

Kaksikielisten sanakirjojen generointi tietokoneella

Preston Palon

Helsinki 28.10.2010

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO – UNIVERSITY OF HELSINKI

Tiedekunta – Faculty		Laitos – Department	
Matemaattis-luonnontieteellinen tiedekunta		Tietojenkäsittelytieteen laitos	
Tekijä – Author			
Preston Palon			
Työn nimi – Title			
Kaksikielisten sanakirjojen generointi tietokoneella			
Oppiaine – Subject			
Tietojenkäsittelytiede			
Työn laji – Level	Aika – Date	Sivumäärä – Number of pages	
	27.10.2010	11 sivua	
Tiivistelmä – Abstract			
<p>Tässä kirjoitelmassa selvitetään kaksikielisten sanakirjojen automaattista generointia tietokoneella. Keskitytään kahteen päämenetelmään: välikielen käyttöön ja rinnakkaistekstien käyttöön. Kummastakin menetelmästä selostetaan niissä kohdattavia ongelmia, näiden ongelmien perinteisiä ratkaisuja sekä uudempiä lähestymistapoja.</p>			
Avainsanat – Keywords			
sanakirja, tietokoneavusteinen, luonnollisen kielen prosessointi			
Säilytyspaikka – Where deposited			
Muita tietoja – Additional information			

Sisältö

1 Johdanto	1
2 Välikielen käyttö	2
2.1 Perusmenetelmät.....	3
2.2 Uudet valintamenetelmät.....	4
3 Rinnakkaistekstien käyttö	5
3.1 Tekstien kohdistaminen.....	5
3.2 Tilastolliset menetelmät.....	7
3.3 Älykkäät menetelmät.....	8
4 Yhteenveto	10
Lähteet	11

1 Johdanto

Sanakirjan kirjoittaminen käsityönä on erittäin työlästä. Tarvetta erilaisille sanakirjoille on yhä enemmän, ja erityisesti tietokoneen käytettäväksi esimerkiksi automaattisessa kielen kääntämisessä tarkoitettujen sanakirjojen pitää olla hyvin laajoja. Niinpä onkin ymmärrettävää, että viime vuosikymmeninä on tutkittu paljon miten tietokoneita voitaisiin käyttää hyväksi sanakirjan tekemisessä. Tulokset ovat lupaavia, ja yhä parempien ja kattavampien sanakirjojen saatavuus näkyy epäsuorasti esimerkiksi automaattisten käännösten laadun jatkuvana paranemisena.

Tietokoneita voidaan käyttää hyväksi sanakirjan teossa monella tasolla, pelkästä tekstinkäsittelystä aina koko prosessin täydelliseen automatisointiin, jolloin tietokone generoi sanakirjan täysin itsenäisesti. Mitä pidemmälle automatisointi viedään, sitä enemmän säästyy aikaa ja rahaa, mutta samalla lopputuloksen laatu heikkenee. Tietokoneen tekemässä sanakirjassa on epätarkkuuksia, virheitä ja usein paljon hyödytöntä informaatiota. Käytännössä usein paras menetelmä onkin, että ensin generoidaan sanakirjan alustava versio tietokoneella, ja sitten se tarkistetaan ja sitä laajennetaan käsityönä.

Sanakirjoja on hyvin monenlaisia. Ne voivat olla yksikielisiä, kaksikielisiä tai monikielisiä. Ne voivat kattaa kielen yleistä sanastoa, tai jotakin erikoissanastoa, esimerkiksi vain verbejä tai jonkin alan erityissanastoa. Sanakirja voi olla tehty ihmisiä varten tai tietokoneen käyttöön. Sanat voidaan luetella perusmuodossaan, jokainen taivutusmuoto erikseen tai jopa useamman sanan fraaseina. Sanat voivat olla aakkosjärjestyksessä, taajuuden mukaan järjestettynä, tai esimerkiksi aakkosjärjestyksessä sanan loppupäästä katsoen. Sanan merkityksen tai käännöksen lisäksi sanasta voidaan antaa esimerkiksi kieliopillista tai foneettista tietoa.

