdesanan asiayhteys eli konteksti voidaan nähdä erilaajuisena. Mikrokonteksti käsittää vain muutaman kohdesanaa lähimpänä olevan sanan. Paikallisessa kontekstissa ikkunan koko on korkeintaan virke. Aihekonteksti käsittää puolestaan useita virkeitä kohdesanan ympärillä. Kaikki nämä erilaajuiset asiayhteydet ovat tärkeässä roolissa sanan merkityksen yksikäsitteistämisiongelman ratkaisemiseksi. Valtaosa yksikäsitteistämisestä käyttää kuitenkin paikallista tai mikrokontekstia. (Turdakov 2010, 312.)

Optimaalisesta ikkunan koosta on eri näkemyksiä. Kaplan (1955, 43) osoitti, että kaksi sanaa monitulkintaisen sanan ympärillä ovat yhtä tehokkaat yksikäsitteistämisessä kuin kohdesanan koko virke. Nykytutkimuksissa kuitenkin todetaan, että yksikäsitteistämisessä tarvittavan asiayhteyden laajuus vaihtelee ja riippuu sovelluksesta (Turdakov 2010, 312). Yarowskyn (1994, 90) mukaan asiayhteyden laajuus saattaa vaihdella monitulkintaisuuden tyypistä riippuen niin, että 3-4 sanaa riittävät ratkaisemaan leksikaalisen monitulkintaisuuden, kun taas semanttinen monitulkintaisuus vaatii 20-50 sanan ikkunan. Gale, Church ja Yarowsky (1993, 429) puolestaan mainitsevat, että yksikäsitteis-

tämisen oikeellisuus paranee laajentamalla kohdesanan kahdentoista naapurisanan asiayhteys sataan sanaan. Myös sanaluokka vaikuttaa yksikäsitteistämässä tarvittavan ikkunan kokoon. Yarowsky (1993, 269-270) osoitti, että substantiivien yksikäsitteistämässä voidaan käyttää laajaa asiayhteyttä, kun taas verbien ja adjektiivien yksikäsitteistämässä asiayhteyden sanojen merkityksellisyys vähenee, kun etäisyys kohdesanaan kasvaa. Juuri mainittu substantiivien yksikäsitteistäminen laajassa asiayhteydessä on ristiriidassa sen kanssa, että edelleen Yarowskyn (1993, 269) mukaan substantiiveja voidaan yksikäsitteistää parhaiten niiden viereisten adjektiivien tai substantiivien perusteella. Näin ollen substantiivien yksikäsitteistämässä ei tarvittaisikaan laajaa asiayhteyttä. Verbit puolestaan johtavat yksikäsitteistämistietoa enemmän objekteista kuin subjekteista, ja adjektiivit saavat lähes kaiken yksikäsitteistämistiedon määrittämistään substantiiveista (Yarowsky 1993, 269). Turdakov (2010, 312) kiteyttää, että tekstin kaikkien sanojen yksikäsitteistäminen vaatii sekä mikro- että aihekontekstia.

Monitulkintaisuudesta johtuvia käännoongelmia on yritetty vähentää monilla eri menetelmillä (Kishida 2005, 437; Zhang & Vines 2005, 1). Yksinkertainen menetelmä on valita vain se käänнос, joka vastaa sanakirjassa ensimmäisenä olevaa merkitystä. Vaihtoehtoisesti voidaan tutkia jokaisen käännoksen esiintymismääriä korpuksessa ja käyttää vain kaikista useimmiten esiintyvää käännoста. Tällaiset yksinkertaiset menetelmät ovat kuitenkin riittämättömiä yksikäsitteistämässä. (Kishida 2005, 439.) Juuri mainittuja menetelmiä kehittyneempiä ovat sanaluokkien jäsentäminen, rinnakkaiscorpusten käyttäminen (Kishida 2005, 439; Pirkola ym. 2001, 218), kohdekielisten sanojen yhteisiintymisen todennäköisyyksien hyödyntäminen tai kyselynlaajennus (Kishida 2005, 439). Yksikäsitteistämässä voidaan käyttää myös rakenteisia kyselyitä (Pirkola ym. 2001, 218), erikoissanakirjoja (Pirkola 1998a, 51) sekä kaksisuuntaista kääntämistekniikkaa (Boughanem, Chrisment & Nassr 2002, 162-163). Myös koneoppimisen menetelmiä hyödynnetään (Jurafsky & Martin 2000, 636). Seuraavaksi tarkastellaan näitä yksinkertaisia menetelmiä kehittyneempiä yksikäsitteistämismenetelmiä.

### **4.3.1 Sanaluokkien jäsentäminen**

Perusidea sanaluokkien jäsentämässä käännoksen yksikäsitteistämiseksi on valita vain ne kohdekieliset käännookset, joilla on sama sanaluokka kuin käännettävässä lähdekielissä sanassa (Kishida 2005, 439). Sanaluokkien jäsentäminen yksikäsitteistämismene-



telmänä edellyttää, että jäsentäminen tehdään sekä lähde- että kohdekielille (Kishida 2005, 439).

Davisin (1996) mukaan sanaluokkiin perustuva yksikäsitteistäminen parantaa kielentvälisen tiedonhaun tuloksellisuutta verrattuna siihen, jos yksikäsitteistämiseksi ei tehtäisi mitään. Myös Ballesteros ja Croft (1998, 69) mainitsevat, että sanaluokkien jäsentämisestä on hyötyä monitulkintaisten sanojen kääntämisessä. Nykyiset yksikäsitteistämismenetelmät hyödyntävät sanaluokkien jäsentämistä muiden menetelmien lisänä (Turdakov 2010, 312).

### **4.3.2 Rinnakkaiskorpusten hyödyntäminen**

Rinnakkaiskorpuksia voidaan hyödyntää käännettyjen sanojen yksikäsitteistämässä (Ballesteros & Croft 1998, 67). Kun sanalla on useita käännöksiä, kukin käännös on yleensä relevantti vain tietyssä asiayhteydessä. Rinnakkaiskorpuksiin perustuvassa yksikäsitteistämässä valitaankin vain ne käännökset, joiden asiayhteys on sama kuin kyselyn asiayhteys. (Boughanem & Nassr 2001, 204.) Ensin käännetään esimerkiksi sanakirjan avulla jokainen lähdekielisen kyselyn sana kohdekielille. Tämän jälkeen haetaan dokumentteja sekä erikseen jokaisella käännetyllä sanalla kohdekielisestä korpuksista että alkuperäisellä kyselyllä lähdekielisestä korpuksista. Lopuksi rinnakkaiskorpusten perusteella valitaan käännös, jolla haettujen kohdekielisten dokumenttien joukko on lähimpänä lähdekielisellä kyselyllä haettujen dokumenttien joukkoa. (Kishida 2005, 440.) Rinnakkaiskorpuksiin perustuva menetelmä epäonnistuu yksikäsitteistämässä, jos korpuksissa ei ole kyselylle ollenkaan tai vain vähän relevantteja dokumentteja (Ballesteros & Croft 1998, 68).

Esimerkiksi Davis (1996) käytti yksikäsitteistämismenetelmänä sekä pelkkiä rinnakkaiskorpuksia että rinnakkaiskorpuksia sanaluokkien jäsentämiseen yhdistettynä. Kyseisen tutkimuksen mukaan pelkkiin rinnakkaiskorpuksiin perustuva yksikäsitteistäminen ei tiedonhaun tuloksellisuuden perusteella tuota hyviä tuloksia. Sanaluokkien jäsentäminen yhdistettynä rinnakkaiskorpusten käyttämiseen puolestaan suoriutuu tiedonhaun tuloksellisuuden perusteella yksikäsitteistämisestä paremmin kuin pelkkä sanaluokkien jäsentäminen. (Davis 1996.) Myös Davisin ja Ogdenin (1997, 95-96) tutkimuksen mukaan rinnakkaiskorpusten käyttämisen ja sanaluokkien jäsentämisen yhdistäminen yksikäsitteistämässä tuottaa käännettyillä kyselyillä tehdyillä tiedonhauilla paremman tu-

loksellisuuden kuin kumpikaan menetelmä yksin. Yksittäisenä yksikäsitteistämismenetelmänä sanaluokkien jäsentäminen on tuloksellisuudeltaan parempi kuin rinnakkaiskorpuksia (Davis & Ogden 1997, 95-96). Juuri mainituista tutkimustuloksista poiketen Boughanemin ja Nassrin (2001, 209) tutkimuksen perusteella korpuksiin perustuva yksikäsitteistäminen antaa kuitenkin hyviä tuloksia. Evaluointi perustui käännettyillä kyselyillä tehtyjen tiedonhakujen tuloksellisuuteen (Boughanem & Nassr 2001, 206-208).

Ballesterosin ja Croftin (1998, 68) mukaan sanojen yhteisesiintymisen todennäköisyyteen perustuva menetelmä on tiedonhaun tuloksellisuuden perusteella parempi yksikäsitteistämismenetelmä kuin pelkkiin rinnakkaiskorpuksiin perustuva menetelmä. Seuraavaksi käsitellään sanojen yhteisesiintymisen todennäköisyyden hyödyntämistä käännettyjen sanojen yksikäsitteistämisessä.

### **4.3.3 Sanojen yhteisesiintymisen todennäköisyyden hyödyntäminen**

Tietoa kyselyn sanojen yhteisesiintymisestä käytetään käännettyjen sanojen yksikäsitteistämisessä (Ballesteros & Croft 1998, 69; Turdakov 2010, 312). Sanojen yhteisesiintymistietoja hyödyntävä menetelmä perustuu rinnakkaiskorpuksiin, koska menetelmä käyttää apunaan korpuksista saatavia yhteisesiintymisten tilastotietoja (Ballesteros & Croft 1998, 64-65). Ballesteros ja Croft (1998, 69) osoittivat, että kyselyn sanojen oikeat käännökset esiintyvät yhdessä kohdekielisisä dokumenteissa, kun taas väärillä käänöksillä ei ole taipumusta yhteisesiintymiseen. Menetelmässä kahden lähdekielisen sanan mahdollisille kohdekielisille vastineille päätellään todennäköisimmät käännökset etsimällä yhteisesiintymismalleja jokaiselle mahdolliselle sanojen muodostamalle käännösparille (Ballesteros & Croft 1998, 65). Ensin lasketaan samankaltaisuuden aste kaikille mahdollisille käännöspareille ja sitten valitaan pari, jolla on suurin samankaltaisuuden aste (Kishida 2005, 440). Esimerkiksi Yarowskyn (1993, 271) mukaan identtisiä kahden sanan yhdistelmiä käytetään samassa merkityksessä 90-99 %:n todennäköisyydellä. Tämän havainnon taustalla on heuristiikka ”yksi merkitys per yhteisesiintymä”, eli keskenään identtisissä sanojen yhteisesiintymisissä sanoilla on sama merkitys. Kyseistä heuristiikkaa käytetään paljon nykytutkimuksissa. (Turdakov 2010, 312.)

Monitulkintaisten sanojen virheelliset käännökset sanojen yhteisesiintymiseen perustuen johtuvat siitä, ettei kyselyssä ole riittävästi asiayhteyttä oikean käännöksen päättelemiseksi (Ballesteros & Croft 1998, 67-68). Ballesteros ja Croft (1998, 68) mainitsevat

tästä esimerkkinä tapauksen, jossa englannin kielelle käännettävä espanjankielinen virke käsitteli suklaan mahdollisia terveysvaikutuksia. Espanjankielinen sana *chocolate* voidaan kääntää englanniksi *chocolate*, *cocoa* tai *blood* (Ballesteros & Croft 1998, 68) eli suklaa, kaakao tai veri. Voidaan olettaa, että sana *health* esiintyy useammin yhdessä sanan *blood* kuin sanan *chocolate* kanssa (Ballesteros & Croft 1998, 68). Sanojen yhteisesiintymisen perusteella valituksi tulee käänös *blood*, vaikka kyseisessä tapauksessa käänös *chocolate* olisi oikea. Jos lähdekielistä kyselyä laajennettaisiin ennen kääntämistä esimerkiksi suklaamerkkiä tarkoittavalla sanalla, monitulkitaisen sanan käännökseksi todennäköisemmin valittaisiin *chocolate* eikä *blood*. (Ballesteros & Croft 1998, 68.) Seuraavaksi käsitellään kyselynlaajennusta yksikäsitteistämismenetelmänä.

#### **4.3.4 Kyselynlaajennus**

Kyselynlaajennuksella kyselyyn lisätyt sanat saattavat tarjota riittävästi asiayhteyttä käännetyn sanan yksikäsitteistämiseksi (Ballesteros & Croft 1998, 68). Ennen käännöstä tehty kyselynlaajennus luo vahvemman perustan käännökselle ja parantaa monitulkitaisen sanan sisältävän haun tuloksellisuutta verrattuna siihen, ettei kyselyä laajennettaisi (Ballesteros & Croft 1998, 66). McNamee ja Mayfield (2002, 165) raportoivat, että ennen kääntämistä tapahtuva kyselynlaajennus on erittäin hyödyllinen yksikäsitteistämisessä, jos käänösresurssien sanastollinen kattavuus on heikko.

Ballesterosin ja Croftin (1998, 69) mukaan kyselyn jälkeen tehtävä kyselynlaajennus parantaa tiedonhaun tuloksellisuutta etenkin, jos sitä käytetään yhteisesiintymismenetelmän kanssa. Myös ennen kääntämistä tehtävä kyselynlaajennus kannattaa yhdistää yhteisesiintymismenetelmään (Ballesteros & Croft 1998, 69). Tästä on esimerkkinä edellä esitetty Ballesterosin ja Croftin (1998, 68) esimerkki suklaan mahdollisia terveysvaikutuksia käsittelevän espanjankielisen virkkeen kääntämisestä englanniksi.

#### **4.3.5 Rakenteisten kyselyiden käyttäminen**

Usein kyselyn tärkeimpiä sanoja ovat sanat, joilla on vain yksi tai kaksi käänösvaihtoehtoa. Sen sijaan niinkään merkityksellisiä eivät ole kyselyn sanat, joilla on monia käänösvaihtoehtoja. (Hull 1997, 90.) Hakusanat, joilla on monia käänösvaihtoehtoja, ja epärelevantit käänösvaihtoehdot ovat hallitsevia kieltenvälisen tiedonhaun rakenteettomissa kyselyissä, joten tärkeiden sanojen vaikutus hakutuloksiin jää vähäiseksi

(Pirkola 1998a, 55). Yksi tapa käsitellä useita käännösvaihtoehtoja onkin käyttää rakenteista kyselyä ja pitää käännöksiä hakuvaiheessa synonyymijoukkona (Pirkola 1998b, 56). Synonyymioperaattoria käytetään kyselyssä ilmaisemaan, että kyseisellä operaattorilla ryhmiteltyjä sanoja pitäisi kohdella yhden sanan ilmaisuina (Pirkola ym. 2001, 219). Synonyymioperaattorin käyttäminen estää ylipainottamasta niitä hakusanoja, joilla on monia käännöksiä ja jotka ovat usein tarpeettomia haussa (Kishida 2005, 442).

Pirkolan (1998b, 56) sekä Ballesterosin ja Croftin (1998, 69) mukaan synonyymioperaattorin käyttäminen auttaa monitulkintaisuusongelmassa. Ballesterosin ja Croftin (1998, 69) tutkimuksen perusteella synonyymioperaattorin käyttäminen on tiedonhaun tuloksellisuuden perusteella parempi yksikäsitteistämismenetelmä kuin sanaluokkien jäsentäminen. Paras tuloksellisuus saavutetaan yhdistämällä nämä kaksi menetelmää (Ballesteros & Croft 1998, 66). Pirkolan (1998b, 56) tutkimuksen perusteella rakenteinen kysely yhdistettynä erikoissanakirjan käyttämiseen yleissanakirjan ohella parantavat kieltenvälisen tiedonhaun tuloksellisuutta. Seuraavaksi tarkastellaan erikoissanakirjan käyttämistä yksikäsitteistämismenetelmänä.

#### **4.3.6 Erikoissanakirjan käyttäminen**

Pirkolan (1998b, 61) tutkimuksessa tarkasteltiin monitulkintaisuusongelmaa sekä sanakirjojen kattavuusongelmaa käyttämällä kääntämisessä yleissanakirjan lisäksi lääketieteellistä erikoissanakirjaa. Yleissanakirjat antavat käännettävälle sanalle useita käännösvaihtoehtoja, kun taas erikoissanakirjat antavat vain yhden tai kaksi vaihtoehtoa. Näin ollen erikoissanakirjaa käytettäessä monitulkintaisuus ei ole yhtä suuri kääntämisiongelma kuin yleissanakirjaa käytettäessä. (Pirkola 1998b, 61.) Pirkola (1998b, 61) mainitsee esimerkkinä, että yleissanakirja antoi sanalle *leikkaus* seitsemän eri käännösvastinetta (*cut, cutting, clipping, operation, editing, section* ja *retrenchment*), kun taas lääketieteellinen erikoissanakirja vain kaksi vastinetta (*operation* ja *surgery*). Erikoissanakirjojen käyttämisen hyödyn luulisi kuitenkin riippuvan siitä, sisältävätkö kyselyt kyseisen erikoisalalan sanoja.

Pirkolan (1998b, 61) mukaan sanakirjaperustaisen kääntämismenetelmän käyttäminen kieltenvälisessä tiedonhaussa tuottaa tiedonhaun tuloksellisuuden perusteella lähes yhtä hyviä tuloksia kuin yksikielinen tiedonhaku. Ehtona on kuitenkin, että kääntämisessä

käytetään sekä rakenteisia kyselyitä että yleissanakirjan lisäksi esikoissanakirjaa (Pirkola 1998b, 61).

### **4.3.7 Kaksisuuntainen kääntämistekniikka**

Boughanem ja muut (2002, 162-163) käyttivät yhtenä yksikäsitteistämismenetelmänä kaksisuuntaista kääntämistekniikkaa, jossa takaisin alkuperäiselle kielelle käännettyä sanaa käytetään käännösvaihtoehtojen karsimisessa. Jos esimerkiksi on tarkoitus kääntää englanninkielinen kysely ranskan kielelle, kaksisuuntaisessa kääntämistekniikassa etsitään ensin kyselyn englanninkielisten sanojen ranskankieliset vastineet englanti-ranska -sanakirjasta. Tämän jälkeen ranska-englanti -sanakirjan avulla jokainen juuri käännetty ranskankielinen sana käännetään takaisin englannin kielelle. Jos joku takaisin englanniksi käännettyistä sanoista vastaa alkuperäisen englanninkielisen kyselyn sanaa, kyseistä sanaa vastaava ranskankielinen käänнос valitaan käännökseksi. (Kishida 2005, 443.)

Boughanemin ja muiden (2002, 167) tutkimuksen perusteella kaksisuuntainen kääntämistekniikka on parempi yksikäsitteistämismenetelmä kuin rinnakkaiskorpuksiin perustuva menetelmä. Tulokset perustuvat käännettyillä kyselyillä tehtyjen tiedonhakujen tuloksellisuuteen (Boughanem ym. 2002, 167).

### **4.3.8 Koneoppimisen menetelmät**

Sanan merkityksen yksikäsitteistäminen voidaan nähdä myös luokitteluongelmana, jossa tekstin jokaiselle sanalle voidaan määritellä luokka eli merkitys (Turakov 2010, 317). Kyseessä on koneoppimisen menetelmä, jossa järjestelmä opetetaan yksikäsitteistämään (Jurafsky & Martin 2000, 636). Koneoppimisessa opetusvaiheen tarkoituksena on luoda tai opettaa luokittelija, joka osaa määrätä oikean merkityksen uusille, järjestelmän ennen näkemättömille sanoille (Jurafsky & Martin 2000, 638). Menetelmässä oleellista on se, että yksikäsitteistämistehtävässä tarvittava tietämys saadaan aineistosta eikä ihmisten tekemän analyysin tuloksena (Jurafsky & Martin 2000, 636).

Koneoppimisen lähestymistavat monitulkintaisen sanan yksikäsitteistämiseksi voivat olla ohjattua tai ohjaamatonta oppimista. Ohjatussa oppimisessa sanan merkitys on opetteluvaiheessa tiedossa (Jurafsky & Martin 2000, 638). Ohjaamattomassa oppimisessa sen sijaan sanan merkitystä ei opetteluvaiheessa tiedetä (Jurafsky & Martin 2000, 644).

## 5 KONEKÄÄNNÖSJÄRJESTELMIEN EVALUOINTI-MITTAREITA

Evaluointi on välttämätöntä, jotta kieltenvälisen tiedonhaun menetelmiä voidaan kehittää (Kishida 2005, 450). Ihmisten tekemä evaluointi on kannattavaa mutta kallista ja aikaa vievää (Papineni ym. 2001, 1). Konekäännösjärjestelmä voidaan evaluoida nopeammin, yksinkertaisemmin ja edullisemmin automaattisilla mittareilla kuin ihmisten tuottamilla arvioinneilla (Lavie & Denkowski 2009, 1). Automaattisen evaluointimittarin kehittäminen on kuitenkin haastavaa, koska käännettävälle tekstille voi olla useita oikeita käännöksiä (Cer, Manning & Jurafsky 2010, 556).

Ensimmäinen konekäännösjärjestelmien automaattinen evaluointimittari oli IBM-ryhmän kehittämä BLEU (Hutchins 2005a, 15). BLEU on viime vuosina ollut käytetyin automaattinen evaluointimittari (Lavie & Denkowski 2009, 1). Vaikka BLEU:ta käytetään laajasti ja se on suuresti edistänyt tilastollisen konekääntämisen kehitystä, sillä on myös monia heikkouksia. METEOR kehitettiin BLEU:n heikkouksien ja puutteiden korjaamiseksi. (Lavie & Denkowski 2009, 3.)

Sekä BLEU että METEOR perustuvat ihmisten tuottamien käännösten käyttämiseen referenssikäännöksinä; konekäännösjärjestelmän tuottamaa käännöstä vertaillaan yhteen tai useampaan referenssikäännökseen (Banerjee & Lavie 2005, 65; Papineni ym. 2001, 2). Perusajatuksena on, että referenssikäännöksen kanssa mahdollisimman samanlainen konekäännös saa korkeat pisteet ja referenssikäännöksestä poikkeava konekäännös saa alhaiset pisteet (Hutchins 2005a, 15).

Seuraavaksi esitellään yleisellä tasolla sekä BLEU:n että METEORin toimintaperiaatteet ja mittareiden oleellimmat erot. Mittareihin tutustutaan, jotta saadaan käsitystä niiden periaatteista ja käyttömahdollisuuksista. Lopuksi esitellään perinteisen tarkkuuden käyttömahdollisuuksia konekääntämisen evaluoinnissa.

### 5.1 BLEU

BLEU (BiLingual Evaluation Understudy) on automaattinen konekäännöksen evaluointimittari, jossa vertaillaan kone- ja referenssikäännösten  $n$ -grammeja ja lasketaan toisiinsa vastaavien  $n$ -grammien lukumääriä (Papineni ym. 2001, 2).  $N$ -grammi on  $n:n$  mer-

kin mittainen osamerkkijono jostain pidemmästä merkkijonosta (Lavie & Denkowski 2009, 2). BLEU:ssa  $n$ -grammilla tarkoitetaan  $n$  määrää sanoja eikä  $n$  määrää merkkejä (Papineni ym. 2001, 2).  $N$ -grammeihin perustuvat tarkkuusarvot lasketaan erikseen yhdestä sanasta  $n$ :ään sanaan ja näistä eri  $n$ :n arvoihin perustuvista tarkkuuksista lasketaan tietyntyyppinen keskiarvo (ks. Lavie & Denkowski 2009, 2-3). Papinenin ja muiden (2001, 4) mukaan paras tulos saadaan  $n$ :n arvolla 4.

BLEU on kieliriippumaton mittari (Papineni ym. 2001, 1), eli sitä voidaan käyttää käännösten evaluoinnissa minkä tahansa kielten välillä. Mittari on myös paikkariippumaton (Papineni ym. 2001, 2), eli vertailtavien sanojen ei tarvitse olla kone- ja referenssikäännöksessä samassa järjestyksessä. BLEU:ssa evaluoinnin perusyksikkönä on virke (Papineni ym. 2001, 3). Mittarin arvo vaihtelee nolasta ykköseen. Arvo on yksi, jos evaluoitava käännös on täysin identtinen referenssikäännöksen kanssa. (Papineni ym. 2001, 6.) BLEU:n on osoitettu korreloivan ihmisten tekemien arvioiden kanssa (Cer ym. 2010, 555; Papineni ym. 2001, 1).

BLEU:ssa otetaan huomioon konekäännöksen pituus verrattuna referenssikäännöksen pituuteen muunnetulla  $n$ -grammitarkkuudella sekä lyhyyssakolla (brevity penalty, BP) (Papineni ym. 2001, 4-5). Muunnettu  $n$ -grammitarkkuus rankaisee liian pitkistä konekäännöksistä (Papineni ym. 2001, 5), eli jos konekäännöksessä on referenssikäännöksessä esiintymättömiä sanoja tai jos jokin sana esiintyy useammin kone- kuin referenssikäännöksessä (Papineni ym. 2001, 4). Lyhyyssakko puolestaan rankaisee lopullista BLEU-arvoa, jos konekäännös on liian lyhyt verrattuna referenssikäännökseen (Papineni ym. 2001, 5) eli jos konekäännöksestä puuttuu referenssikäännöksessä olevia sanoja.

