Sijamuodot haussa –
tarvitseeko kaikkea hakutermien morfologista vaihtelua kattaa?

Kimmo Kettunen

Informaatiotutkimuksen sivuainetutkielma
Lokakuu 2005
Informaatiotutkimuksen laitos
Tampereen yliopisto

TAMPEREEN YLIOPISTO
Informaatiotutkimuksen laitos
KETTUNEN, KIMMO: Sijamuodot haussa – tarvitseeko kaikkea hakutermien morfologista vaihtelua kattaa?

Sivuainetutkielma, 45 s.
Informaatiotutkimus
Lokakuu 2005

Tutkimuksessa selvitettiin, tarvitseeko suomenkielisessä tekstitiedonhaussa kattaa hakutermien morfologista vaihtelua kaikkien sijamuotojen osalta. Suomen substantiivien muotojen runsautta tarkastellaan työssä ensin kieliaineistojen pohjalta. Neljän eri kieliaineiston analyysin perusteella todetaan, että suomen kielen kuusi yleisintä sijamuotoa (nominatiivi, genetiivi, partitiivi, inessiivi, elatiivi ja illatiivi) kattavat teksteissä esiintyvistä substantiivien muodoista jo noin 85 %. Lisäksi todetaan 11,3 miljoonan substantiivin automaattisen morfologisen analyysin avulla, että liitepartikkelit ja omistusliitteet, jotka kasvattavat substantiivien kieliopillisten muotojen laskennallisen määrän suureksi (noin 2000 muotoa), ovat niin harvinaisia, että niiden käsittely tiedonhaussa tuskin on tarpeen.

Kieliaineistojen analyysituloksen perusteella oletettiin, että kattamalla hakutermien muodon muuntelusta vain yleisimmän 3–6 sijamuodon muuntelu (3–12 erilaista hakutermin muotoa) saavutetaan riittävän hyviä hakutuloksia. Oletuksia testattiin kahdessa tekstitietokannassa: TUTKissa ja CLEF 2003 –kokoelmassa osittaistäsmäyttävällä InQuery-hakujärjestelmällä. Verrokkeina rajalliselle sijamuotomenetelmälle olivat lemmaus (perusmuotoistaminen) FINTWOL-ohjelmalla sekä karsinta Snowball-ohjelmalla.

Tekstitietokannoissa tehtyjen hakujen perusteella todettiin, että käyttämällä hauissa yhdeksää erilaista hakutermin muotoa saavutetaan optimaalisin hakutulos. Kahdellatoista hakutermin muodolla saavutetaan hiukan parempi keskitarkkuus, mutta ero yhdeksään muotoon on marginaalinen. Vertailtavilla menetelmillä saavutettujen hakutulosten tilastollisen merkitsevyyden testeissä todettiin, että CLEF 2003 –kokoelmassa tilastollisesti merkitseviä eroja ei ollut kuin lemmauksen ja yhden rajoitetun sijamuotoprosessin välillä. TUTK-kokoelmassa lemmauksen ja rajallisten sijamuotoprosessien erot olivat lähes aina tilastollisesti merkitseviä, mutta muiden menetelmien väliset erot eivät olleet.

Johtopäätökseksi työstä jää, että esitetty rajallinen suomen kielen hakutermien muuntelun kattaminen antaa parhaimmillaan kohtuullisen hyviä hakutuloksia. Menetelmää voi soveltaa myös muihin runsaasti taipuviin kieliin, koska se perustuu sanojen eri sijamuotojen erilaiseen frekvenssiin kielessä ja on siten yleistettävissä.

SISÄLLYS

1 Johdanto 4
2 Suomen substantiivien morfologista tilastoa 9
2.1 Taustaa 9
2.2 Morfologiset tilastot 11
2.3 Sananmuotojen laskennallinen määrä ja todellisten esiintymien määrä tekstiaineistossa 17
2.4 Morfologiset oletukset ja tutkimuskysymykset 19
3 Morfologisten oletusten testaus 20
3.1 Tausta 20
3.2 Vertailtava perustaso 22
3.3 Tulokset 25
3.3.1 TUTKin tulokset 25
3.3.2 Tulokset CLEF 2003 -kokoelmassa 26
3.4 Tulosten tilastollinen testaus 29
3.5 Hakunopeus 30
3.6 Morfologisen ohjelman sanakirjattomuus vai sanakirjallisuus? 31
4 Loppupäätelmät 34
5 Lähdekirjallisuutta 39

1 Johdanto

Suomenkielisten kokotekstidokumenttien haun yhtenä keskeisenä perusongelmana on pidetty suomen kielen sanojen mutkikasta taivutusta (Alkula 2000; Järvelin 1995; Nurminen 1986). Ongelman ratkaisemiseksi on olemassa erilaisia keinoja. Perinteinen ja teknisesti yksinkertaisin keino on ollut hakutermin katkaiseminen ja hakeminen katkaistulla hakutermillä taivutusmuotoisessa indeksissä. Hakujärjestelmän ja hakujen tuloksellisuuden kannalta hakutermin katkaiseminen toimii kohtuullisen hyvin, mutta käyttäjän kannalta ratkaisua ei voi pitää kovin hyvänä, koska sopiva hakutermin katkaisukohta saattaa olla käyttäjälle vaikeasti määriteltävissä. Hakutermien automaattinen katkaisu ilman kieliteknologista prosessointia ei myöskään toimi suomessa yhtä hyvin kuin morfologisesti yksinkertaisemmissa kielissä. (Alkula 2000; Koskenniemi 1983; Kunttu 2004.)

Kieliteknologisia apukeinoja hakutermien muodon muuntelun käsittelyyn on kehitetty useita erilaisia. Vahvin keinoista on *lemmaus*, hakutermien ja dokumenttikokoelman sanojen palauttaminen perusmuotoonsa. Toinen käytetty keino on *hakuvartaloiden muodostaminen* hakutermeistä ja hakuvartaloiden käyttö haussa taivutusmuotoisessa indeksissä. Kolmas keino on *karsinta* (stemming): siinä merkitykseltään yhtenevistä, muodoltaan erilaisista hakutermeistä ja tekstikokoelman sanoista muodostetaan vaihtelemattomia yhteisiä muotoja. Tuloksena ei välttämättä kuitenkaan ole sanan perusmuotoja tai kielitieteellisiä vartaloita. Vielä yhtenä keinona voi pitää hakutermin kaikkien erilaisten muotojen tuottamista ja hakemista niillä. Tätä ei kuitenkaan ole pidetty suomeen soveltuvana menetelmänä, koska haettavien muotojen laskennallinen määrä on suuri ja menetelmä olisi siten tehoton. (Alkula 2000; Galvez, Moya-Anegón & Solana 2005; Koskenniemi 1983, 1985.)

Eri menetelmiä suomen kielen hakutermien taipumisen kattamisessa on testattu erilaisissa hakuympäristöissä. Alkula (2000, nimellä Nurminen 1986) testasi

lemmausta, hakuvartaloiden tuottamista ja hakutermin katkaisemista Boolen hakuympäristössä (Trip-hakujärjestelmä). Alkulan mukaan parhaat hakutulokset saavutettiin lemmaamalla − erityisesti jos käytettiin indeksiä, johon myös tekstin yhdyssanat oli ositettu. Kunttu (2004) puolestaan testasi paljolti Alkulan asetelman kaltaisia keinoja osittaistäsmäyttävässä InQuery-hakujärjestelmässä (Broglio et al. 1995; Callan, Croft & Harding 1992). Kuntun mukaan parhaat tulokset saavutettiin edelleen lemmaamalla ja erityisesti yhdyssanat osiinsa indeksiin jakamalla. Taivutusvartalohakujen ja osittamattomaan perusmuotohakemistoon tehtyjen hakujen väliset erot olivat pienempiä.

Karsintaa ei ole totuttu pitämään suomen kieleen sopivana hakutermien käsittelyn menetelmänä (Koskenniemi 1983, 13), mutta sittemmin on osoitettu, että kielellisesti vajavainenkin karsintaohjelma, kuten vaikka Snowball (Porter 2001), toimii kohtuullisen hyvin myös suomen kielen tekstitiedonhaussa sekä muilla englantia vahvemmin taipuvilla kielillä (Hollink et al. 2004; Mayfield & McNamee 2003). Tampereen yliopiston tiedonhaun laboratoriossa Snowball-karsintaa ovat kokeilleet Airio (2005) ja Kettunen, Kunttu & Järvelin (2005). Airion testeissä käytetyissä CLEF 2003 -kokoelman kyselyissä Snowball selvisi paremmin kuin lemmausohjelma, jos lemmatussa indeksissä ei ollut ositettu yhdyssanoja. Ositettuja yhdyssanoja käytettäessä lemmaus toimi hiukan karsintaa paremmin. Kettusen testeissä Snowball hävisi selvästi sekä lemmaukselle että vartalotuotolle TUTK-kokoelmassa. Selkeää syytä Snowballin erilaiselle menestykselle eri kokoelmissa ei ole voitu osoittaa. Oletettavasti Snowball-ohjelma on kohtalaisen herkkä käytetylle testikokoelmalle ja sen kyselyille.

Myös Tomlinson (2002, 2003) on kokeillut karsintaa suomelle, mutta hänen käytössään on ollut erilainen karsintaohjelma, jonka toiminta perustuu sanakirjan käyttöön toisin kuin Snowballissa. Yleisemminkin karsinta on ollut 1990- ja 2000-luvuilla edelleen paljon käytetty menetelmä kokeellisessa tekstitiedonhaussa, ja erilaisia karsintaohjelmia on toteutettu yli 20:lle hyvinkin erityyppiselle kielelle. Indoeurooppalaisten kielten lisäksi karsintaa on sovellettu muun muassa amharaan

(Alemayehu & Willet 2003), arabiaan **(**Abu-Salem, Al-Omari & Evens 1999), malaijiin (Ahmad, Yussof & Sembok 1996) ja turkkiin **(**Sever & Bitirim 2003). Näyttääkin siltä, että karsinta on edelleen suosittu keino hakutermien morfologisen muuntelun kattamisessa ja menetelmän sovelluskielten määrä on lisääntynyt huomattavasti.

Kaikki mainitut sananmuotojen käsittelytavat toimivat hiukan eri tavoin. Yksi keskeinen tekijä ohjelmissa on, käyttääkö ohjelma suurta sanakirjaa ja säännöstöä vai toimiiko se pelkästään sääntöjen ja mahdollisen pienehkön sanalistan varassa. Sääntöpohjaisia systeemejä järjestelmistä ovat esimerkiksi karsintaohjelma Snowball sekä vartalontuotto-ohjelmat Finstems (Koskenniemi 1985) ja Stemma (Kettunen 1991). Suureen sanakirjaan ja sääntöihin perustuvat esimerkiksi FINTWOL-lemmausohjelma, Krovetzin karsintaohjelma (Krovetz 2000) sekä Tomlinsonsin karsintaohjelma (2002). Sanakirjan käytön tai käyttämättömyyden merkitys ohjelmissa on lähinnä siinä, miten yksinkertainen ja nopea ohjelma on toteuttaa. Kymmeniätuhansia sanoja sanakirjassaan sisältävän ohjelman laatiminen uudelle kielelle ei ole vaadittavan sanastotyön vuoksi kovin nopeaa. Laajan sanakirjan käyttäminen ohjelmassa johtaa myös siihen, että ohjelman sanastollinen kattavuus on helposti epäajantasainen sanakirjasta puuttuvien sanojen vuoksi.[1]

Taulukossa 1 on selkiytetty vartalontuottimen, karsintaohjelman ja perusmuotoistavan ohjelman suhteita kahden tekijän suhteen. Oleelliset tekijät ovat sananmuotojen (tai vartaloiden) analyysi ja generointi (tuottaminen) sekä se, käyttääkö järjestelmä sanakirjaa vai ei. Sanakirjaksi käsitetään tässä yhteydessä jokainen laaja sanalistaus, jossa on kymmeniä tuhansia sanoja sekä niiden mahdollisia kieliopillisia koodauksia tai sitä runsaampia sanakohtaisia tietoja. Myös ”sanakirjattomat” järjestelmät

[1] Sanakirjasta puuttuvien sanojen määrää ja sen mahdollista vaikutusta käsitellään lisää tutkielman jaksossa 4.4.

käyttävät yleensä jonkinlaisia poikkeussanalistoja, mutta näissä listoissa ei ole kuin muutamia satoja sanoja (Kettunen, Kunttu & Järvelin 2005).