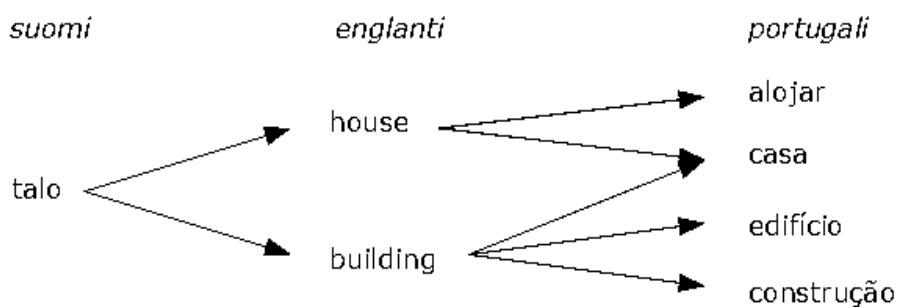
Tässä kirjoitelmassa keskitytään kaksikielisten sanakirjojen automaattiseen generointiin. Tällaisesta sanakirjasta sana haetaan alkukielellä, ja saadaan sitä merkitykseltään mahdollisimman lähellä oleva sana kohdekielellä. Ihmistyön osuus on minimaalinen.

Automaattisessa generoinnissa käytetään kahta perusmenetelmää. Ensiksi tarkastellaan välikielen käyttöä, jolloin uusi sanakirja saadaan tehtyä käyttämällä apukielenä hyväksi jotakin yleistä kolmatta kieltä. Toisena menetelmänä käsitellään rinnakkaistekstien käyttöä. Tässä tekniikassa tarvitaan avuksi samaa tekstiä sekä alku- että kohdekielellä.

2 Välikielen käyttö

Perusajatus välikielen käytössä on yksinkertainen. Jos halutaan tehdä esimerkiksi suomi – portugali sanakirja, ja käytettävissä on suomi – englanti sanakirja ja englanti – portugali sanakirja, niin katsotaan halutun sanan käänнос ensin englanniksi, ja sitten toisesta sanakirjasta englanninkielisen sanan käänнос portugaliksi, ja näin saadaan haluttu sanapari. Tämä menetelmä on erityisen käyttökelpoinen kun alku- ja kohdekieli ovat harvinaisia. Tyypillisesti löytyy kaksikielinen sanakirja jossa toisena kielenä on jokin yleinen Eurooppalainen kieli, kuten englanti, saksa tai espanja, mutta harvinaisemmat kieliparit puuttuvat.

Välikielen käytössä tulee kuitenkin helposti ongelmia. Yhdellä sanalla voi olla toisessa kielessä useampia vastineita, ja tällöin hyvä sanakirja tarjoaa useampia käänносvaihtoehtoja. Sanan merkitys voi myöskin olla toisessa kielessä hieman laajempi tai suppeampi kuin alkukielessä, tai käänноksen merkitys voi erota jonkin verran alkuperäisestä. Välikieltä käytettäessä näitä ongelmia voi tulla joko ensimmäisessä tai toisessa käänноksessä, tai pahimmassa tapauksessa molemmissa, kuten kuva 1 osoittaa.