Saanti on kone- ja referenssikäännöksen yhteisten sanojen osuus referenssikäännöksen sanojen kokonaismäärästä (Lavie & Denkowski 2009, 3). Saannin määrittäminen BLEU:ssa on ongelmallista, koska yhtä konekäännöstä voidaan verrata useampaan kuin yhteen referenssikäännökseen. Näin ollen on vaikea määrittellä saannin laskemisessa tarvittava referenssikäännöksen sanojen kokonaismäärä. (Papineni ym. 2001, 5.) BLEU korvaakin saannin edellä kuvatulla lyhyyssakolla (Lavie & Denkowski 2009, 3). Banerjeen ja Lavien (2005, 67) mukaan saannin mittaamisen puuttuminen on BLEU:n suuri heikkous, eikä keinotekoinen lyhyyssakko riittävästi korvaa tätä puutetta. Myös Cerin ja muiden (2010, 556) mukaan BLEU:n heikkoutena on se, ettei saantia huomioida eli ettei mittari tarkista, puuttuuko konekäännöksestä referenssikäännöksessä olevia sanoja.

Saannin puuttumisen lisäksi BLEU:n heikkoutena pidetään myös sitä, että BLEU:n arvoa laskettaessa täsmäytettävien n-grammien järjestyksellä ei ole väliä, jolloin kone- ja referenssikäännöksen yhteiset sanat voivat olla konekäännöksessä täysin satunnaisessa järjestyksessä (Cer ym. 2010, 556). Esimerkiksi Kettunen (2009, 103) esittää eriävän näkemyksen, jonka mukaan käännöksen sanajärjestyksellä ei ole merkitystä. On kuitenkin eroteltava kaksi erityyppistä kääntämistä: yksittäisistä sanoista koostuvien kyselyiden kääntäminen ja virkkeiden kääntäminen. Virkkeissä sanajärjestyksellä on väliä mutta kyselyissä ei. BLEU:n heikkoutena on myös se, ettei se ota huomioon synonyymeja (Cer ym. 2010, 555). Näin ollen se saattaa rangaista konekäännöstä, jossa esiintyy referenssikäännöksessä olevien sanojen synonyymeja (Cer ym. 2010, 555).

## 5.2 METEOR

METEOR on BLEU:n tavoin automaattinen konekääntämisen evaluointimittari (Lavie & Denkowski 2009, 1). METEOR tulee sanoista *Metric for Evaluation of Translation with Explicit Ordering* (Banerjee & Lavie 2005, 66). Ensimmäinen versio METEORista julkaistiin vuonna 2004 (Lavie & Denkowski 2009, 1).

METEOR perustuu yksittäisten sanojen täsmäyttämiseen konekäännöksen ja yhden tai useamman referenssikäännöksen välillä (Lavie & Denkowski 2009, 2). Täsmäytys ei huomioi pelkästään täysin identtisiä sanoja, vaan se ottaa huomioon myös sanat, joilla on identtinen sanavartalo tai jotka ovat toistensa synonyymeja (Banerjee & Lavie 2005, 66). Yksi METEORin merkittävä parannus aiempiin mittareihin verrattuna onkin eri käännösvariaatioiden huomioiminen (Lavie & Denkowski 2009, 2). Lisäksi toisin kuin tarkkuuteen perustuva BLEU, METEOR painottaa tarkkuuden lisäksi myös saantia. Saannin painottaminen on tärkeää vahvalle korrelaatiolle kone- ja referenssikäännösten välillä. (Lavie & Denkowski 2009, 1.) METEORin heikkoutena kieliriippumattomaan BLEU-mittariin verrattuna on puolestaan se, että METEORia ei ole tarjolla kovinkaan monen kieliparin evaluoinnille (Kettunen 2009, 103).

METEOR-mittarin numeerisen arvon muodostuminen koostuu kahdesta vaiheesta: täsmäytysvaiheesta sekä pisteiden laskemisesta (Banerjee & Lavie 2005, 67-68). Täsmäytysvaiheessa METEOR luo vastaavuudet konekäännöksen ja erikseen jokaisen referenssikäännöksen välille ja valitsee sen referenssikäännöksen, jonka kanssa konekäännöksellä on eniten vastaavuuksia. Jos vastaavuuksien lukumäärien perusteella konekäännös-



tä parhaiten vastaa useampi kuin yksi referenssikäännös, valitaan sen referenssi- ja konekäännöksen vastaavuuksien joukko, jossa on eniten samassa järjestyksessä olevia sanoja. (Banerjee & Lavie 2005, 68.) Kun kaikki kone- ja referenssikäännöksen väliset vastaavuudet on löydetty ja vastaavuusjoukoista on valittu paras, lasketaan varsinaiset METEOR-pisteet (Lavie & Denkowski 2009, 4). METEOR-pisteiden laskeminen edellyttää muun muassa saannin ja tarkkuuden laskemista (Lavie & Denkowski 2009, 4-5). Lisäksi konekäännöksen sanajärjestys otetaan huomioon laskemalla mittarin arvoa vähentävä rangaistus väärästä sanajärjestyksestä (Lavie & Denkowski 2009, 5). BLEU:n tavoin METEORinkin arvo vaihtelee 0 ja 1 välillä (Lavie & Denkowski 2009, 5).

### 5.3 Tarkkuus

Konekäännöksen laatua voidaan evaluoida myös saannilla ja tarkkuudella (Melamed, Green & Turian 2003, 61). Melamedin ja muiden mukaan (2003, 61) näillä mittareilla on korkeampi vastaavuus ihmisten tekemien evaluointien kanssa kuin BLEU:lla. Tässä tutkielmassa tarkastellaan vain tarkkuutta, joka lasketaan jakamalla oikein käännettyjen sanojen lukumäärä käännettyjen sanojen kokonaismäärällä (Melamed ym. 2003, 61). Jurafskyn ja Martinin (2000, 639) mukaan myös sanan merkityksen yksikäsitteistämistäjärjestelmiä voidaan evaluoida tarkkuudella, jolloin lasketaan oikein yksikäsitteistettyjen sanojen prosenttiosuus kaikista evaluoitavista sanoista. Varhainen lähestymistapa konekäännöksen pisteyttämiseen referenssikäännösten avulla olikin laskea yhteisten kone- ja referenssikäännösten sanojen osuus referenssikäännöksen sanojen kokonaismäärästä (Melamed 1995, 187). Myös tarkkuus saa arvonsa välillä 0 ja 1. Mitä parempi käännöslaatu on, sitä lähempänä mittarin arvo on lukua yksi.

Edellä esiteltyjen BLEU:n ja METEORin tavoin tarkkuudenkin käyttäminen konekäännösjärjestelmän evaluointimittarina vaatii tietoa käännettyjen sanojen oikeista käännöksistä (Jurafsky & Martin 2000, 639). Tarkkuuden mittaaminen voi olla vaikeaa, koska eri käännökset sopivat erilaisiin asiayhteyksiin ja useimmissa tapauksissa useampi kuin yksi käännös on oikea (Melamed 1995, 186). Pääongelmana tarkkuuden käyttämisessä konekäännösjärjestelmän evaluoinnissa onkin se, miten löydetään sopiva tapa kone- ja referenssikäännöksen yhteisten käännösten lukumäärän laskemiseksi (Melamed ym. 2003, 61).

## 6 TUTKIMUSASETELMA

Kieltenvälisen tiedonhaun konekäännösjärjestelmien evaluointitutkimuksissa vaikuttaa olevan pääasiassa kolmenlaista käytäntöä. Yhtenä vaihtoehtona on evaluoida pelkästään käännöksen laatua jollain konekäännösjärjestelmien evaluointiin kehitetyllä mittarilla, esimerkiksi BLEU:lla tai METEORilla, tai useammalla evaluointimittarilla samanaikaisesti (ks. Carpuat & Wu 2007a, 48-50; Kit & Wong 2008, 299; Toral ym. 2011, 14; Zens ym. 2002, 28-30). Toisena vaihtoehtona on evaluoida pelkästään käännettyillä kyselyillä tehtyjen tiedonhakujen tuloksellisuutta, jolloin evaluointimittareina käytetään esimerkiksi keskitarkkuutta tai keskitarkkuuksien keskiarvoa (ks. Ballesteros & Croft 1998, 66; Oard 1998, 480-481; Wu ym. 2008, 75-76). Kolmas vaihtoehto on puolestaan se, että evaluoidaan sekä käännösten laatua että tiedonhaun tuloksellisuutta (ks. Kettunen 2009, 104-106; Zhu & Wang 2006, 596-599). Tutkimuksen tarkoituksesta ja tutkimuskysymyksistä riippuu, millaista tietoa tarvitaan ja miten evaluointi kannattaa toteuttaa. On myös muistettava, että käännöksen laatu ja tiedonhaun tuloksellisuus liittyvät toisiinsa (Zhu & Wang 2006, 593).

Evaluointitutkimuksissa vertaillaan usein eri järjestelmiä tai menetelmiä. Esimerkiksi Wu ja muut (2008) vertailivat Google Translate -konekäännösjärjestelmää ja sanakirjaperustaista kääntämistä ja Kettunen (2009) kahtatoista verkossa toimivaa konekäännösjärjestelmää sekä lyhyitä ja pitkiä kyselyitä. Toral ja muut (2011) vertailivat CoSyne-konekäännösjärjestelmää neljään verkossa toimivaan konekäännösjärjestelmään sekä eri konekääntämisen kääntämismenetelmiä. Kit ja Wong (2008) vertailivat puolestaan lakitekstien kääntämisessä kuutta konekäännösjärjestelmää eri kieliyhdistelmillä ja Ballesteros ja Croft (1998) monitulkintaisten sanojen yksikäsitteistämismenetelmiä.

Kuten aiemmin mainittiin, käännösten laatu vaikuttaa kieltenvälisen tiedonhaun hakutuloksiin (Huang ym. 2007, 81), ja käännösten laadusta voidaan ennustaa tiedonhaun tuloksellisuutta (Kettunen 2009, 104). Tässä tutkimuksessa keskitytäänkin ainoastaan käännösten laadun evaluointiin. Koska konekäännösjärjestelmien käännöksissä on edelleen paljon puutteita ja heikkouksia (Manning & Schütze 2000, 463), tarkasteltava ongelma halutaan rajata ja keskitytään pelkästään monitulkintaisten sanojen käännösten laatuun. Yksikäsitteisten sanojen käännösten laatua tarkastellaan vain, jotta saadaan tietoa eri konekäännösjärjestelmien käännösten laadun perustasosta ja jotta voidaan ver-

tailla monitulkintaisten sanojen käännosten laatua yksikäsitteisten sanojen käännösten laatuun. Esimerkiksi Ballesterosin ja Croftin (1998, 69) sekä Kishidan (2005, 439) mukaan sanojen monitulkintaisuus on keskeinen ongelma kieltenvälisessä tiedonhaussa.

Seuraavaksi määritellään tämän tutkimuksen tutkimuskysymykset. Sen jälkeen esitellään testatut konekäännösjärjestelmät. Järjestelmäesittelyiden jälkeen kuvaillaan, miten aineisto kerättiin. Lopuksi raportoidaan, miten aineistoa analysoitiin.

## 6.1 Tutkimuskysymykset

Tämän tutkimuksen tutkimusongelmana on, kuinka virheettömästi konekäännösjärjestelmät kääntävät englanninkielisiä monitulkintaisia sanoja suomeksi. Tutkimuskysymykset ovat seuraavat

1. Onko virkeyhteudessa esiintyvien yksikäsitteisten ja monitulkintaisten sanojen käännösten laadussa eroja?
2. Onko monitulkintaisten sanojen käännösten laadussa eroja testattavien konekäännösjärjestelmien välillä?
3. Onko monitulkintaisten sanojen käännösten laadussa eroja erilaajusten asiayhteyksien välillä?
4. Onko monitulkintaisten sanojen käännösten laadussa eroja sääntöperustaisten ja tilastollisten konekäännösjärjestelmien välillä?
5. Minkä tyyppisiä virheitä konekäännösjärjestelmät tekevät virkeyhteudessa esiintyvien monitulkintaisten sanojen kääntämisessä?

Kyseessä on vertaileva evaluointitutkimus, joka suoritetaan ilman käyttäjää. Tutkimuksessa keskitytään pelkästään siihen, kuinka virheettömästi konekäännösjärjestelmät kääntävät tarkasteltavat yksikäsitteiset ja monitulkintaiset sanat. Tutkimuksessa ei olla kiinnostuneita siitä, ovatko virkkeiden muut sanat käännetty oikein tai ovatko virkkeet kokonaisuudessaan käännetty oikein. Näin ollen tutkimuksessa ei tarkastella sanojen taivutusmuotoja, käännettyjen virkkeiden sanajärjestystä eikä koko virkkeen ymmärrettävyyttä.

## 6.2 Testatut konekäännösjärjestelmät

Tutkimusongelman perusteella testattujen konekäännösjärjestelmien piti kääntää englannista suomeksi. Tämä kielirajaus vähensi huomattavasti mahdollisten konekäännösjärjestelmien määrää, koska suomen kieli ei ole vaihtoehtona kovinkaan monessa konekäännösjärjestelmässä. Myös Toralin ja muiden (2001, 17) tutkimuksessa käytettyjen konekäännösjärjestelmien ensimmäisenä valintakriteerinä luonnollisesti oli se, että järjestelmien piti kääntää testattavilla kielipareilla. Konekääntämistä saattaakin rajoittaa se, ettei järjestelmiä ole tarjolla läheskään kaikille kielipareille (Zhu & Wang 2006, 594). Esimerkiksi Babelfish tai Systran eivät sisällä suomen kieltä.

Jotta voitiin vastata toiseen tutkimuskysymykseen eli vertailla eri konekäännösjärjestelmien käännösten laatua, oli testattava useampaa kuin yhtä konekäännösjärjestelmää. Jotta puolestaan voitiin vastata neljänteen tutkimuskysymykseen eli vertailla eri konekäännös menetelmien käännösten laatua, osan testattavista konekäännösjärjestelmistä piti olla esimerkiksi sääntöperustaisia ja osan esimerkiksi tilastollisia. Testattujen konekäännösjärjestelmien valitsemisen edellytyksenä oli siis se, että konekäännösjärjestelmistä piti olla tarjolla riittävästi tietoa, jotta voitiin päätellä, mitä kääntämismenetelmää kukin järjestelmä käyttää.

Testatut konekäännösjärjestelmät olivat seuraavat viisi järjestelmää

- TeemaPoint (TP) ([www.teemapoint.com](http://www.teemapoint.com))
- Sunda ([www.sunda.fi/kaantaja.html](http://www.sunda.fi/kaantaja.html))
- SDL Free Translation (SDL) ([www.freetranslation.com](http://www.freetranslation.com))
- Google Translate (Google) ([translate.google.com](http://translate.google.com))
- Bing<sup>TM</sup> Translator Beta (Bing) ([www.microsofttranslator.com](http://www.microsofttranslator.com)).

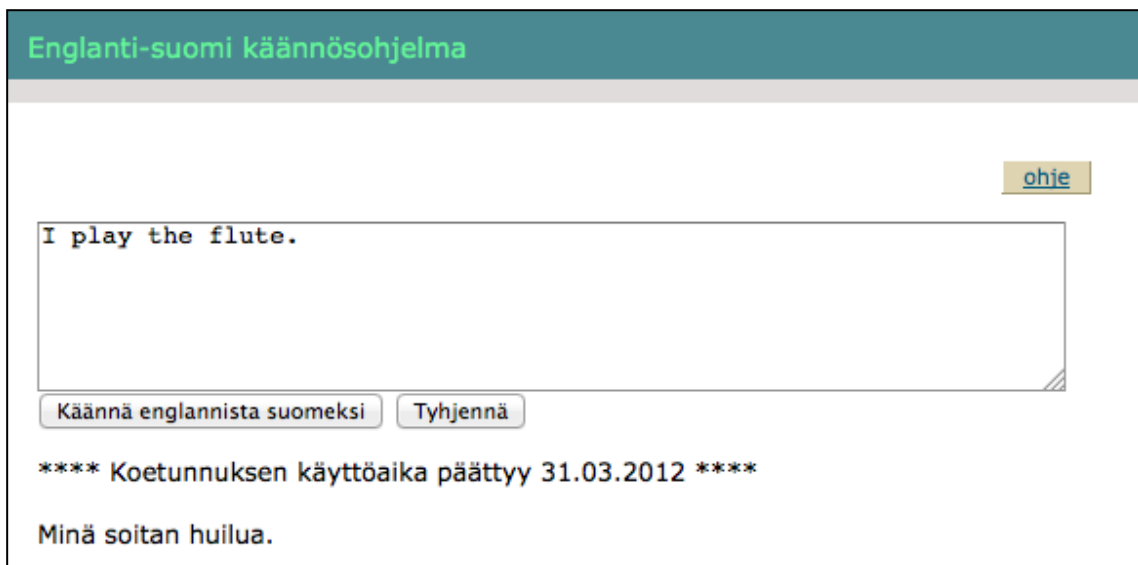
Suluissa on taulukoissa ja kuvissa järjestelmistä käytettävät lyhenteet sekä konekäännösjärjestelmien verkko-osoitteet.

Seuraavaksi esitellään testatut järjestelmät. Yksityiskohtaista tietoa järjestelmien toimintatavoista ei esimerkiksi järjestelmien verkkosivuilla paljasteta, koska kyseessä ovat liikesalaisuudet. Ei esimerkiksi ole tietoa, mitä monitulkintaisten sanojen yksikäsitteistämismenetelmiä testatut konekäännösjärjestelmät käyttävät. Konekäännösjärjestelmien esittelyn jälkeen tehdään yhteenvetoa järjestelmien ominaisuuksista.

## 6.2.1 TeemaPoint

TeemaPoint on suomalainen konekäännösjärjestelmä. Järjestelmä on maksullinen, mutta sen verkkosivuilta voi tilata kolmeksi kuukaudeksi ilmaisen testitunnuksen. TeemaPointista voi käyttää verkko- tai työasemaversiota. (TeemaPointin verkkosivut 2011.) Tässä tutkimuksessa käytettiin verkkoversiota. TeemaPointilla voi kääntää ainoastaan englannista suomeksi (TeemaPointin verkkosivut 2011), joten järjestelmä on kaksikielinen ja yksisuuntainen. TeemaPoint tarjoaa myös muita palveluita, esimerkiksi sanaluokkien jäsentämistä (TeemaPointin verkkosivut 2011).

TeemaPoint kääntää kerralla enintään 3000 merkin mittaisen tekstin. Jos jollekin englanninkieliselle sanalle ei ole sanastossa käännösvastinetta tai järjestelmä ei pysty valitsemaan kyseiseen asiayhteyteen sopivaa käännösvaihtoehtoa, järjestelmä merkitsee kääntämättä jääneen sanan hakasulkuihin. (TeemaPointin verkkosivut 2011.) Esimerkiksi virkkeen ”*He could feel the beat of her heart.*” käännös on ”*Hän voisi tuntea hänen sydämen [beat].*”. Kuvassa 4 on TeemaPointin käyttöliittymä esimerkikäännöksellä ”*I play the flute.*”.



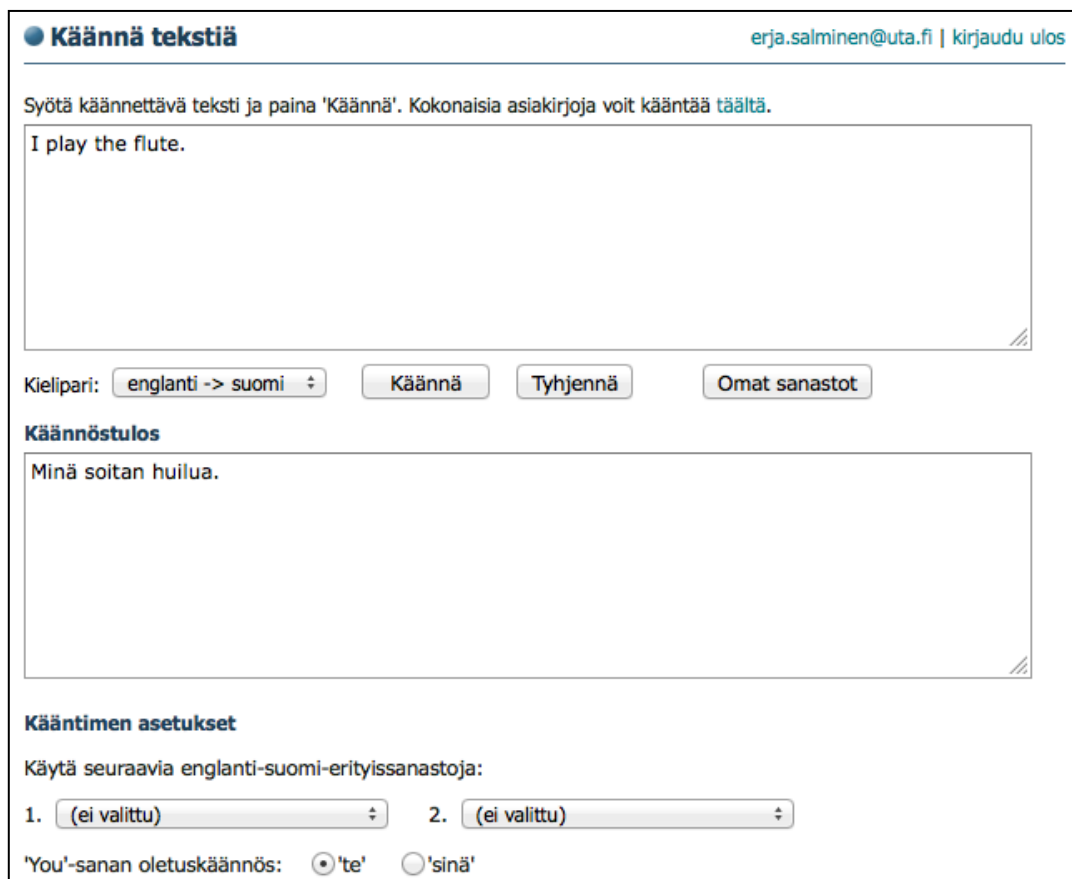
Kuva 4. TeemaPoint-konekäännösjärjestelmän käyttöliittymä

TeemaPoint on sääntöperustainen konekäännösjärjestelmä, mikä on pääteltävissä järjestelmän verkkosivuilla olevasta järjestelmän kuvauksesta. Verkkosivuilla mainitaan, että järjestelmä tekee ennen kääntämistä syntaktista jäsentämistä, esimerkiksi sanaluokkien tunnistamista. Tämän jäsentämisvaiheen mainitaan olevan kääntämisen aikaa vievin prosessi. (TeemaPointin verkkosivut 2011.) Järjestelmä käyttää WordNetiä (WordNet

2011) yhtenä sanastotietokantana. Järjestelmässä on yli 50 000:lle WordNetin englanninkieliselle sanalle suomenkieliset vastineet. Lisäksi järjestelmä suomentaa englanninkielisten sanojen käyttöesimerkkejä. (TeemaPointin verkkosivut 2011.)

TeemaPoint käyttää yksikäsitteistämismenetelmänä ainakin sanaluokkien jäsentämistä. Järjestelmän verkkosivuilla kuitenkin mainitaan, että sanojen monitulkintaisuuden ratkaisemisessa tarvitaan sanaluokkien jäsentämisen lisäksi myös muita menetelmiä, jotta oikea merkitys pystytään valitsemaan asiayhteyden perusteella. (TeemaPointin verkkosivut 2011.)

## 6.2.2 Sunda



The screenshot shows the Sunda web interface for text translation. At the top, it says "Käännä tekstiä" and "erja.salminen@uta.fi | kirjaudu ulos". Below this, there is a text input field containing "I play the flute." and a button "Käännä". Below the input field, there is a dropdown menu for "Kielipari:" set to "englanti -> suomi", and buttons for "Tyhjennä" and "Omat sanastot". Below this, there is a section "Käännöstulos" containing the translated text "Minä soitan huilua." Below this, there is a section "Kääntimen asetukset" with the text "Käytä seuraavia englanti-suomi-erityissanastoja:" and two dropdown menus for selecting specific words, both currently set to "(ei valittu)". At the bottom, there is a radio button selection for "'You'-sanan oletuskäännös:" with options "'te'" (selected) and "'sinä'".

Kuva 5. Sunda-konekäännösjärjestelmän käyttöliittymä

Sunda on TeemaPointin tavoin suomalainen konekäännösjärjestelmä (Sundan verkkosivut 2011). Järjestelmällä voi kääntää joko suomesta englanniksi tai englannista suomeksi (Sundan verkkosivut 2011), joten järjestelmä on kaksikielinen ja -suuntainen. Kehitteillä on myös käännösmahdollisuus englannista ruotsiksi (Sundan verkkosivut 2011). Kuvasta 5 ilmenee Sundan käyttöliittymä käännettäessä esimerkkivirke ”*I play the flute.*”.

Sunda tarjoaa useita käännöstuotteita niin yksityis- kuin yrityskäyttöön. Tässä tutkimuksessa testattiin Sunda Toolia, joka on yksityiskäyttöön tarkoitettu tekstien ja asiakirjojen käännösjärjestelmä. Sunda Toolia voi käyttää sekä verkko- että työasemaversiona. (Sundan verkkosivut 2011.) Tässä tutkimuksessa testattiin verkkoversiota.