Taulukko 1. Erilaisia sanojen kieliteknologisia käsittelyohjelmia ja niiden toimintatapa.

	Sanakirja käytössä	**Ei sanakirjaa**
Analyysi	TWOL (Koskenniemi 1983) Morfo (Jäppinen & Ylilammi 1986) Krovetzin karsintaohjelma (2000) Tomlinsonin karsintaohjelma (2002)	Lovinsin karsintaohjelma (1968) Porterin karsintaohjelma (Porter 1980) Snowball (Porter 2001)
Generointi	TWOL (Koskenniemi 1983)	Hahmotin (Alkula 2000) FinStems (Koskenniemi 1985) Stemma (Kettunen 1991)

Kolmantena oleellisena erottavan tekijänä voidaan pitää sitä, millaisia sananmuotoja ohjelmien analyysin tai tuotoksen jälkeen syntyy. Vartalontuottimet ja perusmuotoistavat ohjelmat palauttavat *kielitieteellisesti motivoituja* sananmuotoja tai sanavartaloita. Karsintaohjelman jäljiltä syntyvät käsitellyt sananmuodot saattavat olla kielitieteellisesti perusteltuja, mutta yleensä ne ovat vain riittävästi typistettyjä merkkijonoja, jotka kattavat mahdollisimman paljon merkitykseltään yhteenkuuluvien, muodoltaan erilaisten sanojen yhteistä taipumista. Tuotetut ”vartalot” eivät usein ole minkään sanan kielitieteellisiä vartaloita tai perusmuotoja, niiden ainoa motivaatio on tiedonhaullinen. (Galvez, Moya-Anegón & Solana 2005.)

Kaikki esitetyt ratkaisut hakutermien morfologisen muuntelun kattamisesta perustuvat siihen, että vaihtelevista sanoista tuotetaan tai analysoidaan kaikki mahdolliset esiintyvät muodot, myös harvinaiset. Lemmausohjelma analysoi kaikki sananmuodot ja tuottaa niistä perusmuodon, karsintaohjelma pyrkii tuottamaan

yhtenäisen ja vaihtelemattoman muodon kaikista samantyyppisistä sanoista[2]. Hakuvartaloilla puolestaan kyetään hakemaan kaikki sanan erilaiset vaihtelevat muodot.

On kuitenkin tunnettua, että kielten sananmuotojen esiintymät eivät ole tasaisesti jakautuneet. Joitain muotoja esiintyy huomattavasti enemmän kuin toisia, jotkut ovat äärimmäisen harvinaisia (Baayen 2001; Manning & Schütze 1999; Niemikorpi 1990, 1991). Erilaisten sananmuotojen jakauma ei kuitenkaan ole satunnainenkaan, vaan noudattaa tilastollisia todennäköisyyksiä ja pohjautuu myös sanojen merkitykseen. Esimerkiksi Karlsson (1986) osoittaa muutamalla yksinkertaisella aineistopohjaisella sanaesimerkillä, että suomen sanojen sijamuotojen esiintymät ovat sidoksissa sanojen merkitykseen. Paikannimi *Helsinki* saa esimerkiksi nominatiivin ja genetiivin lisäksi kohtalaisen runsaasti paikallissijaisia esiintymiä, kun taas henkilöön viittaava nimi *Martti* esiintyy lähinnä vain nominatiivissa. Samantapaiseen päätelmään päätyvät esimerkiksi Kostić, Markovic ja Baucal (2003) serbian analyysissaan. Heidän näkemyksensä selkeän aineiston tulkinnassa on kuitenkin yllättävän epävarma: he vain olettavat, että sanan semantiikan täytyy olla perimmäinen syy eri sijamuotojen frekvenssin vaihteluun.

Tarkoitukseni on tässä informaatiotutkimuksen sivuainetutkielmassa tutkia, miten vain osittainen suomenkielisten hakutermien taivutuksen kattaminen vaikuttaa hakutuloksiin osittaistäsmäyttävässä InQuery-hakujärjestelmässä. Suomen kielen sananmuotojen runsaus tarjoaa tähän hyvän mahdollisuuden, ja työssä käytetty menettelytapa sopii myös muihin runsaasti taipuviin kieliin. Vastaavanlainen kielen keskeisten sijamuotojen analyysi voidaan toteuttaa helposti kielikohtaisesti ja keskeisten sijamuotojen riittävyys tiedonhaussa voidaan testata.

Ennen tutkielman tutkimuskysymysten esittämistä ja niiden testaamista suomen kielen sanojen sijamuotojen jakaumaa lähestytään kielellisten tilastojen ja niiden

[2] Poikkeuksena ehkä kuitenkin ns. kevyt karsintaohjelma (light stemmer), joka käsittelee vain osan erilaisista kieliopillisista päätteistä ja tunnuksista. Snowballin suomen kielen karsintaohjelman lähdekoodin perusteella ohjelmaa voi osittain pitää kevyenä karsintaohjelmana, sillä ohjelma ei edes pyri käsittelemään kaikkia kieliopillisia päätteitä ja tunnuksia. Syy tähän voi kuitenkin olla se, että ohjelman tekijä ei ole tuntenut suomen kieltä, vaan on perustanut kuvauksensa kielioppeihin ja saamiinsa kommentteihin.

analyysin avulla. Tilastojen avulla osoitetaan erilaisten sijamuotojen jakauma, ja tältä pohjalta muotoillaan oletukset siitä, mitä muotoja haussa kannattaisi kattaa

2 Suomen substantiivien morfologista tilastoa

2.1 Taustaa

Tarkastelen tässä luvussa pääasiassa suomen substantiivien sijamuotojakaumaa. Osa jakaumista esitetään sananmuotojen konkreettisella esiintymätasolla, osa tyyppitasolla, erilaisten sananmuotojen määränä. Saadut tulokset ovat silti hyvin samansuuntaisia, ja esiintymä- ja tyyppijakaumien samankaltaisuus nähdäkseni lähinnä vahvistaa jakaumien todistusvoimaa.

Ennen substantiivien tilastotietoja esittelemistä perustelen lyhyesti sijamuotojen käsittelyn substantiivikeskeisyyttä. Tiedonhaun peruslähtökohtana hakutermejä valittaessa ja annettaessa on yleensä se, että haut suoritetaan substantiiveilla (Baeza-Yates & Ribeiro-Neto 1999, 169). Vaikka asiaa ei tiedonhakukirjallisuudessa varsinaisesti perustellakaan, tätä voidaan pitää sekä käytännöllisesti että ontologisesti perusteltuna: maailman olioihin viitataan substantiiveilla, ja niinpä substantiivit myös kantavat suuren osan kielellisistä merkityksistä (Lyons 1977). Lisäksi substantiivit ovat kielen avoimin sanaluokka, johon syntyy jatkuvasti lisää sanoja. Yksinkertainen korpustilasto vahvistaa substantiivien suuren osuuden teksteissä. Taulukossa 2 on substantiivien suhteellinen osuus kolmessa eri aineistossa. TUTKin (Sormunen 2000, 59; 53 893 lehtiartikkelia kolmesta eri sanomalehdestä vuosilta 1988–1992) ja HUT-korpuksen (Creutz & Linden 2004, 6; 32 miljoonan sananmuodon aineisto kirjoista, sanomalehdistä ja STT:n uutisista) osalta analyysit perustuvat FINTWOLilla suoritettuihin analyysiajoihin. Suomen kielen taajuussanaston (Saukkonen, Haipus, Niemikorpi & Sulkala 1979) aineistoon sanaluokat on merkitty käsin. TUTKin osalta luvussa ovat kaikki FINTWOLin antamat substantiivitulkinnat, HUT-korpuksen kohdalla on eroteltu yksiselitteiset ja kaikki substantiivitulkinnat.

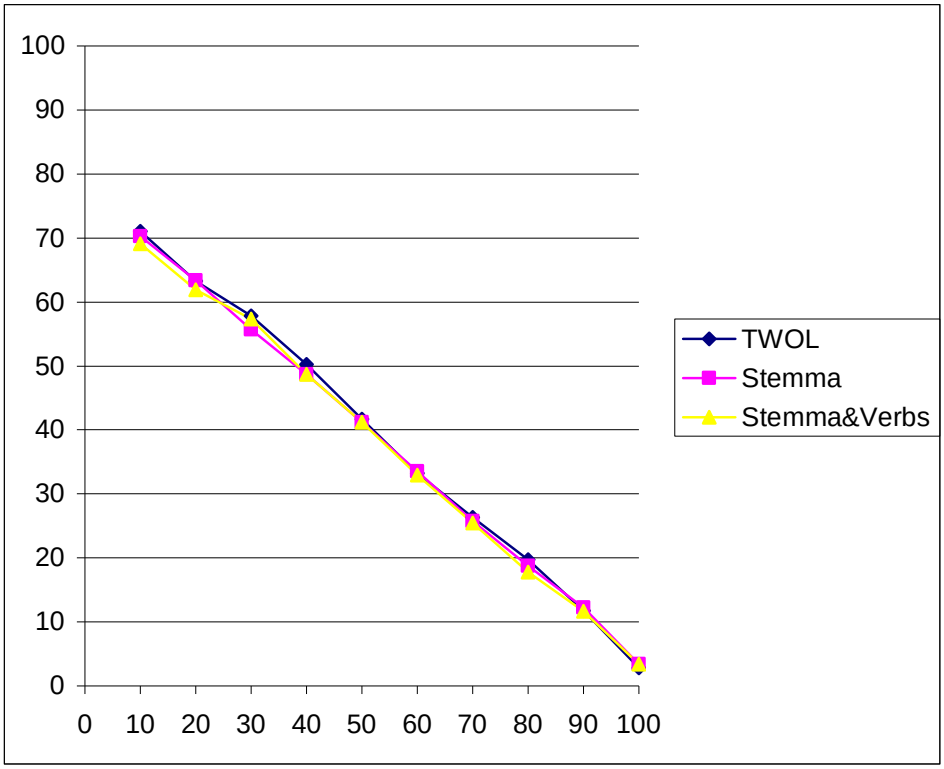
Taulukko 2. Substantiivien prosentuaalisia osuuksia kolmessa eri aineistossa.

Aineisto	**Substantiivit**	**Sanoja yhteensä**	**Prosentti-osuus**
TUTK (tyypit)			
Kaikki substantiivitulkinnat	514 795	783 983	65,66
HUT-korpus (esiintymät)			
yksiselitteiset substantiivi-tulkinnat	11 339 099	32 017 012	35,42
kaikki substantiivitulkinnat	14 745 103	32 017 012	46,05
Saukkonen et al. (1979, 16−19)			
koko aineisto (esiintymät)	146 335	408 301	35,83
koko aineisto (tyypit)	32 863	43 670	75,25
lehdet (esiintymät)	40 371	106 376	37,95
lehdet (tyypit)	13 235	18 782	70,47

Taulukosta ilmenee, että tulkinnasta riippuen substantiivit muodostavat 35–75 % tekstin sanoista. Yleisimmät muut teksteissä esiintyvät sanaluokat Saukkosen ja kumppaneiden mukaan (1979, 16–17) ovat verbit − esiintymätasolla 24,33 % koko aineistossa − ja adjektiivit, 9,22 % koko aineistossa. Sen jälkeen tulevat pronominit, 8,42 % koko aineistossa, ja adverbit ja konjunktiot - molemmat yhteensä 14,8 % koko aineistossa. Pronomineja, adverbeja ja konjunktioita käsitellään tiedonhaussa yleensä sulkusanoina, eli ne poistetaan hakutermien joukosta kokonaan. Adjektiivit ja verbit otetaan mukaan hakuun, mutta niiden vaikutus hakutulokseen on melko vähäinen.

