KIELITEKNOLOGIASTA SUOMENKIELISTEN TEKSTIEN TUTKIMISESSA

Mikko Lounela, Kotimaisten kielten tutkimuskeskus

Kieliteknologian ja kielentutkimuksen suhde on monitahoinen. Kielentutkimus on kieliteknologialle välttämätön osa perustutkimusta, ja kieliteknologia antaa perinteiselle kielentutkimukselle uusia kvantitatiivisia näkökulmia. Kieliteknologian käyttöön soveltuva lingvistinen pohjatyö on kallista ja huonosti näkyvää, ja se vaatii poikkiteellistä osaamista. Sama pätee kieliteknologian käyttöön kielentutkimuksen apuvälineenä. Olemassaolevat tieteen rahoitusmallit eivät tue tämänkaltaista työtä, mikä johtaa siihen, että työkalut pienellä kielialueella eivät useinkaan ole riittävän hyviä tieteellisen kielentutkimuksen pohjaksi.

Artikkeli syventyy kieliteknologian ja korpuslingvistiikan käyttöön tekstilingvistiikan apuvälineenä. Se esittelee mallin, jonka mukaisesti suomenkielisiä tekstiaineistoja voidaan valmistella tieteelliseen tutkimukseen kelpaavalle tasolle ja käyttää kvantitatiivisen tiedon eristämiseen tekstiaineistoista. Se esittelee mallin käyttöä vertailemalla Aamulehden paikallisuutisten ja presidentti Kekkosen uudenvuodenpuheiden kvantitatiivisia kieliopillisia piirteitä.

Avainsanat: Korpuslingvistiikka, kieliteknologia, kvantitatiivinen tekstintutkimus.

TAUSTAA

Tietokonelingvistiikka ja kieliteknologia ovat olleet 1960-luvulta lähtien ajoittain tutkimusrahoituksen painopisteessä. 1960-luvulla Yhdysvalloissa toivottiin nopeita läpimurtoja ja niistä seuraavia (sotilaalliseen käyttöön sopivia) sovelluksia. Konekäännösohjelmia ja automaattisia tiivistelmän tekijöitä pidettiin potentiaalisina kylmän sodan aseina. Optimismi ei kuitenkaan ollut perusteltua, ja pettymykset johtivat tietokonelingvistiikan rahoituksen vähenemiseen pariksikymmeneksi vuodeksi. 1980- ja 1990-luvuilla Euroopassa nousi tarve saada yhdentyvän Euroopan byrokratiasta kumpuavat päätökset ja ohjeet mahdollisimman tehokkaasti ja edullisesti kaikenkielisten Euroopan päätöksentekijöiden, päätösten toimeenpanijoiden ja kansalaisten ulottuville. Kielelliset sovellukset nähtiin myös merkittävänä potentiaalisena kansalaisten yhdenvertaisuuden edistäjänä tietoteknisessä maailmassa. Kieliteknologia nousi jälleen jonon kärkipaikoille, kun rahaa jaettiin.

Seurauksena oli eurooppalaisia jättihankkeita, esimerkkeinä 1990-luvun lopun saksalainen Verbmobil, jonka tavoitteena on ollut automaattisen tulkkausvälineen kehittäminen keskeisille Euroopan kielille, ja samoihin aikoihin toteutettu eurooppalainen kieliaineistohanke Le-parole, jonka tarkoituksena oli kerätä vertailukelpoista tutkimusmateriaalia EU-kielistä teknisiä sovelluksia varten. Monesti suuret kansalliset ja EU-tasoiset hankkeet ovat jääneet tuloksiltaan ja vaikutuksiltaan suunniteltua vaatimattomammiksi. 2000-luvulla globalisoituva maailma kaipaa kaupan ja talouden tarpeisiin ohjelmistoja, jotka mahdollistavat mahdollisimman

Kirjoittajan yhteystiedot: Mikko Lounela
Kotimaisten kielten tutkimuskeskus
Sörnäisten rantatie 25, 00500 Helsinki
p. 09 7315 245, mikko.lounela@kotus.fi

nopean ja edullisen tiedonsiirron kielieroista riippumatta. Yleisimpiin Internetin hakukoneisiin sisältyy nykyään automaattisia kääntäjiä maailman suurimmille kielille.