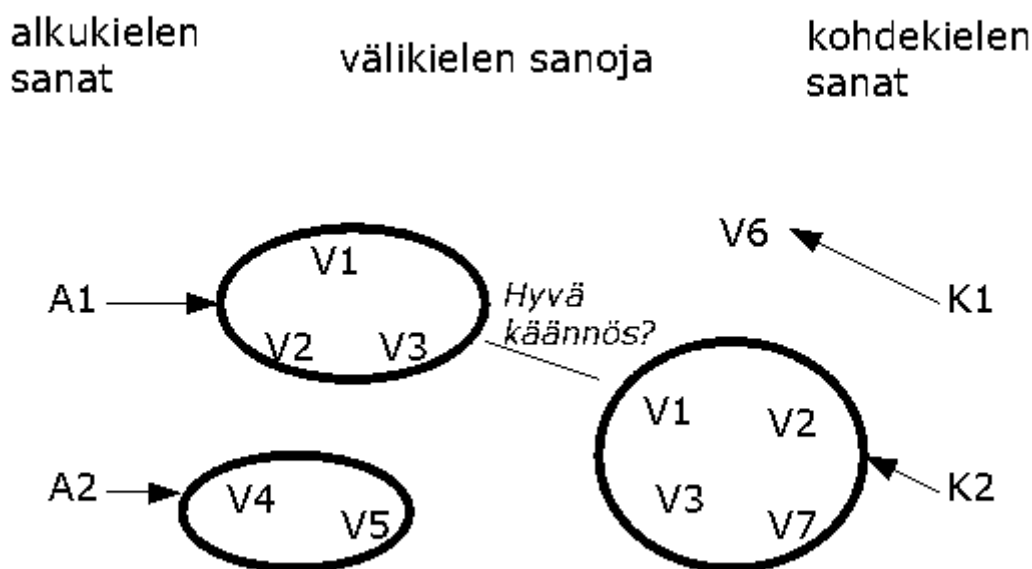
Kuva 1: Sanoilla voi olla monta vastinetta.

Englannille on tyypillistä, että merkitykseltään sopivia sanoja löytyy useita. Tässä esimerkissä sanat **house** ja **building** on otettu **talon** vastineeksi. Kun tästä edetään portugaliin, saadaan sanan **house** käänноkseksi esimerkiksi **alojar**, mikä tarkoittaa lähinnä **majoittaa**, eli se ei millään tavalla vastaa suomalaista sanaa **talo**, mutta on kylläkin sanan **house** yksi merkitys. Vastaavasti sanalle **building** on löytynyt käänнос **construção**, mikä olisi suomeksi esimerkiksi

rakentaminen, mutta ei **talo**. Kummastakin englantilaisesta käännöksestä on edetty portugalin sanaan **casa**, ja tämän perusteella voisi arvata että ehkä se olisi tässä tapauksessa sopiva käännös. Se tosiaankin sattuu olemaan paras vastine suomen sanalle **talo**, mutta tämä ei selvästikään ole luotettava tai aina toimiva menetelmä parhaan käännöksen valitsemiseksi.

2.1 Perusmenetelmät

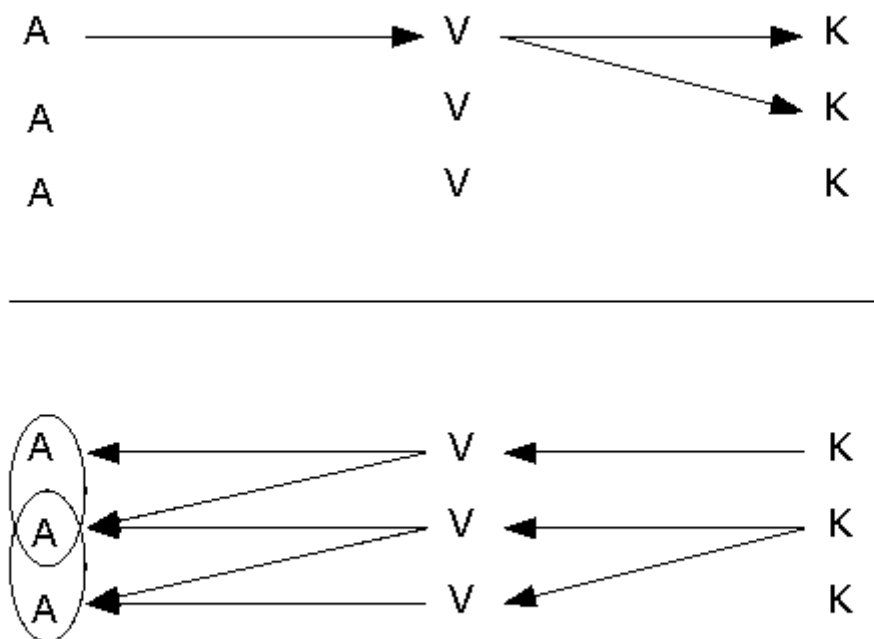
Miten tämä parhaan käännöksen valintaongelma sitten voidaan ratkaista? Sjöbergh [Sjö05] esittää menetelmän jossa ei edetäkään alkukielestä välikielen kautta kohdekielen, vaan sekä alkukielestä että kohdekielestä välikielen. Koska sanoille tyypillisesti löytyy useampia käännöksiä, saadaan alkukielen ja kohdekielen sanoja vastaamaan joukko välikielen sanoja. Tämän jälkeen tutkitaan missä joukoissa on eniten päällekkäisyyttä, ja tällä perusteella valitaan alkukielen sanoille sopivin kohdekielen käännös. Kuva 2 havainnollistaa menetelmää.