Niinpä onkin hyvin perusteltua, että tiedonhaut tehdään enimmäkseen vain

substantiiveja käyttäen. Kokeellisesti näkemystä voi vielä vahvistaa esimerkiksi kuvalla 1, jossa suomen kielen vartalomuotohaussa on katettu myös verbien kaikki vaihtelevat vartalot: hakujen keskitarkkuudet huononivat hiukan, kun hakutermien verbien taipuminen oli katettu. Verbien taipumisen kattaminen ei siis tuonut hakutulokseen parannusta vaan pienen, joskin merkityksettömän, huononnuksen.



Kuva 1**.** Tarkkuus- ja saantikäyrät**,** liberaali relevanssitaso, TUTK. TWOL = lemmaus, Stemma = vartalotuotto, Stemma & Verbs = vartalotuotto nominaalisten sanaluokkien lisäksi verbeille. Liberaalissa relevanssitasossa on otettu mukaan kaikki relevantit dokumentit alkuperäisistä relevanssitasoista 1–3.

2.2 Morfologiset tilastot

Suomen kielen sijamuotojen jakaumista on kielitieteellisessä kirjallisuudessa julkaistu

kolme merkittävämpää tilastoa. Karlsson (1983, 308) esittää seuraavat sijamuotojen jakaumat suomen kielen nominaalisille sanoille. Aineistoina on neljä erilaista 5 000 saneen tekstiaineistoa. Taulukkoon 3 on otettu vain sijamuotojen prosenttiosuudet, keskeisimmät sijat on lihavoitu.

Taulukko 3. Suomen nominaalisten sanojen sanaluokkajakauma Karlssonin (1983, 308) mukaan.

Sijamuoto	**Prosenttiosuus**	**Tärkeimpien sijojen kumulatiivinen prosenttiosuus**
1) Nominatiivi (NOM)	29,5	
2) Genetiivi (GEN)	20,3	
3) Akkusatiivi	3,1	
4) Pronominien -t akkusatiivi	0,1	
5) OSMA (gen./akk./instr.)	0,8	
6) Partitiivi (PTV)	13,7	1,2 ja 6 (kieliopilliset sijat) = **63,5 %**
7) Essiivi	2,6	
8) Translatiivi	2,2	
9) Inessiivi (INE)	7,1	
10) Elatiivi (ELA)	4,4	
11) Illatiivi (ILL)	6,3	9–11 (sisäpaikallisijat) = **17,8 %.**
12) Adessiivi	4,4	
13) Ablatiivi	1,0	
14) Allatiivi	2,3	
15) Abessiivi	0,2	
16) Komitatiivi	0,1	
17) Instruktiivi	1,8	
18)		Kieliopilliset sijat ja sisä-paikallissijat (1–6, 9–11) = **85, 3 %**

Niin kutsutut kieliopilliset sijat, nominatiivi, genetiivi ja partitiivi, muodostavat aineistossa jo 63,5 % kaikista sijojen esiintymistä. Jos taulukon marginaaliset muodot riveiltä 3–5 sisällytetään lukuun, muotojen yhteinen kattavuus aineistossa on jo 67,5 %, yli kaksi kolmasosaa. Lopuista sijoista 8 on paikallissijoja: sisäpaikallissijat ovat riveillä 10–12, ulkopaikallissijat riveillä 13–15 ja

yleispaikallissijat riveillä 7–8. Niiden osuus on 30,3 %, josta sisäpaikallissijat muodostavat 17,8 %. Kieliopilliset sijat yhdessä sisäpaikallissijojen kanssa muodostavat aineistossa siis 85, 3 % sijamuotojen esiintymistä.

Samanlaisia jakaumia esiintyy myös muissa aineistoissa riippumatta aineiston koosta. Isossa suomen kieliopissa (Hakulinen et al. 2005, 1180) on eritelty Lauseopin arkiston aineiston kaikkien substantiivien sijajakauma 64 391 sananmuodon osalta. Jakauman prosenttiosuudet ovat taulukossa 4.

Taulukko 4. Substantiivien sijamuotojakauma Lauseopin arkiston aineistossa (Hakulinen et al. 2005, 1180).

Sijamuoto	**Prosenttiosuus**	**Kumulatiivinen prosenttiosuus**
1) Nominatiivi (NOM)	**26,3**	
2) Genetiivi (GEN)	**27,4**	
3) Partitiivi (PTV)	**16,2**	1–3 = **69,9 %**
4) Essiivi	2,6	
5) Translatiivi	1,6	
6) Inessiivi (INE)	**6,7**	
7) Elatiivi (ELA)	**5,3**	
8) Illatiivi (ILL)	**6,3**	6–8 = **18,3 %**
9) Adessiivi	3,6	
10) Ablatiivi	0,9	
11) Allatiivi	2,4	
12) Abessiivi	0,1	
13) Komitatiivi	0,1	
14) Instruktiivi	0,5	
15)		1–3, 6–8 = **88,2 %**

Kuusi yleisintä sijaa muodostavat tässä aineistossa siis 88, 2 prosenttia sijamuotojen esiintymistä.

Räsänen (1979, 24) esittelee lukuisia pieniä, erityyppisiä aineistoja, jotka ovat kukin kooltaan noin 2 000 sananmuotoa. Kolmessa aineistossa on asiatyylistä tekstiä, ja

näiden tekstien kuuden yleisimmän sijan prosenttiosuudet ovat taulukossa 5 (yhteensä 6 562 sananmuotoa).

Taulukko 5. Yleisimpien sijojen sijamuotojakaumat Räsäsen (1979) kirjakielisissä asiatekstiaineistoissa.

Sijamuoto	**Prosenttiosuus**	**Kumulatiivinen prosenttiosuus**
1) Nominatiivi (NOM)	25,9	
2) Genetiivi (GEN)	18,9	
3) Partitiivi (PTV)	15,6	1–3 = **60,4 %**
4) Inessiivi (INE)	5,9	
5) Elatiivi (ELA)	4,1	
6) Illatiivi (ILL)	7,8	4–6 = **17,8 %**
		1–6 **= 78,2 %**

Räsäsen muut aineistot ovat pääasiassa kaunokirjallista tekstiä, mutta niidenkin sijamuotojakaumat ovat varsin samanlaisia: nominatiivi, genetiivi ja partitiivi sekä inessiivi, elatiivi ja illatiivi muodostavat myös niissä 70–85 % sijamuodoista.

On tunnettua, että erilaisiin kielellisten ilmiöiden frekvensseihin vaikuttavat paljon käytetty aineisto ja sen koko (esimerkiksi Baayen 1993, 2001; Biber 1993a, 1993b). Tähän mennessä on havaittu varsin pienillä aineistoilla, että suomen kielen substantiivien sijamuotojakaumat ovat tietynlaisia. Ennen varsinaisten päätelmien tekemistä ja päätelmistä tehtävien tiedonhakuun liittyvien oletusten muodostamista onkin syytä tutkia sijamuotojakaumaa vielä kahdella erillisellä suurehkolla sana-aineistolla.

Tampereen yliopiston tiedonhaun laboratorion testikokoelma TUTK muodostuu 53 893 lehtiartikkelista vuosilta 1988–1992. Sananmuotoja, sanojen konkreettisia esiintymiä, on aineistossa Sormusen (2000) mukaan noin 12,5 miljoonaa. Tekstitietokannan sanojen frekvenssi-indeksistä tekemäni laskelman mukaan

sananmuotoja on yhteensä 12 109 779. Sananmuototyyppejä, erilaisia sananmuotoja, tässä määrässä on 719 011.

Sananmuototyyppien sijamuodoista olen tehnyt jakaumalaskelman seuraavasti. Sananmuototyypit analysoitiin FINTWOLilla ja FINTWOLin analyyseista laskettiin kaikki ne tulkinnat, joissa oli sanaluokkana N eli substantiivi. Lukema sisältää siis myös analyysin monitulkintaiset tapaukset. Tulokset ovat taulukossa 6.

Taulukko 6. TUTK-aineiston substantiivien tyyppien sijamuotojakauma-arvio FINTWOL-analyysin perusteella.

Sijamuoto	**Sijamuodon esiintymät (sanatyypit)**	**Prosentti-osuus**	**Kumulatiivinen prosenttiosuus**
1) Nominatiivi (NOM)	135 241	26,27	
2) Genetiivi (GEN)	109 385	21,24	
3) Partitiivi (PTV)	80 158	15,57	1–3 = **63,08 %**
4) Essiivi	11 263	2,18	
5) Translatiivi	11 460	2,22	
6) Inessiivi (INE)	31 007	6,02	
7) Elatiivi (ELA)	41 778	8,11	
8) Illatiivi (ILL)	38 392	7,45	6–8 = **21,58 %**
9) Adessiivi	23 467	4,55	
10) Ablatiivi	8 468	1,64	
11) Allatiivi	19 511	3,79	
12) Abessiivi	555	0,1	
13) Komitatiivi	1 586	0,30	
14) Instruktiivi	2 524	0,48	
Yhteensä	**514 795**	**99,92 %**	1–3, 6–8 = **84,66 %**

Tässä aineistossa kieliopilliset sijat ja paikallissijat muodostavat siis yhteensä noin 84,7 prosenttia aineiston substantiivien sijamuodoista, kun aineistoa tarkastellaan sananmuototyyppien tasolla ottaen huomioon kaikki mahdolliset, myös monitulkintaiset, substantiivitulkinnat jotka FINTWOL antaa.

Helsingin teknillisessä korkeakoulussa eri lähteistä kootun 32 miljoonan sananmuodon aineistosta (Creutz & Linden 2004) on ajettu vastaavanlainen FINTWOL-analyysi, mutta vain ohjelman antamista yksitulkintaisista substantiivitapauksista sananmuototasolla. Tästä aineistosta olen koonnut taulukkoon 7 tilastotiedot substantiivien kuuden yleisimmän sijamuodon suhteen (Creutz 2005).

Taulukko 7. Sijamuotokaumat 11,3 miljoonan substantiivin aineistossa sananmuototasolla.

Sijamuoto	**Määrä** (N = 11 339 099)	**Prosenttiosuus**
NOM	3 758 334	33,14
GEN	2 900 884	25,58
PTV	1 428 117	12,59
INE	819 333	7,23
ILL	593 513	5,23
ELA	520 101	4,59
Yhteensä	10 020 282	88, 36 %

Tässä aineistossa kuuden keskeisen sijamuodon yhteisosuus substantiivien joukossa oli noin 88,4 prosenttia.

Kootusti taulukot 3–7 kertovat siis yhdensuuntaisesti, että suomen kielen 14 morfologisesta sijamuodosta kuusi sijamuotoa muodostaa teksteissä noin 78–88 % kaikista sijojen esiintymistä. Tästä lukemasta voidaan alustavasti olettaa, että tiedonhaun kannalta vain nämä sijamuodot olisivat myös oleellisia katettavia hakutermien muodon vaihtelussa.

Ennen kuin esitän lopulliset tutkimukseni tutkimuskysymykset, esitän vielä suomen kielen substantiivien muotojen laskennallisen määrän ja niiden tosiasiallisten esiintymien välisen eron, joka tukee lisää tässä työssä tehtäviä oletuksia.