Kieliteknologiaa siis on tarvittu moneen ja rahoitettu avokätisestikin. Rahoitus puolustushallinnoissa, EU:ssa ja kansallisesti on kuitenkin yleensä ollut sovellushakuista. Rahalle on haluttu vastineeksi näkyvä, mieluummin kaupalliseen levitykseen kelpaava tuote. Perustutkimus – analyysialgoritmien ja kielen koneelliseen käsittelyyn soveltuvien kielioppien kehittäminen – on saanut hoitaa itsensä joko yliopistoissa ja tutkimuslaitoksissa niukentuvilla budjettivaroilla tai kaupallisesti markkinoiden ehdoilla.

Perustutkimuksen lisäksi tuotekehityksen varjoon on (ainakin Suomessa) jäänyt kieliteknologian käyttö kielentutkimuksen apuvälineenä. Kieliresurssien ja korpuslingvististen menetelmien systemaattinen kehittäminen jää helposti luonnontieteellis-teknisen ja humanistisen tutkimuksen välimaastoon, ja siksi sille on hankala löytää rahoittajaa. Lopputulos ei ole myytävä tuote, ja tuotokset vaatisivat ylläpitoa, jota projektipainotteiset rahoitusmallit eivät tue. Niinpä esimerkiksi suomen kielen automaattiset jäsentimet eivät ole sen tasoisia, että niitä voisi suoraan käyttää tieteellisen tutkimuksen analyysityökaluina. Kunnollisesti toteutettu lingvistinen pohjatyö on liian kallista kannattaakseen kaupallisesti näin pienellä markkina-alueella. Sovellettavien kielioppien ja vastaavien pohjatyökalujen parissa tehtävä työ on myös usein liian teknistä humanisti-kielentutkijoille ja liian humanistista kieliteknologeille ja tietotekniikan asiantuntijoille.

Kieliteknologian käytöllä kielentutkimuksen apuvälineenä tarkoitan tässä kielellisten (morfologisten tai syntaktisten) analysaattoreiden ja niiden tuottamaa tietoa käsittelevien laskentaohjelmien käyttämistä tutkijan apuna mahdollisimman tarkan ja monipuolisen kvantitatiivisen tiedon saamiseksi aineistoista. Kieliteknologiaa hyödyntävä kvantitatiivinen tutkimusprosessi voidaan jakaa neljän osaan. Se sisältää yleensä aineiston valinnan, aineiston valmistelun koneellisesti ymmärrettävään muotoon, automaattisen tunnuslukujen laskennan ja tunnuslukujen ja tekstien tulkinnan. Näistä humanisti-kielentutkijan osaamisen ulkopuolelle jäävät yleensä kaksi keskimmäistä vaihetta, kun taas kahta muuta ei pätevästi pysty hoitamaan kukaan muu. Kyseessä on siis lähes välttämättä yhteistyö eri alojen osaajien välillä. Tätä yhteistyömallia suomalaisen kielitieteen yksinpuurtamisen perinne ei tunnu tukevan.

Vaikeata siis on kvantitatiivisella kielentutkimuksella. Kotimaisten kielten tutkimuskeskuksessa (Kotuksessa) on kuitenkin viime vuosina kehitelty morfologisesti merkattujen tekstiaineistojen malli ja puoliautomaattinen prosessi tällaisten aineistojen tuottamiseksi tekstien tutkijoiden tarpeisiin (Lehtinen & Lounela, 2004). Malli on toteutettu niillä työkaluilla, joita suomen kielelle on saatavissa tai itse tehtävissä. Esittelen tässä artikkelissa kehittämäämme mallia, ja siihen liittyviä analyysiohjelmia (Lounela, 2005). Näytän esimerkein, miten mallia ja ohjelmistoa voi käyttää tekstijoukkojen kvantitatiivisten ominaisuuksien vertailussa.