Kuva 2: Välikielen sanajoukoista etsitään päällekkäisyyksiä.

Tanaka ja Umemura [TaU94] taas ovat ratkaisseet ongelman niin, että kun on löydetty joitakin kohdekielen sanoja jotka ovat mahdollisia käännöksiä, tehdään käänteinen tarkistus, eli palataan välikielen kautta takaisin alkukieleen ja katsotaan miten lähelle alkuperäistä sanaa päästiin. Käänteisessä tarkistuksessa saadaan jokaista käännösehdokasta vastaamaan joukko alkukielen sanoja (kuva 3). Alkuperäinen alkukielen sana ei välttämättä esiinny tässä joukossa ollenkaan,

tai se voi esiintyä yhden tai useamman kerran. Jos tällä perusteella ei voida valita parasta käännöstä, voidaan apuna käyttää esimerkiksi synonyymilistoja. Käänteisessä tarkistuksessa saatuihin alkukielen sanajoukkoihin lisätään niiden sanojen kaikki synonyymit, ja lasketaan miten monta kertaa alkuperäinen sana nyt esiintyy joukossa.



Kuva 3: Ylimmästä kohdekielen sanasta päästään takaisin alkuun.

2.2 Uudet valintamenetelmät

Viimeaikoina on ilmestynyt paljon kehittyneempiä menetelmiä parhaan käännöksen valintaan. Nämä menetelmät käyttävät hyväkseen kielitieteellistä informaatiota.

Varga, Yokoyama ja Hashimoto [VYH09] käyttävät ensin välikielimenetelmää tavalliseen tapaan, ja turvautuvat sitten lopullisen sanaparin valinnassa WordNet tietokantaan. WordNet sisältää paljon tietoa sanojen merkityksestä ja niiden välisistä semanttisista suhteista. Menetelmässä pyritään valitsemaan se sanapari, jolla on suurin ”semanttinen päällekkäisyys”. Tätä varten tutkitaan alkukielen sanan ja kohdekielen käännöskandidaattien synonyymejä, antonyymejä, sekä muuta WordNetistä löytyvää tietoa, ja monimutkaisen kaavan avulla valitaan

se käännös joka näyttäisi olevan semanttisesti lähinnä alkukielen sanaa. Tekijät katsovat osoittaneensa että WordNetin avulla saavutetaan parempi tarkkuus kuin pelkkien sanakirjojen avulla.

Otero ja Campos [Oti10] alottavat myös tavalliseen tapaan generoimalla sanapareja kahden kaksikielisen sanakirjan avulla, Mutta parhaan käännöksen valintaan käytetään tietoa siitä, minkälaisissa konteksteissa sanoja käytetään. He analysoivat automaattisesti sekä alkukielen että kohdekielen tekstejä joihin on lisätty tietoa sanojen kielipiillisestä roolista. Kun selviää minkälaisissa konteksteissa alkukielen sana esiintyy, voidaan olettaa että paras käännös on se sana joka esiintyy kohdekielen teksteissä samanlaisimmissa konteksteissa. Jos esimerkiksi alkukielen sana on substantiivi joka esiintyy usein tietyn verbin subjektina, paras käännös todennäköisesti myös nähdään usein vastaavan verbin subjektina.

3 Rinnakkaistekstien käyttö

Välikielen käyttö onnistuu vain jos saatavilla on riittävän laadukkaat sanakirjat yhteiselle kolmannelle kielelle. Jos sellaisia ei ole, voidaan sanakirja yrittää generoida rinnakkaisteksteistä. Tämä edellyttää sitä että käytettävissä on lähdemateriaalia jossa sama teksti on sekä alku- että kohdekielellä. Nämä menetelmät siis muistuttavat Rosetta kiven käyttöä.