Sunda Toolia voi testata ilmaiseksi Sundan verkkosivuilla kahdenkymmenen käännöksen verran. Muuten konekäännösjärjestelmän käyttäminen on maksullista. Tuotteen voi ostaa käyttöön kolmeksi kuukaudeksi tai vuodeksi kerrallaan. (Sundan verkkosivut 2011.) Tätä tutkimusta varten saatiin kahdeksi viikoksi ilmainen testitunnus.

Sunda on TeemaPointin tavoin sääntöperustainen järjestelmä, koska kääntäminen perustuu kielellisten rakenteiden perusteelliseen analyysiin sekä sanastoon (Sundan verkkosivut 2011). Sundan englanti-suomi -järjestelmä tunnistaa yli 160 000 sanaa ja sanontaa, ja se sisältää paljon myös erityisalojen sanoja liittyen esimerkiksi tietotekniikkaan, talouteen tai terveyteen ja biologiaan. Erityissanasto voidaan ottaa käyttöön järjestelmän käyttöliittymässä. Perusversiota kehittyneemmissä versioissa voidaan käyttää myös omia sanastoja. (Sundan verkkosivut 2011.) Nämä valintamahdollisuudet ilmenevät kuvasta 5. Tässä tutkimuksessa ei käytetty omia eikä erityissanastoja. Kuvasta 5 ilmenee myös se, että sanalle *you* voidaan valita oletuskäännökseksi *te* tai *sinä*. Tässä tutkimuksessa oletusvalinnan annettiin olla *te*, koska yksikkö- tai monikkomuodolla ei oletettu olevan vaikutusta tutkimustuloksiin.

### 6.2.3 SDL Free Translation



Kuva 6. SDL Free Translation -konekäännösjärjestelmän käyttöliittymä

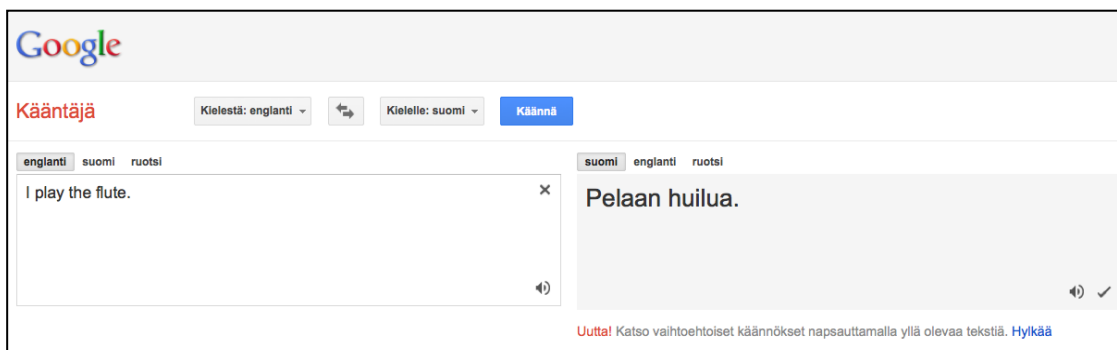
SDL Free Translation on verkossa käytettävä ilmainen konekäännösjärjestelmä (SDL Free Translationin verkkosivut 2011). Järjestelmä on monikielinen ja kaksisuuntainen.

Järjestelmässä on runsas määrä kieliä, mutta esimerkiksi suomenkielistä tekstiä voidaan kääntää vain englanniksi, kun taas englanninkielistä tekstiä voidaan kääntää 34 eri kielelle (SDL Free Translationin verkkosivut 2011). Kuvassa 6 on SDL Free Translation -järjestelmän käyttöliittymä esimerkkikäännöksellä ”*I play the flute.*”.

SDL Free Translation on sääntöperustainen konekäännösjärjestelmä (Toral ym. 2011, 17), mikä on pääteltävissä myös järjestelmän verkkosivuilla annetuista ohjeista. Ohjeissa kehoitetaan lopettamaan virkkeet pisteeseen ja käyttämään kieliopillisesti ja oikeinkirjoitukseltaan virheetöntä kieltä (SDL Free Translationin verkkosivut 2011). Järjestelmän kerrotaan analysoivan lauseita kielioppisääntöjen perusteella. Sivulla mainitaan myös, että järjestelmä ratkaisee monitulkintaisen sanan käännsvaihtoehdoista todennäköisimmän kielioppisääntöjen ja kontekstianalyysin avulla. (SDL Free Translationin verkkosivut 2011.) Yksikäsitteistämässä siis ilmeisesti käytetään ainakin sanaluokkien jäsentämistä. SDL Free Translationin verkkosivujen ohjeissa kuitenkin kehoitetaan välttämään monitulkintaisia sanoja, joiksi mainitaan useampaa kuin yhtä sanaluokkaa edustavat sanat (SDL Free Translationin verkkosivut 2011).

#### 6.2.4 Google Translate

Google Translate on ilmainen verkossa käytettävä tilastollinen konekäännösjärjestelmä (Wu ym. 2008, 71). Järjestelmän verkkosivuilla mainitaan, että kun Google Translate luo käänöksen, se etsii malleja ihmisten kääntämistä dokumenteista ja määrittää niiden avulla sopivan käänöksen. Järjestelmän sanotaan kattavan 57 kieltä, vaikka käyttöliittymässä on valittavana 63 kieltä. (Google Translaten verkkosivut 2011.) Järjestelmä on monikielinen ja kaksisuuntainen.



Kuva 7. Google Translate -konekäännösjärjestelmän käyttöliittymä

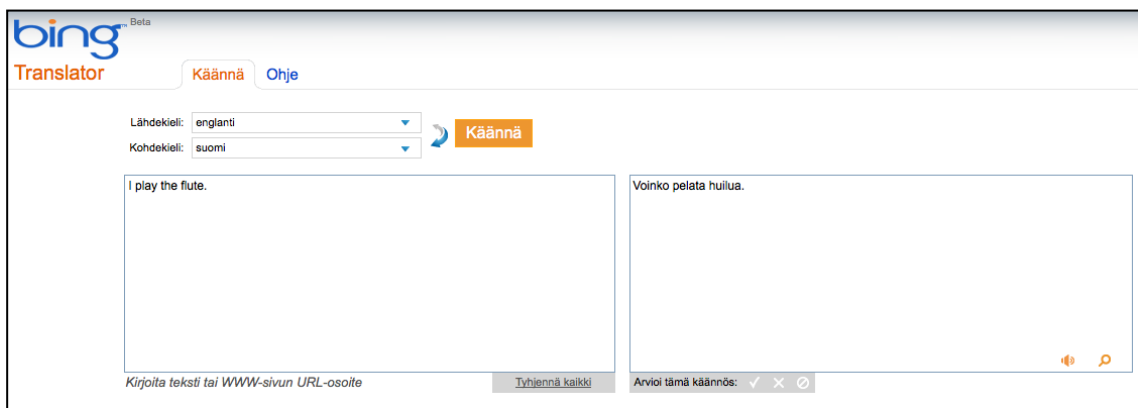
Kuvassa 7 on Google Translate -konekäännösjärjestelmän käyttöliittymä. Esimerkivirkkeenä on jälleen ”*I play the flute.*”.



Google Translate -konekäännösjärjestelmässä on lisäominaisuutena mahdollisuus tarkastella ja valita käännetylle sanalle vaihtoehtoisia käännöksiä. Lisäksi käännöksen voi arvioida asteikolla hyödyllinen, hyödytön, loukkaava. (Google Translaten verkkosivut 2011.)

### 6.2.5 Bing™ Translator Beta

Bing™ Translator Beta on Google Translaten tavoin verkossa toimiva ilmainen tilastollinen konekäännösjärjestelmä (Toral ym. 2011, 17). Järjestelmän taustalla on Microsoft Translator® -käännöspalvelu. Koska kyseessä on järjestelmän kokeiluversio, järjestelmä on vielä kehittä- ja testausvaiheessa. (Bing Translatorin verkkosivut 2011.) Järjestelmä on monikielinen ja kaksisuuntainen. Tällä hetkellä järjestelmä sisältää 37 kieltä (Bing Translatorin verkkosivut 2011). Kuvassa 8 on Bing™ Translator Beta -konekäännösjärjestelmän käyttöliittymä esimerkkivirkkeellä ”I play the flute.”.



Kuva 8. Bing™ Translator Beta -konekäännösjärjestelmän käyttöliittymä

Bing™ Translator Beta -konekäännösjärjestelmässä on Google Translaten tavoin mahdollisuus arvioida ehdotettu käännös. Lisäksi käännöksellä voi suoraan tehdä haun Bing-hakukoneella. (Bing Translatorin verkkosivut 2011.)

### 6.2.6 Yhteenvetoa testatuista konekäännösjärjestelmistä

Seuraavaksi kootaan yhteen testattujen konekäännösjärjestelmien ominaisuuksia. Kuten edellä olevista konekäännösjärjestelmien käyttöliittymien kuvista voidaan havaita, kaikissa muissa käyttöliittymissä paitsi TeemaPointissa on kaksi tekstikenttää: toiseen kirjoitetaan tai liitetään käännettävä teksti ja toiseen tulostuu käännetty teksti. TeemaPointissa on tekstikenttä ainoastaan käännettävälle tekstille, ja käännös tulostuu tämän teks-

tikentän alapuolelle. Google Translate eroaa muista järjestelmistä siinä, että kyseisessä järjestelmässä voidaan valita pikakääntämistilan, jossa käännetty teksti tulostuu sitä mukaa kuin käännettävää tekstiä kirjoitetaan. Muiden järjestelmien käyttöliittymissä on erikseen klikattava kääntämispainiketta, kun käännös halutaan tulostaa tekstikenttään.

Taulukkoon 1 on koottu testattujen konekäännösjärjestelmien ominaisuuksia. Taulukosta näkyy, onko konekäännösjärjestelmä yksi- vai kaksisuuntainen ja kaksi- vai monikielinen. Siitä ilmenee myös, onko konekäännösjärjestelmä kääntämismenetelmältään tilastollinen vai sääntöperustainen. Lisäksi taulukosta näkyy, onko konekäännösjärjestelmä suomalainen ja onko sen käyttäminen ilmaista. Tässä konekäännösjärjestelmää pidetään maksullisena, jos vain testitunnukset ovat maksuttomat tai järjestelmää voi käyttää ilmaiseksi vain tietyin rajoittein.

Taulukko 1. Testattujen konekäännösjärjestelmien ominaisuuksia

	TP	Sunda	SDL	Google	Bing
Käännösten suunnat	yksi-suuntainen	kaksi-suuntainen	kaksi-suuntainen	kaksi-suuntainen	kaksi-suuntainen
Kielten määrä	kaksi-kielinen	kaksi-kielinen	monikielinen	monikielinen	monikielinen
Kääntämismenetelmä	sääntöperustainen	sääntöperustainen	sääntöperustainen	tilastollinen	tilastollinen
Suomalainen vai ulkomaalainen yhtiö	suomalainen	suomalainen	ulkomaalainen	ulkomaalainen	ulkomaalainen
Maksullisuus	maksullinen	maksullinen	ilmainen	ilmainen	ilmainen

## 6.3 Aineiston kerääminen

Aineiston kerääminen oli kaksivaiheinen. Ensimmäisessä vaiheessa luotiin käännettävät englanninkieliset virkkeet. Toisessa vaiheessa virkkeet käännettiin testattavilla konekäännösjärjestelmillä. Seuraavaksi esitellään kumpaakin näistä aineistonkeruuvaiheista.

### 6.3.1 Käännettävien virkkeiden luominen

Kieltenvälisen tiedonhaun tutkimuksissa käytetään usein valmiita testikokoelmia (ks. esim. Kettunen 2009, 104; Wu ym. 2008, 74) tai muuten valmiita tekstejä, esimerkiksi lakitekstejä (Kit & Wong 2008, 299) tai uutisaiheisia tekstejä (Toral ym. 2011, 15). Koska tämän tutkimuksen tutkimuskysymykset asettivat tiettyjä rajoituksia ja ehtoja käännettäville virkkeille, virkkeet luotiin itse.

Tutkimusongelman perusteella jokaisessa virkkeessä piti esiintyä monitulkintainen sana. Kolmannen tutkimuskysymyksen perusteella saman monitulkintaisen sanan piti puolestaan esiintyä erilaajuisissa asiayhteyksissä. Monissa kieltenvälisen tiedonhaun tutkimuksissa on vertailtu lyhyitä ja pitkiä kyselyitä (ks. esim. Kettunen 2009, 104; Wu ym. 2008, 74). Tässä tutkimuksessa käytettiin kolmea erilaajuista asiayhteyttä. Yhtenä asiayhteytenä oli se, ettei asiayhteyttä ollut, eli monitulkintaiset sanat käännettiin yksittäisinä sanoina. Kaksi muuta erilaajuista asiayhteyttä oli virkkeen lyhyt ja pitkä versio, jotka muodostivat virkeparin. Käännettäviä virkkeitä rajoitti se, että jo virkeparin lyhyen version piti olla kokonainen virke. Näin ollen esimerkiksi testikokoelmien hakuaiheiden otsikoita ei sellaisenaan olisi voitu käyttää. Koska käännösyksikkönä oli virke, kontekstin ikkunalähestymistavan perusteella kyseessä oli paikallinen konteksti (ks. Turdakov 2010, 312).

Virkkeiden luominen aloitettiin monitulkintaisten sanojen etsimisellä. Alun perin tarkasteltaviksi monitulkintaisiksi sanoiksi oli tarkoitus ottaa Senseval-3:n englanninkielisten monitulkintaisten sanojen sanastotestissä käytetyt monitulkintaiset sanat (Senseval-3 aineisto 2004). Kyseisiä sanoja oli 57. Sanojen merkitykset oli rajoitettu vain tiettyyn sanaluokkaan eikä läheskään kaikille sanoille ollut riittävän yksitulkintaisia merkityksiä edes ihmisen käännettäväksi. Tämän takia Senseval-3:n sanoista valittiin vain 25. Koska 25 monitulkintaista sanaa ei koettu riittävän kattavaksi, monitulkintaisia sanoja etsittiin lisää niin, että lopulta tarkasteltavia sanoja oli 50. Sanat on listattu liitteeseen 1.

Monitulkintaisia sanoja valittiin seuraavien periaatteiden mukaan: Sanan eri merkitysten piti olla riittävän yksikäsitteisiä niin, ettei ihmisen ollut vaikea erottaa näitä eri merkityksiä toisistaan. Sanan ympärille oli pystyttävä luomaan yksinkertaisesti ja ymmärrettävästi sekä virkkeen lyhyt että pitkä versio niin, että pidemmässä versiossa monitulkintainen sana saisi yksikäsitteistämistä varten lisävihjeitä. Oli myös suotavaa, jos monitulkintaisen sanan merkitykset edustivat eri sanaluokkia. Monitulkintaisten sanojen valinnassa ja niiden merkitysten määrittelyssä käytettiin apuna WordNetin hakutoimintoa (<http://wordnetweb.princeton.edu/perl/webwn>).

Monitulkintaisten sanojen valitsemisen jälkeen luotiin englanninkieliset virkkeet. Apuna käytettiin Senseval-3:n tiedostossa (Senseval-3 aineisto 2004) olevia WordNetin määritelmiä sekä edellä mainitulla WordNetin hakutoiminnolla saatavia sanojen merkitysten määritelmiä ja esimerkkejä. Ideoita virkkeisiin haettiin myös muualta internetistä.

Jokaisesta virkkeestä luotiin siis kaksi versiota: lyhyt ja pitkä. Virkkeet muodostettiin niin, että pitkä versio alkoi aina kyseisen virkeparin lyhyellä versiolla. Tällä haluttiin taata, että kaikkia virkkeitä ja näin ollen tarkasteltavan monitulkitaisen sanan asiayhteyttä laajennettiin samalla periaatteella. Esimerkiksi Kettusen (2009, 112) tutkimuksessa erilaajuisia asiayhteyksiä vertailtaessa käytettiin käännettävien kyselyiden lyhyessä ja pitkässä versiossa täysin samaa alkua. Kyseisessä tutkimuksessa lyhyenä versiona oli pelkkä hakuaiheen otsikko ja pitkänä versiona otsikko ja tämän jatkona ollut kuvailuosan alku (Kettunen 2009, 112). Tämän tutkimuksen virkepareista osa luotiin niin, että ensin löydettiin idea lyhyeen versioon, jota itse jatkettiin. Osa virkepareista puolestaan luotiin niin, että ensin löydettiin idea pitkään versioon, josta loppu jätettiin pois lyhyen version saamiseksi. Käännettäviä virkkeitä ei ollut mahdollista jatkaa täysin samalla sanamäärällä, koska virkkeiden piti olla ymmärrettäviä, oikeita virkkeitä. Luonnollisesti myös monitulkitaisen sanan paikka virkkeessä vaihteli virkkeestä riippuen.

Virkkeiden pitkiin versioihin pyrittiin tuomaan lisää asiayhteyttä monitulkitaisen sanan merkitykseen liittyvillä sanoilla, esimerkiksi ”*I play the flute.*” ja ”*I play the flute in the orchestra at the concert hall.*” verrattuna virkepariin ”*I play football.*” ja ”*I play football in the football team and we have a game at the stadium every week.*”. Joissain tapauksissa asiayhteyttä laajennettiin selittämällä monitulkitaista sanaa, esimerkkinä virkepari ”*Results are expressed as a mean.*” ja ”*Results are expressed as a mean, which is the average of a set of values and which is calculated by summing values together and dividing by the number of terms.*”. Virkettä voitiin laajentaa myös käyttämällä monitulkitaisen sanan synonyymeja kuten virkeparissa ”*He could feel the beat of her heart*” ja ”*He could feel the beat of her heart and her heart rate was fast.*”.

Virkkeet muodostettiin niin, että tarkasteltava monitulkitainen sana oli aina perusmuodossa. Esimerkiksi sanasta *play* ei virkkeissä käytetty muotoja *playing*, *played* tai *plays*. Perusmuodossa olevasta sanasta konekäännösjärjestelmä ei suoraan voinut saada tietoa käännettävän monitulkitaisen sanan sanaluokasta, jolloin sanan muoto ei voinut auttaa sanan merkityksen yksikäsitteistämässä. Tällä kontrolloitiin sitä, että kaikki monitulkitaiset sanat olivat ainakin tältä osin samanarvoisessa asemassa yksikäsitteistämistehävässä.

Jotta voitiin vertailla virkkeissä esiintyvien yksikäsitteisten ja monitulkitaisien sanojen käänntösten laatueroja, jokaisessa virkkeessä piti esiintyä myös yksikäsitteinen sana.

Tiedonhaun näkökulmasta tarkasteltavien yksikäsitteisten sanojen oli järkevintä olla substantiiveja. Käännösten laadun selvittämisen näkökulmasta tarkasteltavien yksikäsitteisten sanojen piti puolestaan olla joka virkeparissa eri sanoja. Jos luodussa virkeparissa oli useampi kuin yksi yksikäsitteinen substantiivi, tarkasteltava yksikäsitteinen sana valittiin numeroimalla vaihtoehdot ja arpomalla Excelin satunnaislukugeneraattorilla yksi näistä yksikäsitteisistä sanoja vastaavista luvuista. Tarkasteltavia yksikäsitteisiä sanoja olivat esimerkiksi *poem*, *politicians*, *armchair*, *juice*, *king*, *bookkeeper*, *flute*, *employee*, *mathematics*, *football*, *summer* ja *teacher*.

Kun käännettävät virkkeet oli luotu, kaksi suomalaista yliopistossa englantia opiskellutta henkilöä oikoluki ne. Toinen oikolukijoista opettaa Suomessa englantia peruskoulun luokille 3-6 mutta asuu tällä hetkellä tilapäisesti Yhdysvalloissa. Toinen oikolukijoista käyttää muuten työssään englantia. Oikolukemisessa oli kaksi tavoitetta: tarkastaa, tarkoittivatko monitulkintaiset sanat suomeksi sitä, mitä niiden oli merkitty tarkoittavan ja olivatko englanninkieliset virkkeet kielellisesti oikein. Tämän jälkeen virkkeet tarkasti Minneapolisin yliopistossa opettajana ja assistenttina toimiva syntyperäinen yhdysvaltalainen. Kyseinen henkilö on opettanut Minneapolisin yliopistossa suomea, ruotsia sekä tanskaa ja on itse opiskellut samaisessa yliopistossa kaikkia pohjoismaisia kieliä sekä saksaa. Virkkeiden tarkastamisen tarkoituksena oli varmistaa, olivatko virkkeet sujuvaa ja virheetöntä englantia ja voitaisiinko tarkasteltavia monitulkintaisia sanoja luonnollisesti käyttää ilmaistuissa yhteyksissä. Korjausten perusteella muutamaa virkeparia muokattiin tai luotiin kokonaan uusi. Muutama virkepari jätettiin pois, koska niissä monitulkintaisen sanan käyttäminen vaikutti keinotekoiselta; oikeasti sama asia ilmaistaisiin jollain toisella sanalla. Lisäksi virkkeistä korjattiin kielioppivirheitä, muun muassa lisättiin artikkeleita. Korjausten ja muutosten jälkeen virkeitä ei enää oikolukenukukaan ulkopuolinen tarkastaja.

Lopullisia käännettäviä virkepareja oli 195, eli virkeitä oli kaiken kaikkiaan 390. Esimerkiksi Kettusen (2009, 106) konekäännösjärjestelmiä ja eripituisia kyselyitä vertailevassa tutkimuksessa käytettiin 56 kyselyä. Saman tutkimuksen erillisessä lisätarkastelussa käytettiin 60 paria eripituisia kyselyitä (Kettunen 2009, 112). Wun ja muiden (2008, 74) Google Translate -konekäännösjärjestelmää ja sanakirjaperustaista kääntämistä vertailevassa evaluointitutkimuksessa käytettiin puolestaan 44 englanninkielistä kyselyä. Näin ollen tämän tutkimuksen aineiston pitäisi laajuudeltaan olla riittävän kat-

tava. Toisaalta aineisto on vähäinen verrattuna esimerkiksi Toralin ja muiden (2011, 15) tutkimuksen aineistoon, jossa käännettäviä virkkeitä oli tuhat.

Yhdestä monitulkintaisesta sanasta oli 2-7 virkeparia niin, että kunkin sanan yhdestä merkityksestä oli vain yksi virkepari. Jos tietty konekäännösjärjestelmä tarjoaisi tietylle englanninkieliselle monitulkintaiselle sanalle aina saman käännöksen ja jos tästä tietyistä merkityksestä olisi käännettävänä enemmän kuin yksi virkepari, konekäännösjärjestelmä vaikuttaisi toimivan paremmin, kuin se oikeasti toimisikaan. Tästä samasta syystä tarkasteltavat yksikäsitteiset sanat olivat jokaisessa virkeparissa eri sanoja. Tarkasteltavaa monitulkintaista sanaa kohti oli keskimäärin 4 virkeparia. Virkeparien lyhyissä versioissa oli keskimäärin 6,3 ja pitkissä 14,1 sanaa. Sanoiksi laskettiin kaikki sanat, eli esimerkiksi artikkeleita tai prepositioita ei jätetty huomioimatta. Esimerkiksi Wun ja muiden (2008, 74) tutkimuksessa lyhyet kyselyt olivat pituudeltaan keskimäärin 4, keskipitkät 27 ja pitkät 127 sanaa. Kettusen (2009, 104) tutkimuksessa käännettävien hakuaiheiden otsikoiden keskimääräinen pituus oli puolestaan 3,7 sanaa ja otsikon ja kuvailuosan yhdistelmien 18,8 sanaa (Kettunen 2009, 104). Näihin lukumääriin verrattuna tämän tutkimuksen lyhyen ja pitkän version sanamäärien pituuserot olivat vähäisemmät. Toisaalta lyhyiden virkkeiden pituutta hakuaiheiden otsikoihin verrattuna lisäsi se, että tässä tutkimuksessa virkkeen lyhyenkin version piti olla kokonainen virke.

### **6.3.2 Sanojen ja virkkeiden kääntäminen konekäännösjärjestelmillä**

Varsinainen aineisto kerättiin kääntämällä englanninkieliset sanat ja virkkeet konekäännösjärjestelmillä. Sanat ja virkkeet käännettiin niin, että alkuperäisten englanninkielisten sanojen tai virkkeiden taulukosta kopioitiin sana tai virke kerrallaan kaikkiin testattuihin konekäännösjärjestelmiin ja kunkin konekäännösjärjestelmän tuottama käännös kopioitiin käännöstuloksena samaisiin taulukoihin. Suurin osa käännettävistä virkkeistä päättyi pisteeseen ja muutama kysymysmerkkiin. Esimerkiksi Kettusen (2009, 104) tutkimuksessa kaikki käännettävät kyselyt päättyivät pisteeseen, kun ne syötettiin konekäännösjärjestelmään.