PUOLIAUTOMAATTISESTI MERKATTU AINEISTO

Tekstiaineiston puoliautomaattisen merkkaamisen neljästä vaiheesta (tekstijoukon valinta, tekstien puoliautomaattinen valmistelu tutkimusaineistoksi, valmistellun tekstin automaattinen analyysi ja analyysin tulkitseminen) viimeinen vaihe suoritetaan täysin ihmisvoimin. Siinä automaattisen analyysin tuottamia lukuja ja listoja tutkitaan ja verrataan. Luvut ja listat saattavat osoittaa teks-

teissä piileviin kiinnostaviin ominaisuuksiin, joita voidaan sitten edelleen tutkia syventymällä kvalitatiivisesti itse teksteihin. Tämä vaihe on kuitenkin minun erikoisalani ulkopuolella, joten tyydyn esittelemään prosessin kolmea ensimmäistä vaihetta.

Vaihe 1: Tekstijoukon valinta

Analyysin ensimmäisessä vaiheessa tutkija valitsee joukon tekstejä tutkittavakseen. Tähän vaiheeseen yleensä liittyy myös alustavien hypoteesien ja tutkimusongelmien muodostaminen. Katson kuitenkin tätä vaihetta lähinnä tekniseltä kannalta ja jätän muun tämän katsauksen ulkopuolelle.

Mallimme mukaisessa kvantitatiivisessa analyysissa aineiston valmistelu on melko työlästä, joten tekstijoukon tulisi olla kooltaan kohtuullinen mutta edustava. Tavoiteltava koko riippuu muun muassa tutkimuskysymyksistä, tekstien pituudesta ja tekstien keskinäisestä vaihtelevuudesta. Yleisimpien sanojen jakaumasta miljoona sanaa sanomalehtitekstiä ei anna juuri sen kummempia tuloksia kuin kymmenen tuhatta sanaa. Jotakin muuta ominaisuutta laskettaessa tilanne voi olla aivan toinen. Suurin lopullisen aineiston tekstien määrään vaikuttava seikka on kuitenkin se työpanos, joka tekstien valmisteluun voidaan panna. Kokemus osoittaa, että esimerkiksi pro gradu -työn tekijä, joka itse valmistelee materiaalinsa, voi maksimissaan koodata työnsä yhteydessä noin sadan tuhannen sanan aineiston. Suuremmissa hankkeissa, joissa on mahdollisuus käyttää tutkimusapulaisia, määrä voi luonnollisesti olla paljon suurempikin.

Vaihe 2: Aineiston valmistelu

Aineiston valmistelu on Kotuksen mallin mukaisen analyysin työläin vaihe. Teksteihin merkitään kappale- ja virkerajat, otsikot ja kunkin sanan morfologiset ominaisuudet. Näitä tehtäviä varten on olemassa automaattisia työkaluja, mutta osa työstä on tehtävä itse tai ainakin jokaisen vaiheen lopputulos on tarkistettava huolellisesti. Käyn aineiston valmistelun eri vaiheet läpi käyttäen esimerkkinä virkettä *Pahimmat vaikeudet ovat olleet valtiontaloudessa* presidentti Kekkosen uudenvuodenpuheesta vuodelta 1964.

Aluksi, ennen varsinaista käsittelyä, tekstit muunnetaan elektroniseen yleiseen tekstimuotoon esimerkiksi skannaamalla tai tallentamalla tekstinkäsittelyohjelmasta. Tämän jälkeen merkkaamisprosessi voi alkaa.

Prosessin ensimmäisessä vaiheessa aineistoon merkitään muun muassa otsikoiden, kappaleiden ja virkkeiden alut ja loput (head-, p- ja s-merkinnöillä). Tämä voidaan joskus tehdä automaattisesti, mutta koska esimerkiksi lyhenteiden jäljessä olevat pisteet sotkevat virkkeiden tunnistusta, tulos on aina tarkistettava. Esimerkkivirkkeemme näyttää ensimmäisen vaiheen jälkeen tältä:

```
<p>
 <s>Pahimmat vaikeudet ovat olleet valtionta-
loudessa.</s>
 [...]
</p>
```