2.3 Sananmuotojen laskennallinen määrä ja todellisten esiintymien määrä tekstiaineistossa

Suomen kielen erilaisten sanamuotojen määrä esitetään kielitieteellisessä kirjallisuudessa yleensä toisiinsa liittyvien morfeemien kombinaatioiden laskelmana. Näin päädytään laskennallisesti suureen kieliopillisten sananmuotojen määrään. Substantiiveilla on mahdollista olla noin 2000 erilaista muotoa, adjektiiveilla noin 6600, verbeillä noin 12 000 (Karlsson 1983). Tässä työssä oleellisten suomen substantiivien mahdollisten eri muotojen laskelma on taulukossa 8 esitetty Karlssonin (1983, 356) mukaan.

Taulukko 8. Suomen kielen substantiivien sanamuotojen laskennallinen määrä Karlssonin (1983) mukaan. Luvussa ei ole mukana vaihtoehtoisia muotoja.

Sananmuodon mahdolliset kieliopilliset morfeemit	Luku	Sija	Omistusliite	Liitepartikkeli	**Yhteensä (luvut kerrottuina keskenään)**
Mahdollisten morfeemien määrä	2	13	6	12	1 872

Kuten taulukosta havaitaan, oleellisin vaikuttava tekijä substantiivien erilaisten muotojen laskennallisesti suureen määrään ovat omistusliitteet ja erityisesti liitepartikkelit (2*13***6*12** = 1872). Sananmuotojen laskennallista määrää suuresti kasvattavien liitepartikkelien todellisten esiintymien määrää onkin syytä tarkastella empiirisesti riittävän suuressa aineistossa, koska tutkittavat kielen piirteet ovat oletusarvoisesti suhteellisen harvaan esiintyviä.

Taulukossa 9 on esitetty jo aiemmin käytetyn Helsingin teknillisen korkeakoulun (HUT) aineistosta FINTWOLilla analysoidun 11 339 099 yksikäsitteisen substantiivitulkinnan omistusliitteet ja liitepartikkelit sekä niiden prosenttiosuus.

Taulukko 9. Omistusliitteiden ja liitepartikkelien esiintymämäärät Creutzin ja Lindenin (2004) aineistossa, Creutz (2005). Pienimmät prosenttiosuudet on jätetty laskematta.

Liitepartikkelit	**N**	**Prosenttiosuus**
-kin	28 578	0,25
-kAAn	9 005	0,08
-hAn	2 888	---
-kO	940	---
-pA	470	---
-kOs	90	---
Omistusliite		
3. persoona, yksikkö ja monikko	208 167	1,83
1. persoona, yksikkö	15 598	0,14
1. persoona, monikko	9 283	---
2. persoona, yksikkö	2 071	---
2. persoona, monikko	852	---

Kuten taulukon 9 luvuista näkyy, vain 3. persoonan omistusliite on niin yleinen (1,83 %), että sen voi arvella esiintyvän jokseenkin usein teksteissä. Seuraavaksi yleisin on liitepartikkeli –kin, mutta sen prosenttiosuus aineistossa on vain 0,25. Suuri osa omistusliitteistä ja liitepartikkeleista esiintyy aineistossa erittäin harvoin.

Esitettyjen eri kieliaineistoihin perustuvien laskelmien pohjalta näyttääkin siltä, että suomen substantiivien muotojen laskennallinen rikkaus ei ole kovin hyvä peruste sille, että kaikki eri hakutermin vaihtelevat muodot tulisi kattaa suomenkielisessä tekstitiedonhaussa. Ensinnäkin sananmuotojen suuresta laskennallisesta määrästä vastaavat erityisesti erilaiset sanan liitteet ja niiden mahdolliset yhdistelmät. Laskennallisesti kombinaatioiden määrä on suuri, mutta itse liitepartikkeleiden ja omistusliitteiden esiintymäfrekvenssit todellisissa sananmuodoissa ovat suurehkossakin aineistossa erittäin matalia, kuten taulukko 9 selvästi osoittaa.

Toisekseen suomen suhteellisen runsaasta sijamuotomäärästä (tässä 13) erottuu selvästi kuusi sijamuotoa, jotka eri aineistolähteiden ja laskelmien mukaan kattavat noin 78–88 % sijamuotojen esiintymistä asiatyylisissä teksteissä.

2.4 Morfologiset oletukset ja tutkimuskysymykset

Luvuissa 2.2 ja 2.3 esitettyjen analyysien valossa voi nyt muodostaa oletukset siitä, mitä sijamuotoja kannatta käyttää, jos suomenkielisen tekstitietokannan nominaalisten hakutermien muoto-opillisesta vaihtelusta halutaan kattaa vain merkittävin osa. Taulukkoon 10 on koottu tilastotietojen pohjalta muodostetut hakuprosessit ja selitetty niiden kattamien hakutermien sijamuodot. Hakuprosesseissa on otettu erikseen huomioon sijamuotojen yksikkö- ja monikkomuodot, koska niiden vaikutus hakuun on myös selvästi vaikuttava tekijä.

Taulukko 10. Kieliaineistojen analyysin pohjalta tehdyt oletukset tiedonhaussa katettaviksi sijamuodoiksi. Red = reduced. (* = luku ei sisällä rinnakkaismuotoja, joita voi olla monikon genetiivissä ja partitiivissa sanatyyppikohtaisesti).

Hakuprosessissa katetut hakutermien muodot	**Hakumuotojen hakutermi-kohtainen määrä**	**Hakuprosessin nimi**
NOM-GEN-PTV, pelkät yksikkömuodot	3	Red_Min1
NOM-GEN-PTV, yksikkö- ja monikkomuodot	6	Red_Min2
NOM-GEN-PTV, yksikkö- ja monikkomuodot; INE-ELA-ILL, vain yksikkömuodot	9*	Red_Max1
NOM-GEN-PTV, INE-ELA-ILL, yksikkö- ja monikkomuodot	12*	Red_Max2

Kuten taulukosta 10 nähdään, syntyy haussa katettavia hakutermikohtaisia muotoja enimmilläänkin vain noin 12, ei suinkaan esimerkiksi 26 (2*13 sijaa) tai 2000, kuten sananmuotojen laskennallisen määrän arvioista voitaisiin päätellä. Nykyisenkaltaisissa hakujärjestelmissä ja tehokkaissa tietokoneissa hakutermien määrä tällaisenaan on riittävän alhainen, jotta ne kaikki voidaan hyvin hakea.

Tämän työn tutkimuskysymyksiksi muodostuvat seuraavat kaksi kysymystä:

- Saavutetaanko rajallisella suomen hakutermien muodon kattamisella riittävän hyviä hakutuloksia?
- Jos menetelmällä ylipäänsä saavutetaan hyviä hakutuloksia, mikä on paras yhdistelmä katettujen sijamuotojen ja hakutermien määrän suhteen?

3 Morfologisten oletusten testaus

3.1 Tausta

Tiedonhaun laboratoriomallissa on tyypillistä testata erilaisia hakutapoja vakioiduissa testikokoelmissa valitun evaluointikriteerin mukaan. Yleisimmin käytetty evaluointikriteeri on hakujen saanti ja tarkkuus. (Baeza-Yates & Ribeiro-Neto 1999, 74.) ”Hakutuloksen saanti kuvaa hakutuloksen osumien suhdetta kaikkiin relevantteihin dokumentteihin”, ja se esitetään usein prosenttilukuna välillä 0–100. Hakutuloksen tarkkuus puolestaan ”kuvaa hakutuloksen osumien suhdetta kaikkiin löydettyihin dokumentteihin”, ja sitä kuvataan myös prosenttilukuna 0–100. (Järvelin 1995, 55–56.) Usein hakujen saanti- ja tarkkuustulokset esitetään hakujen keskimääräistä tuloksellisuutta kuvaavana käyränä, johon on yhdistetty yksittäisten hakujen tulokset (Järvelin 1995, 60; Baeza-Yates & Ribeiro-Neto 1999, 76–77; Meadow, Boyce & Kraft 2000, 322−325).

Testit edellisessä luvussa esitetyille tutkimuskysymyksille on suoritettu Tampereen

yliopiston tiedonhaun laboratorion kahdessa tekstitietokannassa: TUTKissa ja CLEF 2003 -kokoelmassa. TUTK-tekstitietokannasta käytettiin 30:a kyselyä kannan 35 mahdollisesta kyselystä. TUTKissa on käytössä moniarvoinen dokumenttien relevanssiluokitus, josta tässä työssä on muodostettu kolme erilaista relevanssitasoa: liberaali, normaali ja tiukka. Liberaalissa relevanssitasossa hakuun otetaan mukaan kaikki tulokset, jotka ovat vähänkin relevantteja (alkuperäiset relevanssitasot 1–3). Normaalissa relevanssitasossa mukaan otetaan vain erittäin relevantit ja relevantit dokumentit (alkuperäiset relevanssitasot 2–3). Tiukalla relevanssitasolla mukaan otetaan vain erittäin relevantit dokumentit (alkuperäinen relevanssitaso 3). (Kunttu 2004; Sormunen 2000.)

CLEF 2003 -kokoelmassa kyselyitä on 60. Kokoelmassa käytetään vain yhtä relevanssitasoa (Peters 2003). Molemmissa kokoelmissa on jokseenkin sama määrä dokumentteja: TUTKissa 53 893 (Sormunen 2000), CLEF 2003:ssa 55 344 (Airio 2005).

Kyselyt on TUTKin hakuaiheista muodostettu osin automaattisesti, osin manuaalisesti. Kettunen (2005) esittää menetelmän, jossa hakutermien vaihtelu katetaan lisäämällä hakuvartaloiden jälkeen kuhunkin vartaloon mahdollisesti liittyvä sijapääte säännöllisillä lausekkeilla. Kyselytiedostot on tässä menetelmässä tuotettu puoliautomaattisesti: ensin vartalo-ohjelmalla ja Unix-skripteillä on tuotettu hakutermeistä niiden vartalot sekä InQuery-kyselyn alustava rakenne. Sen jälkeen hakutiedostoa on editoitu käsin tarvittavien hakutermin muotojen kattamiseksi.

Tässä työssä menetelmää sovellettiin niin, että aiemmin tuotettuja kyselyiden pohjatiedostoja editoitiin uudestaan kullekin prosessille sopivaksi. Ensin tuotettiin vähiten sijamuotoja käyttävä prosessi Red_Min1 poistamalla aiemmista kyselytiedostoista turhat sijamuodot. Jatkossa tätä tiedostoa editoitiin vaihe vaiheelta niin, että kutakin prosessia varten syntyi kyselytiedosto, jossa oli kunkin prosessin vaatima määrä hakutermien sijamuotoja. Hakuaiheiden substantiiveista ja adjektiiveista on tehty kyselytiedostoon hakuprosessien mukaiset muodot, verbeistä

on kyselytiedostossa vain perusmuoto

CLEF 2003 -kokoelmaa varten luotiin kyselytiedostot editoimalla käsin neljä kyselytiedostoa, joissa oli prosessien mukaiset sijamuodot hakutermeistä.

Kyselyiden muokkaaminen osin käsin hakuaiheista *simuloi* hakutermien vaihtelun käytöstä haussa. Oikeassa tuotannollisessa järjestelmässä hakutermien vaihtelevat muodot tuotettaisiin automaattisesti käyttäjän antamasta hakutermin perusmuodosta.

Vertailukohtana rajalliselle hakutermien muodon vaihtelun kattamiselle käytetään lemmausta ja karsintaa. Lemmauksen on aiemmin osoitettu toimivan suomen kieltä käsittelevistä menetelmistä parhaiten osittaistäsmäytysympäristössä (Airio 2005; Kettunen, Kunttu & Järvelin 2005; Kunttu 2004). Karsinnan suhteen on saavutettu hiukan ristiriitaisia tuloksia: Airion (2005) testeissä Snowball-karsinta menestyi lähes yhtä hyvin kuin lemmaus FINTWOLilla CLEF 2003 –kokoelmassa, mutta TUTKissa karsinnalla saavutettiin huonompia tuloksia (Kettunen, Kunttu & Järvelin 2005). Lisäksi kyselyt on ajettu myös kokonaan käsittelemättömillä, suoraan hakuaiheesta otetuilla hakutermeillä.