## **6.4 Aineiston analysointi**

Tutkimuskysymysten perusteella oltiin kiinnostuneita vain siitä, kuinka virheettömästi eri konekäännösjärjestelmät käänsovät tarkasteltavat yksikäsitteiset ja monitulkintaiset

sanat. Näin ollen ei ollut mielekästä eikä tarkoituksenmukaista käyttää esimerkiksi suomenkin kieleen soveltuvaan BLEU-mittaria, jossa tarkastellaan koko virkkeen käännöksen oikeellisuutta. BLEU ottaa huomioon virkkeen kaikkien sanojen käännökset, sanojen taivutusmuodot sekä konekäännöksen sanojen lukumäärän verrattuna referenssikäännöksen sanojen lukumäärään (Papineni ym. 2001, 4-5). Käännöstuloksia analysoitaessa riitti vertailla tarkasteltavien yksikäsitteisten ja monitulkintaisten sanojen perusmuotoja niiden oikeiden suomenkielisten käännösten perusmuotoihin.

Tässä tutkimuksessa evaluointimittarina käytettiin tarkkuutta, joka huomioi ainoastaan oikeiden käännösten osuuden kaikista käännöksistä (Melamed ym. 2003, 61). Kuten aiemmin mainittiin, Melamedin ja muiden (2003, 61) mukaan tarkkuutta voidaan käyttää konekäännöksen laadun evaluointimittarina. Myös Jurafskyn ja Martinin (2000, 639) mukaan sanan merkityksen yksikäsitteistämistä voidaan evaluoida tarkkuudella, jolloin lasketaan oikein yksikäsitteistettyjen sanojen osuus kaikista evaluoitavista sanoista. Koska jokaisessa virkkeessä tarkasteltiin vain yhden yksikäsitteisen ja yhden monitulkintaisen sanan käännöksiä, referenssikäännöksiä ei tehty koko virkkeistä vaan ainoastaan tarkasteltavista sanoista. Melamed ja muut (2003, 61) mainitsevat, että tarkkuuden laskemisessa usein on ongelmana löytää sopiva tapaa laskea oikeiden käännösten lukumäärä. Koska tässä tutkimuksessa jokaisesta virkkeestä pisteytettiin vain kaksi erillistä sanaa, oikeiden käännösten lukumäärän laskeminen ei ollut ongelmallista.

Aineiston analysointi koostui kolmesta vaiheesta. Ensimmäisessä vaiheessa pisteytettiin käännökset. Toisessa vaiheessa laskettiin käännöksille annettujen pisteiden perusteella tutkimuskysymysten vaatimat erilaiset tarkkuusarvot ja prosenttiosuudet. Kolmannessa vaiheessa suoritettiin tilastolliset merkitsevyytestit. Seuraavaksi raportoidaan kaikki nämä analysointivaiheet.

#### **6.4.1 Käännösten pisteyttäminen**

Käännösten pisteyttäminen koostui kolmesta vaiheesta. Ensimmäisenä pisteytettiin tarkasteltavien yksikäsitteisten ja monitulkintaisten sanojen käännösten oikeellisuus. Toisena merkittiin, olivatko kunkin virkeparin lyhyessä ja pitkässä virkkeessä esiintyvän monitulkintaisen sanan käännöksen saamat pisteet keskenään samat vai oliko pisteissä eroja. Kolmantena pisteytettiin virheelliset käännökset niiden virhetyyppien perusteella. Seuraavaksi esitellään pisteyttämisen periaatteita.

Perusidea virkeyhteydessä esiintyneiden monitulkintaisten sanojen käännosten pisteyttämisessä oli seuraava: Jos tarkasteltava sana oli käännetty oikein, käänнос sai yhden pisteen. Referenssikäännoksessä saattoi olla useita oikeita vaihtoehtoja monitulkintaiselle sanalle. Konekäänнос tulkittiin oikeaksi, jos se oli mikä tahansa näistä oikeista käännosvaihtoehdoista. Esimerkiksi sanan *argument* käännoksiksi hyväksyttiin tietyssä virkeparissa *kiista* tai *väittely*, sanan *direct* käännoksiksi *suunnata* tai *kohdistaa* ja sanan *lose* käännoksiksi *hävittää* tai *kadottaa*. Jos tarkasteltavan sanan käänнос kuului referenssikäännoksen kanssa samaan johdosperheeseen, käännoskelle annettiin 0,5 pistettä. Esimerkiksi jos sana *watch* olisi pitänyt virkeyhteydessä kääntää sanaksi *vartija*, käännookset *vartioida* ja *vartiointi* saivat 0,5 pistettä. Virkeyhteydessä esiintyvän monitulkintaisen sanan käänнос sai nolla pistettä, jos tarkasteltava sana oli käännetty väärin, sana oli jäänyt englanninkieliseksi, sana puuttui kokonaan konekäännoksesta tai sanan käännoستä ei voinut tunnistaa konekäännoksesta. Tarkasteltava yksikäsitteinen sana sai puolestaan oikeasta käännoksesta yhden pisteen ja väärästä käännoksesta nolla pistettä.

Yksittäin käännettyjen monitulkintaisten sanojen käännookset pisteytettiin erikseen jokaisen virkeparin määräämässä merkityksessä. Näin saatiin tieto siitä, oliko yksittäin käännetty sana käännetty oikein tietyn asiayhteyden tarkoittamalla tavalla. Jos siis esimerkiksi konekäännosjärjestelmä käänsi yksittäisen sanan *play* sanaksi *pelata*, annettiin yksittäisen sanan käännoskelle yksi piste sen virkeparin merkityksessä, jossa sana *play* tarkoitti *pelata*. Yksittäisen sanan käänнос sai puolestaan 0,5 pistettä sen virkeparin merkityksessä, jossa sanan *play* oikea käänнос kuului sanan *pelata* kanssa samaan johdosperheeseen. Muiden sanan *play* virkeparien merkityksessä käänнос *pelata* tulkittiin vääräksi, ja käänнос sai nolla pistettä.

Konekäännosjärjestelmät poikkesivat toisistaan sen suhteen, kuinka monta käännosvaihtoehtoa ne ehdottivat yksittäin käännetylle monitulkintaiselle sanalle. TeemaPoint, SDL Free Translation sekä Bing<sup>TM</sup> Translator Beta ehdottivat yksittäiselle sanalle vain yhden käännoksen, mikä on tyypillistä konekääntämisessä (ks. McNamee & Mayfield 2002, 159; Wu ym. 2008, 74). Google Translate sen sijaan esitti yhden ensisijaisen käännoksen mutta listasi myös vaihtoehtoisia käännoksia, joita saattoi olla jopa 30. Vain ensisijainen käänнос pisteytettiin. Myös Sunda esitti lähes kaikille sanoille useamman kuin yhden käännosvaihtoehdon. Nämä vaihtoehdot eivät kuitenkaan olleet minkäänlaisessa paremmuusjärjestyksessä, joten käännoksista ei voitu valita tarkasteltavaksi vain yhtä. Sundan käännookset pisteytettiinkin tarkasteltavan monitulkintaisen sanan jokaisen



virkeyhteyden merkityksessä kaavalla  $m/n$ , jossa  $m$  tarkoittaa yksin käännetyn sanan tietyn virkeyhteyden merkityksessä oikein olevien käänntösten lukumäärää ja  $n$  järjestelmän antamien käänntösten kokonaismäärää kyseiselle yksin käännetylle sanalle. Jos esimerkiksi yksin käännetyn monitulkitaisen sanan käänntöksistä tietyssä virkeyhteydessä yksi käänntöksistä olisi täysin oikein, kaksi käänntöstä kuuluisi oikean käänntöksen kanssa samaan johdosperheeseen ja tarkasteltavalle sanalle olisi kaiken kaikkiaan kaksitoista käänntövaihtoehtoa, Sundan saamat pisteet kyseisen yksittäin käännetyn sanan käänntöksille kyseisen virkeyhteyden merkityksessä laskettaisiin  $(1 + 0,5 + 0,5) / 12 \approx 0,167$ .

Taulukko 2. Esimerkki yksittäin käännettyjen monitulkitaisien sanojen käänntöksistä ja pisteyttämisestä

Käännettävä sana	Sunda	Google	
wash	1. pesu, pyykki, kohina, loiske, kuohut, ilmavirta, pyörre, laveeraus ( <i>subst</i> ) 2. pestä ( <i>verbi</i> )	pestä	
Käännettävä virke	Merkitys virkkeessä	Pisteet: Sunda	Pisteet: Google
1. I <b>wash</b> the stain.	pestä	$(0,5 + 1) / 9 \approx 0,167$	1
2. My coat needs a dry <b>wash</b> .	pesu	$(1 + 0,5) / 9 \approx 0,167$	0,5
3. I heard the <b>wash</b> of the waves.	kohina, loiske	$2 / 9 \approx 0,222$	0
4. A pail may <b>wash</b> up.	huuhtoutua	$0 / 9 \approx 0$	0
	Keskiarvo	0,139	0,375

Taulukossa 2 on esimerkit Sundan ja Google Translaten yksittäin kääntämisen sanan *wash* pisteyttämisestä. Tarkastellaan ensin Sundan saamat pisteet. Ensimmäisessä virkeyhteydessä *wash*-sanana oikea käänntös oli *pestä*. Sundan *wash*-sanana käänntöksissä oli tämä oikea käänntös sekä oikean käänntöksen kanssa samaan johdosperheeseen kuuluva sana *pesu*. Näin ollen yksittäin käännetty sana sai kyseisen virkeyhteyden merkityksessä yhden pisteen käänntöksestä *pestä* ja 0,5 pistettä käänntöksestä *pesu*. Näiden pisteiden summa jaettiin yhdeksällä, koska Sunda ehdotti kaiken kaikkiaan yhdeksän käänntövaihtoehtoa sanalle *wash*. Toisessa virkeyhteydessä Sundan käänntös *pesu* sai yhden pisteen ja käänntös *pestä* 0,5 pistettä. Sundan kolmanteen virkeyhteyteen liittyvät pisteet tulivat siitä, kun yksittäin käännetyn *wash*-sanana käänntöksissä oli kaksi täysin oikeaa käänntöstä kyseiseen virkeyhteyteen, *kohina* ja *loiske*. Koska Sundan käänntöksissä ei ollut viimeisen virkeyhteyden merkitystä *huuhtoutua*, yksittäin käännetty sana sai viimeisen virkeyhteyden merkityksessä nolla pistettä. Kun yksittäin käännetyn sanana pisteet oli annettu sanana jokaisen virkeyhteyden merkityksessä, pisteistä laskettiin sanakohtainen keskiarvo.

Kuten taulukosta 2 ilmenee, Google Translate käänsi yksittäisen sanan *wash* sanaksi *pestä*. Ensimmäisen virkeyhteyden merkityksessä käänös *pestä* sai yhden pisteen ollessaan oikea käänös. Toisen virkeyhteyden merkityksessä käänös *pestä* sai puolestaan 0,5 pistettä kuuluessaan oikean käänöksen kanssa samaan johdosperheeseen. Kahdessa viimeisessä virkeyhteydessä käänös sai nolla pistettä, koska kyseisessä virkeyhteydessä sanan *wash* oikea merkitys oli jotain muuta kuin *pestä* tai sen johdosperheeseen kuuluva sana. Kun Google Translaten käänös *pestä* oli pisteytetty kaikkien virkeyhteyksien merkityksissä, pisteistä laskettiin sanakohtainen keskiarvo. Sanakohtaisen keskiarvon laskeminen teki Sundan ja muiden järjestelmien pisteistä vertailukelpoisemmat kuin summan laskemisella, koska kunkin tarkasteltavan monitulkitaisen sanan merkitys oli keskiarvon ansiosta pisteissä edustettuna yhtäläisellä painolla.

Taulukko 3. Esimerkkejä monitulkitaisien sanojen yksiselitteisesti pisteytettävistä käänöksistä

soittaa ( <i>huilu</i> )	I <b>play</b> the <i>flute</i> .	Yks	Mon
TP	Minä <b>soitan</b> <i>huilua</i> .	1	1
Sunda	Minä <b>soitan</b> <i>huilua</i> .	1	1
SDL	<b>Leikin</b> <i>huilua</i> .	1	0
Google	<b>Pelaan</b> <i>huilua</i> .	1	0
Bing	Voinko <b>pelata</b> <i>huilua</i> .	1	0
vapauttaa ( <i>lehmä</i> )	They <b>free</b> the <i>cows</i> twice a day.	Yks	Mon
TP	He <b>vapauttavat</b> <i>lehmät</i> kahdesti päivää.	1	1
Sunda	Ne <b>vapauttavat</b> <i>lehmät</i> kahdesti päivässä.	1	1
SDL	Ne <b>vapauttavat</b> <i>lehmät</i> kaksi kertaa päivässä.	1	1
Google	He <b>vapaa</b> <i>lehmät</i> kahdesti päivässä.	1	0,5
Bing	Ne <b>vapaasti</b> <i>lehmä</i> kahdesti päivässä.	1	0,5

Suurin osa käänöksistä oli yksiselitteisiä pisteyttäviä. Taulukossa 3 on esimerkkejä yksiselitteisesti pisteytettävistä käänöksistä. Esimerkkivirkkeet ovat virkeparien lyhyitä virkkeitä. Taulukossa harmaalla pohjalla on käännettävä virke sekä tarkasteltavien sanojen referenssikäännökset niin, että ensin mainitaan tarkasteltavan monitulkitaisen sanan oikea käänös ja tämän jälkeen suluissa tarkasteltavan yksikäsitteisen sanan oikea käänös. Englanninkielisen virkkeen alapuolella seuraavalla viidellä rivillä on kunkin konekäännösjärjestelmän ehdottama käänös. Tarkasteltava monitulkitainen sana on alleviivattu ja tummennettu sekä käännettävästä virkkeestä että käänöksistä. Tarkasteltava yksikäsitteinen sana on puolestaan alleviivattu ja kursivoitu. Oikeanreunimmaisissa sarakkeissa näkyvät yksikäsitteisten ja monitulkitaisien sanojen käänöksille annetut

pisteet. Taulukon toisesta esimerkistä ilmenee myös tilanne, jossa käänös sai 0,5 pistettä ollessaan oikean käänöksen kanssa samaan johdosperheeseen kuuluva sana.

Epäselvissä ja tulkinnanvaraisissa tapauksissa pisteyttämisen tarkastivat englanninkieliset käännettävät virkkeetkin oikolukeneet kaksi suomalaista henkilöä. Ensin toinen näistä henkilöistä tarkasti tulkinnanvaraisten käännösten pisteytykset. Jos kyseinen henkilö ja tutkimuksen tekijä olivat pisteytyksestä eri mieltä, ratkaiseva päätös kysyttiin toiselta suomalaiselta englannin kielen osaajalta. Näin varmistettiin, ettei pisteytys ollut täysin subjektiivista.

Taulukko 4. Esimerkkejä monitulkintaisten sanojen tulkinnanvaraisten käännösten pisteyttämisestä

syke, lyönti ( <i>sydän</i> )	He could feel the <b>beat</b> of her <i>heart</i> .	Yks	Mon
TP	Hän voisi tuntea hänen <i>sydämen</i> [ <b>beat</b> ].	1	0
Sunda	Hän saattoi tuntea <i>sydämensä</i> <b>lyöntiä</b> .	1	1
SDL	Hän voisi tuntea hänen <i>sydämensä</i> <b>lyönniltä</b> .	1	1
Google	Hän pystyi tuntemaan <b>voittaa</b> hänen <i>sydämensä</i> .	1	0
Bing	Hän voisi tuntea hänen <i>sydämen</i> <b>syke</b> .	1	1
voittaa, lyödä, hakata, ( <i>Ruotsi</i> )	Finland will <b>beat</b> <i>Sweden</i> .	Yks	Mon
TP	Suomi <b>lyö</b> <i>Ruotsin</i> .	1	1
Sunda	Suomi <b>lyö</b> <i>Ruotsin</i> .	1	1
SDL	Suomi <b>hakkaa</b> <i>Ruotsin</i> .	1	1
Google	Suomi <b>voittaa</b> <i>Ruotsin</i> .	1	1
Bing	Suomi <b>voittaa</b> <i>Ruotsi</i> .	1	1

Taulukossa 4 on esimerkkejä tulkinnanvaraisista käännöksistä ja niiden pisteyttämisestä. Tulkinnanvaraisia käännöksiä olivat esimerkiksi sanojen kuvainnolliset merkitykset eli monitulkintaisuuden tyypiltään semanttisesti monitulkintaiset sanat (ks. Turdakov 2010, 310). Esimerkiksi taulukon ensimmäisessä esimerkissä oikeaksi käännökseksi oli ajateltu *syke* mutta pisteytyksessä myös merkitys *lyönti* hyväksyttiin. Toisessa esimerkissä sanan *beat* oikeaksi käännökseksi oli alun perin ajateltu pelkästään *voittaa* mutta suomeksi saatetaan käyttää myös kuvainnollisia merkityksiä *lyödä* tai jopa *hakata*. Näin ollen virkkeiden luomisvaiheessa ei ollut tullut mieleen kaikkia mahdollisia oikeiksi tulkittavia käännöksiä, mutta pisteytysvaiheessa oikeiksi käännöksiksi hyväksyttiin perustellusti muitakin kuin alun perin oikeiksi suunniteltuja käännöksiä.

Pisteyttämisen vaiheessa käänöksen pisteiden lisäksi tulostaulukkoon merkittiin, olivatko monitulkintaisen sanan käänöksen pisteet samat vai toisistaan eroavat virkeparin lyhy-

essä ja pitkässä virkkeessä. Jos pisteet olivat paremmat pitkässä kuin lyhyessä virkkeessä, tulostaulukon viimeiseen sarakkeeseen merkittiin plusmerkki (+). Jos pisteet olivat paremmat virkeparin lyhyessä kuin pitkässä virkkeessä, sarakkeeseen merkittiin miinusmerkki (-). Jos taas käännösten pisteissä ei ollut eroa lyhyen ja pitkän virkeyhteyden välillä, sarakkeeseen merkittiin nolla (0). Plus- ja miinusmerkkejä käytettiin, koska monissa tutkimuksissa on mainittu, että konekäännöksen laatu on parempaa pitkässä kuin lyhyessä asiayhteydessä (ks. esim. Oard & Hackett 1998, 1; Wu ym. 2008, 75). Taulukossa 5 on esimerkkejä käännösten piste-eroista lyhyessä ja pitkässä virkkeessä sekä erojen merkitsemisestä. Sarakkeet *Y* ja *M* tarkoittavat yksikäsitteisten ja monitulkintaisten sanojen käännöksille annettuja pisteitä ja sarake *E* lyhyessä ja pitkässä virkkeessä esiintyvän monitulkintaisen sanan käännöksen pisteiden eroa edellä kuvatulla tavalla. Taulukko toimii samalla mallina käytetystä tulostaulukosta.

Taulukko 5. Esimerkkejä erilaajuisissa virkkeissä esiintyneiden yksikäsitteisten ja monitulkintaisten sanojen käännösten pisteyttämisestä

ranta ( <i>kanootti</i> )	They pulled the <i>canoe</i> up on the <b>bank</b> .	Y	M	They pulled the <i>canoe</i> up on the <b>bank</b> of the river, sat on the bank and watched the currents.	Y	M	E
TP	He vetivät <i>kanootin</i> [up] <b>pankilla</b> .	1	0	He [pulled] <i>kanootilla</i> ylös joen <b>ranta</b> , istuivat [bank] ja katsoivat virrat.	1	1	+
Sunda	Ne vetivät <i>kanootin</i> ylös <b>pankilla</b> .	1	0	Ne vetivät <i>kanootin</i> ylös joen <b>rannalla</b> , istuivat pankilla ja katsoivat virtoja.	1	1	+
SDL	Ne vetivät <i>kanootin</i> ylös <b>pankissa</b> .	1	0	Ne vetivät <i>kanootin</i> ylös joen <b>pankissa</b> , joka on istunut pankissa ja katsoi virtoja.	1	0	0
Google	He vetivät <i>kanootti</i> ylös <b>pankki</b> .	1	0	He vetivät <i>kanootti</i> ylös joen, istui pankki ja katseli virtaukset.	1	0	0
Bing	Ne vedetään <i>kanootti</i> <b>pankin</b> .	1	0	Ne veti <i>kanootti</i> joen <b>rannalla</b> , istui pankin ja merkinantovirrat seurattavat.	1	1	+
alaikäinen ( <i>olut</i> )	We don't sell <i>beer</i> to a <b>minor</b> .	Y	M	We don't sell <i>beer</i> to a <b>minor</b> , only to adults.	Y	M	E
TP	Me emme myy <i>olutta</i> [ <b>minor</b> ].	1	0	Me emme myy <i>olutta</i> [ <b>minor</b> ], vain aikuisille.	1	0	0
Sunda	Me emme myy <i>olutta</i> <b>alaikäiselle</b> .	1	1	Me emme myy <i>olutta</i> <b>alaikäiselle</b> vain aikuisille.	1	1	0
SDL	Emme myy <i>olutta</i> <b>vähäinen</b> .	1	0	Emme myy <i>olutta</i> <b>vähäinen</b> vain aikuisille.	1	0	0
Google	Emme myy <i>olutta</i> <b>alaikäiselle</b> .	1	1	Emme myy <i>olutta</i> <b>pieni</b> , vain aikuisille.	1	0	-
Bing	Meillä ei myydä <i>oluen</i> <b>alaikäinen</b> .	1	1	Meillä ei myydä <i>oluen</i> <b>alaikäinen</b> vain aikuisille.	1	1	0

Virkeyhteydessä käännettyjen monitulkintaisten sanojen virheellisten käännösten perusteella muodostettiin kuusi virhetyyppiä. Ensimmäisessä virhetyypissä tarkasteltava sana oli jäänyt kääntämättä, eli se esiintyi käännöksessä samana englanninkielisenä sanana kuin käännettävässä virkkeessä. Toisessa virhetyypissä käännetty sana kuului oikean käännöksen kanssa samaan johdosperheeseen mutta ei ollut täysin oikea käännös. Kol-

mas virhetyyppi oli puolestaan se, että käännökseksi oli valittu väärä monitulkintaisen sanan useista käännösvaihtoehdoista. Neljännessä virhetyypissä tarkasteltava sana oli käännetty, mutta käänös ei ollut mikään tarkasteltavan sanan käännösvaihtoehdoista. Viidennessä virhetyypissä tarkasteltavan sanan käänös puuttui kokonaan käännettystä virkkeestä. Kuudes virhetyyppi sisälsi muita kuin juuri luokiteltuja virheitä. Esimerkkinä näistä muista virheistä ovat Bing<sup>TM</sup> Translator Betan tekemät virheet, joissa tarkasteltava sana ei ollut suomea eikä englantia. Bing<sup>TM</sup> Translator Beta käänsi esimerkiksi virkkeen ”I play football.” virkkeeksi ”Naglararô akó ng football.”. Taulukossa 6 on esimerkki kunkin virhetyypin pisteyttämisestä. Sarakkeet vastaavat juuri kuvattuja virhetyyppejä 1-6.

Taulukko 6. Esimerkkejä monitulkintaisten sanojen käännösten virhetyypien pisteyttämisestä

Käännettävät virkkeet	Käännetyt virkkeet	1	2	3	4	5	6
Results are expressed as a <b>mean</b> .	Tulokset on ilmaistu <b>[mean]</b> .	1	0	0	0	0	0
My coat needs a dry <b>wash</b> .	Takkini on kuiva <b>pestä</b> .	0	1	0	0	0	0
All essays <b>treat</b> politics.	Kaikki esseitä <b>hoitoon</b> politiikassa.	0	0	1	0	0	0
I just want to <b>lie</b> on the sofa, because I'm sick.	Haluan vain <b>oikeudenkäyntikulujen</b> sohva, koska olen sairas.	0	0	0	1	0	0
The pilot had to <b>bank</b> the aircraft.	Pilotti oli lentokoneen.	0	0	0	0	1	0
I <b>play</b> football.	Naglararô akó ng football.	0	0	0	0	0	1

#### 6.4.2 Tarkkuusarvojen ja prosenttiosuuksien laskeminen

Kun kaikki konekäännökset oli pisteytetty, laskettiin pisteistä erilaisia tarkkuuksia ja prosenttiosuuksia tutkimuskysymysten vaatimilla tavoilla. Laskut toteutettiin Excelillä.

Konekäännösjärjestelmien vertailua varten laskettiin tarkkuus seuraavalla tavalla: Jokaiselle tarkasteltavalle monitulkintaiselle sanalle laskettiin sanakohtaiset yhteispisteet monitulkintaisen sanan käännösten pisteistä yksittäisenä sanana sekä lyhyissä ja pitkissä virkkeissä käännettyinä. Tämän jälkeen nämä kolmen erilaajuisen asiayhteyden yhteen lasketut sanakohtaiset pisteet laskettiin järjestelmäkohtaisesti yhteen ja jaettiin oikeiden käännösten lukumäärällä. Oikeiden käännösten lukumäärä oli 440, koska yksittäin käännettyjä monitulkintaisia sanoja oli 50 ja tarkasteltavia monitulkintaisten sanojen merkityksiä sekä lyhyissä että pitkissä virkkeissä kummassakin 195. Taulukossa 7 on esimerkki TeemaPointin sanakohtaisten pisteiden ja tarkkuuden laskemisesta.