Sopivia tekstejä on aika paljon saatavilla. Esimerkiksi kansainväliset organisaatiot ja monikansalliset yritykset joutuvat kääntämään materiaalia monille eri kielille. Yhdistyneet Kansakunnat (YK) on yksi hyvä lähdemateriaalin tuottaja, sillä joitakin YK:n julkaisuja on käännetty valtavalle määrälle eri kieliä, myös erittäin harvinaisille kielille.

Rinnakkaistekstejä voidaan käyttää hyväksi monella eri tavalla. Tässä kirjoituksessa tutustutaan ensin tilastolliseen menetelmään, ja sitten tutkitaan älykkäämpiä tekniikoita. Aivan ensiksi pitää kuitenkin miettiä sitä, miten erikieliset tekstit saadaan kohdistettua niin että tiedetään mitkä kohdat vastaavat toisiaan.

3.1 Tekstien kohdistaminen

Rinnakkaisteksteistä pystytään ehkä suoraankin tekemään joitakin johtopäätöksiä, esimerkiksi sanojen taajuuksia laskemalla. Paljon pidemmälle kuitenkin päästään jos erikieliset tekstit pystytään kohdistamaan niin että tiedetään mitkä kappaleet, lauseet ja jopa jotkin yksittäiset sanat vastaavat toisiaan. Nyrkkisääntönä voidaan sanoa että mitä parempi kääntäjä tekstin on kääntänyt, sen hankalampaa kohdistaminen on. Hyvä kääntäjä voi muuttaa lauseita ja sanoja

paljonkin päästäkseen parhaaseen lopputulokseen, kun taas huono kääntäjä kääntää tekstiä sana kerrallaan. Kuitenkin esimerkiksi ihmisten ja paikkojen nimet sekä monet tekniset termit ovat usein samoja tai lähes samoja kummassakin tekstissä.

Kay ja Röscheisen [KaR93] ovat esitelleet klassisen kohdistamisalgoritmin, joka ei toimiakseen tarvitse mitään ulkopuolista tietoa kummastakaan kielestä. Menetelmässä on kaksi osaa: ensin huolehditaan sanojen taivutusmuodoista, ja sitten kohdistaminen suoritetaan iteroimalla.

Sanojen taivuttaminen aiheuttaa kohdistuksessa isoja ongelmia. Tietty sana voi esiintyä esimerkiksi alkukielessä aina samassa muodossa, mutta kohdekielessä sitä vastaakin useita eri muotoja. Kohdistamisen kannalta olisikin edullista, jos sanat voitaisiin esittää perusmuodossaan. Tätä ei kuitenkaan saa viedä liian pitkälle. Esimerkiksi jokin pääte voi tehdä sanasta verbin, ja toinen pääte substantiivin. On tärkeää että tällaista informaatiota ei kadoteta, sillä se taas aiheuttaisi ongelmia kohdistuksessa.

Menetelmässään Kay ja Röscheisen pyrkivät tekemään aika karkeita päätelmiä sanojen alkuun ja loppuun mahdollisesti lisättävistä osista, mutta he jättävät huomioimatta sanan perusosassa tapahtuvat muutokset. He tallentavat sanat puurakenteisiin. Jokaisella kirjaimella alkaville sanoille on oma puunsa, ja alkukirjain on puun juuri. Juuresta sitten edetään kohti lehtiä kirjain kerrallaan, ja jokaisessa solmussa on tietoa siitä, miten monta sanaa kulkee solmun kautta. Sitten he käyttävät yksinkertaista funktiota löytämään kirjainyhdistelmiä joita siinä kielessä mahdollisesti lisätään sanojen alkuun tai loppuun. Menetelmä ei ole erityisen tarkka, vaan pikemminkin kompromissi.