3.2 Vertailtava perustaso

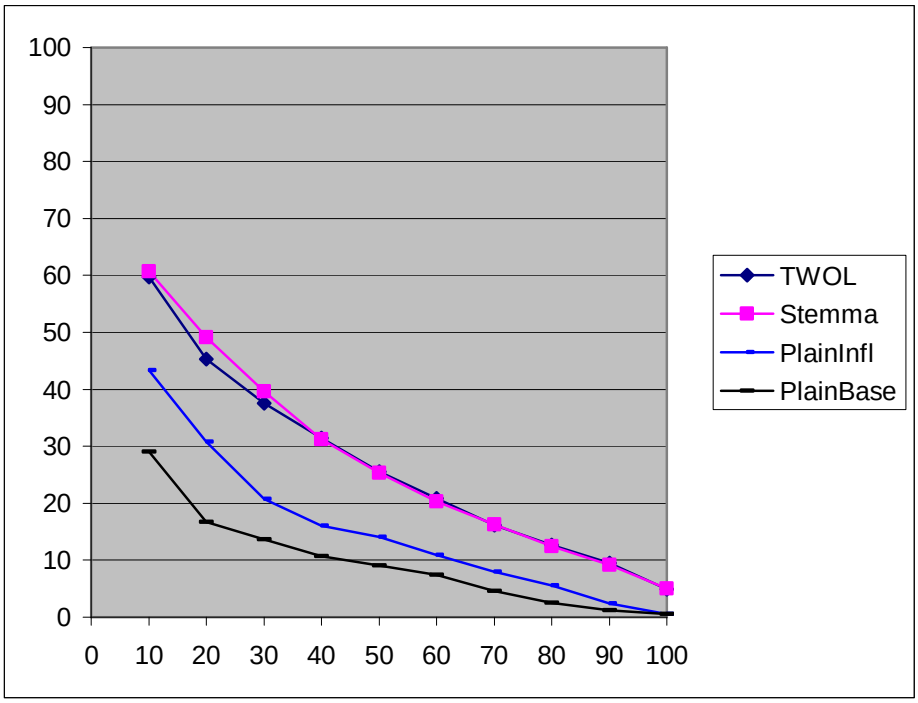
Kansainvälisissä tiedonhakuartikkeleissa näyttää muodostuneen yleiseksi tavaksi, että erilaisten hakutermien morfologisen muuntelun kattamisen menetelmien vertailukohdaksi otetaan kokonaan käsittelemättömien hakutermien antama perustulos (esim. Braschler & Ripplinger 2004; Hollink et al. 2004; Mayfield & McNamee 2003). Tapa pohjautunee alalla alkujaan vallinneeseen englanninkielisen tekstitiedonhaun perinteeseen. Englannin tavoin vähän taipuville kielille menettelyä voikin pitää realistisena, mutta enemmän taipuvien kielten kanssa menettely vaikuttaa kyseenalaiselta. Näin saadaan kyllä osoitetuksi hyviä saannin ja tarkkuuden suhteellisia suorituskykyparannuksia, mutta jos vertailukohta on jo valmiiksi alhainen, on suorituskyvyn selkeä parantuminen osin ilmiselvää vahvasti taipuvassa

kielessä. Tällaisissa koejärjestelyissä ei siis suoranaisesti osoiteta perustasoon vertailtavien menetelmien hyvää suorituskykyä, vaan hyvä suorituskyky ennalta valittuun matalaan ja osin itsestään selvään perustasoon nähden.

Esimerkiksi Mayfield ja McNamee (2003) saavuttavat kokonaan käsittelemättömillä suomen hakutermeillä noin 20 % keskitarkkuuden, joka on kuitenkin vain noin 63 prosenttia testien parhaan menetelmän, n-grammien, keskitarkkuudesta. Tomlinson (2003) saa keskitarkkuudeksi käsittelemättömillä hakutermeillä 30.1 %, mikä puolestaan on vain noin 54 % parhaasta keskitarkkuudesta, joka saavutetaan sanakirjaa käyttävällä karsintaohjelmalla. Heikohkoon perushakutulokseen nähden saavutetaan siis suurehkoja suhteellisia parannuksia, joilla hieman näennäisen oloisesti perustellaan oman menetelmän suorituskykyä.

Runsaasti taipuvien kielten hakutermien käsittelyssä olisi nähdäkseni mielekkäämpää ottaa menetelmien vertailussa vertailukohdaksi mahdollisimman hyvä hakutulos, ei mahdollisimman huonoa. Tässä työssä erilaisia keskenään kilpailevia menetelmiä verrataan parhaisiin tuloksiin, käytännössä yleensä lemmaamalla saavutettuihin hakutuloksiin. Kokonaan käsittelemättömien sanojen tuloksia esitellään taulukoissa myös, mutta erilaisten menetelmien antamia tuloksia verrataan kuitenkin vain parhaisiin saavutettuihin tuloksiin.

Kokonaan käsittelemättömien hakutermien käytössä haun verrokkina on toinenkin periaatteellinen ongelma. Kettunen (2005, alla alkuperäisistä kuvista vain Kuva 2 muokattuna) osoittaa, että kokonaan käsittelemättömillä hakutermeillä saanti ja tarkkuus runsaasti taipuvassa kielessä vaihtelevat jo sen mukaan, mitä hakutermien muodot kyselyssä sattuvat olemaan. Jos kyselyyn laitetaan hakuaiheen sanat sellaisinaan – mikä on normaali laboratoriotestauksen lähtökohta – tulos on selvästi parempi kuin silloin, jos hakutermit perusmuotoistetaan ja kysely tehdään taivutusmuotoiseen hakemistoon.



Kuva 2. Käsittelemättömien hakutermien tulokset taivutusmuotoisessa suomenkielisessä indeksissä (TUTK). Lyhyet kyselyt, normaali relevanssitaso. TWOL = lemmaus, Stemma = hakuvartalot, PlainInfl = käsittelemättömät hakutermit suoraan hakuaiheesta, PlainBase = hakuaiheen hakutermit perusmuotoistettuina.

Käytännön hakujärjestelmää, esimerkiksi www-hakukonetta, käytettäessä käyttäjä olisi valinnan edessä: koska www-hakukoneet eivät tarjoa sanojen katkaisua, eri muotojen tuottamista tai lemmausta käytettäväksi, käyttäjä joutuu antamaan hakutermin joko perusmuodossaan tai jossain satunnaisessa taivutetussa muodossa hakuun. Kyselyn suorittaja ei voisi siis tietää, missä muodossa annettuna hakutermi antaa parhaita tuloksia eikä hakujärjestelmä auttaisi kyselyn hakutermien muotojen käsittelyssä mitenkään. Oletettavaa olisi, että käyttäjä antaa hakutermistä perusmuodon, joka siis tuottaa selkeästi heikomman tuloksen taivutusmuotoisesta indeksistä haettaessa.

3.3 Tulokset

3.3.1 TUTKin tulokset

InQuery-ajoissa saadut hakutulokset TUTK-kokoelmassa ovat taulukossa 11.

Taulukko 11. FINTWOL, Snowball sekä rajoitettu hakutermien sijojen kattaminen TUTKissa. Erot on ilmoitettu absoluuttisina prosenttiyksikköinä FINTWOLiin verrattuna.

	Liberaali relevanssi Keskitarkkuus saantitasoilla (%)	**Normaali relevanssi** Keskitarkkuus saantitasoilla (%)	**Tiukka relevanssi** Keskitarkkuus saantitasoilla (%)
FINTWOL	37.8	35.0	24.1
Red_Max2	32.7 (-5.1)	30.0 (-5.0)	21.4 (-2.7)
Red_Max1	32.4 (-5.4)	29.6 (-5.4)	21.3 (-2.8)
Red_Min2	30.9 (-6.9)	28.0 (-7.0)	21.0 (-3.1)
Snowball	29.8 (-8.0)	27.7 (-7.3)	20.0 (-4.1)
Red_Min1	26.4 (-11.4)	23.9 (-11.1)	18.9 (-5.2)
Plain (sanat hakuaiheen muodoissa)	19.6 (-18.2)	18.9 (-16.1)	12.4 (-11.7)

Tuloksista nähdään, että hakutermien kolmen yleisimmän sijamuodon pelkkien yksikkömuotojen kattaminen haussa (Red_Min1) ei anna kovin hyvää hakutulosta. Hakutulos jää keskitarkkuudeltaan 5.2–11.4 prosenttiyksikköä FINTWOLilla saaduista parhaista keskitarkkuuksista, vaikka ylittääkin selvästi käsittelemättömien hakutermien tulokset. Erot tasoittuvat selvästi, kun mukaan otetaan myös samojen sijamuotojen monikkomuodot (Red_Min2). Hakutulokset jäävät enää 3.1–6.9 prosenttiyksikköä FINTWOLin tuloksista. Normaalilla ja liberaalilla relevanssitasolla tulokset parantuvat vielä noin 1,5 prosenttiyksikköä, kun sijoihin lisätään kolmen sisäpaikallissijan yksikkömuodot (Red_Max1). Tiukalla relevanssitasolla tulos parantuu tällöin vain 0.3 prosenttiyksikköä edelliseen prosessiin verrattuna. Kattavin prosessi, Red_Max2, sisältää kuusi sijamuotoa monikossa ja yksikössä, siis 12 kunkin

hakutermin erilaista muotoa ynnä mahdolliset monikon partitiivin ja genetiivin rinnakkaismuodot. Prosessin antamat tulokset parantavat keskitarkkuutta Red_Max1-prosessiin verrattuna vain 0,1–0,4 prosenttiyksikköä. Ainakaan TUTKin tässä kyselyjoukossa sisäpaikallissijojen monikkomuotojen kattaminen ei siis tuota enää oleellista parannusta hakujen keskitarkkuuteen.

Kolme rajallisista sijamuotomenetelmistä saavuttaa paremman keskitarkkuuden kuin Snowball-karsinta TUTKissa. Erot Snowballiin eivät ole kuitenkaan kovin suuria.

Parhaiksi rajallisiksi sijamuotoprosesseiksi TUTKissa jäävät siis Red_Max2 ja Red_Max1. Niiden keskinäinen ero on vain 0.1–0.4 prosenttiyksikköä eri relevanssitasoilla, joten erot ovat täysin marginaalisia. Vähemmän hakutermien muotoja käyttävä prosessi Red_Min2 puolestaan jää näistä kahdesta prosessista noin 1.5 prosenttiyksikköä normaalilla ja liberaalilla relevanssitasolla, mutta vain 0.3–0.4 prosenttiyksikköä tiukalla relevanssitasolla.

3.3.2 Tulokset CLEF 2003 -kokoelmassa

Tulokset CLEF 2003 -kokoelmassa ovat taulukossa 12.