Taulukko 7. Esimerkki käännosten sanakohtaisten pisteiden ja tarkkuuden laskemisesta

	Sana	Yksin	Lyhyt	Pitkä	Summa
1.	appear	0,33	2	2	4,33
2.	argument	0,67	2	2	4,67
...	...	...	...	...	...
50.	win	0,33	1	1	2,333
Yhteensä					211,71
Tarkkuus ( 211,71 / (50+195+195) )					0,481

Menetelmäkohtaiset tarkkuudet kolmen mittaustavan keskiarvona laskettiin niin, että ensin tiettyä menetelmää edustavien järjestelmien sanakohtaisista pisteistä laskettiin keskiarvot. Sen jälkeen nämä keskiarvot laskettiin menetelmäkohtaisesti yhteen ja jaettiin saatu summa käännosten kokonaismäärällä eli 440:lla. Sanakohtaisista pisteistä ei voitu käyttää menetelmäkohtaista summaa vaan oli käytettävä keskiarvoa, koska testattuja sääntöperustaisia ja tilastollisia konekäännösjärjestelmiä oli eri määrä. Menetelmäkohtaiset tarkkuudet erilaajuisten asiayhteyksien vertailua varten laskettiin muuten juuri kuvatulla periaatteella, mutta myös lyhyissä ja pitkissä virkkeissä käännettyjen sanojen sanakohtaiset pisteet perustuivat sanan eri merkityksistä saatujen pisteiden keskiarvoon eivätkä summaan. Keskiarvoa käytettiin, jotta virkeyhteydessä käännettyjen sanojen sanakohtaisia pisteitä voitiin vertailla keskiarvoon perustuviin yksittäisten sanojen sanakohtaisiin pisteisiin. Näiden sanakohtaisten pisteiden keskiarvojen summat jaettiin 50:llä, koska tarkasteltavia sanoja oli 50. Taulukossa 8 on esimerkki menetelmäkohtaisten tarkkuusarvojen laskemisesta. KA-sarakkeissa ilmenevät sanakohtaisten pisteiden menetelmäkohtaiset keskiarvot.

Taulukko 8. Esimerkki käännosten menetelmäkohtaisten pisteiden ja tarkkuuden laskemisesta

		Sääntöperustaiset				Tilastolliset		
	Sana	TP	Sunda	SDL	KA	Google	Bing	KA
1.	appear	4,33	4,33	2,33	3,67	4,33	2,17	3,25
2.	argument	4,67	6,40	4,67	5,24	5,33	4,67	5,0
...	...	...	...	...	...	...	...	...
50.	win	2,33	2,25	2,17	2,25	2,33	2,33	2,33
Yhteensä					214,18	Yhteensä		195,55
Tarkkuus (214,18 / 440)					0,487	Tarkkuus (195,55 / 440)		0,444

Erilajuisissa asiayhteyksissä tehtyjen käännosten laadun vertailua varten tarkkuudet laskettiin eri tavoin kuin edellä esitetyissä tarkkuuksien laskemisissa. Monitulkintaisten sanojen käännosten tarkkuudet laskettiin niin, että ensin laskettiin yhteen kunkin järjes-

telmän sanakohtaiset pisteet erikseen kaikissa kolmessa erilaajuisessa asiayhteydessä käännettyjen sanojen pisteistä. Kaikki nämä sanakohtaiset pisteet perustuivat kunkin monitulkintaisen sanan eri merkitysten pisteiden keskiarvoon. Tarkkuuden laskemiseksi erilaajuisien asiayhteyksien sanakohtaisten pisteiden summa jaettiin 50:llä, koska tarkasteltavia monitulkintaisia sanoja oli 50. Taulukossa 9 on esimerkki TeemaPointin käännosten tarkkuuden laskemisesta erilaajuisissa asiayhteyksissä käännettyistä käännöksistä. Kaikista järjestelmäkohtaisista erilaajuisien asiayhteyksien tarkkuuksista laskettiin myös keskiarvot, jotta saatiin kaikkien järjestelmien keskimääräiset tarkkuudet erikseen yksittäisinä sanoina sekä lyhyessä ja pitkässä virkkeessä käännettyjen sanojen käännöksistä.

Taulukko 9. Esimerkki erilaajuisissa asiayhteyksissä esiintyneiden monitulkintaisten sanojen käännosten pisteiden ja tarkkuuden laskemisesta

		Yksin	Lyhyt	Pitkä
1.	appear	0,33	0,67	0,67
2	argument	0,67	0,67	0,67
...	...	...	...	...
50.	win	0,33	0,33	0,33
	Yhteensä	10,21	26,43	27,59
	Tarkkuus	$10,21/50 \approx 0,204$	$26,43/50 \approx 0,529$	$27,59/50 \approx 0,552$

Tarkkuudet laskettiin myös virkeyhteydessä esiintyneiden yksikäsitteisten ja monitulkintaisten sanojen käännosten laadun vertailua varten. Tätä vertailua varten tarkkuuksien laskemisessa ei eroteltu lyhyessä ja pitkässä virkkeessä esiintyneiden sanojen käännosten pisteitä. Ensin laskettiin järjestelmäkohtaisesti yhteen kaikkien tarkasteltavien sanojen käännosten pisteet ja jaettiin saadut summat 390:llä, koska sekä lyhyissä että pitkissä virkkeissä esiintyi 195 tarkasteltavaa sanaa. Järjestelmäkohtaisista tarkkuuksista laskettiin myös keskiarvot, jotta saatiin keskimääräiset tarkkuudet siitä, kuinka hyvin konekäännösjärjestelmät keskimäärin käänsivät virkeyhteydessä esiintyviä yksikäsitteisiä ja monitulkintaisia sanoja.

Tarkkuuksien lisäksi laskettiin prosenttiosuudet kaikille kuudelle virhetyypille. Järjestelmäkohtaisten prosenttiosuuksien laskemiseksi ensin laskettiin sanakohtaisesti virhetyypeittäin yhteen kunkin järjestelmän virkeyhteydessä tekemien käännosten virhepisteet, jotta saatiin selville kunkin järjestelmän eri virhetyyppien kokonaispisteet. Tämän jälkeen laskettiin yhteen virhetyyppikohtaiset summat järjestelmäkohtaisesti, jotta saatiin selville kunkin järjestelmän virheiden kokonaispisteet. Varsinainen virhetyyppien

osuus laskettiin virhetyypeittäin osuutena kunkin järjestelmän kaikista virheistä. Keskimääräiset virhetyyppien osuudet kaikkien järjestelmien virheistä laskettiin laskemalla yhteen kaikkien järjestelmien virhepisteet virhetyypeittäin ja jakamalla nämä summat kaikkien järjestelmien kokonaisvirheiden määrällä. Taulukossa 10 on esimerkki TeemaPointin Väärä vaihtoehto -virhetyypin prosenttiosuuden laskemisesta. TeemaPointin kokonaisvirhepisteet olivat taulukossakin ilmenevät 192.

Taulukko 10. Esimerkki virhetyyppien prosenttiosuuksien laskemisesta

		Väärä vaihtoehto (Lyhyt)	Väärä vaihtoehto (Pitkä)	Väärä vaihtoehto (Yhteensä)
1.	appear	1	1	2
2.	argument	1	1	2
...	...	...	...	...
50.	win	2	2	4
Yhteensä				101
Prosenttiosuus (101 / 192)				53%

Prosenttiosuudet laskettiin myös virkeparien erilaajuisissa virkkeissä käännettyjen monitulkintaisten sanojen käännösten laatueroille. Näiden prosenttiosuuksien perusteella voitiin vertailla, olivatko monitulkintaisten sanojen käännösten laatu sama vai toisistaan poikkeava virkeparien lyhyissä ja pitkissä virkkeissä.

### 6.4.3 Tilastollisten merkitsevyydestien suorittaminen

Käännösten laatuerojen ja virhetyyppien jakaumien erojen tilastollinen merkitsevyys selvitetään tilastollisilla testeillä. Tilastollisessa päättelyssä arvioidaan sattuman todennäköisyyttä (Laininen 2000, 25). Jos sattuma voi suurella todennäköisyydellä aiheuttaa havaitun poikkeaman, poikkeama selitetään sattuman aiheuttamaksi (Laininen 2000, 25). Tällöin mittaustuloksissa havaitut erot eivät ole tilastollisesti merkitseviä (Laininen 2000, 25), eikä mittaustuloksissa ilmeneviä eroja voida pitää luotettavina tuloksina. Jos sattuma voi aiheuttaa havaitun poikkeaman vain pienellä todennäköisyydellä, poikkeamaa pidetään tilastollisesti merkitseväenä eikä sattuman aiheuttamana (Laininen 2000, 25), ja mittaustuloksissa ilmeneviä eroja voidaan pitää luotettavina tuloksina. Sattuman mahdollisuutta mittaavan todennäköisyyden arvo on nimeltään p-arvo (Laininen 2000, 26). Vertailtavien mittaustulosten välistä eroa sanotaan tilastollisesti melkein merkitseväksi, jos p-arvo on alle 0,05, tilastollisesti merkitseväksi, jos p-arvo on alle



0,01 ja tilastollisesti erittäin merkitseväksi, jos p-arvo on alle 0,001 (Manninen & Ylén 2000, 68).

Käännösten laatuerojen tilastollinen merkitsevyys selvitettiin Friedmanin kaksisuuntaisella järjestyslukutestillä (Siegel & Castellan 1988, 174). Friedmanin testiä käytetään nimenomaan tilanteissa, joissa vertaillaan useampaa kuin kahta järjestelmää (Hull 1993, 334). Peruseriaatteena Friedmanin testissä on ensin vertailla keskenään kaikkia testattavia järjestelmiä. Jos järjestelmien välillä ilmenee tilastollisesti merkitseviä eroja, vertaillaan järjestelmiä pareittain. (Kettunen 2007, 42.)

Friedmanin testi on epäparametrinen testi (Siegel & Castellan 1988, 174). Epäparametrisissa testeissä ei ole oletuksia populaation jakaumasta (Siegel & Castellan 1988, 3). Friedmanin testiä voidaan käyttää, kun aineisto on vähintään järjestysasteikollista (Siegel & Castellan 1988, 174). Tämän tutkimuksen käännösten pisteistä koostuva aineisto oli järjestysasteikollista, koska käännös sai pisteikseen 0, 0,5 tai 1. Friedmanin testin yhtenä ehtona on myös se, että vertailtavissa otoksissa on oltava sama määrä tapauksia (Siegel & Castellan 1988, 175). Tässä tutkimuksessa tämäkin ehto toteutui. Friedmanin merkitsevyydestit laskettiin SPSS-ohjelmalla. Testit laskettiin edellä kuvattujen monitulkintaisten sanojen käännösten erityyppisten sanakohtaisten pisteiden perusteella.

Virhetyyppien osuuksien jakaumien erojen tarkastelussa käytettiin  $\chi^2$ -testiä (chi-square test). Testi kuuluu Friedmanin testin tavoin epäparametrisiin testeihin (Laininen 2000, 48).  $\chi^2$ -testiä käytetään, kun aineisto koostuu luokitteluasteikollisen muuttujan frekvenssiarviosta (Siegel & Castellan 1988, 111). Virhetyyppi oli luokitteluasteikollinen muuttuja, koska virhetyypit luokiteltiin luokkiin 0 tai 1 virheen esiintymisen perusteella. Tietyntyypinen virhe siis esiintyi tai ei esiintynyt.  $\chi^2$ -testillä voidaan tutkia, onko havaintoaineisto peräisin tietystä jakaumasta (Laininen 2000, 48) tai eroavatko tarkasteltavat ryhmät toisistaan tietyn ominaisuuden suhteen (Siegel & Castellan 1988, 111).  $\chi^2$ -testiä varten aineisto järjestetään ristiintaulukkoon (Siegel & Castellan 1988, 111). Testin laskeminen perustuu havaittujen ja odotettujen frekvenssien välisiin eroihin (Laininen 2000, 49). Odotettujen frekvenssin laskeminen perustuu havaittuihin frekvensseihin (ks. Siegel & Castellan 1988, 112-113).  $\chi^2$ -testiä voidaan käyttää, jos enintään 20 %:ssa ristiintaulukon soluista odotettu frekvenssi on vähemmän kuin viisi ja jos minkään solun odotettu frekvenssi ei ole vähemmän kuin yksi (Siegel & Castellan 1988, 123). Jos nämä testin vaatimukset eivät täyty, luotuja kategorioita voidaan yhdistää, jotta odotettu-

jen frekvenssien arvoja saataisiin kasvatettua (Siegel & Castellan 1988, 123). Myös  $\chi^2$ -testit laskettiin SPSS-ohjelmalla.

Luvussa 7.5 esiteltävien järjestelmäkohtaisten virhetyyppien osuuksien epätasaisten jakaumien perusteella haluttiin tarkastella TeemaPointin virhetyyppien jakaumaa suhteessa kaikkien muiden järjestelmien virhetyyppien jakaumiin sekä Bing<sup>TM</sup> Translator Betan virhetyyppien jakaumaa muiden kuin TeemaPointin jakaumiin. Kaikilla kuudella virhetyypillä laskettaessa  $\chi^2$ -testin ehto siitä, että enintään 20 %:ssa ristiintaulukon soluista odotettu frekvenssiarvo saa olla pienempi kuin viisi, ei toteutunut. Ehdon toteutumattomuuden takia tilastollisten merkitsevyydestien laskemiseksi alkuperäiset virhetyypit yhdistettiin virheiden esiintymisten perusteella kolmeen virhetyyppiin. TeemaPointin vertailussa kaikkiin muihin järjestelmiin ensimmäisen virhetyypin muodosti alkuperäinen virhetyyppi 1, toisen alkuperäinen virhetyyppi 3 ja kolmannen loput alkuperäisistä virhetyypeistä. Jotta voitiin vertailla Bing<sup>TM</sup> Translator Betan virhetyyppien jakaumaa muihin kuin TeemaPointin virhetyyppien jakaumaan, virhetyyppejä yhdistettiin niin, että ensimmäisen virhetyypin muodosti alkuperäinen virhetyyppi 3, toisen virhetyyppi 4 ja kolmannen loput virhetyypeistä.  $\chi^2$ -testin ehdot toteutuivat juuri mainituilla virhetyyppien yhdistelmillä sekä TeemaPointin ja muiden järjestelmien vertailussa että Bing<sup>TM</sup> Translator Betan ja loppujen kolmen järjestelmän vertailussa.

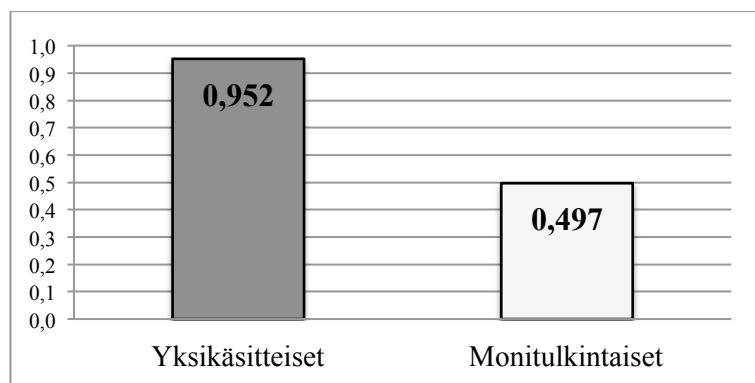
## 7 TULOKSET

Tässä luvussa raportoidaan tutkimustulokset tutkimuskysymysten perusteella. Ensimmäisenä vertaillaan virkeyhteydessä esiintyneiden yksikäsitteisten ja monitulkintaisten sanojen käännösten laatueroja. Toisena tarkastellaan, oliko monitulkintaisten sanojen käännösten laadussa eroja eri konekäännösjärjestelmien välillä. Kolmantena vertaillaan, oliko monitulkintaisten sanojen käännösten laadussa eroja erilaajuisten asiayhteyksien välillä. Neljäntenä keskitytään puolestaan siihen, oliko monitulkintaisten sanojen käännösten laadussa eroja sääntöperustaisten ja tilastollisten konekäännösjärjestelmien välillä. Lopuksi raportoidaan, minkä tyyppisiä virheitä järjestelmät tekivät virkeyhteydessä esiintyneiden monitulkintaisten sanojen kääntämisessä.

Tuloksina raportoidaan vain ne, joissa tilastollisten merkitsevyydestien perusteella vertailtavien mittaustulosten välisen eron p-arvo oli pienempi kuin 0,05. Näin ollen tuloksina raportoidaan vähintään tilastollisesti melkein merkitsevät tulokset (ks. Manninen & Ylén 2000, 68). Jos mittaustulosten väliset erot läpäisivät tilastolliset merkitsevyydestit, tarkasteltiin tarkkuusarvojen eroja vertailemalla arvojen erojen suuruutta. Sparck Jonesin (1974, 397) mukaan vähintään viiden prosenttiyksikön eroa pidetään pienenä ja vähintään kymmenen prosenttiyksikön eroa huomattavana. Tähän näkemykseen viitataan jatkossa SJ-sääntönä. Kuvien tarkkuusarvot on pyöristetty kolmen desimaalin tarkkuudella ja taulukoiden prosenttiluvut kokonaisluvuiksi. Kaikki arvot on kuitenkin laskettu tarkoilla arvoilla.

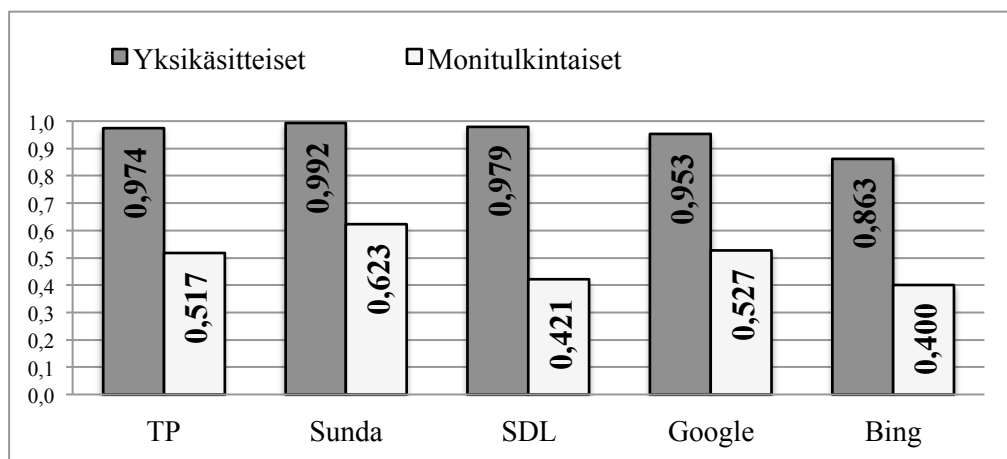
### 7.1 Käännöslaadun erot yksikäsitteisten ja monitulkintaisten sanojen välillä

Yksikäsitteisten sanojen käännösten laatua tarkastellaan perustasona, johon monitulkintaisten sanojen käännösten laatua voidaan vertailla. Yksikäsitteisten ja monitulkintaisten sanojen käännösten tarkkuudet ilmenevät kuvasta 9. Yksikäsitteisten sanojen käännösten keskimääräinen tarkkuus kaikissa virkkeissä oli 0,952 ja monitulkintaisten sanojen 0,497. SJ-säännön perusteella näiden tarkkuuksien välinen ero oli huomattava. Näin ollen virkeyhteydessä esiintyneiden monitulkintaisten sanojen kääntäminen oli haasteellisempaa kuin virkeyhteydessä esiintyneiden yksikäsitteisten sanojen kääntäminen.



Kuva 9. Virkeytydessä esiintyneiden yksikäsitteisten ja monitulkintaisten sanojen käänntösten tarkkuudet järjestelmien tarkkuuksien keskiarvona laskettuna

Jokaisen järjestelmän käänntöksissä yksikäsitteisten ja monitulkintaisten sanojen käänntösten välinen laatuero oli SJ-sääntön perusteella huomattava. Järjestelmäkohtaiset yksikäsitteisten ja monitulkintaisten sanojen käänntösten tarkkuudet ilmenevät kuvasta 10. Yksikäsitteisten sanojen käänntösten tarkkuudet olivat väliltä 0,863 ja 0,992, kun taas monitulkintaisten sanojen käänntösten tarkkuudet olivat väliltä 0,400 ja 0,623. Suurin ero virkeytydessä esiintyneiden yksikäsitteisten ja monitulkintaisten sanojen käänntösten laadussa oli SDL Free Translationin käänntöksissä ja pienin puolestaan Sunda käänntöksissä.

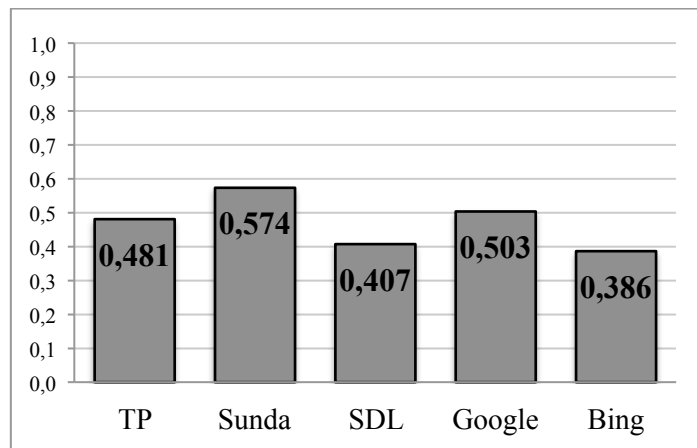


Kuva 10. Virkeytydessä esiintyneiden yksikäsitteisten ja monitulkintaisten sanojen käänntösten tarkkuudet järjestelmäkohtaisesti

Yksikäsitteisten sanojen käänntösten laadusta saadaan tietoa kunkin konekäänntösjärjestelmän käänntösten laadun yleisestä tasosta. Pelkkien yksikäsitteisten sanojen käänntöslaadun perusteella TeemaPoint, Sunda ja SDL Free Translation käänntivät laadukkaasti, kun taas Bing<sup>TM</sup> Translator Beta käännti näitä järjestelmiä huomattavasti heikommin.

## 7.2 Käännöslaadun erot konekäännösjärjestelmien välillä

Seuraavaksi tarkastellaan, oliko monitulkintaisten sanojen käännösten laadussa eroja testattujen konekäännösjärjestelmien välillä. Järjestelmäkohtaiset tarkkuudet laskettiin sanakohtaisesti huomioiden monitulkintaisten sanojen käännökset yksittäisinä sanoina ja sekä lyhyessä että pitkässä virkkeessä käännettyinä. Tarkkuudet ilmenevät kuvasta 11.



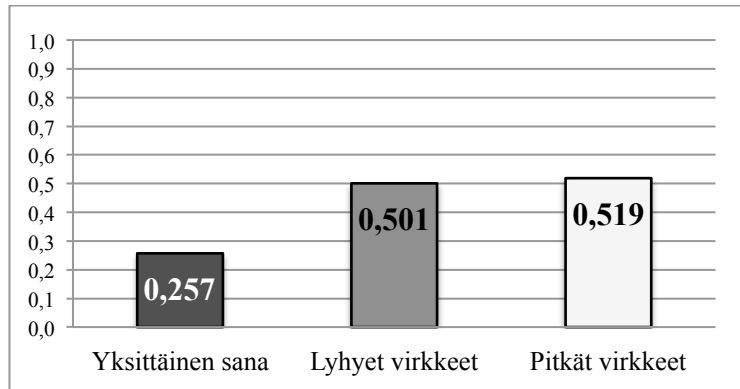
Kuva 11. Järjestelmäkohtaiset monitulkintaisten sanojen käännösten tarkkuudet kolmen mittaustavan keskiarvolla laskettuna

Järjestelmien keskinäisissä vertailuissa järjestelmien käännöslaadussa ilmeni eroja. Sunda tarkkuudella 0,574 ja Google Translate tarkkuudella 0,503 erottuivat SDL Free Translationista ja Bing<sup>TM</sup> Translator Betasta, joiden tarkkuudet olivat 0,407 ja 0,386. SJ-säännön perusteella Sundan paremmuus sekä SDL Free Translationiin että Bing<sup>TM</sup> Translator Betaan verrattuna oli huomattava. Vastaavasti Google Translaten ero SDL Free Translationiin verrattuna oli pieni ja Bing<sup>TM</sup> Translator Betaan verrattuna huomattava. Näin ollen Sunda ja Google Translate olivat parhaat järjestelmät ja SDL Free Translation sekä Bing<sup>TM</sup> Translator Beta huonoimmat. TeemaPoint jäi näiden parhaimpien ja huonoimpien järjestelmien välimaastoon tarkkuudella 0,481.

## 7.3 Käännöslaadun erot erilaajusten asiayhteyksien välillä

Seuraavaksi tarkastellaan, oliko monitulkintaisten sanojen käännösten laadussa eroja erilaajusten asiayhteyksien välillä. Vertailtavana oli kolme erilaajusta asiayhteyttä. Yksittäin käännettyillä monitulkintaisilla sanoilla ei ollut asiayhteyttä lainkaan. Kaksi muuta erilaajusta asiayhteyttä olivat virkkeiden lyhyet ja pitkät versiot. Kuvasta 12 il-

menevät monitulkintaisten sanojen käännösten keskimääräiset tarkkuudet erilaajuisissa asiayhteyksissä kaikkien järjestelmien tarkkuuksien keskiarvoina laskettuna.