Kun taivutusmuotoja on vähennetty, päästään käyttämään kohdistusalgoritmia. Algoritmi pyörii silmukassa, yrittää löytää sanapareja, ja niiden avulla kohdistaa yhä enemmän ja enemmän lauseita. Hieman yksinkertaistettuna se toimii näin:

- 1. Kohdista muutama lause tekstien alusta ja lopusta.**
- 2. Yritä löytää kohdistetuista lauseista sanapareja. (Esiintyykö lauseissa samoja sanoja? Miten monta kertaa missäkin lauseessa?)**
- 3. Lisää löytyneet sanaparit sanaparitaulukkoon.**
- 4. Kohdista sanaparitaulukon avulla lisää lauseita.**
- 5. Jos kaikkia lauseita ei ole vielä kohdistettu, palaa kohtaan 2.**

Yleensä tarvitaan vain muutama iteraatio kaikkien lauseiden kohdistamiseen. Kohdistettujen lauseparien lisäksi tästä algoritmista saataisiin tulokseksi sanaparitaulukko, mutta se ei olisi kovin luotettava.

3.2 Tilastolliset menetelmät

Yksinkertaisin tapa sanaparien löytämiseen rinnakkaisista teksteistä on tilastollisten menetelmien käyttö. Voidaan esimerkiksi tutkia sanojen taajuuksia koko tekstissä ja/tai pienemmissä osissa, ja voidaan mitata mitkä sanat esiintyvät lähellä toisiaan ja miten usein. Yksinkertainen perusmenetelmä on niinsanotun kosinifunktion käyttö.

$$\text{kosini}(A, K) = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Tässä kaavassa:

A = alkukielen sana

K = kohdekielen sana

a = Niiden kohdistettujen lauseparien lukumäärä joissa sekä A että K esiintyy.

b = Niiden kohdistettujen lauseparien lukumäärä joissa vain A esiintyy.

c = Niiden kohdistettujen lauseparien lukumäärä joissa vain K esiintyy.

Näin saadaan kosiniarvo kaikille sanapareille, ja se kohdekielen sana jolla on korkein tulos tietyn alkukielen sanan kanssa valitaan käännökseksi.

Echizen, Araki ja Momouchi [EAM06] antavat esimerkin joka osoittaa että tilastolliset menetelmät eivät toimi kovin hyvin jos lähdeaineistoa ei ole riittävästi. Oletetaan että haluamme löytää englannin sanalle **parcel** Japanin kielisen vastineen, ja meillä on käytössä seuraavat kolme englanti – japani lauseparia:

1. *Your parcel is on the table – teburu ni anata no kozutsumi ga ari masu.*
2. *Do you know where mine is – watashi no kozutsumi ga doko ni aru ka shiri masen ka?*
3. *It's under the table – sore wa teburu no shita desu yo.*

Ensimmäisen lauseen perusteella näyttäisi siltä että oikea käännös voisi olla joko **teburu** tai **kozutsumi**. Lasketaan molemmille pareille kosini. Kumpikin saa täsmälleen saman arvon, 0.71. Oikea vastaus olisi ollut **kozutsumi**. Tilastollisia menetelmiä ei siis kannata käyttää jos rinnakkaistekstiä on vain vähän käytettävissä.

3.3 Älykkäät menetelmät

Tilastollisten menetelmien voima siis on rajallinen. Avuksi voidaan ottaa tekniikoita, jotka käyttävät hyväkseen enemmän tietoa kielten lauseiden rakenteesta.

Yksi esimerkki on järjestelmä joka oppii sääntöjä joiden avulla sanapareja voidaan löytää kohdistettujen lauseparien osista [EAM06]. Se yrittää löytää lausepareja joissa on sekä samoja osia että erilaisia osia. Palataan luvun 3.2 esimerkissä esiintyneeseen englantia – japani lausepariin:

Your parcel is on the table – teburu ni anata no kozutsumi ga ari masu.

Jos löytyy muitakin lausepareja joissa esiintyvät samat osat **your** ja **anata no**, mutta niitä seuraa aina eri sanoja, voidaan olettaa että pätee sääntö: **your @ - anata no @**, missä **your** ja **anata no** ovat samaa tarkoittavia fraaseja ja @ on muuttuja. Soveltamalla sitten tätä sääntöä lausepariin saadaan sanan **parcel** käännökseksi **kozutsumi**.