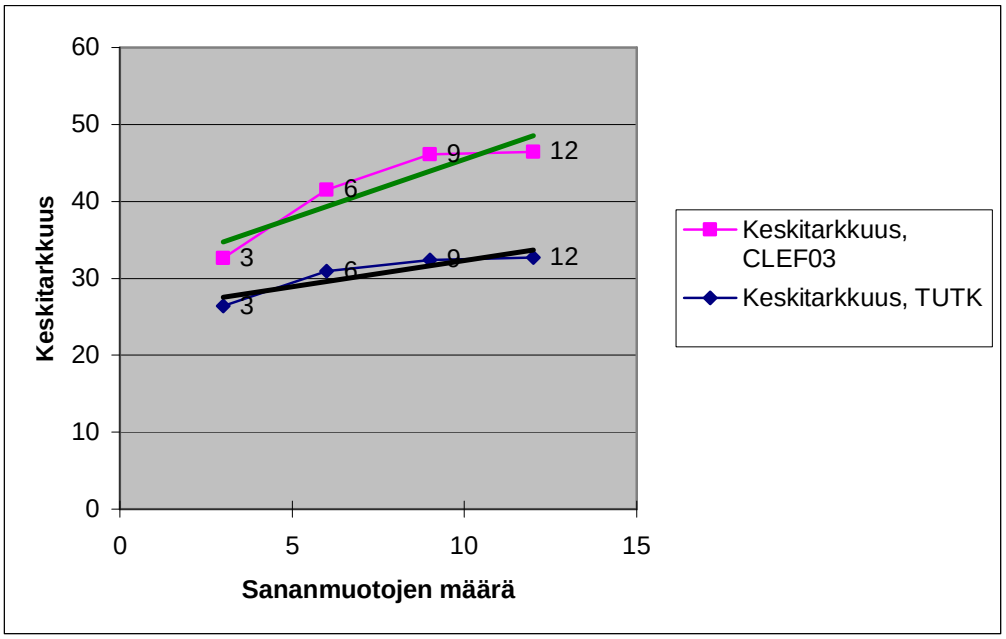
Taulukko 12. CLEF 2003 -kokoelman tulokset. Fintwolin, Snowballin ja Plainin tulokset ovat Airiolta (2005).

Menetelmä	**Keskitarkkuus saantitasoilla (%)**
FINTWOL	
-ositetut yhdyssanat	50.5
-osittamattomat yhdyssanat	47.0 (-3.5)
Snowball	48.5 (-2.0)
Red_Max2	46.4 (-4.1)
Red_Max1	46.1 (-4.4)
Red_Min2	41.5 (-9.0)
Red_Min1	32.6 (-17.9)
Plain (sanat hakuaiheen muodoissa)	31.0 (-19.5)

CLEF 2003 -kokoelmassa saadut tulokset poikkeavat jonkin verran TUTKin tuloksista. Suurin ero on siinä, että Snowball-karsinnalla saadaan tässä kokoelmassa selvästi paremmat tulokset kuin TUTKissa. TUTKissa rajoitettu sijamuotojen kattaminen menestyy Snowballia paremmin, mutta CLEF 2003:ssa Snowball menestyy rajoitettuja sijamuotoprosesseja paremmin sekä paremmin kuin FINTWOL, jos yhdyssanoja ei ole ositettu indeksiin. Parhaat rajoitetut sijamuotoprosessit jäävät kuitenkin FINTWOLista vain 0,6–0,9 prosenttiyksikköä silloin, kun FINTWOLille on käytetty indeksiä, jossa yhdyssanoja ei ole ositettu.

Molempien kokoelmien tulokset ovat kuitenkin sikäli yhdensuuntaisia, että ne osoittavat parhaiden rajoitettujen sijamuotoprosessien tuottavan jokseenkin vertailukelpoisia hakutuloksia parhaisiin menetelmiin (FINTWOL, Snowball) verrattuna. Red_Max1 ja Red_Max2 menestyvät molemmissa kokoelmissa hyvin. Red_Max2 saavuttaa TUTKissa 86,5 prosenttia parhaasta hakutuloksesta liberaalilla relevanssitasolla, CLEF 2003:ssa 91,9 %. Red_Max1 puolestaan saavuttaa TUTKissa 85,7 prosenttia parhaasta hakutuloksesta liberaalilla relevanssitasolla, CLEF 2003:ssa 91,3.

Kuvaan 3 on yhdistetty molemmissa kokoelmissa saavutetut kaikkien rajallisten sijamuotomenetelmien keskitarkkuudet ja käytetyt hakutermien määrät. TUTKin osalta kuvaan on piirretty vain liberaalin relevanssitason keskitarkkuuksien käyrät.



Kuva 3. Rajallisten sananmuotoprosessien saavuttaman keskitarkkuuden ja käytettyjen hakutermien muotojen välinen suhde. Kuvassa oleva paksumpi viiva on lineaarinen trendiviiva.

Käyristä näkee selkeämmin saman asian kuin keskitarkkuuksien taulukkoesityksestä: hakujen keskitarkkuudet nousevat tasaisesti yhdeksään hakutermin muotoon saakka, mutta sen jälkeen keskitarkkuuden kasvu hidastuu selvästi. Tämän tulkinnalle on kaksi mahdollisuutta: joko keskitarkkuus ei tule juuri nousemaan riippumatta siitä, mitä muotoja yhdeksän muodon jälkeen käytetään, tai sitten Red_Max2:ssa käytetyt sisäpaikallissijojen monikkomuodot eivät ole parhaat mahdolliset lisämuodot. Tämän selvittäminen vaatisi ainakin kahden uuden prosessin tekemistä. Sijamuotojen jakaumia taulukoista 3–8 tutkimalla seuraavat kaksi sijamuotoparia näyttävät kiinnostavimmilta: adessiivin ja ablatiivin yksikkömuodot tai translatiivin ja essiivin yksikkömuodot. Näitä ei kuitenkaan lähdetä enää empiirisesti testaamaan tässä työssä, koska on oletettavaa, etteivät keskitarkkuudet tule nousemaan enää suuresti. On ilmeistä, että hakutermien morfologisen muuntelun kattamisella saavutettavat parhaat keskitarkkuudet eri menetelmillä InQuery-hakuympäristössä ovat ne, joita ovat saavuttaneet Kunttu (2004), Airio (2005) ja Kettunen, Kunttu & Järvelin (2005).

TUTKissa parhaat saavutetut keskitarkkuudet ovat noin 35–38 %, CLEF 2003:ssa 46–50 %.

3.4 Tulosten tilastollinen testaus

Parhaiten menestyneitä rajallisia sijamuotoprosesseja Red_Max1, Red_Max2 ja Red_Min2 verrattiin tilastollisesti TUTK-kokoelmassa ja CLEF 2003 -kokoelmassa FINTWOLiin ja Snowballiin. Tilastotestinä käytettiin Friedmanin merkitsevyystestiä (Conover 1980, 299–302), koska keskenään vertailtavia menetelmiä oli useita. Tilastotestien tulokset ovat taulukoissa 13 ja 14. Taulukkoihin on koottu tilastotestien tuloksista vain ne vertailut, joissa oli tilastollisesti merkitseviä eroja ($p = 0.05$, $p = 0.01$ tai $p = 0.001$).

Taulukko 13. Erojen tilastollisen merkitsevyyden testaus parhaiden menetelmien osalta TUTKissa.

	Liberaali relevanssitaso	**Normaali relevanssitaso**	**Tiukka relevanssitaso**
FINTWOL >	Tilastollisesti erittäin merkitsevä ero kaikkiin muihin menetelmiin ($p < 0.001$)	Tilastollisesti erittäin merkitsevä tai merkitsevä ero kaikkiin muihin menetelmiin	Tilastollisesti erittäin merkitsevä tai merkitsevä ero kaikkiin muihin menetelmiin paitsi Red_Max1:een
Red_Max2 >	---	Snowball $p = 0.02$	---
Red_Max1 >	---	Snowball $p = 0.01$	---
Red_Max1 >	---	Red_Min2 $p = 0.03$	---

Taulukko 14. Erojen tilastollisen merkitsevyyden testaus parhaiden menetelmien osalta CLEF 2003 –kokoelmassa.

Menetelmä	
FINTWOL: ositetut yhdyssanat >	Red_Min2 p = 0.005
FINTWOL: osittamattomat yhdyssanat >	Red_Min2 p = 0.02

Tilastollisesti merkitseviä eroja menetelmien välillä oli siis enemmän TUTKissa. Osaksi tämä voi johtua kokoelmasta käytetyn hakuaihemäärän pienuudella (30), osaksi kyselyiden erilaisuudesta verrattuna CLEF 2003 –kokoelman kyselyihin. Lopputulokseksi jää kuitenkin, että työssä esitellyt uudet menetelmät menestyivät paremmin CLEF 2003 –kokoelmassa.

3.5 Hakunopeus

Hakutermien täysien taivutettujen muotojen rajallisen määrän käyttämisessä haussa on myös muita vaikuttavia tekijöitä kuin menetelmällä saavutettava saannin ja tarkkuuden taso. Kettunen, Kunttu & Järvelin (2005) sekä Kettunen (2005) esittelevät hakuvartaloiden käyttöä hiukan eri menetelmin. Osittaistäsmäyttävässä hakujärjestelmässä haun toteuttaminen hakuvartaloita käyttäen johtaa helposti hitaisiin hakuihin, koska hakuvartalot täsmäävät teksti-indeksissä suureen määrään myös täysin epärelevantteja sanoja. Asiaa voi korjata rajoittamalla osumia menetelmällä, jonka Kettunen (2005) esittää. Silti haut eivät tälläkään menetelmällä ole kovin nopeita ja tekstitietokannan indeksin läpi käyminen on haun hitain vaihe. Kokonaisia taivutettuja hakutermejä käytettäessä haku indeksistä on kuitenkin nopeaa, eikä käytettyjen hakutermien määrän vaihtelu (3–12 hakutermiä) tunnu vaikuttavan haun nopeuteen juurikaan. Käytännöllisiä hakusovelluksia varten menetelmä vaikuttaa siis varsin lupaavalta, koska siinä yhdistyvät sekä hyvä haun keskitarkkuus että riittävä hakunopeus.

3.6 Morfologisen ohjelman sanakirjattomuus vai sanakirjallisuus?

Tutkielman johdannossa käsiteltiin lyhyesti erilaisten hakutermien muodon vaihtelua käsittelevien ohjelmien eroja ja yhtenevyyksiä. Yhtenä keskeisenä erontekona pidettiin sanakirjan käyttämistä tai käyttämättömyyttä. Tässä työssä esitelty lähestymistapa pohjautuu sanakirjattomaan hakutermien keskeisten sijamuotojen tuottamiseen. Testeissä verrokkina käytetty FINTWOL puolestaan on suurta sanakirjaa käyttävä ohjelma.

Laajoihin sanakirjoihin perustuvat kieliteknologiset ohjelmat törmäävät käytännön sovelluksissa siihen, että analysoitavissa teksteissä tai annetuissa kyselyissä on aina sanoja, jotka eivät sisälly niiden sanakirjoihin. Useimmiten tällaiset sanat ovat erilaista teknistä sanastoa, nimiä (paikannimet, tuotenimet, henkilönnimet jne.), vierasperäisiä sanoja, kaavoja tai kaavoja ja numeroita sisältäviä sanoja, kirjakielestä poikkeavia sanoja ja lyhenteitä (Sampson 1989). Vastaavasti myös väärin kirjoitetut sanat ovat yleisiä teksteissä. Tiedonhaussa apuna käytettävälle kielelliselle ohjelmalle tuntemattomat sanat tunnetaan sanakirjasta puuttuvina sanoina (OOV, out of vocabulary words, Grefenstette 2000, 4). Niiden käsittelyyn on käytetty erilaisia menetelmiä tiedonhaussa, tyypillistä on käyttää esimerkiksi erilaisia sumean täsmäytyksen menetelmiä (Kraaij 2004, 63). Vastaavasti myös TWOL-tyyppinen sanakirjapohjainen morfologinen analysoija voidaan sovittaa käsittelemään paremmin sanakirjasta puuttuvia sanoja kirjoittamalla analyysisäännöt ja niiden automaatit ”joustaviksi” (ks. esimerkiksi Alegria et al. 2002; Koskenniemi 1996; Oflazer 1996). Jäppinen ja Ylilammi (1986) puolestaan esittivät jo 1980-luvulla morfologisen analysoijan sanakirjasta puuttuvien sanojen analysoinnin Morfo-ohjelmassa. Tämäntyyppisen menetelmän yhtenä ongelmana voidaan pitää sitä, että sääntöjen joustavuus johtaa vastaavasti myös ehdotettujen analyysien monitulkintaisuuden kasvuun (Koskenniemi 1996).