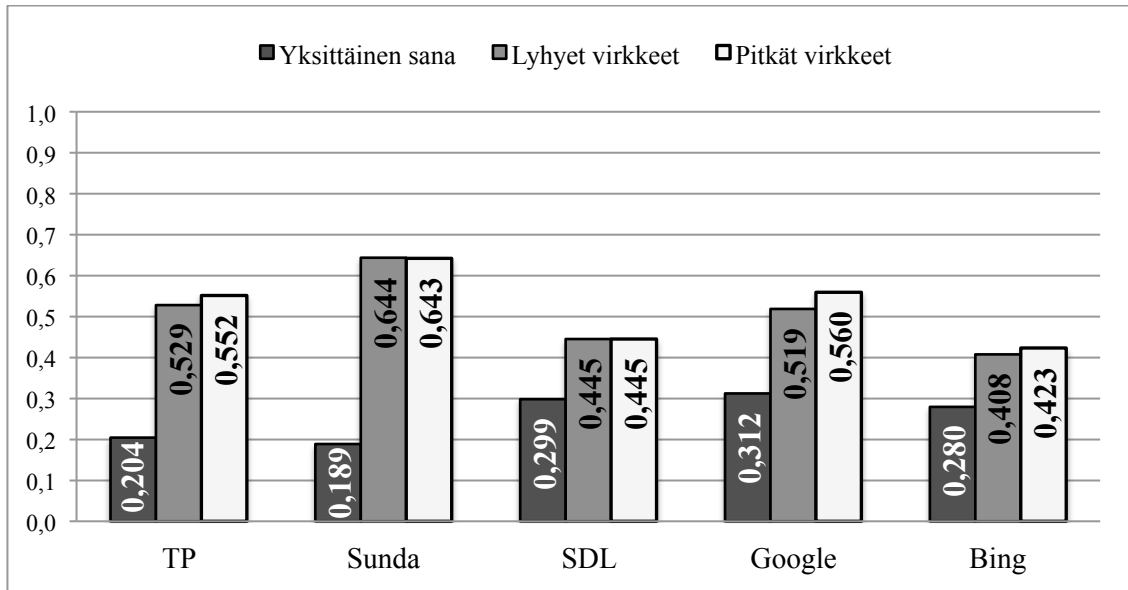
Kuva 12. Monitulkintaisten sanojen käännösten tarkkuudet erilaajuisissa asiayhteyksissä kaikkien järjestelmien tarkkuuksien keskiarvoina laskettuna

Parhaat tarkkuudet olivat virkeyhteydessä käännettyillä monitulkintaisilla sanoilla. Pitkissä virkkeissä käännettyjen sanojen tarkkuus oli 0,519 ja lyhyissä virkkeissä 0,501. Huonoin tarkkuus oli yksittäisinä sanoina käännettyjen monitulkintaisten sanojen käännöksillä tarkkuudella 0,257. SJ-säännön perusteella kummassa tahansa virkeyhteydessä käännettyjen monitulkintaisten sanojen käännöslaatu oli huomattavasti parempi kuin yksittäin käännettyjen sanojen. Sen sijaan lyhyessä ja pitkässä virkkeessä käännettyjen monitulkintaisten sanojen käännöslaadussa ei ollut eroa, jonka mitaustulos läpäisisi tilastollisen merkitsevyydestin.

Järjestelmäkohtaisessa tarkastelussa erilaajuisissa asiayhteyksissä tehtyjen käännösten laatueroja voidaan tarkastella kahdesta näkökulmasta: Voidaan vertailla monitulkintaisten sanojen käännösten laatua järjestelmäkohtaisesti, kun sanat on käännetty yksittäisinä sanoina tai erilaajuisissa virkkeissä. Toisaalta voidaan vertailla keskenään eri järjestelmien tietynlaajuisessa asiayhteydessä tekemien käännösten laatua.

Kuvasta 13 ilmenevät monitulkintaisten sanojen käännösten järjestelmäkohtaiset tarkkuudet kolmessa erilaajuisessa asiayhteydessä. Kunkin järjestelmän käännöksissä lyhyessä sekä pitkässä virkkeessä käännettyjen monitulkintaisten sanojen käännöslaadun paremmuus yksittäin käännettyjen sanojen käännöslatuun verrattuna oli SJ-säännön perusteella huomattava. Sen sijaan minkään järjestelmän monitulkintaisten sanojen käännösten laadussa ei ollut tilastollisen merkitsevyydestin läpäisseitä eroja, kun tarkasteltava sana käännettiin lyhyessä tai pitkässä virkkeessä. Näin ollen jokaisen järjestelmän käännöksissä monitulkintaisten sanojen käännöslaatu oli huomattavasti parempaa vir-

keyhteydessä kuin yksittäisinä sanoina käännettyinä, mutta minkään järjestelmän käännöslaadussa erilaajuisten virkeyhteyksien välillä ei ollut eroja.



Kuva 13. Monitulukintaisten sanojen käännösten järjestelmäkohtaiset tarkkuudet erilaajuisissa asiayhteyksissä

Lyhyissä ja pitkissä virkkeissä käännettyjen monitulukintaisten sanojen käännösten laatueroja voidaan tarkastella myös vertailemalla, kuinka suuressa osassa virkeparien käännöksistä käännöslaatu oli pitkässä virkkeessä laadukkaampaa kuin lyhyessä ja kuinka suuressa osassa virkeparien käännöksistä käännöslaatu oli lyhyessä virkkeessä laadukkaampaa kuin pitkässä. Voidaan myös tarkastella, kuinka suuressa osassa virkeparien käännöksistä käännösten laatu oli sama virkeparin lyhyessä ja pitkässä virkkeessä. Taulukkoon 11 on koottu monitulukintaisten sanojen käännösten laatuero virkeparien lyhyissä ja pitkissä virkkeissä. Taulukon alimmalla rivillä on järjestelmäkohtaisten prosenttiosuuksien keskiarvot. Koska prosenttiosuudet on pyöristetty kokonaisluvuiksi, pyöritetyillä arvoilla laskettuna prosenttiosuuksien summasta ei välttämättä tule tasan 100 %. Tarkoilla arvoilla laskettuna summa kuitenkin on tasan 100 %.

Kaikkien järjestelmien keskiarvojen perusteella vain 2 %:ssa virkepareista käännösten laatu oli pitkässä virkkeessä parempi kuin lyhyessä, ja 1 %:ssa laatu oli parempi lyhyessä kuin pitkässä virkkeessä. Sen sijaan 97 %:ssa virkepareista käännösten laatu oli sama virkelaajuudesta riippumatta. Myös järjestelmäkohtaisesta tarkastelusta huomataan, että kunkin järjestelmän käännösten laadussa ei juurikaan ollut eroja lyhyen tai pitkän virkeyhteyden välillä. Tämänkin tarkastelutapa osoittaa, että selvässä enemmistössä moni-

tulkintaisten sanojen käännot laadussa ei ollut eroja lyhyessä tai pitkässä virkeyhteys-  
dessä tehtyjen käännot välillä.

Taulukko 11. Monitulkintaisten sanojen käännot laatuero virkeparien lyhyissä ja pitkissä virkkeissä

	Pitkässä virkkeessä parempi laatu kuin lyhyessä	Lyhyessä virkkeessä parempi laatu kuin pitkässä	Ei eroa lyhyessä ja pitkässä virkkeessä	Yhteensä
TP	3 %	0 %	97 %	100 %
Sunda	1 %	1 %	99 %	101 %
SDL	0 %	0 %	100 %	100 %
Google	6 %	3 %	92 %	101 %
Bing	3 %	2 %	95 %	100 %
Keskiarvo	2 %	1 %	97 %	100 %

Kaikkien järjestelmien käännösladun keskinäisissä vertailuissa yksittäin käännettyjen  
monitulkintaisten sanojen käännösladussa ilmeni eroja. Kuten kuvasta 13 huomataan,  
Google Translaten ja SDL Free Translationin paremmuus Sundaan verrattuna oli SJ-  
säännön perusteella huomattava ja Bing™ Translator Betan paremmuus Sundaan verrat-  
tuna pieni. Google Translaten tarkkuus oli 0,312, SDL Free Translationin 0,299, Bing™  
Translator Betan 0,280 ja Sundan 0,189. Google Translaten paremmuus TeemaPointiin  
verrattuna oli myös huomattava ja SDL Free Translationin paremmuus TeemaPointiin  
verrattuna pieni. TeemaPointin tarkkuus oli 0,204. Näin ollen yksittäin käännettyjen  
monitulkintaisten sanojen kääntämisessä Google Translate ja SDL Free Translation  
erottuivat parhaimpina järjestelminä TeemaPointiin ja Sundaan verrattuna. Sunda puo-  
lestaan erottui huonoimpana järjestelmänä kaikista muista paitsi TeemaPointista.

Kun Sundan ja muiden järjestelmien yksittäin kääntämien monitulkintaisten sanojen  
käännot laatu vertaillaan, on muistettava, että Sundan käännösten pisteytys poikkesi  
muiden järjestelmien käännösten pisteytyksestä. Pisteytysten ero johtui siitä, että Sunda  
tarjosi yksittäin käännettyille sanoille monta käännösvaihtoehtoa. Vertailuongelma liit-  
tyy kuitenkin vain Sundan ja muiden järjestelmien väliseen vertailuun; Sundaa lukuun  
ottamatta kaikkien muiden järjestelmien keskinäisissä vertailuissa ei vastaavaa ongel-  
maa ole.

Myös lyhyissä virkkeissä käännettyjen monitulkintaisten sanojen käännösten laadussa  
ilmeni eroja. Kuten kuvasta 13 ilmenee, Sundan paremmuus kaikkiin muihin järjestel-  
miin verrattuna oli SJ-säännön perusteella huomattava. Sundan tarkkuus oli 0,644.

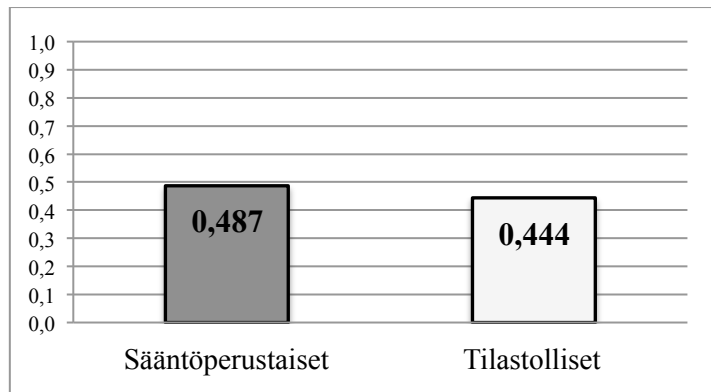


Google Translaten paremmuus SDL Free Translationiin verrattuna oli pieni ja Bing™ Translator Betaan huomattava. Google Translaten tarkkuus oli 0,519, SDL Free Translationin 0,445 ja Bing™ Translator Betan 0,408. Myös TeemaPointin paremmuus Bing™ Translator Betaan oli huomattava, mutta TeemaPointin ero SDL Free Translationiin ei läpäissyt tilastollista merkitsevyydestä. TeemaPointin tarkkuus oli 0,529. Näin ollen lyhyissä virkkeissä käännettyjen monitulkintaisten sanojen käännoslaadun perusteella Sunda oli selvästi paras järjestelmä. Bing™ Translator Beta erottui huonoimpana kaikista muista paitsi SDL Free Translationista. SDL Free Translation puolestaan erottui Sundaa ja Google Translatea huonompana.

Kuten kuvasta 13 ilmenee, myös pitkissä virkkeissä käännettyjen monitulkintaisten sanojen käännosten laadussa ilmeni eroja. Pitkissä virkkeissä käännettyjen monitulkintaisten sanojen käännosten tarkkuuden perusteella Sunda oli paras järjestelmä arvolla 0,643. Toiseksi parhaat järjestelmät olivat Google Translate tarkkuudella 0,560 ja TeemaPoint tarkkuudella 0,552. Sundan paremmuus sekä Google Translateen että TeemaPointiin verrattuna oli SJ-säännön perusteella pieni. Huonoimmat järjestelmät olivat SDL Free Translation tarkkuudella 0,445 ja Bing™ Translator Beta tarkkuudella 0,423. Sundan, Google Translaten sekä TeemaPointin paremmuus SDL Free Translationiin sekä Bing™ Translator Betaan verrattuna oli huomattava. Näin ollen Sunda oli paras järjestelmä ja SDL Free Translation sekä Bing™ Translator Beta huonoimmat; Sunda oli parempi kuin neljä muuta järjestelmää ja Google Translate sekä TeemaPoint parempia kuin kaksi huonointa järjestelmää.

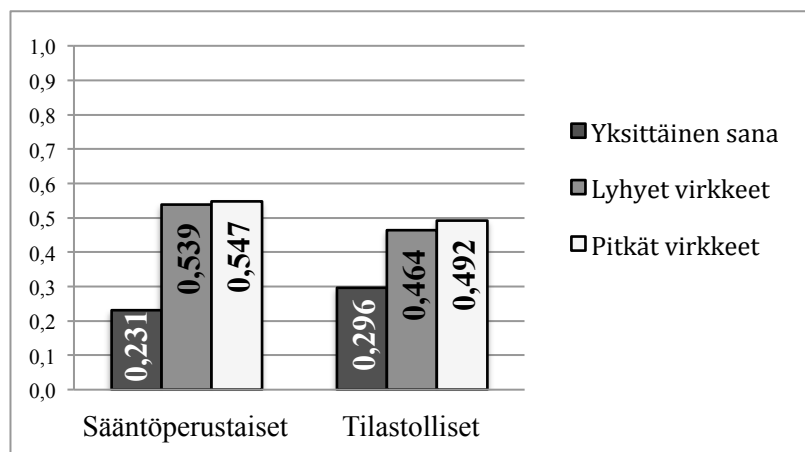
#### **7.4 Käännoslaadun erot kääntämismenetelmien välillä**

Seuraavaksi tarkastellaan, oliko monitulkintaisten sanojen käännosten laadussa eroja sääntöperustaisten ja tilastollisten konekäännösjärjestelmien välillä. TeemaPoint, Sunda ja SDL Free Translation ovat sääntöperustaisia ja Google Translate ja Bing™ Translator Beta tilastollisia. Kuvasta 14 ilmenevät monitulkintaisten sanojen käännosten tarkkuudet menetelmäkohtaisesti. Tarkkuudessa huomioitiin monitulkintaisten sanojen käännosten oikeellisuus käännettäessä sanat yksittäisinä sanoina sekä osana lyhyitä ja pitkiä virkeitä. Sääntöperustaisten konekäännösjärjestelmien käännosten keskimääräinen tarkkuus oli 0,487 ja tilastollisten 0,444. SJ-säännön perusteella kääntämismenetelmien tarkkuuksien välillä ei ollut eroa.



Kuva 14. Monitulkintaisten sanojen käänntösten menetelmäkohtaiset tarkkuudet kolmen mittaustavan keskiarvona

Kuvassa 15 on monitulkintaisten sanojen käänntösten menetelmäkohtaiset tarkkuudet kolmessa erilaajuisessa asiayhteydessä. Kummallakin menetelmällä monitulkintaisten sanojen käänntösten laatu oli paras virkeyhteydessä käänntetyillä sanoilla ja heikoin yksittäisinä sanoina käänntetyillä sanoilla. Sääntöperustaisilla konekäänntösjärjestelmillä yksittäin käänntettyjen monitulkintaisten sanojen käänntösten tarkkuus oli 0,231, kun taas lyhyissä virkkeissä käänntettyjen monitulkintaisten sanojen tarkkuus oli 0,539 ja pitkissä virkkeissä 0,547. Tilastollisten konekäänntösjärjestelmien käänntösten vastaavat tarkkuudet olivat 0,296, 0,464 ja 0,492. Kummallakin menetelmällä virkeyhteydessä käänntettyjen sanojen käänntöslaadun paremmuus yksittäin käänntettyjen sanojen käänntöslaatuun verrattuna oli SJ-säännön perusteella huomattava. Näin ollen käänntettiin monitulkintaisten sanat kummalla tahansa menetelmällä, niin yksittäin käänntettyjen sanojen käänntöslaatu oli selvästi heikompaa kuin virkeyhteydessä käänntettyjen sanojen. Käänntöslaadussa ei kuitenkaan ollut eroja erilaajuisissa virkeyhteyksissä tehtyjen käänntösten välillä.



Kuva 15. Monitulkintaisten sanojen käänntösten menetelmäkohtaiset tarkkuudet erilaajuisissa asiayhteyksissä

Yksittäin käännettyjen sanojen käänöksissä tilastollisten konekäännösjärjestelmien tarkkuus oli parempi kuin sääntöperustaisten. SJ-säännön perusteella ero oli pieni. Sen sijaan sääntöperustaisten järjestelmien tarkkuus oli tilastollisia järjestelmiä parempi niin lyhyissä kuin pitkissä virkkeissä käännettyjen monitulkintaisten sanojen käänöksissä. Kummatkin näistä eroista olivat pieniä. Näin ollen tilastolliset konekäännösjärjestelmät käänsivät sääntöperustaisia järjestelmiä laadukkaammin yksin esiintyneitä monitulkintaisia sanoja, kun taas sääntöperustaiset järjestelmät käänsivät tilastollisia järjestelmiä laadukkaammin virkeytydessä esiintyneitä monitulkintaisia sanoja.

## 7.5 Monitulkintaisten sanojen käänösten virhetyypit

Seuraavaksi tarkastellaan, minkä tyyppisiä virheitä konekäännösjärjestelmät tekivät virkeytydessä esiintyneiden monitulkintaisten sanojen kääntämisessä. Taulukosta 12 ilmenevät kunkin virhetyypin prosenttiosuudet järjestelmäkohtaisesti sekä kaikkien järjestelmien osuuksien keskiarvoina. Taulukon prosentit on pyöristetty kokonaisluvuiksi. Pyöristettyjen osuuksien summista ei välttämättä tule tasan 100 %, mutta tarkoilla arvoilla laskettuina prosenttiosuuksista tulee yhteensä tasan 100 %. Myös prosenttiosuuksien keskiarvot on laskettu alkuperäisillä tarkoilla arvoilla eikä taulukossa olevilla pyöristetyillä arvoilla. Virhetyypin Muu keskiarvo pyöristyisi nolllaksi, mutta taulukkoon on merkitty keskiarvoksi < 1 %, jotta keskiarvo erottuu tilanteesta, jossa jotain virhetyypistä ei esiintynyt käänöksissä ollenkaan.

Taulukko 12. Virhetyypien prosenttiosuudet kaikista virheistä

	TP	Sunda	SDL	Google	Bing	Keskiarvo
Jäänyt kääntämättä	44 %	1 %	0 %	4 %	3 %	10 %
Oikea johdosperhe	4 %	5 %	4 %	9 %	4 %	5 %
Väärä vaihtoehto	53 %	91 %	95 %	85 %	78 %	80 %
Ei mikään vaihtoehtoista	0 %	3 %	1 %	2 %	10 %	3 %
Ei esiintynyt käänöksessä	0 %	0 %	0 %	1 %	4 %	1 %
Muu	0 %	0 %	1 %	0 %	1 %	< 1 %
Yhteensä	101 %	100 %	101 %	101 %	100 %	100 %

Virhetyypien osuudet jakautuivat epätasaisesti kaikkien järjestelmien keskimääräisissä virheissä. Virhetyypien keskiarvojen tarkastelun perusteella yleisin virhe oli se, että monitulkintaisen sanan käänökseksi oli valittu väärä vaihtoehto tarkasteltavan monitulkintaisen sanan useasta käänösvaihtoehdosta. Kyseistä virhetyypistä edusti selvä

enemmistö, 80 %, kaikista virheistä. Toiseksi yleisin virhetyyppi keskiarvojen perusteella oli se, että tarkasteltava sana oli jäänyt kokonaan kääntämättä. Kyseisen virhetyypin osuuksien keskiarvo oli 10 %.

Kaikkien muiden kuin juuri mainittujen yleisimpien virhetyyppien osuudet olivat vähäisiä. Viisi prosenttia keskimääräisistä virheistä selittyi sillä, että käännetty sana kuului oikean käännöksen kanssa samaan johdosperheeseen mutta ei ollut merkitykseltään täysin oikea. Kolme prosenttia keskimääräisistä virheistä selittyi sillä, että tarkasteltavan monitulkintaisen sanan käännökseksi oli valittu joku muu kuin mikään kyseisen monitulkintaisen sanan käännösvaihtoehtoista. Yksi prosentti keskimääräisistä virheistä johtui puolestaan siitä, ettei käännettävää monitulkintaista sanaa esiintynyt käännöksessä. Alle prosentti virheistä johtui muista syistä.

Kunkin järjestelmän virheistä enemmistö selittyi sillä, että monitulkintaisen sanan käännökseksi oli valittu väärä sanan käännösvaihtoehtoista. Eri järjestelmien virhetyyppien jakaumat kuitenkin osin poikkesivat toisistaan. Etenkin TeemaPointin jakauma poikkesi kaikkien muiden järjestelmien jakaumista. TeemaPointin virheistä väärä käännösvaihtoehto selitti pienemmän osuuden virheistä kuin muiden järjestelmien virheistä: TeemaPointin virheistä 53 % selittyi väärällä käännösvaihtoehdolla, kun SDL Free Translationin kyseinen prosenttiosuus oli 95% ja Sundan 91 %, Google Translaten 85 % ja Bing<sup>TM</sup> Translator Betan 78 %. TeemaPointin jakauma poikkesi muiden järjestelmien virhetyyppien jakaumista myös siinä, että TeemaPointin virheistä 44 % selittyi sanan kääntämättä jäämisellä. Muiden järjestelmien vastaava osuus oli enintään 4 %.

Bing<sup>TM</sup> Translator Betan virhetyyppien jakaumassa oli eroja verrattuna muiden kuin TeemaPointin jakaumaan. Bing<sup>TM</sup> Translator Betan virheistä 10 % johtui siitä, että monitulkintaisen sanan käännös ei ollut mikään käännösvaihtoehtoista. Muilla järjestelmillä kyseinen virhetyyppi oli vähäisempi kuin Bing<sup>TM</sup> Translator Betalla, enintään 3 %. Lisäksi 4 % Bing<sup>TM</sup> Translator Betan virheistä selittyi sillä, että tarkasteltavan sanan käännös puuttui koko käännöksestä. Bing<sup>TM</sup> Translator Betan lisäksi toisen tilastollisen konekäännösjärjestelmän, Google Translaten, virheistä kyseinen virhetyyppi selitti vain 1 %:n virheistä. Sääntöperustaisten järjestelmien käännöksissä esiintyi aina tarkasteltavan monitulkintaisen sanan käännös. SDL Free Translation ja Bing<sup>TM</sup> Translator Beta olivat puolestaan ainoat järjestelmät, joilla oli Muu-virhetyyppiin kuuluvia käännöksiä, vaikkakin vain yhden prosentin osuudella.

## 8 POHDINTA

Tämän tutkimuksen tarkoituksena oli evaluoida konekäännösjärjestelmien käännöslaatua, kun käännetään englanninkielisiä monitulkintaisia sanoja suomeksi. Tutkimuskysymyksiä oli viisi. Ensimmäisenä selvitettiin, onko virkeyhteudessa esiintyvien yksikäsitteisten ja monitulkintaisten sanojen käännösten laadussa eroja. Toisena tarkasteltiin, onko monitulkintaisten sanojen käännösten laadussa eroja eri konekäännösjärjestelmien välillä. Kolmannen tutkimuskysymyksen perusteella haluttiin puolestaan selvittää, onko monitulkintaisten sanojen käännösten laadussa eroja erilaajuisten asiayhteyksien välillä. Neljäntenä tutkimuskysymyksenä oli se, onko monitulkintaisten sanojen käännösten laadussa eroja sääntöperustaisten ja tilastollisten konekäännösjärjestelmien välillä. Lisäksi tarkasteltiin, minkä tyyppisiä virheitä konekäännösjärjestelmät tekevät virkeyhteudessa esiintyvien monitulkintaisten sanojen kääntämisessä.

Testattavat konekäännösjärjestelmät olivat TeemaPoint, Sunda, SDL Free Translation, Google Translate ja Bing<sup>TM</sup> Translator Beta. Kääntämismenetelmältään Google Translate ja Bing<sup>TM</sup> Translator Beta ovat tilastollisia ja loput kolme järjestelmää sääntöperustaisia. Aineistona oli 50 englanninkielistä monitulkintaista sanaa ja näiden sanojen ympärille luotua 195 virkeparia, jotka sisälsivät virkkeistä lyhyen ja pitkän version. Pitkä virke alkoi aina virkeparin lyhyellä versiolla. Evaluointimittarina käytettiin tarkkuutta. Virkkeiden käännösten oikeellisuuteen ja ymmärrettävyyteen ei tarkasteltavien yksikäsitteisten ja monitulkintaisten sanojen lisäksi kiinnitetty huomiota. Tarkkuuden laske- mista varten aineisto pisteytettiin niin, että tarkasteltavan yksikäsitteisen ja monitulkintaisten sanan oikeasta käännöksestä annettiin 1 piste ja väärästä 0 pistettä. Monitulkintaisten sanan käännös sai 0,5 pistettä, jos se kuului oikean käännöksen kanssa samaan johdosperheeseen mutta ei ollut täysin oikea käännös. Pisteiden perusteella laskettiin tutkimuskysymyksiin vastaamisessa tarvittavia tarkkuuksia. Lisäksi virkeyhteudessa esiintyneiden monitulkintaisten sanojen käännösten virheet luokiteltiin kuuteen virhetyyppiin ja käännökset pisteytettiin myös näiden virhetyyppien perusteella. Käännösten laatuerojen tilastollinen merkitsevyys selvitettiin Friedmanin testillä ja virhetyyppien jakaumien erojen tilastollinen merkitsevyys  $\chi^2$ -testillä.