Toisessa vaiheessa teksti ajetaan morfologisen analysaattorin läpi. Käyttämämme analysaattori on Lingsoft OY:n Fintwol, joka antaa tekstin jokaiselle sanalle sen kaikki mahdolliset tulkinnat. Fintwol ei pysty antamaan sanoille tulkintoja niiden kontekstin perusteella. Niinpä sellaiset sananrajat ylittävät piirteet kuin liittoaikamuodot perfekti ja pluskvamperfekti jäävät tulkitsematta, samoin kuin lauseenjäsenet predikaattiverbi, subjekti ja objekti. Alla olevassa esimerkissä *olleet*-sana on saanut kolme tulkintaa, joista aineiston valmistelijan on valittava oikea.

```
<p>
<s>
<w lemma="paha" norm="pahimmat"
type="A" msd=" SUP NOM PL ">Pahim-
mat</w>
<w lemma="vaikeus" norm="vaikeudet"
type="N" msd=" DA-US NOM PL ">vaikeu-
det</w>
<w lemma="olla" norm="ovat" type="V"
msd=" COP PRES ACT PL3 ">ovat</w>
<w lemma="olla" norm="olleet" type="V" msd="
COP PAST ACT NEG PL ">olleet</w>
<w lemma="olla" norm="olleet" type="PCP2"
msd=" COP ACT POS NOM PL ">olleet</w>
<w lemma="ollut" norm="olleet" type="A"
msd=" COP ACT PCP2 POS NOM PL ">ol-
leet</w>
<w lemma="valtion#talous"
norm="valtiontaloudessa" type="N" msd="
INE SG ">valtiontaloudessa</w>
<w lemma="." norm="." type="PUNCT"
msd=" FULLSTOP ">.</w>
</s>
[...]
</p>
```

Tämän muotoisesta tekstistä aineiston valmistelijana toimiva ihminen poistaa analyyseista kontekstissaan väärät, niin että kullekin sanalle jää ainoastaan yksi (mahdollisimman oikea) tulkinta. Tämän jälkeen sanoihin lisätään merkintöjä. Esimerkiksi yllä mainitut liittoaikamuodot saavat molempiin osiinsa merkinnän *function=" P "* tai *function=" PL "*. Lisäksi erillisellä merkitsimellä merkitään sanat, joita ei haluta mukaan analyysiin esimerkiksi sen vuoksi, että Fintwolin varastosta ei ole löytynyt niille tässä kontekstissa kelvollista tulkintaa. Käsityövaiheen jälkeen esimerkkiteksti on valmis käytettäväksi ja näyttää tältä:

```
<p>
<s>
<w lemma="paha" norm="pahimmat"
type="A" msd=" SUP NOM PL ">Pahim-
mat</w>
<w lemma="vaikeus" norm="vaikeudet" type="N"
msd=" DA-US NOM PL ">vaikeudet</w>
<w lemma="olla" norm="ovat" type="V" msd="
COP PRES ACT PL3 " function=" P ">ovat</
w>
<w lemma="olla" norm="olleet" type="PCP2"
msd=" COP ACT POS NOM PL " function="
P ">olleet</w>
<w lemma="valtion#talous"
norm="valtiontaloudessa" type="N" msd="
INE SG ">valtiontaloudessa</w>
<w lemma="." norm="." type="PUNCT"
msd=" FULLSTOP ">.</w>
</s>
[...]
</p>
```

Suomen kielelle on olemassa myös sellaisia analysaattoreita, jotka näkevät sanarajan yli ja poistavat morfologisen analyysin jälkeen ei-toivotut tulkinnat itse. Tällaisia ovat Connexor Oy:n Machinese Syntax ja Kielikone Oy:n Finmorfo. Nämä jäsentimet tekevät kuitenkin virheitä, joiden etsiminen ja korjaaminen olisi varmastikin yhtä työlästä ja ehkä epävarmempaa kuin kaikkien väärien tulkintojen poistaminen käsityönä. Käsin tehtyjen korjaus- ja tarkistuskierrosten jälkeen tekstiaineisto on valmis automaattiseen tunnuslukujen ja listojen tuottamiseen.