Sampson (1989) on analysoinut, miten normaalin laajahkon ja hyvin tehdyn elektronisen sanakirjan sanasto vertautuu todelliseen tekstimateriaalin. Hän vertasi

Oxford Advanced Learner's Dictionary of Current English –sanakirjan (68 742 sanakirjavientiä) kattavuutta pieneen aineistoon, joka oli otettu Lancaster-Oslo/Bergen-korpuksesta (LOB). Sampsonin otoksessa oli noin 4.5 % LOB:n sanoista, 45 622 sananmuotoa. 1477:ää sananmuotoa ei löytynyt sanakirjasta, mikä on 3.24 %. Kun puuttuvista sanoista poistettiin kaikki ennalta arvattavimmat luokat (nimet, vierasperäiset sanat jne.), jäljelle jäi 417 sanatyyppiä, joita sanakirja ei tuntenut. Noin puolet niistä oli substantiiveja. Vaikka Sampsonin saamat lukemat ovat yllättävänkin matalia, osoittavat ne selvästi ongelman olemassaolon. Huomattava on myös Sampsonin osoittama substantiivien suuri osuus tuntemattomissa sanoissa, sillä nimenomaan substantivit ovat tiedonhaussa keskeinen sanaluokka ja toisaalta ne ovat myös kielen avoimin sanaluokka, joka kasvaa jatkuvasti. Kraaij'n (2004, 63) pienessä hollannin tuntemattomien sanojen otoksessa erisnimet ja yhdyssanat ovat suurimmat tuntemattomien sanojen luokat, ja mitä ilmeisimmin tuntemattomat yhdyssanat ovat nimenomaan substantiiveja.

Lemmausohjelman sanakirjan puutteita koskevat samat rajoitukset kuin julkaistua sanakirjaa, voivatpa puutteet lemmausohjelman sanakirjassa olla suurempiakin. Saadakseni arvion puuttuvien sanojen osuudesta FINTWOLin sovelluskäytössä, koostin yhteen tietoja FINTWOLille tuntemattomista sanoista kahdessa eri tekstilähteessä. Ensinnäkin analysoin FINTWOLin sanakirjasta puuttuvien sanojen määrää analysoimalla TUTKin taivutusmuotoisen indeksin kaikki sanatyypit FINTWOlilla twol-r –F –moodissa. Tällöin FINTWOL tulostaa vain ne sananmuodot, joille se ei löydä sanakirjastaan perusmuotoa. Creutz ja Linden (2004, 6) ovat puolestaan tehneet saman laajemmalle aineistolle, jossa on noin 1,7 miljoonaa sananmuototyyppiä (32 miljoonan saneen aineistosta). Creutz (2005) on myös tuottanut tuntemattomien sanojen määrän sananmuototasolla. FINTWOLille tuntemattomien sanojen määrät näissä kahdessa aineistossa on koottu taulukkoon 15.

Taulukko 15. FINWOLille tuntemattomat sanat kahdessa eri tekstimateriaalissa sananmuotojen ja sananmuototyyppien tasolla.

Tekstitietokanta	TUTK	Creutz & Linden 2004
1. Sananmuototyyppejä tekstitietokannassa	719 011	Noin 1,7 miljoonaa
2. FINTWOLille tuntemattomien sananmuototyyppien määrä	120 633	Noin 300 000
3. FINTWOLille tuntemattomien sananmuototyyppien prosenttiosuus aineistossa	16,77 %	17,64 %
4. Sananmuotojen määrä tekstitietokannassa	12 109 779	32 017 012
5. Sananmuotojen ja sananmuototyyppien suhde tekstitietokannassa	16,84	n. 18,83
6. FINTWOLILLE tuntemattomien sananmuotojen määrä	495 938 [3]	1 716 668
7. FINTWOLille tuntemattomien sananmuotojen prosenttiosuus	4,09 %	5, 36 %

Molemmat tekstiaineistot, TUTK ja Creutz & Linden (2004), antavat samansuuntaisia tuloksia sekä sananmuototasolla että sananmuotojen tyyppien tasolla. Kun myös aineistot ovat tekstityypeiltään samankaltaisia – sanomalehtitekstiä ja uutisia, myös materiaalia kirjoista – (Creutz ja Linden 2004, 6; Sormunen 1994) antavat tulokset kohtalaisen tarkan arvion siitä, mikä määrä sanoja FINTWOLin sanastosta puuttuu, kun ohjelmaa käytetään käytännön sovelluksen osana. Tietty osa puuttuvista sanoista tosin on aina väärin kirjoitettuja sananmuotoja, joskin kirjoitusvirheiden prosentuaalisen osuuden arviointi on vaikeaa muuten kuin otoksella. Kraaij'n (2004, 63) pienessä hollannin tuntemattomien sanojen otoksessa kirjoitusvirheitä oli tuntemattomista sanoista 10 %.

Taulukon 13 tuloksissa on syytä on kiinnittää huomiota siihen, että sananmuotojen ja sananmuototyyppien puutosprosentit ovat erilaisia. Tämä johtuu siitä, että sananmuotojen ja sananmuototyyppien suhde tekstissä on erilainen (type-token ratio,

[3] Tuntemattomien sananmuotojen määrä perustuu Eija Airion toukokuussa 2005 ajamaan analyysiin.

Baayen 2001, 25– ; Schütze & Manning 1999, 22). Tiedonhaun kannalta oleellista on se, miten käyttäjän antamat hakutermit tulkitaan. Ilmeistä on, että käyttäjän antamia hakutermejä on pidettävä enemmän sananmuotoina kuin sananmuototyyppeinä. Niinpä sanakirjaan perustuvan lemmausohjelman sanakirjasta puuttuvien sanojen tuottamat puutteellisuudet haussa asettunevat lähemmäs sananmuotojen prosenttilukua kuin sananmuototyyppien.

Myös tekstityypit vaikuttavat siihen, miten hyvin sanakirja kattaa tekstin sanat. Esimerkiksi sanomalehtiartikkeleissa ja kirjoissa on yleensä vähemmän kirjoitusvirheitä, epästandardia kieltä ja uudissanoja kuin vaikkapa web-teksteissä, mutta suuri osa web-teksteistä on silti kielellisesti tavanomaista (Kilgariff & Grefenstette 2003). Niinpä onkin oletettavaa, että esimerkiksi web-hakukoneessa FINTWOLin tyyppinen lemmausohjelma ei menestyisi yhtä hyvin kuin sanomalehtitekstitietokannassa.

Hakutermien puuttuminen morfologisen analysoijan sanakirjasta on yksi varteenotettava tekijä arvioitaessa ohjelmia, joita käytetään tiedonhaussa hakutermien vaihtelun käsittelyyn. Tässä työssä esitetty lähestymistapa perustuu hakutermien sanakirjattomaan käsittelyyn, ja se on ainakin periaatteessa vähemmän altis sanakirjasta puuttuvien sanojen aiheuttamille puutoksille hakutuloksissa.

4 Loppupäätelmät

Tässä työssä esiteltiin siis, miten suomenkieliset tekstitiedonhaut käyttäytyvät InQuery-osittaistäsmätysjärjestelmässä, jos hakutermin muodon vaihtelusta ei katetakaan sanan kaikkia sijamuotoja, vaan vain tietty osa, joka perustuu tietoon sanojen sijamuotojen yleisestä tilastollisesta jakaumasta. Saatujen tulosten mukaan menetelmää voi pitää kaikin puolin kohtuullisen hyvänä.

Korpusanalyysilla saatuja keskeisiä sijamuotoja koskevia oletuksia ja niiden

perusteella tehtyjä tutkimuskysymyksiä testattiin kahdessa eri tekstikokoelmassa, TUTKissa ja CLEF 2003 -kokoelmassa. Kokoelmissa tehtyjen testien tulokset olivat hieman erilaisia: TUTKissa ero käsittelemättömiin hakutermeihin oli jo kolmen yleisimmän sijamuodon yksikkömuodoilla kohtalainen, kun CLEF 2003 -kokoelmassa samalla tasolla ero oli hyvin pieni. Myös Snowball-karsinnan saavuttama keskitarkkuustaso ylittyi TUTKissa nopeammin, CLEF 2003:ssa tasoa ei saavutettu ollenkaan millään määrällä sijamuotoja. Toisaalta Snowball saavuttaakin CLEF 2003 –kokoelmassa selvästi parempia tuloksia kuin TUTK-kokoelmassa (Airio 2005; Kettunen, Kunttu & Järvelin 2005). FINTWOLin saavuttamia keskitarkkuuksia ei saavuteta millään rajallisista sijamuotomenetelmistä, mutta erot FINTWOLiin eivät ole tilastollisesti merkitseviä CLEF 2003 -kokoelmassa kuin Red_Min2-menetelmällä. TUTKissa rajallisten sijamuotoprosessien erot FINTWOLiin olivat tilastollisesti merkitseviä kaikilla relevanssitasoilla.

Hakujen käytännön toteutusten kannalta näyttää siltä, että keskeisten sijamuotojen haku tietokannan indeksistä toimii huomattavasti nopeammin kuin vartaloiden tai vartaloiden ja säännöllisten lausekkeiden käyttö, joka on osoittautunut melko hitaaksi indeksihaun suhteen (Kettunen 2005; Kettunen, Kunttu & Järvelin 2005). Tällä puolestaan on merkitystä käytännön järjestelmien kannalta. Niinpä voikin päätellä, että tässä työssä esitetty rajallinen hakutermien muotojen muuntelun kattamisen menetelmä olisi kokeilemisen arvoinen myös tuotannollisessa hakujärjestelmässä. Sen saavuttama keskitarkkuus ei jää kovin paljon lemmaamalla saavutetuista tuloksista eikä hakuvartaloista, mutta toisaalta systeemi toimii erityisesti hakuvaiheessa huomattavasti nopeammin kuin vartalohaut. Hakutermien rajallisen eri sijamuodoissa olevien muotojen tuottavan ohjelman tekeminen ei ole kovinkaan mutkikasta, esimerkkejä taivutettuja sananmuotoja tuottavista ohjelmista suomea varten ovat WGEN (Koskenniemi 1985), Finnmorf (Holman 1988) ja FORMO (Lassila 1988).

Hakutermien erilaisia muuntelun kattamisen menetelmien etuja ja haittoja on syytä

kuitenkin tarkastella laajemmin kuin vain hakujen saannin ja tarkkuuden kannalta. Harman (1991) luettelee kolme erilaista hyötyä, joita hakutermien morfologisella käsittelyllä saavutetaan:

- helppokäyttöisyys käyttäjän näkökulmasta (hakujärjestelmä huolehtii hakutermien morfologisen muuntelun käsittelemisestä),
- tallennustilan säästö (hakuindeksin pieneneminen),
- parantunut hakutulos (hakujen parantunut keskitarkkuus).

Näiden lisäksi oleellinen vaikuttava tekijä on myös käytettyjen morfologisten keinojen sanastollinen kattavuus (luku 3.6). Lisäksi vertailuun voidaan ottaa muita, erityisesti hakujen suorituskykyyn liittyviä, kriteereitä. Tällaisia ovat esimerkiksi hakujen suoritusaika ja tietokannan indeksien muodostamiseen kuluva aika (Järvelin 1995, 52 - 53). Taulukossa 16 on vertailtu tässä työssä käsiteltyjä tai mainittuja menetelmiä näihin kuuteen kriteeriin.

Taulukko 16. Eri menetelmien vertailu Harmanin esittämien kolmen kriteerin ja kolmen muun kriteerin suhteen. 1 = lemmaus, 2 = hakuvartalot, 3 = Snowball-karsinta, 4 = hakutermin keskeisten sijamuotojen tuottaminen, 5 = käsittelemättömät hakutermit.