Tässä luvussa pohditaan käytettyjen menetelmien pätevyyttä ja luotettavuutta sekä tulosten luotettavuutta. Myös ideoita mahdollisille jatkotutkimuksille esitellään.

## 8.1 Menetelmien pohdinta

Tämän tutkimuksen menetelmiin sisältyivät aineistonkeruumenetelmät sekä aineiston analysointimenetelmät. Aineisto kerättiin valitsemalla tarkasteltavat monitulkintaiset sanat, luomalla käännettävät virkkeet, satunnaisesti poimimalla virkkeistä tarkasteltavat yksikäsitteiset sanat ja kääntämällä yksittäiset monitulkintaiset sanat sekä virkkeet testattavilla konekäännösjärjestelmillä. Aineiston analyysi puolestaan kattoi tarkasteltavien yksikäsitteisten ja monitulkintaisten sanojen käännösten pisteyttämisen, virhetyyppien pisteyttämisen, pisteiden, tarkkuuksien ja prosenttiosuuksien laskemisen sekä tilastollisten merkitsevyytestien tekemisen.

Monitulkintaisia sanoja ei valittu täysin satunnaisesti vaan tiettyjen periaatteiden mukaisesti. Täysin satunnaista sanojen poimimista ei koettu mahdolliseksi eikä järkeväksi, koska yksiselitteisesti ymmärrettävien virkkeiden luominen asetti rajoituksia monitulkintaisille sanoille. Tarkasteltavien monitulkintaisten sanojen lukumäärää voidaan pitää riittävänä ainakin Senseval-3:n monitulkintaisten sanojen sanastotestiin verrattuna. Myös virkeparien lukumäärää voidaan pitää riittävänä. Kun esimerkiksi tarkasteltiin järjestelmäkohtaisia kolmeen mittaustapaan perustuvia tarkkuusarvoja, tieto käännösten laadusta saatiin 440 käännöksestä.

Aineistona käytetyt virkkeet olisivat olleet aidompia, jos niitä ei olisi luotu itse vaan olisi käytetty esimerkiksi jonkin testikokoelman valmiita tekstejä. Virkkeiden lyhyen ja pitkän version luominen asetti kuitenkin rajoituksia sille, millaisia virkkeitä aineistona voitiin käyttää. Toisaalta kun virkkeet luotiin itse, ne olivat edes osin systemaattisesti muodostettuja. Pitkä virke alkoi aina virkeparin lyhyellä virkkeellä. Lisäksi tarkasteltava monitulkintainen sana oli aina perusmuodossa. Tällä taattiin, ettei konekäännösjärjestelmä saanut sanan muodosta vihjeitä monitulkintaisen sanan merkityksen yksikäsitteistämistä varten. Luoduista virkkeistä saatiin vielä aidompia, kun syntyperäinen englannin kielen osaaja oikoluki käännettävät virkkeet. Tarkastamisen perusteella keinotekoisia virkkeitä saatiin muokattua tai poistettua ja kielioppivirheitä korjattua. Koska monet tutkimuskysymykset perustuivat vertailuun, oleellisinta oli kuitenkin se, että käännettävät virkkeet olivat kaikille konekäännösjärjestelmille tai kääntämismenetelmille samat.

Virkkeiden luomisessa pyrittiin siihen, että kukin tarkasteltavista yksikäsitteisistä sanoista ja monitulkintaisten sanojen merkityksistä esiintyi vain yhdessä virkeparissa. Tällä haluttiin taata se, ettei konekäännösjärjestelmä saisi virheellisen hyvää tarkkuutta, jos se kääntäisi oikein usein esiintyvän sanan tai merkityksen mutta ei välttämättä muita tarkasteltavia sanoja tai merkityksiä. Suurimassa osassa tapauksia tämä tavoite toteutui, mutta muutamassa tapauksessa monitulkintaisen sanan tietty merkitys hyväksyttiin pisteytysvaiheessa oikeaksi useammassa kuin yhdessä virkkeessä, koska konekäännös oli oikea kuvainnollisessa mielessä. Esimerkiksi sanan *beat* virkkeistä kahdessa hyväksyttiin käännökset *lyödä* sekä *hakata*, koska yhdessä virkkeessä ne alun perinkin piti hyväksyä ja toisessa ne piti hyväksyä kuvainnollisen merkityksen takia. Yksikäsitteisten sanojen käännösten laadun vertailussa tulosten luotettavuutta tukee se, että jos virkkeessä oli useampi kuin yksi vaihtoehto tarkasteltavaksi yksikäsitteiseksi sanaksi, tarkasteltava sana valittiin satunnaislukugeneraattorilla.

Tulosten luotettavuutta Sundan vertailussa muihin järjestelmiin saattaa heikentää se, että muista järjestelmistä poiketen Sunda antoi lähes kaikille yksittäin käännettyille monitulkintaisille sanoille monta samanarvoista käännösvaihtoehtoa. Näin ollen Sundan yksittäisten sanojen käännösten pisteytys poikkesi muiden järjestelmien käännösten pisteytyksestä. Tässä tutkimuksessa käytetyllä pisteiden laskutavalla pyrittiin mahdollisimman tasapuoliseen pisteytykseen ja näin ollen myös vertailtavuuteen. Silti jollain toisella laskutavalla saatettaisiin saada erilaisia tuloksia. Erilaisia tuloksia saatettaisiin saada myös silloin, jos oikean käännöksen johdosperheeseen kuuluvat käännökset eivät saisikaan yhtään pistettä. Tässä tutkimuksessa oikean käännöksen kanssa samaan johdosperheeseen kuuluvalla sanalle annettiin 0,5 pistettä, koska kieltenvälisessä tiedonhaussa oikean käännöksen kanssa samaan johdosperheeseen kuuluvat käännökset voivat tuottaa relevantimpia hakutuloksia kuin johdosperheeseen kuulumattomat sanat.

Monitulkintaisten sanojen käännösten laadun vertailun luotettavuutta erilaajuisissa virkkeissä saattaa vähentää se, että virkeparin lyhyen ja pitkän virkkeen pituuksissa ei välttämättä ollut riittävän suurta eroa. Jatkotutkimuksena voitaisiinkin luoda virkkeistä vielä kolmannet versiot, jotka olisivat selvästi pidempiä kuin tämän tutkimuksen pitkät versiot. Yhtenä vaihtoehtona olisi käyttää paikallista kontekstia laajempaa aihekontekstia, jolloin tarkasteltavan monitulkintaisen sanan asiayhteys ei rajoittuisi yhteen virkkeeseen (Turdakov 2010, 312).

TeemaPointin käännosten tulosten luotettavuutta saattaa heikentää se, että käännettävien virkkeiden luomisessa käytettiin WordNetiä, jota TeemaPoint käyttää apunaan kääntämisessä. Näin ollen osa käännettävistä virkkeistä saattoi olla tutumpia TeemaPointille kuin muille järjestelmille. Toisaalta TeemaPoint on sääntöperustainen järjestelmä, joten aineiston perusteella ei tapahdu oppimista niin kuin tilastollisilla konekäännösjärjestelmillä (ks. Carl ym. 2000, 225).

Suurin osa käännoksistä oli yksiselitteisiä pisteyttä. Epäselvien ja tulkinnanvaraisten käännosten pisteytyksen subjektiivisuutta vähensi se, että mielipide pyydettiin vähintään toiselta käännettävät virkkeetkin oikolukeneelta suomalaiselta englannin kielen osaajalta. Kahden tai tarvittaessa kolmen ihmisen tekemillä arvioinneilla taattiin, ettei pisteytys ollut täysin subjektiivista. Silti on huomioitava, että vaikka useampi kuin yksi henkilö on miettinyt käännosten pisteytyksiä, ihmisten tekemät arvoinnit ovat aina subjektiivisia; jotkut toiset olisivat saattaneet päätyä joissain tapauksissa erilaisiin pisteytyksiin ja näin myös erilaisiin tuloksiin. Pisteytyksen ja evaluoinnin luotettavuutta tukee kuitenkin se, kun luvussa 6.4 on tarkasti kuvattu ja esimerkein esitetty, miten aineisto pisteytettiin ja miten tarkkuudet ja prosenttiosuudet laskettiin.

Menetelmien pätevyyttä puoltaa evaluointimittariksi valittu tarkkuus, jolla käännosten laadun evaluointi saatiin kohdennettua vain tarkasteltavien yksikäsitteisten ja monitul-kintaisten sanojen käännoksiin. Esimerkiksi BLEU olisi evaluoinut käännoksista asioita, jotka eivät liity tämän tutkimuksen tutkimuskysymyksiin. Koska mittarina ei käytetty BLEU:n kaltaista mittaria, jossa referenssikäännös on kokonaisuudessaan luotava etukäteen, vielä pisteyttämismvaiheessa oli mahdollista hyväksyä konekäännös, jota ei alun perin oltu osattu ajatella mahdollisena käännosvaihtoehtona. Tämä lisää pisteytyksen luotettavuutta.

## **8.2 Tulosten pohdinta**

Tämän tutkimuksen tulokset vahvistavat, että monitul-kintaisten sanojen kääntäminen konekäännösjärjestelmillä on selvästi haastavampaa kuin yksikäsitteisten sanojen kääntäminen. Myös järjestelmäkohtaisessa tarkastelussa yksikäsitteisten ja monitul-kintaisten sanojen käännosten laadun välinen ero on huomattava. Monitul-kintaisten sanojen kääntämisen onkin sanottu olevan yksi kieltenvälisen tiedonhaun suurimmista haasteita (Bal-leston & Croft 1998, 69; Kishida 2005, 439).



Chen ja Geyn (2004, 150) mukaan konekääntäminen ei ole ihanteellinen menetelmä kyselyiden kääntämiselle, ellei kyselyitä ole muotoiltu kieliopillisesti oikeiksi virkkeiksi. Konekäännösjärjestelmät on yleensä optimoitu kääntämään virkkeitä eikä niinkään yksittäisiä sanoja (Zhu & Wang 2006, 599). Tässä tutkimuksessa tarkasteltava monitulkintainen sana oli osa kieliopillisesti oikeaa, kokonaista virkettä, mutta käännöslaatu oli silti heikkoa. On kuitenkin muistettava, että tässä tutkimuksessa tarkasteltiin haastavien monitulkintaisten sanojen kääntämistä eikä esimerkiksi koko virkkeen kääntämistä. Seuraavaksi voitaisiin tutkia, onko monitulkintaisten sanojen käännösten laadussa eroja käännettäessä kokonainen virke tai erillisistä sanoista muodostuva kysely. Käännettävä virke päättyisi pisteeseen mutta kysely ei. Esimerkkinä virkkeestä voisi olla ”*I play the flute in the orchestra.*” ja kyselystä *play flute orchestra*. Ideana olisi selvittää, tarvitseeko monitulkintaisen sanan oikea käännös virkerakenteen ympärilleen vai riittävätkö pelkät merkityksestä vihjeitä antavat sanat. Kieltenvälisessä tiedonhaussa kun on oleellista, että konekäännösjärjestelmät osaisivat kääntää myös monitulkintaisia sanoja, jotka eivät ole osana virkettä, koska harvoin käyttäjän hakukoneeseen syöttämä kysely on virke. Esimerkiksi yksittäisistä sanoista koostuvan kyselyn sanojen sanajärjestyksellä ei pitäisi olla vaikutusta haun tuloksellisuuteen. Onkin oleellista erottaa kaksi eri kääntämistarkoitusta kieltenvälisessä tiedonhaussa: yksittäisistä sanoista koostuvan kyselyn kääntäminen ja virkkeiden tai tekstin kääntäminen. Tiedonhaun määritelmästä riippuen tiedonhakuna voidaan pitää pelkästään sitä, kun käyttäjä kääntää virkkeen konekäännösjärjestelmällä.

Tässä tutkimuksessa monitulkintaisen sanan sisältävät virkkeet olivat yksinkertaisia perusvirkeitä yksinkertaisilla lauserakenteilla. Yhtenä jatkotutkimusideana olisi selvittää, kuinka laadukkaasti konekäännösjärjestelmät kääntävät monitulkintaiset sanat osana rakenteeltaan monimutkaisia virkeitä. Voitaisiin tarkastella myös aineistona olleiden virkkeiden käännöslaatua koko virkkeen kääntämisen osalta eikä ainoastaan monitulkintaisten sanojen kääntämisen osalta.

Kaikkien käännösten tarkkuuksien perusteella monitulkintaisten sanojen kääntämisessä Sunda ja Google Translate ovat parhaat järjestelmät. Huonoimmat järjestelmät ovat puolestaan SDL Free Translation ja Bing<sup>TM</sup> Translator Beta. TeemaPoint jää parhaimpien ja huonoimpien järjestelmien välimaastoon. Järjestelmien keskinäisessä paremmuusjärjestyksessä on eroja riippuen tarkasteltavan sanan asiayhteyden laajuudesta. Sunda on paras järjestelmä ja Bing<sup>TM</sup> Translator Beta sekä SDL Free Translation huonoimmat,

kun monitulkintaiset sanat käännetään pitkissä virkkeissä. Sunda on paras ja Bing<sup>TM</sup> Translator Beta huonoin, kun monitulkintaiset sanat käännetään lyhyissä virkkeissä. Sen sijaan kun monitulkintaiset sanat käännetään yksittäisinä sanoina, parhaat järjestelmät ovat Google Translate sekä SDL Free Translation ja huonoimmat TeemaPoint sekä Sunda. Suurimpana erona virkeyhteydessä tehtyjen käännösten tarkkuuksien mukaiseen paremmuusjärjestykseen on se, että SDL Free Translation ja Bing<sup>TM</sup> Translator Beta ovat kärkipäässä ja Sunda sekä TeemaPoint häntäpäässä. TeemaPointin yksittäin kääntämien sanojen käännöslaatu vähentää se, kun järjestelmä ei ehdota kaikille yksittäin käännettyille monitulkintaisille sanoille käännöstä. Sundan yksittäin kääntämien sanojen tarkkuutta vähentää se, kun järjestelmä tarjoaa lähes kaikille sanoille monta käännösvaihtoehtoa. Näin ollen käännösvaihtoehtojen joukossa on enemmän kuhunkin asiayhteyteen kuuluvia vääriä kuin oikeita käännöksiä.

Sundassa yksittäisten sanojen kääntämisellä on ilmeisesti tarkoitettu sanakirjatyypistä kääntämistä. Tarkoituksena ei ehkä olekaan tarjota vain yhtä käännösvaihtoehtoa, vaan käyttäjälle on haluttu esittää kaikki mahdolliset sanan käännökset. Kieltenvälisessä tiedonhaussa monen eri käännösvaihtoehdon sisällyttäminen kyselyyn saattaa joko heikentää tai parantaa hakutuloksia. Kuten sanakirjaperustaisen menetelmän esittelyn yhteydessä mainittiin, monien käännösvaihtoehtojen sisällyttäminen kyselyyn saattaa lisätä monitulkintaisuusongelmaa (ks. Ballesteros & Croft 1996, 800; Ballesteros & Croft 1997, 84). Pirkolan (1998b, 56) rakenteisen kyselyn malli kuitenkin auttaa tässä monitulkintaisuusongelmassa. Toisaalta taas ainoastaan yhden käännösvaihtoehdon kyselyyn sisällyttämisen vaarana on, että oikea käännös kadotetaan (Airio 2009, 30).

Tässä tutkimuksessa testatuista konekäännösjärjestelmistä osa oli samoja kuin Toralin ja muiden (2011, 17) tutkimuksessa testatuista järjestelmistä. Yhteisiä järjestelmiä olivat Google Translate, SDL Free Translation sekä Bing<sup>TM</sup> Translator Beta. Toralin ja muiden (2011,19) tutkimuksen perusteella Google Translate oli paras järjestelmä. Myös tässä tutkimuksessa kaikkien käännösten perusteella Google Translate oli paras niistä järjestelmistä, jotka olivat yhteisiä Toralin ja muiden (2011, 19) tutkimuksen konekäännösjärjestelmien kanssa. Tässä tutkimuksessa Bing<sup>TM</sup> Translator Betan käännöslaatu poikkeaa Toralin ja muiden (2011, 19) tutkimustuloksista. Kyseisessä tutkimuksessa Bing<sup>TM</sup> Translator Beta oli käännösten laadun perusteella selvästi parempi kuin SDL Free Translation (Toral ym. 2011, 19), kun taas tämän tutkimuksen perusteella näiden kahden järjestelmän käännöslaatu on suunnilleen yhtä heikkoa.

Edellä olevia vertailuja tehtäessä on kuitenkin muistettava, että Toralin ja muiden (2011, 13) tutkimuksessa käännettiin saksasta, italiasta ja hollannista englanniksi, kun taas tässä tutkimuksessa käännettiin englannista suomeksi. Lisäksi käännettävä aineisto ja tutkimuskysymykset poikkesivat toisistaan. Toralin ja muiden (2011, 13) tutkimuksessa käännettävä aineisto oli uutisaineistoa, eivätkä tarkasteltavina olleet pelkät monitulkintaiset sanat.

Kit ja Wong (2008, 312) mainitsevat, että konekäännösjärjestelmien kääntämisosat eri kieliparille ovat ikään kuin pienoiskonekäännösjärjestelmiä. Ne eivät välttämättä toimi täysin samalla tavalla tai samalla teknologialla, eikä niillä välttämättä ole samanarvoisia kielellisiä resursseja käytössä. Nämä resurssierot saattavat aiheuttaa eroja eri kieliyhdistelmien käännöslaadun välille. (Kit & Wong 2008, 312.) Testattujen konekäännösjärjestelmien resursseissa on todennäköisesti eroja. Toisaalta suomalaiset maksulliset Sunda ja TeemaPoint voivat keskittyä ulkomaalaisia, monikielisiä järjestelmiä paremmin nimenomaan englannin- ja suomenkielisten käännösten kehittämiseen. Toisaalta taas ulkomaalaisilla jättyyhtiöillä, kuten Googlella ja Microsoftilla, luulisi olevan enemmän resursseja käytössään kuin suomalaisia järjestelmiä kehittäville yhtiöillä. Esimerkiksi Toral ja muut (2011, 18) mainitsevat, että evaluointitutkimuksen tuloksia tarkasteltaessa on huomioitava, että Bingin konekäännösjärjestelmää ja Google Translatea on kehitetty useiden vuosien ajan ja niillä on ollut käytössään valtava määrä resursseja. Hieman ristiriitaista juuri mainitun kommentin kanssa on se, että Bing<sup>TM</sup> Translator Beta on vasta kokeiluversio. Kokeiluversio saattaa osin selittää kyseisen konekäännösjärjestelmän huonoja tuloksia tässä tutkimuksessa.

Vaikka samankaan järjestelmän eri kieliyhdistelmillä tehtyjä käännöksiä ei voida täysin luotettavasti vertailla keskenään, jatkotutkimuksena voitaisiin kuitenkin kääntää tämän tutkimuksen aineisto muillekin kuin suomen kielelle. Testattavia järjestelmiä saataisiin tätä tutkimusta enemmän, koska suomen kieli ei rajoittaisi mahdollisten konekäännösjärjestelmien valintaa.

Tässä tutkimuksessa tarkasteltavana oli kolme erilaajuista asiayhteyttä: monitulkintaisen sanojen kääntäminen yksittäisinä sanoina sekä osana lyhyitä ja pitkiä virkkeitä. Yksittäisinä sanoina käännettyinä monitulkintaisten sanojen käännösten laatu on huomattavasti heikompaa kuin lyhyissä tai pitkissä virkkeissä käännettyinä. Tulos on itsestään selvä, koska ihminenkin ei ilman monitulkintaisen sanan asiayhteyttä tiedä, pitäisikö

yksin esiintyvä sana *play* suomentaa sanaksi *pele*, *leikki* vai *näytelmä*. Kieltenvälisessä tiedonhaussa saattaa kuitenkin esiintyä tilanteita, joissa käyttäjä hakee vain yksittäisellä monitulkintaisella sanalla ja saa hakutuloksena paljon tarkoittamaansa merkitykseen liittymättömiä dokumentteja.

Tämän tutkimuksen perusteella monitulkintaisten sanojen kääntäminen pitkissä virkeyhteysissä ei ole laadukkaampaa kuin lyhyissä. Tämä tutkimustulos poikkeaa monesta aiemmasta tutkimuksesta, joiden perusteella konekäännösjärjestelmän käännökset ovat pitkissä asiayhteyksissä laadukkaampia kuin lyhyissä, koska konekääntäminen tarvitsee oikean käännöksen päättämiseksi tietoa sanan asiayhteydestä (ks. Chen & Gey 2004, 150; Oard & Hackett 1998, 1). Sen sijaan tutkimustulos on samansuuntainen Kettusen (2009, 112) tutkimustulosten kanssa, koska Kettusen mukaan lyhyiden ja pitkien kyselyiden käännösten laadussa ei juurikaan ole eroja. Tutkimustuloksia vertailtaessa on kuitenkin muistettava, että tässä tutkimuksessa tarkasteltiin pelkästään monitulkintaisten sanojen käännösten laatua eikä koko virkkeen tai kyselyn käännöslaatua, joka monissa tutkimuksissa on ollut tarkastelun kohteena. Myös aineisto ja evaluointimittarit ovat edellä mainituissa tutkimuksissa muut kuin tässä tutkimuksessa käytetyt virkkeet ja mittarina käytetty tarkkuus.

Tämän tutkimuksen perusteella sekä sääntöperustaisilla että tilastollisilla konekäännösjärjestelmillä käännettäessä monitulkintaisten sanojen kääntäminen yksittäisinä sanoina on huomattavasti heikkolaatuisempaa kuin niiden kääntäminen virkeyhteydessä. Kummankaan kääntämismenetelmän käännöslaadussa ei ole eroja monitulkintaisten sanojen kääntämisessä lyhyessä tai pitkässä virkkeessä.

Tämän tutkimuksen kaikkien käännösten perusteella sääntöperustaisien ja tilastollisten konekäännösjärjestelmien käännösten laadussa ei ole eroa monitulkintaisten sanojen kääntämisessä. Näin ollen tämän tutkimuksen tulokset poikkeavat selvästi Toralin ja muiden (2011, 19) tutkimustuloksista, joiden mukaan viidestä testatusta konekäännösjärjestelmästä tilastolliset järjestelmät olivat kaikilla kielipareilla ja kaikilla evaluointimittareilla mitattuina laadultaan parempia kuin sääntöperustaiset. Tässä tutkimuksessa järjestelmien keskinäisessä paremmuusjärjestyksessä sääntöperustaiset ja tilastolliset konekäännösjärjestelmät ovat sen sijaan keskenään sekaisin. Jälleen on kuitenkin huomioitava, että käytetyt kieliyhdistelmät ja aineisto sekä tutkimuskysymykset poikkesivat Toralin ja muiden (2011, 14) tutkimuksen vastaavista. Lisäksi kääntämismenetelmän

ohella monitulkintaisten sanojen käännösten laatuun vaikuttaa sanan merkityksen yksikäsitteistämässä käytettävät menetelmät.

Tilastolliset konekäännösjärjestelmät kääntävät sääntöperustaisia järjestelmiä laadukkaammin yksin esiintyviä monitulkintaisia sanoja, kun taas sääntöperustaiset järjestelmät kääntävät tilastollisia järjestelmiä laadukkaammin virkeyhteydessä esiintyviä monitulkintaisia sanoja. Tulos on osin yllättävä, koska tilastollisten menetelmien luulisi tarvitsevan enemmän tietoa asiayhteydestä kuin sääntöperustaisten järjestelmien. Toisaalta taas tilastolliset konekäännösjärjestelmät käyttävät sanojen esiintymisiin perustuvia tilastotietoja apunaan käännöksen päättämässä, joten luulisi, että yksittäiselle sanalle tarjotaan käännökseksi todennäköisemmin sanan yleinen kuin harvinainen käännös. Käännytyissä virkkeissä esiintyi enimmäkseen monitulkintaisten sanojen yleisiä peruserkityksiä, joten yleisinkin käännösvaihtoehto varmasti esiintyi. Sääntöperustaisten konekäännösjärjestelmien yksittäin käännettyjen sanojen käännösten laatua vähentää se, kun TeemaPoint ei ehdota käännöstä kaikille yksittäin käännettäville sanoille ja Sunda puolestaan ehdottaa useita käännösvaihtoehtoja.

Konekäännösjärjestelmien tekemät virheet monitulkintaisten sanojen kääntämisessä voidaan luokitella kuuteen virhetyyppiin. Selvästi yleisin virhe virkkeessä esiintyvien monitulkintaisten sanojen kääntämisessä on se, että monitulkintaisen sanan käännökseksi valitaan väärä vaihtoehto sanan useasta käännösvaihtoehdosta. Toiseksi yleisin virhetyyppi on se, että monitulkintainen sana jää kokonaan kääntämättä. Loppuja virhetyyppejä edustaa selvä vähemmistö, eikä näiden virhetyyppien välillä ole keskenään suuria eroja. Näihin harvinaisiin virhetyyppeihin kuuluvat seuraavat virheet: Käännös on oikean käännöksen kanssa samaan johdosperheeseen kuuluva mutta merkitykseltään väärä käännös, käännöksenä ei ole mikään monitulkintaisen sanan käännösvaihtoehdoista tai monitulkintaisen sanan käännös puuttuu koko käännöksestä. Kuudes virhetyyppi sisältää jo mainittuihin virhetyyppeihin kuulumattomia virheitä, joita esiintyy kaikista vähiten.