Vaihe 3: Tunnuslukujen ja listojen tuottaminen

Valmis tekstiaineisto sisältää kaikki alkuperäisen tekstin sanat siinä muodossa, jossa ne lähteessä esiintyvät. Lisäksi tekstiin on lisätty sanojen perusmuodot ja niiden morfologisista ominaisuuksista kertovia merkitsimiä. Näitä tiedonpalasia yhdistelemällä ja niiden

esiintymistaajuuksia laskemalla on mahdollista saada kuvaavaa ja tarkkaa tietoa tutkittavan tekstijoukon laskettavista ominaisuuksista. Olen tätä varten kirjoittanut neljä erillistä tietokoneohjelmaa, jotka analysoivat kuvatulla tavalla valmisteltua tekstiä. Nämä ohjelmat eristävät tietoja tekstien yleisistä ominaisuuksista, verbimaailmasta, nominimaailmasta ja sanastosta. Yhden ohjelman tulostama tieto antaa osittaisen kuvan tekstistä. Yhdessä ne kuvaavat kohteena olevaa tekstijoukkoa melko monipuolisesti. Kukin ohjelma tuottaa valikoiman taajuuslistoja ja tunnuslukuja. Valikoima voisi olla aivan erinäköinenkin – juuri nämä listat ja luvut ovat olleet hyödyllisiä Kotuksen tekstintutkijoiden tutkimuksissa (esimerkiksi Heikkinen ym., 2005).

Tekstin yleisiä ominaisuuksia ovat monet keskimääräiset pituudet: Tekstien keskipituus virkkeinä, lauseina ja sanoina, virkkeiden pituus lauseina ja sanoina ja lauseiden pituus sanoina. Näiden lukujen lisäksi ohjelma eristää tekstistä taajuuslistat yleisimmistä välimerkeistä, sanaluokista, omistusliitteistä, yhdyssanojen sanarajojen määristä ja tekstien sanojen yleisimmistä perusmuodoista ja sananmuodoista.

Verbimaailman kuvauksessa keskitytään luonnollisesti niihin tekstin ominaisuuksiin, joita verbit kantavat. Näitä ovat pääluokka, tapaluokka, aikamuoto, persoona ja infinitiivityypit. Lisäksi verbejä luotaava ohjelmamme tuottaa taajuuslistat partisiipeista ja niiden sanaluokkajakaumista, lauseenvastikkeista ja verbien yleisimmistä perusmuodoista ja sananmuodoista. Ohjelma myös laskee semanttisten, kieliopillisten ja finiittisten verbien määrät tekstijoukossa. Nominimaailman ominaisuuksia kuvaavan ohjelman tuloksia ovat nomineihin lasketujen sanaluokkien jakauma, sijamuotojakauma, vertailumuotojakauma, lukujen jakauma, omistusliitteiden jakauma, nominaalisten yhdyssanojen sananosien määrien jakauma ja nominien sananmuoto- ja perusmuotojakaumat.

Neljäs ohjelma vie meidät hieman lähemmäs itse tekstiä. Sanastoanalysaattori tekee kustakin tekstijoukosta sanaluokittaiset sananmuotojen ja perusmuotojen taajuuslistat.

Käyn esimerkkien avulla läpi nämä neljä näkökulmaa tekstiin. Kustakin näkökulmasta olen valinnut yhden (mahdollisesti osittaisen) taajuuslistan ja mahdollisesti joitakin siihen liittyviä tunnuslukuja. Koska luvut ja listat ilman vertailukohtaa eivät kerro paljoakaan, vertailen kahta erilaista tekstijoukkoa. Materiaaleina käytän Aamulehden paikallisuutisista koostettua n. 13 000 sanan tekstijoukkoa vuodelta 2003 ja Kekkosen kaikkia uudenvuodenpuheita hänen presidenttikaudeltaan. Uudenvuodenpuheet muodostavat noin 17 500 sanan materiaalin. Molemmat tekstijoukot ovat osia suuremmista kokoelmista.