Güvenir ja Cicekli [GuC98] esittävät samantapaisen menetelmän kielen kääntämisessä käytettäväksi. Heidän menetelmässään opitaan esimerkkien avulla malleja joita voidaan käyttää automaattiseen kielen kääntämiseen, ja samalla saadaan tulokseksi sanapareja. Heidän esimerkissään englantia – turkki rinnakkaistekstistä on löytynyt kaksi hyvin samantapaista lausetta:

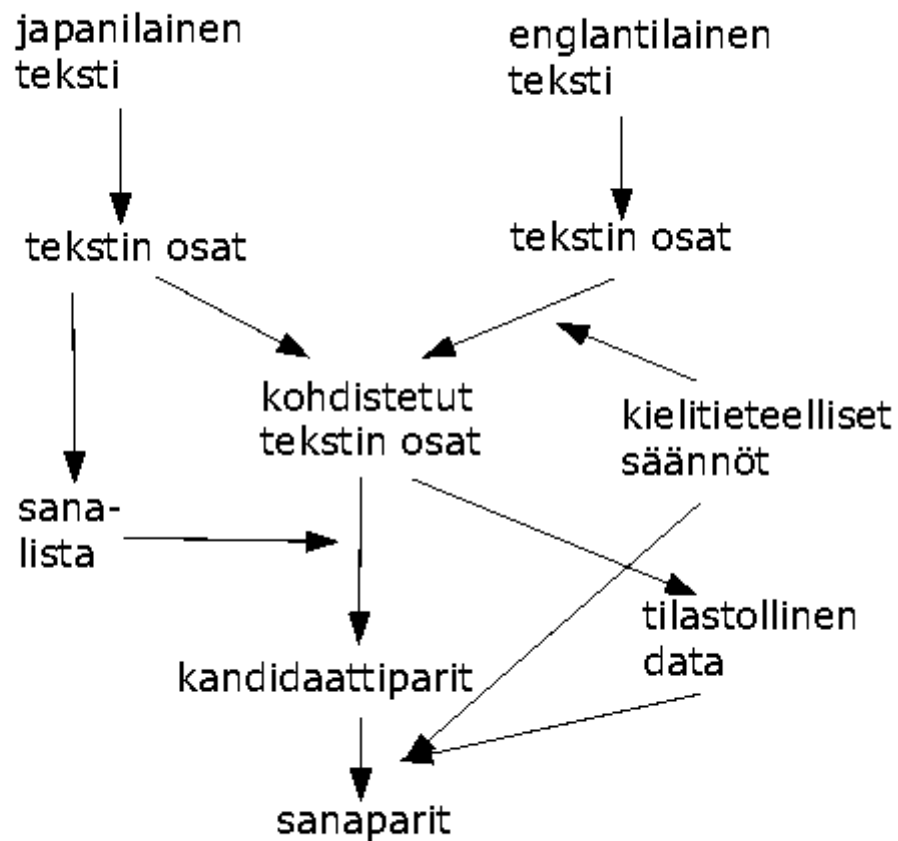
1. *I took a ticket from Mary – Mary'den bir bilet aldim*
2. *I took a pen from Mary – Mary'den bir kalem aldim*

Samanlaiset kohdat on alleviivattu. Tämän esimerkin perusteella opitaan yksi kääntämismalli ja kaksi sanaparia:

1. **I took a A1 from Mary – Mary'den bir K1 aldim**
2. **ticket – bilet**
3. **pen – kalem**

Opitut mallit ja sanaparit päätyvät sitten automaattisen kielenkääntöohjelman käytettäväksi. Tekijöiden mukaan heidän menetelmänsä muistuttaa ihmisten tapaa oppia kieliä.

Mitä enemmän sanaparien etsimisessä otetaan huomioon kielitieteellisiä lainalaisuuksia, sen vaikeampaa menetelmästä on tehdä kaikille kielille soveltuva. Toisaalta, jos hyväksytään että järjestelmä toimii vain tietyille kieliparille, se voidaan hienovirtittää hyvinkin tarkaksi. Esimerkiksi Kumano ja Hirakawa [KuH94] ovat kehittäneet melko hienostuneen menetelmän joka on tarkoitettu japani – englantia sanaparien löytämiseen. Järjestelmän osat ovat kuvassa 4.