Eri konekäännösjärjestelmien tekemien virhetyyppien osuuksien jakaumien välillä on eroja. Kunkin järjestelmän virheistä enemmistö selittyy sillä, että monitulkintaisen sanan käännökseksi valitaan väärä kyseisen sanan käännösvaihtoehdoista. TeemaPointin virhetyyppien jakauma kuitenkin poikkeaa kaikkien muiden järjestelmien jakaumista. Väärä käännösvaihtoehto selittää TeemaPointin virheistä selvästi pienemmän osuuden

kuin muiden järjestelmien virheistä. TeemaPoint erottuu myös siinä, että sen virheistä selvästi enemmän selittyy sanan kääntämättä jäämisellä kuin muiden järjestelmien virheistä. Bing™ Translator Betan virhetyyppien jakaumassa on puolestaan eroja verrattuna muiden kuin TeemaPointin jakaumaan: Bing™ Translator Betan virheistä selittyy muita järjestelmiä enemmän sillä, ettei monitulkintaisen sanan käänös ole mikään sanan käänösvaihtoehtoista tai että monitulkintaisen sanan käänös puuttuu käänöksestä.

Myös pieni osa toisen tilastollisen konekäännösjärjestelmän, Google Translaten, käänösten virheistä selittyy sillä, että tarkasteltavan sanan käänös puuttuu käännetystä virkkeestä. Sen sijaan ainakin tämän tutkimuksen perusteella sääntöperustaisten järjestelmien käänöksistä on aina tunnistettavissa, mikä sana käänöksessä vastaa käännettävän virkkeen tarkasteltavaa monitulkintaista sanaa. Menetelmien erot tässä virhetyypissä saattavat selittyä sillä, kun sääntöperustaiset menetelmät jäsentävät käännettävän tekstin ennen kääntämistä.

Kaiken kaikkiaan tämän tutkimuksen tulokset osoittavat, että ainakaan testattujen konekäännösjärjestelmien käänöksiin monitulkintaisen sanojen kääntämisessä ei voida luottaa. Kaikkien konekäännösjärjestelmien kaikkien käänösten keskimääräinen tarkkuus monitulkintaisten sanojen kääntämisessä on 0,470. Tämä tarkoittaa sitä, että on todennäköisempää, että monitulkintainen sana käännetään väärin kuin oikein. Kitin ja Wongin (2008, 320) mukaan konekäännösjärjestelmien käänöslaatu onkin yhä kaukana tyydyttävästä, puhumattakaan täydellisestä ja julkaisukelpoisesta. Tämän tutkimuksen tulokset vahvistavat, että ainakin monitulkintaisten sanojen kääntämisessä konekäännöksen tarjoama käänös on vain raakakäännös, joka käyttäjän on itse osattava muokata. Käyttäjän on siis itse tiedettävä, että virke ”*I play the flute.*” ei tarkoita ”*Pelaan huilua.*” eikä ”*Esitän huilua.*”.

## LÄHTEET

- Airio, E. (2008). Who benefits from CLIR in web retrieval? *Journal of Documentation*, 64, (5), 760-778. Saatavilla Emerald-tietokannasta: <<http://www.emeraldinsight.com/>> (käytetty 17.10.2011).
- Airio, E. (2009). Morphological problems in IR and CLIR: applying, linguistic methods and approximate string matching tools. (Acta Electronica Universitatis Tamperensis 842). Väitöskirja. Tampereen yliopisto, Informaatiotutkimuksen ja interaktiivisen median laitos. <<http://acta.uta.fi/pdf/978-951-44-7708-9.pdf>> (käytetty 28.9.2011).
- Arnold, D. (2003). Why translation is difficult for computers. Teoksessa H. Somers (toim.), *Computers and translation. A translator's guide*, 119-142. Amsterdam: Benjamins.
- Aurélien, M., Rafik, M., & Langlais, P. (2008). Explorations in using grammatical dependencies for contextual phrase translation disambiguation. Teoksessa *Proceedings of 12th annual conference of the European Association for Machine Translation*, Hampuri, Saksa, 22.-23.9.2008, 114-119. <<http://www.mt-archive.info/EAMT-2008-Max.pdf>> (käytetty 3.4.2011).
- Ballesteros, L., & Croft, B. (1996). Dictionary methods for cross-lingual information retrieval. Teoksessa W. Roland, & T. Helmut (toim.), *Database and Expert Systems Applications*. (Lecture Notes in Computer Science 1134, 791-801). Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 10.10.2011).
- Ballesteros, L., & Croft, B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. Teoksessa *Proceedings of the 20th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Philadelphia, Yhdysvallat, 27.-31.7.1997, 84-91. Saatavilla ACM Digital Library -tietokannasta: <<http://dl.acm.org/>> (käytetty 23.9.2011).

- Ballesteros, L., & Croft, B. (1998). Resolving ambiguity for cross-language retrieval. Teoksessa *Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Melbourne, Australia, 24.-28.8.1998, 64-71. Saatavilla ACM Digital Library - tietokannasta: <<http://dl.acm.org/>> (käytetty 22.9.2011).
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Teoksessa *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*, Ann Arbor, Yhdysvallat, 25.-30.6.2005, 65-72. <[aclweb.org/anthology-new/W/W05/W05-09.pdf](http://aclweb.org/anthology-new/W/W05/W05-09.pdf)> (käytetty 18.10.2011).
- Bennett, S., & Gerber, L. (2003). Inside commercial machine translation. Teoksessa H. Somers (toim.), *Computers and translation. A translator's guide*, 173-190. Amsterdam: Benjamins.
- Bing Translatorin verkkosivut. (2011). <[www.microsofttranslator.com](http://www.microsofttranslator.com)>
- Boughanem, M., & Nassr, N. (2001). Mercure at CLEF-1. Teoksessa C. Peters (toim.), *Cross-language information retrieval and evaluation*. (Lecture Notes in Computer Science 2069, 202-209). Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 14.10.2011).
- Boughanem, M., Chrisment, C., & Nassr, N. (2002). Investigation on disambiguation in CLIR: aligned corpus and bi-directional translation-based strategies. Teoksessa M. Braschler, J. Gonzalo, & M. Kluck (toim.), *Evaluation of Cross-Language Information Retrieval Systems*. (Lecture Notes in Computer Science 2406, 158-168). Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 17.10.2011).
- Braschler, M., & Schäuble, P. (2000). Using corpus-based approaches in a system for multilingual information retrieval. *Information Retrieval*, 3, (3), 273-284. Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 30.9.2011).



- Brunning, J., de Gispert, A., & Byrne, W. (2009). Context-dependent alignment models for statistical machine translation. Teoksessa *Proceedings of Human Language Technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Yhdysvallat, 31.5-5.6.2009, 110-118. <<http://www.mt-archive.info/NAACL-HLT-2009-Brunning.pdf>> (käytetty 1.4.2011).
- Carl, M., Pease, C., Iomdin, L., & Streiter, O. (2000). Towards a dynamic linkage of example-based and rule-based machine translation. *Machine Translation*, 15, (3), 233-257. Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 8.12.2011).
- Carpuat, M., & Wu, D. (2007a). How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. Teoksessa *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, Skövde, Ruotsi 7.-9.9.2007, 43-52. <<http://www.mt-archive.info/TMI-2007-Carpuat.pdf>> (käytetty 1.4.2011).
- Carpuat, M., & Wu, D. (2007b). Context-dependent phrasal translation lexicons for statistical machine translation. Teoksessa *Proceedings of MT Summit XI*, Kööpenhamina, Tanska, 10.-14.9.2007, 73-80. <<http://www.mt-archive.info/MTS-2007-Carpuat.pdf>> (käytetty 2.4.2011).
- Cer, D., Manning, C.D., & Jurafsky, D. (2010). The best lexical metric for phrase-based statistical MT system optimization. Teoksessa *Proceedings of 2010 annual conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, Los Angeles, Yhdysvallat, 2.-4.6.2010, 555-563. <<http://www.mt-archive.info/NAACL-HLT-2010-Cer-1.pdf>> (käytetty 20.10.2011).
- Chen, A., & Gey, F. (2004). Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7, (1-2), 149-182. Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 30.9.2011).

- Croft, B., Metzler, D., & Strohman, T. (2010). *Search Engines: Information Retrieval in Practice*. Boston: Addison-Wesley.
- Davis, M. (1996). New experiments in cross-language text retrieval at NMSU's computing research lab. Teoksessa *Proceedings of TREC-5*, Gaithersburg, Yhdysvallat, 20.-22.11.1996.  
<[http://trec.nist.gov/pubs/trec5/t5\\_proceedings.html](http://trec.nist.gov/pubs/trec5/t5_proceedings.html)> (käytetty 10.10.2011).
- Davis, M., & Ogden, W. C. (1997). QUILT: implementing a large-scale cross-language text retrieval system. Teoksessa *Proceedings of the 20th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Philadelphia, Yhdysvallat, 27.-31.7.1997, 92–98. Saatavilla ACM Digital Library -tietokannasta: <<http://dl.acm.org/>> (käytetty 10.10.2011).
- Dorr, B., Hovy, E., & Levin, L. (2006). Machine Translation: Interlingual methods. Teoksessa K. Brown (toim.), *Encyclopedia of language & linguistics*, 7, 383-394. Oxford: Elsevier.
- Fujii, A., & Ishikawa, T. (2000). Applying machine translation to two-stage cross-language information retrieval. Teoksessa *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas*, Cuernavaca, Meksiko, 10.-14.10.2000, 13-24. Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 30.9.2011).
- Gale, W.A., Church, K.W., & Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26, (5-6), 415-439. Saatavilla SpringerLink-tietokannasta:  
<<http://www.springerlink.com/>> (käytetty 10.10.2011).
- Gaspari, F. (2006). The added value of free online MT services: confidence boosters for linguistically-challenged Internet users, a case study for the language pair Italian-English. Teoksessa *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Yhdysvallat,

8.-12.8.2006, 46-55 <<http://www.mt-archive.info/AMTA-2006-Gaspari.pdf>> (käytetty 29.12.2011).

Google Translaten verkkosivut. (2011). <[translate.google.com](http://translate.google.com)>

Hedlund, T. (2003). *Dictionary-based cross-language information retrieval*. (Acta Universitatis Tamperensis 962). Väitöskirja. Tampereen yliopisto, Informaatiotutkimuksen laitos. <<http://acta.uta.fi/pdf/951-44-5790-0.pdf>> (käytetty 30.9.2011).

Huang, R., Sun, L., Li, J., Pan, L., & Zhang, J. (2007). ISCAS in CLIR at NTCIR-6: experiments with MT and PRF. Teoksessa *Proceedings of NTCIR-6 Workshop*, Tokio, Japani, 15.-18.5.2007, 81-84. <<http://research.nii.ac.jp/ntcir/ntcir-ws6/OnlineProceedings/NTCIR/55.pdf>> (käytetty 14.10.2011).

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. Teoksessa *Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Pittsburgh, Yhdysvallat, 27.6.-1.7.1993, 329-338. Saatavilla ACM Digital Library - tietokannasta: <<http://dl.acm.org/>> (Käytetty 13.1.2012).

Hull, D. (1997). Using structured queries for disambiguation in cross-language information retrieval. Teoksessa *Working notes of the AAAI Spring Symposium of Cross-Language Text and Speech Retrieval*, Stanford, Yhdysvallat 24.-26.3.1997, 84-98. <[www.aaai.org/Papers/Symposia/Spring/1997/SS-97-05/SS97-05-012.pdf](http://www.aaai.org/Papers/Symposia/Spring/1997/SS-97-05/SS97-05-012.pdf)> (käytetty 29.12.2011).

Hutchins, J. (2005a). *Machine translation: a concise history*. <<http://www.hutchinsweb.me.uk/history.htm>> (käytetty 11.10.2011).

Hutchins, J. (2005b). Towards a definition of example-based machine translation. Teoksessa *Proceedings of the 2nd Workshop on Example-Based Machine Translation*, Phuket, Thaimaa, 16.9.2005, 63-70. <<http://www.mt-archive.info/MTS-2005-Hutchins.pdf>> (käytetty 1.4.2011).

- Hutchins, J., & Somers, H. L. (1992). *An introduction to machine translation*. Lontoo: Academic Press. <<http://www.hutchinsweb.me.uk/IntroMT-TOC.htm>> (käytetty 12.10.2011).
- Kaplan, A. (1955). An experimental study of ambiguity and context. *Mechanical Translation*, 2, (2), 39-46. <<http://www.mt-archive.info/MT-1955-Kaplan.pdf>> (käytetty 16.10.2011).
- Kettunen, K. (2007). Reductive and generative approaches to morphological variation of keywords in monolingual information retrieval. (Acta Electronica Tampereensis 654). Väitöskirja. Tampereen yliopisto, Informaatiotutkimuksen laitos. <<http://acta.uta.fi/pdf/978-951-44-7088-2.pdf>> (käytetty 22.4.2012).
- Kettunen, K. (2009). Facing the machine translation Babel in CLIR - can MT metrics help in choosing CLIR resources? Teoksessa M. A. Klopotek, A. Przepiórkowski, S.T. Wierzchon, & K. Trojanowski (toim.), *Recent advances in intelligent information systems*, 103-116. <<http://www.mt-archive.info/IIS-2009-Kettunen.pdf>> (käytetty 15.9.2011).
- Kilgarriff, A., & Palmer, M. (2000). Introduction to the special issue on SENSEVAL. *Computers and the Humanities*, 34, (1-2), 1-13. Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 7.12.2011).
- Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. *Information Processing and Management*, 41, (3), 433-455. Saatavilla EBSCO-tietokannasta: <<http://www.ebsco.com>> (käytetty 14.9.2011).
- Kit, C., & Wong, T. M. (2008). Comparative evaluation of online machine translation systems with legal texts. *Law Library Journal*, 100, (2), 299-321. <[http://www.aallnet.org/products/pub\\_llj\\_v100n02/2008-16.pdf](http://www.aallnet.org/products/pub_llj_v100n02/2008-16.pdf)> (käytetty 29.12.2011).
- Knight, K. (1997). Automating knowledge acquisition for machine translation. *AI Magazine*, 18, (4), 81-96. <<http://www.mt-archive.info/AIMag-1997-Knight.pdf>> (käytetty 4.12.2011).

- Knight, K., & Koehn, P. (2003). SMT Tutorial. <<http://www.statmt.org/>> (käytetty 27.11.2011).
- Laininen, P. (2000). *Tilastollisen analyysin perusteet*. 597. 3. painos. Helsinki: Otatieto.
- Lavie, A., & Denkowski, M. (2009). The METEOR metric for automatic evaluation of machine translation. Hollanti: Kluwer Academic Publishers.  
<<http://www.cs.cmu.edu/~alavie/METEOR/pdf/meteor-mtj-2009.pdf>> (käytetty 16.10.2011).
- Lehtokangas, R., Airio, E., & Järvelin, K. (2004). Transitive dictionary translation challenges direct dictionary translation in CLIR. *Information Processing & Management*, 40, (6), 973-988.  
<[http://www.info.uta.fi/tutkimus/fire/archive/RL\\_EA\\_KJ\\_IPM\\_2004.pdf](http://www.info.uta.fi/tutkimus/fire/archive/RL_EA_KJ_IPM_2004.pdf)> (käytetty 28.12.2011).
- Lopez, A. (2008). Statistical machine translation. *Computing Surveys*, 40, (3), 1-49.  
Saatavilla ACM Digital Library -tietokannasta: <<http://dl.acm.org/>> (käytetty 3.12.2011).
- Manninen, P., & Ylén, M. (2001). *Tilastollisen päättelyn käytäntö. Tilastotiedettä sovel-tajalle*. 2. painos. Tampere: TI-Tutkimuspalvelu.
- Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language pro-cessing*. 3. painos. Cambridge: MIT Press.
- McNamee, P., & Mayfield, J. (2002). Comparing cross-language query expansion tech-niques by degrading translation resources. Teoksessa *Proceedings of the 25th annual International ACM SIGIR conference on Research and De-velopment in Information Retrieval*, Tampere, Suomi, 11.-15.8.2002, 159-166. Saatavilla ACM Digital Library -tietokannasta: <<http://dl.acm.org/>> (käytetty 30.9.2011).
- Melamed, I. D. (1995). Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. Teoksessa *Proceedings of the 3rd workshop on Very Large Corpora*, Cambridge, Yhdysvallat, 30.6.1995, 184-198.

<<http://www.mt-archive.info/VLC-1995-Melamed.pdf>> (käytetty 3.12.2011).

Melamed, I. D., Green, R., & Turian, J. P. (2003). Precision and recall of machine translation. Teoksessa *Proceedings of the 2003 annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Kanada, 27.5.-1.6.2003, 61-62. Saatavilla ACM Digital Library -tietokannasta: <<http://dl.acm.org/>> (käytetty 4.12.2011).

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys*, 41, (2), 1-69. Saatavilla ACM Digital Library -tietokannasta: <<http://dl.acm.org/>> (käytetty 1.4.2011).

Nie, J. Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. Teoksessa *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Berkeley, Yhdysvallat, 15.-19.8.1999, 74-81. Saatavilla ACM Digital Library -tietokannasta: <<http://dl.acm.org/>> (käytetty 30.9.2011).

Oard, D., & Dorr, B. (1996). A survey of multilingual text retrieval. *Technical Report UMIACS-TR-96-19*. Marylandin yliopisto. <<http://terpconnect.umd.edu/~oard/pdf/cstr3615.pdf>> (käytetty 30.9.2011).

Oard, D. (1998). A comparative study of query and document translation or cross-language information retrieval. Teoksessa D. Farwell, L. Gerber, & E. Hovy (toim.), *Machine Translation and the Information Soup*. (Lecture Notes in Computer Science 1529, 472-483). Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 14.10.2011).

Oard, D., & Hackett, P. (1998). Document translation for cross-language text retrieval at the University of Maryland. Teoksessa *Proceedings of TREC-6*, Gaithersburg, Yhdysvallat, 19.-21.11.1997, 1-10. <[http://academiccommons.columbia.edu/download/fedora\\_content/download/ac:133485/CONTENT/trec97.pdf](http://academiccommons.columbia.edu/download/fedora_content/download/ac:133485/CONTENT/trec97.pdf)> (käytetty 30.9.2011).

- Pirkola, A. (1998a). Kyselyrakenteiden ja erikoissanakirjan vaikutus sanakirjakäännökseen perustuvassa kieltenvälisessä tiedonhaussa. *Informaatiotutkimus*, 17, (3), 48-58. Saatavilla Tieteelliset verkkolehdet -tietokannasta: <<http://ojs.tsv.fi/index.php/index/index>> (käytetty 27.9.2011).
- Pirkola, A. (1998b). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. Teoksessa *Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Melbourne, Australia, 24.-28.8.1998, 55-63. <[http://www.info.uta.fi/tutkimus/fire/archive/the\\_effects\\_of\\_query.pdf](http://www.info.uta.fi/tutkimus/fire/archive/the_effects_of_query.pdf)> (käytetty 10.1.2012).
- Pirkola, A., Hedlund, T., Keskustalo, H., & Järvelin, K. (2001). Dictionary-based crosslanguage information retrieval: problems, methods, and findings. *Information Retrieval*, 4, (3-4), 209-230. <[http://www.info.uta.fi/tutkimus/fire/archive/dictionary\\_based.pdf](http://www.info.uta.fi/tutkimus/fire/archive/dictionary_based.pdf)> (käytetty 30.9.2011).
- Rosembat, G., Gemoets, D., Browne, A. C., & Tse, T. (2003). Machine translation supported cross-language information retrieval for a consumer health resource. Teoksessa *Proceedings of annual American Medical Informatics Association Symposium*, Washington, Yhdysvallat, 8.-12.11.2003, 564-568. <[http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480254/pdf/amia2003\\_0564.pdf](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480254/pdf/amia2003_0564.pdf)> (käytetty 30.9.2011).
- SDL Free Translation verkkosivut. (2011). <[www.freetranslation.com](http://www.freetranslation.com)>
- Senseval-3 aineisto. (2004). <<http://www.senseval.org/senseval3/tasks.html>> (käytetty 6.11.2011).
- Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, 14, (2), 113-157. Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 1.12.2011).

- Sparck Jones, K. (1974). Automatic indexing. *Journal of Documentation*, 30, (4), 393-432.
- Sundan verkkosivut. (2011). <[www.sunda.fi/kaantaja.html](http://www.sunda.fi/kaantaja.html)>
- Talvensaari, T. (2008a). *Comparable corpora in cross-language information retrieval*. (Acta Electronica Universitatis Tamperensis 779). Väitöskirja. Tampereen yliopisto, Tietojenkäsittelytieteiden laitos. <<http://acta.uta.fi/pdf/978-951-44-7490-3.pdf>> (käytetty 4.10.2011).
- Talvensaari, T. (2008b). Effects of aligned corpus quality and size in corpus-based CLIR. Teoksessa C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, R. W. White (toim.), *Advances in Information Retrieval*. (Lecture Notes in Computer Science 4956, 114-125). Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 30.9.2011).
- Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., & Laurikkala, J. (2008). Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11, (5), 427-445. Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 10.10.2011).
- TeemaPointin verkkosivut. (2011). <[www.teemapoint.com](http://www.teemapoint.com)>
- Toral, A., Gaspari, F., Naskar, S. K., & Way, A. (2011). Comparative evaluation of re-search vs. online MT systems. Teoksessa *Proceedings of the 15th International Conference of the European Association for Machine Translation*, Leuven, Belgia, 30.-31.5.2011, 13-20. <<http://www.ccl.kuleuven.be/EAMT2011/proceedings/pdf/eamt2011proceedings.pdf>> (käytetty 29.12.2011).
- Turdakov, D. Y. (2010). Word sense disambiguation methods. *Programming and Computer Software*, 36, (6), 309-326. Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 14.9.2011).
- Vauquois, B. (1976). Automatic translation - a survey of different approaches. *Statistical Methods in Linguistics*, 127-135. <<http://www.mt-archive.info/SMIL-1976-Vauquois.pdf>> (käytetty 13.12.2011).



- WordNet. (2011). WordNetin verkkosivu <<http://wordnet.princeton.edu/>> (käytetty 4.12.2011).
- Wu, D., He, D., Ji, H., & Grishman, R. (2008). A study of using an out-of-box commercial MT system for query translation in CLIR. Teoksessa *Proceeding of the 2nd ACM workshop on Improving non English web searching*, Napa Valley, Yhdysvallat, 30.8.2008, 71-76. Saatavilla ACM Digital Library -tietokannasta: <<http://dl.acm.org/>> (käytetty 23.11.2011).
- Yarowsky, D. (1993). One sense per collocation. Teoksessa *Proceedings of the Workshop on Human Language Technology*, Plainsboro, Yhdysvallat, 21.-24.3.1993, 266-271. Saatavilla ACM Digital Library -tietokannasta: <<http://dl.acm.org/>> (käytetty 14.10.2011).
- Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. Teoksessa *Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics*, New Mexico, Yhdysvallat, 27.-30.6.1994, 88-95. <<http://www.mt-archive.info/ACL-1994-Yarowsky.pdf>> (käytetty 14.10.2011).
- Zens, R., & Ney, H. (2004). Improvements in phrase-based statistical machine translation. Teoksessa *Proceedings of the 2004 annual conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Boston, Yhdysvallat, 2.-7.5.2004. <<http://www.mt-archive.info/HLT-NAACL-2004-Zens.pdf>> (käytetty 27.11.2011).
- Zens, R., Och, F. J., & Ney, H. (2002). Phrase-based statistical machine translation. Teoksessa M. Jarke, J. Koehler, & G. Lakemeyer (toim.), *KI 2002: Advances in Artificial Intelligence*. (Lecture Notes in Computer Science 2479, 18-32). Saatavilla SpringerLink-tietokannasta: <<http://www.springerlink.com/>> (käytetty 27.11.2011).
- Zhang, Y., & Vines, P. (2005). Using the web for translation disambiguation: RMIT University at NTCIR-5 Chinese-English CLIR. Teoksessa *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access*

*Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokio, Japani, 6.-9.12.2005, 1-6.  
<<http://www.mt-archive.info/NTCIR-2005-Zhang.pdf>> (käytetty 3.4.2011).

Zhu, J., & Wang, H. (2006). The effect of translation quality in MT-based cross-language information retrieval. Teoksessa *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17.-21.7.2006, 593-600. <<http://www.mt-archive.info/Coling-ACL-2006-Zhu.pdf>> (käytetty 1.4.2011).

# LIITE 1: TARKASTELTAVAT MONITULKINTAISET

## SANAT

appear	minor
argument	miss
arm	note
ask	order
atmosphere	paper
audience	party
bank	play
beat	present
can	race
check	show
degree	sign
difference	solid
direct	space
eat	spill
express	stamp
fall	string
flat	suit
free	suspend
interest	tender
lie	train
light	treat
lose	trip
mark	wash
match	watch
mean	win