TEKSTIJOUKKOJEN KVANTITATIIVISTA VERTAILUA

Yleisen ohjelman tuottamien lukujen ja listojen joukosta tutustutaan tekstijoukkojen yleisimpien sanaluokkien jakaumaan (Taulukko 1). Taulukosta hahmottuu ensi vilkaisulla kuva, jonka mukaan substantiiveja on kummassakin joukossa runsas kolmannes tai vajaa puolet sanoista, verbejä noin puolet substantiivien määrästä ja adjektiiveja, adverbeja, konjunktioita ja pronomineja vaihtelevassa järjestyksessä noin joka kymmenes sana tai sitä vähemmän. Silmiinpistäviä eroja paikallisuutisten ja presidentti Kekkosen puheiden välillä näkyy ainakin substantiivien ja verbien yhteismäärässä, joka uutisissa on 63 % ja presidentin puheissa 52,8 %. Vastaavasti presidentin puheissa jää enemmän tilaa muille sanaluokille, etenkin adjektiiveil-

le. Taulukon tulkintaan voi vaikuttaa myös se tieto, että uutismateriaalin virkkeet ovat keskimärin 10,7 sanan ja 1,6 lauseen pituisia, kun presidentti Kekkosen virkkeissä on keskimäärin 14,8 sanaa ja 1,9 lausetta.

Verbimaailman näkökulmasta olen valinnut esiteltäväksi persoonamuotojen jakauman. Siinä kolmansien persoonien osuus on kiistattomasti hallitseva molemmissa tekstijoukoissa, samoin yksikön kolmannen persoonan johtava asema. Huomiota kiinnittää myös ensimmäisten persoonien lähes 14 prosentin yhteenlaskettu osuus persoonamuodoista presidentti Kekkosen puheissa verrattuna Aamulehden alle kolmen prosentin lukemaan. Tämä tuo hyvin esille kaksi yleistä piirrettä, jotka olemme havainneet taajuuslistoja tutkiessamme. Ensinnäkin, kuten jo edellä on mainittu, huomaamme kiinnostavia ominaisuuksia usein vasta verratessamme tekstijoukon lukuja toisen tekstijoukon vastaaviin lukuihin. Toiseksi, kiinnostavat erot eivät useinkaan löydy taajuuslistojen kaikkein yleisimpien jäsenten jakaumasta, vaan monesti heti niiden jäljessä olevien, suhteellisen yleisten jäsenten jakaumista.

Nomineja ovat ohjelman tulkinnan mukaan substantiivit, adjektiivit, numeraalit ja pronominit. Kun vaihdamme näkökulmamme nominien maailmaan, voimme valita vertailtavaksi ominaisuudeksi sijamuotojen jakauman. Siinä, toisin kuin yleisen näkökulman yhteydessä, kärjen jälkeen tuleva joukko on aineistoissa hyvin samannäköinen, mutta kaksi yleisintä sijamuotoa profiloivat tekstijoukot erilaisiksi. Aamulehden uutisissa nominatiiveja on melko tarkkaan kolmannes kaikista nominien sijamuodoista, ja genetiivi on toisella sijalla hieman vajaan neljänneksen osuudella. Presidentti Kekkosen puheissa ero on pienempi, runsaat neljä prosenttia, mutta kärjessä onkin genetiivi. Nominien osuus sanoista on uutismateriaalissa 58,3 % ja presidentti Kekkosen puheissa 57,7 %, eli melko tarkkaan samaa luokkaa.

Sanastonäkökulma vie meidät hieman lähemmäs varsinaisia tekstejä ja sitä, mistä niissä puhutaan. Verbit kertovat meille mitä teksteissä tehdään, ja substantiivien lista sen, mitä toimijoita tai toiminnan kohteita teksteistä löytyy. Katsomme kuitenkin nyt konjunktioiden taajuuslistaa, joka kertoo meille jota-

Taulukko 1. Yleisimmät sanaluokat Aamulehden paikallisuutisissa ja presidentti Kekkosen puheissa.

Aamulehti			**Kekkonen**		
Sanaluokka	Lukumäärä	Osuus %	Sanaluokka	Lukumäärä	Osuus %
Substantiivi	5626	43,2	Substantiivi	6269	35,8
Verbi	2578	19,8	Verbi	2980	17,0
Adjektiivi	923	7,1	Adjektiivi	2046	11,7
Adverbi	919	7,1	Konjunktio	1275	7,3
Konjunktio	785	6,0	Pronomini	1261	7,2
Pronomini	726	5,6	Adverbi	1259	7,2

Taulukko 2: Verbien persoonamuodot Aamulehden paikallisuutisissa ja presidentti Kekkosen puheissa.