Kuva 4: Kehittynyt järjestelmä japani - englanti sanaparien löytämiseen.

Rinnakkaistekstit paloitellaan osiksi, joiden ei välttämättä tarvitse olla lauseita. Tekstit kohdistetaan käyttämällä hyväksi kielitieteellistä tietoutta Japanin ja Englannin kielistä. Japanin tekstin osien avulla muodostetaan alkukielen sanalista. Kohdistetuista tekstin osista generoidaan kandidaatti sanapareja, ja myöskin kerätään tilastollista informaatiota. Kielitieteellisten sääntöjen ja tilastollisen informaation avulla kandidaattipareista sitten valitaan lopulliset japani – englanti käännökset. Tekijät kertovat että tällä menetelmällä on saatu lupaavia tuloksia patenttiteksteistä löytyvien vaikeiden termien kääntämisessä.

4 Yhteenveto

Yhä suurempien ja parempien sanakirjojen tarve kasvaa edelleen, ja niitä tarvitaan myös harvinaisempien kielten välillä. Erityisesti automaattisessa kielenkääntämisessä on tarvetta erittäin kattaville sanakirjoille. Tässä kirjoitelmassa on tarkasteltu sanakirjojen automaattista generointia.

Kaksi sanakirjojen generoinnissa käytettyä päämenetelmää ovat välikielten käyttö ja rinnakkaistekstien käyttö. Luvussa 2 pohdittiin ensiksi välikielen käyttämisestä aiheutuvia ongelmia. Sitten selostettiin miten nämä ongelmat on perinteisesti pyritty ratkaisemaan. Luvun lopussa esitettiin uudempia ja parempia menetelmiä sanaparien valinnan ongelmaan. Luvussa 3 käsiteltiin rinnakkaistekstien käyttöä. Ensiksi selostettiin miten kahdella kielellä kirjoitettu sama teksti voidaan kohdistaa. Sitten käsiteltiin tilastollisia menetelmiä sanaparien löytämiseksi, ja lopuksi älykkäämpiä menetelmiä joissa turvaudutaan esimerkiksi oppimiseen ja kielitieteellisen informaation hyödyntämiseen.

Tutkimus tällä alalla on ollut aktiivista, ja esimerkiksi verrattaessa uusia automaattisia kielenkääntöohjelmia ensimmäisiin vapaasti saatavilla olleisiin ohjelmiin, huomataan että kehitys on ollut valtavan nopeaa.

Lähteet

- EAM06 Echizen-ya, H., Araki, K., Momouchi, Y., Automatic extraction of bilingual word pairs using inductive chain learning in various languages. *Information Processing and Management* 42 (2006), pages 1294-1315.
- GuC98 Guvenir, H., Cicekli, I., Learning translation templates from examples. *Information Systems*, Vol. 23, No. 6, 1998.
- KaR93 Kay, M., Roscheisen, M., Text - translation alignment. *Computational Linguistics*, volume 19, number 1, 1993.
- KuH94 Kumano, A., Hirakawa, H., Building an MT dictionary from parallel texts based on linguistic and statistical information. R & D Center, Toshiba Corporation, 1994.
- NiN04 Niessen, S., Ney, H., Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, volume 30, number2, 2004.
- OtJ10 Otero, P. G., Jose, R. P. C., Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. *CICLing 2010, LNCS 6008*, pages 473 – 483.
- Sjö05 Sjöbergh, J., Creating a free digital Japanese – Swedish lexicon. KTH Nada, Stockholm, Sweden, 2005.
- TaU94 Tanaka, K., Umemura, K., Construction of a bilingual dictionary intermediated by a third language. Division of Engineering, University of Tokio, Japan, 1994.
- VYH09 Varga, I., Yokoyama, S., Hashimoto, C., Dictionary generation for less-frequent language pairs using WordNet. *Literary and Linguistic Computing*, volume 24, number 4, 2009.