Aamulehti			**Kekkonen**		
Persoonamuoto	Lukumäärä	Osuus %	Persoonamuoto	Lukumäärä	Osuus %
Yksikön 3.	1358	81,1	Yksikön 3.	1351	70,9
Monikon 3.	263	15,7	Monikon 3.	280	14,7
Yksikön 1.	25	1,5	Yksikön 1.	132	6,9
Monikon 1.	20	1,2	Monikon 1.	131	6,9
Yksikön 2.	6	0,4	Yksikön 2.	10	0,5
Monikon 2.	2	0,1	Monikon 2.	2	0,1

Taulukko 3: Nominien yleisimmät sijamuodot Aamulehden paikallisuutisissa ja presidentti Kekkosen puheissa.

Aamulehti			**Kekkonen**		
Sijamuoto	Lukumäärä	Osuus %	Sijamuoto	Lukumäärä	Osuus %
Nominatiivi	2467	32,9	Genetiivi	2731	28,5
Genetiivi	1758	23,5	Nominatiivi	2307	24,1
Partitiivi	1127	15,0	Partitiivi	1423	14,8
Inessiivi	476	6,4	Inessiivi	632	6,6
Illatiivi	426	5,7	Illatiivi	595	6,2

Taulukko 4: Yleisimmät konjunktiot Aamulehden paikallisuutisissa ja presidentti Kekkosen puheissa.

Aamulehti			**Kekkonen**		
Konjunktio	Lukumäärä	Osuus %	Konjunktio	Lukumäärä	Osuus %
Ja	322	41,0	Ja	553	43,4
Että	120	15,3	Että	260	20,4
Tai	62	7,9	Kuin	108	8,5
Jos	47	6,0	Kun	80	6,3
Mutta	47	6,0	Mutta	70	5,5
Kun	36	4,6	Sekä	59	4,6
Kuin	35	4,5	Jos	34	2,7
Sillä	23	2,9	Tai	23	1,8
Sekä	20	2,5	Vaikka	22	1,7

kin siitä, miten tekstit on rakennettu, ja ehkä myös sitä, miten varmoina asiat teksteissä ilmaistaan. Huomiota voi kiinnittää esimerkiksi siihen, että *tai-* ja *jos*-konjunktioiden yhteinen osuus Aamulehden uutisissa on noin neljätoista prosenttia, kun se presidentti Kekkosen uudenvuodenpuheissa jää neljään ja puoleen prosenttiin. Huomaamme myös, että yleisimmän *ja*-konjunktion suhteelliset osuudet ovat lähellä toisiaan, vaikka lukumääräisesti presidentti Kekkosen puheissa näitä sanoja on lähes kaksin verroin Aamulehden uutisiin verrattuna. Huomattava ero absoluuttisissa lukumäärissä selittyy sillä, että uudenvuodenpuheista koostettu aineisto on jonkin verran isompi ja konjunktiot ovat siinä kaikkiaan hieman yleisempiä.

Edellä olevat esimerkit kertovat meille jotakin siitä, miten kvantitatiivinen analyysi voi auttaa kielentutkijaa ohjaamalla hänen huomionsa tekstijoukkoja erottaviin ominaisuuksiin. Lukuja ja listoja voi käyttää myös toisella tavalla, tekstejä lukiessa syntyneiden työhypoteesien testaamiseen. Vakioraportteja tuottavien ohjelmien lisäksi voimme joskus ohjelmoida uusia työkaluja erityisongelmia varten. Joka tapauksessa on syytä muistaa, että pelkkiin lukuihin ja listoihin perustuvat johtopäätökset siitä, mitä tekstissä ilmaistaan ja mitä jätetään ilmaisematta, voivat mennä pahastikin metsään. Presidentti Kekkosen puheiden *tai-* ja *jos*-konjunktioiden vähäinen osuus saattaa merkitä sitä, että puheissa ilmaistaan vain vähän epävarmuutta verrattuna uutisiin, mutta yhtä hyvin epävarmuus saatetaan ilmaista siellä muilla keinoilla.

Kokemuksemme mukaan kieliteknologiaan perustuva tekstien kvantitatiivinen analyysi ei ole helppoa eikä se vapauta tutkijaa ajattelemasta itse. Aineistojen valmistelu on työlästä, joskin se myös antaa uuden näkökulman tekstin, jos tutkija tekee työn itse. Ohjelmien tuottamat luvut eivät myöskään ole valmista tutkimustulosta, vaan tutkimuksen apuvälineitä. Tulkitsevan tekstintutkijan on joka tapauksessa varmistettava hypoteesinsa menemällä itse tekstiin ja tutkimalla sitä suurennuslasilla.

KIRJALLISUUTTA

Niiden, jotka haluavat perehtyä tarkemmin tekstien puoliautomaattiseen merkkaukseen, kannattaa lukea Lehtisen ja Lounelan artikkeli (2004). Esimerkeissä käytetyt morfologiset merkitsimet selityksineen ovat Lingsoft OY:n verkkosivulla (Lingsoft). Uutismateriaaliin perustuvasta tekstilingvistisestä tutkimuksesta on esimerkki artikkelissa Heikkinen ym. (2005). Morfologisesti analysoidun tekstin perusteella tehtäviä laskelmia käsittelee Lounela (2005). Sanaluokan käsitteen ongelmia suomen kielen automaattisessa ja puoliautomaattisessa analyysissa käsittelee Heikkisen ja Lounelan artikkeli (2006). Muita esimerkkejä suomalaisesta kvantitatiivisesta lingvistiikasta ovat muun muassa Hakulinen ym. (1996) ja Saukkonen (2001).

LÄHTEET

Hakulinen, A. & Karlsson, F. & Vilkuna, M (1996). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus* Helsinki: Helsingin yliopiston yleisen kielitieteen laitos.

Heikkinen, V. & Lehtinen, O. & Lounela, M. (2005). Lappeenrantalaismies löi toista nenään baarissa, Uutisia ja uutisia. Teoksessa V. Heikkinen (toim.), *Tekstien arki, tutkimusmatkoja jokapäiväisiin merkityksiimme,* (s. 231–258). Helsinki: Gaudeamus.

Heikkinen, V. & Lounela, M. (2006). *Sanaluokka automaattisen analyysin kategoriana.* 31. Kielitieteen päivät Tallinnassa.

Lehtinen, O. & Lounela, M. (2004). A model for composing and (re-)using text materials for linguistic research. Teoksessa M. Nenonen (toim.), *Papers from the 30th Finnish Conference of Linguistics,* (s. 73–78). Joensuu: University of Joensuu.

Lingsoft: *Tags (Partial list)* HTTP://www2.lingsoft.fi/doc/fintwol/intro/tags.html. Lingsoft OY, Helsinki. Viitattu 25. 9. 2006

Lounela, M. (2005). Exploring morphologically analysed text material. Teoksessa A. Arppe ym. (toim.), *Inquiries into words, constraints and contexts. Festschrift in the honour of Kimmo Koskenniemi on his 60th birthday.* Helsinki: Gummerus.

Saukkonen, P. (2001) *Maailman hahmottaminen teksteinä. Tekstirakenteen ja tekstilajien teoriaa ja analyysia.* Helsinki: Yliopistopaino.

OF QUANTITATIVE ANALYSIS IN EXPLORING FINNISH TEXTS

Mikko Lounela, Research Institute for the Languages of Finland

The relation between linguistic research and language technology is a complex one. Linguistic research is basic research for language technology, and language technology offers linguists new points of view. Linguistic research for language technology is expensive and not visible, and it requires multidisciplinary skills. The same is true for using language technology in linguistic research. The models of science funding do not support this type of effort.

The article goes into using language technology and corpus linguistics in text linguistics. It introduces a model of preparing text materials to be used in linguistic research of scientific level, and using them to describe linguistic properties of different text sets. It demonstrates the model by comparing properties of local news in the Finnish newspaper Aamulehti with properties of new year's speeches given by president Urho Kekkonen.

Key words: Corpus linguistics, language technology, quantitative text linguistics.