Menetelmä	1	2	3	4	5
Helppokäyttöisyys	+	+	+	+	+/-
Tallennustilan säästäminen	+	-	+	-	-
Parantunut hakutulos	+	+	+	+	-
Sanastollinen kattavuus	-	+	+	+	+
Hakujen suoritusaika	+	-	+	+	+
Tietokannan indeksien muodostamiseen käytetty aika	-	+	-	+	+
Yhteensä	4	4	5	5	3

Kun kaikkia tekijöitä verrataan karkealla tasolla, eri menetelmien vahvuudet asettuvat melko tasaisiksi. Hakujen keskitarkkuuden suhteen parhaiten yleensä menestyvän lemmauksen käyttökelpoisuutta heikentävät seuraavat tekijät:

- laajan, päivitystä vaativan sanakirjan käyttö ohjelmassa,
- lemmausohjelman sanakirjasta puuttuvien sanojen tuottamat ongelmat (Alkula 2000; Koskenniemi 1996),
- lemmausohjelman pitempi toteuttamisaika, jos kielelle ei ole jo olemassa lemmainta,
- hakuindeksin vaatimat erilliset perusmuotoistusajot (Galvez, Moya-Anegón & Solana 2005).

Hakuvartaloiden ja hakutermin keskeisten sijamuotojen käyttäminen haussa välttyy

kaikilta näiltä haitoilta. Sanakirjatonta karsintaohjelmaa koskevat tästä listasta vain erilliset hakuindeksin ajot.

Yksiselitteistä vastausta parhaaseen hakutermien vaihtelun kattamisen menetelmään ei siis ole. Lopputulokseksi jää pikemminkin, että hakutermien morfologista vaihtelua voidaan suomenkielisessä tekstitiedonhaussa käsitellä sellaisillakin menetelmillä, joita ei ole totuttu pitämään kielen mutkikkuuden vuoksi sopivina. Tämä koskee erityisesti yksinkertaista karsintaa ja hakutermien kokonaan taivutettujen muotojen tuottamista tässä työssä esitetyssä rajallisessa muodossa.

5 Lähdekirjallisuutta

Abu-Salem H., Al-Omari M. & Evens M.W. 1999. Stemming methodologies over individual query words for an Arabic information retrieval system. Journal of the American Society for Information Science 50, 524−529

Ahmad, F., Yusoff M. & Sembok, T. 1996. Experiments with a stemming algorithm for Malay words. Journal of the American Society for Information Science 47(12), 909−918

Airio E. 2005. Word normalization and decompounding in mono- and bilingual IR. Information Retrieval, ilmestyy.

Alegria, I., Aranbaze M., Ezeiza, A. & Urizar. R. 2002. Robustness and customisation in an analyzer/lemmatiser for Basque. http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1019570039/publikoak/robust2.pdf. Viitattu 15.8.2005.

Alemayehu, N. & Willet, P. 2003. The effectiveness of stemming for information retrieval in Amharic. Program, 37 (4), 254−259.

Alkula R. 2000. Merkkijonoista suomen kielen sanoiksi. Acta Universitatis Tamperensis 763. http://acta.uta.fi/pdf/951-44-4886-3.pdf. Viitattu 30.3. 2004.

Alkula R. 2001. From plain character strings to meaningful words: producing better full text databases for inflectional and compounding languages with morphological analysis software. Information Retrieval 4 (3−4), 195−208.

Baayen, H.R 1993. Statistical Models for Word Frequency Distribution: a Linguistic Evaluation. Computers and the Humanities 26, 347−363.

Baayen, H.R. 2001. Word Frequency Distributions. Dordrecht: Kluwer Academic Publishers.

Baeza-Yates, R. & Ribeiro-Neto, B. 1999. Modern Information Retrieval. USA: Addison Wesley.

Biber, D. 1993a. Representativeness in Corpus Design. Literary and Linguistic Computing 8 (4), 243−257.

Biber, D. 1993b. Using Register-diversified Corpora for General Language Studies. Computational Linguistics 19 (2), 219−241.

Braschler, M. & Ripplinger, B. 2004. How Effective is Stemming and Decompounding for German Text Retrieval? Information Retrieval 7 (3−4), 291−316.

Broglio J., Callan J., Croft B. and Nachbar D. 1995. Document retrieval and routing using the INQUERY system. Teoksessa Proceedings of the Third Text Retrieval Conference (TREC-3), Gaithesburg, MD: National Institute of Standards and Technology, special publication 500-225, 29−38.

Callan J., Croft B. & Harding S. 1992. The INQUERY retrieval system. Teoksessa Proceedings of the Third International Conference on Databases and Expert Systems Applications. Berlin: Springer Verlag, 78−84.

Conover, W.J. 1980. Practical Nonparametric Statistics. Toinen painos. New York: John Wiley & Sons.

Creutz, M. 2005. Kaksi sähköpostia, 17. 5. 2005.

Creutz, M. & Linden, K. 2004. Morpheme Segmentation Gold Standards for Finnish and English. Publications in Computer and Information Science. Report A77. Espoo:

Helsinki University of Technology.

Galvez, C., Moya-Anegón, F. & Solana, V. H. 2005. Term conflation methods in information retrieval. Non-linguistic and linguistic approaches. Journal of Documentation, 61 (4), 520−547.

Grefenstette, G. 2000. The Problem of Cross-language information retrieval. Teoksessa G. Grefenstette (toim.), Cross-Language Information Retrieval. Toinen painos. Kluwer Academic Publishers, 1−9.

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V. Heinonen T.R. & Alho, I. 2004. Iso suomen kielioppi. Helsinki: Suomalaisen Kirjallisuuden Seura.

Harman, D. 1991. How effective is Suffixing? Journal of the American Society for Information Science 42 (1), 7−15.

Hollink V., Kamps J., Monz C. & de Rijke, M. 2004. Monolingual document retrieval for European languages. Information Retrieval 7 (1−2), 33−52.

Holman, E. 1988. Finnmorf: A computerized research tool for students of Finnish morphology. Computers and the Humanities 22 (3), 165−172.

Jäppinen, H. & Ylilammi, M. 1986. Associative Model of Morphological Analysis: an Empirical Inquiry. Computational Linguistics 12 (4), 257−272.

Järvelin, K. 1995. Tekstitiedonhaku tietokannoista. Espoo: Suomen ATK-Kustannus.

Karlsson, F. 1983. Suomen kielen äänne- ja muotorakenne. Helsinki: WSOY.

Karlsson, F. 1986. Frequency Considerations in Morphology. Zeitsschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 39 (1), 19−28.

Kettunen K. 1991. Doing the stem generation with Stemma. Teoksessa Jussi Niemi (toim.) Papers from the Eighteenth Finnish Conference of Linguistics. Joensuu: Kielitieteellisiä tutkimuksia, Joensuun yliopisto, N:o 24, 80−97

Kettunen K., Kunttu T. & Järvelin K. 2005. To stem or lemmatize a highly inflectional language in a probabilistic IR environment? Journal of Documentation, 61(4), 476−496.

Kettunen, K. 2005. Developing an automatic linguistic truncation operator for best-match retrieval in inflected word form text database indexes. Lähetetty Journal of Information Science -lehteen 5.8.2005.

Kilgariff, A. & Grefenstette, G. 2003. Introduction to the Special Issue on the Web as Corpus. Computational Linguistics 29 (3), 333−347.

Koskenniemi, K. 1983. Two-Level Morphology: a General Computational Model for Word-form Recognition and Production. Publications of the Department of General linguistics, No. 11. Helsinki: University of Helsinki.

Koskenniemi, K. (1985). A system for generating Finnish inflected word forms. Karlsson, F. (toim.), Computational morphosyntax. Report on research 1981− 84. Publications of the Department of General linguistics, University of Helsinki. No. 13, 63−80.

Koskenniemi, K. 1996. Finite state morphology and information retrieval. Natural Language Engineering 2 (4), 331− 336.

Kostic, A. , T. Markovic & A. Baucal 2003. Inflectional Morphology and word meaning: Orthogonal or co-implicative cognitive domains. Teoksessa R. H. Baayen & R. Schreuder (toim.) Morphological Structure in Language Processing. Trends in

Linguistics, Studies and Monographs 151. Berlin: Mouton de Gruyter, 1−43.

Kraaij W. 2004. Variations on Language Modeling for Information Retrieval. Haag: CTIT Ph. D. series No. 04-62.

Lassila, E. 1988. Suomen kielen sanamuodot taivuttava ohjelma FORMO. Teoksessa Mäkelä et al. (toim.) STeP-88. Invited Papers. Contributed Papers: Applications. Espoo, 118−126.

Lyons, J. 1977. Semantics 2. Cambridge: Cambridge University Press.

Manning, C.D. & Schütze, H. 1999. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press.

Mayfield J. & McNamee P. 2003. Single N-gram stemming. Teoksessa Proceedings of Sigir2003, The Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 415−416.

McNamee, P. & Mayfield, J. (2004). Cross-Language retrieval using HAIRCUT for CLEF 2004. http://www.clef-campaign.org/2003/WN_web/00.2%20-%20intro.pdf. Viitattu 1.9. 2005.

Meadow, C.T., Boyce, B. R. & Kraft, D.H. 2000. Text Information Retrieval Systems. Toinen painos. San Diego: Academic Press.

Niemikorpi, A. 1990. Suomen kielen sanaston frekvenssianalyysia. Proceedings of the University of Vaasa. Research Papers, 150. Vaasa: Vaasan yliopisto.

Niemikorpi, A. 1991. Suomen kielen sanaston dynamiikkaa. Acta Wasaensia 26. Vaasa: Vaasan yliopisto.

Nurminen R. 1986. Suomen kielen sanamuotoja tulkitsevien ohjelmien hyödyntäminen tiedonhakujärjestelmissä. Tutkimuksia 386. Espoo: Valtion teknillinen tutkimuslaitos.

Oflazer, K. 1996. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. Computational Linguistics 22 (1), 73−89.

Peters, C. 2003. Introduction to the CLEF 2003 working notes. http://www.clef-campaign.org/2003/WN_web/00.2%20-%20intro.pdf. Viitattu 1.9. 2005.

Porter M.F. (1980). An algorithm for suffix stripping. Program 14 (3), 130-137.

Porter M.F. 2001. Snowball: a language for stemming algorithms. http://snowball.tartarus.org/texts/introduction.html. Viitattu 28.11. 2003.

Räsänen, S. 1979. Havaintoja suomen sijojen frekvensseistä. Sananjalka 21, 17−43.

Sampson, G. 1989. How fully does a machine-usable dictionary cover English text? Literary and Linguistic Computing 4 (1), 29−35.

Saukkonen, P., Haipus, M., Niemikorpi A. & Sulkala, H. 1979. Suomen kielen taajuussanasto. Helsinki: WSOY.

Sever, H. & Bitirim, Y. 2003. FindStem: Analysis and evaluation of a Turkish stemming algorithm. Teoksessa Nascimento et al.(toim.) String Processing and Information Retrieval. 10th International Symposium, SPIRE 2003, 238−251.

Sormunen E. 1994. Vapaatekstihaun tehokkuus ja siihen vaikuttavat tekijät sanomalehtiaineistoa sisältävässä tekstikannassa. VTT julkaisuja 790. Espoo: Valtion teknillinen tutkimuskeskus, tietopalvelu.

Sormunen E (2000). A method for measuring wide range performance of Boolean queries in full-text databases. Acta Universitatis Tamperensis 748. Tampere: Tampereen yliopisto.

The Finnish stemming algorithm.
http://snowball.tartarus.org/algorithms/finnish/stemmer.html.Viittattu 28.11. 2003

Tomlinson S. 2002. Experiments in 8 European languages with Hummingbird SearchServer™ at CLEF 2002.
http://clef.iei.pi.cnr.it:2002/workshop2002/WN/26.pdf. Viitattu 28.4.2004.

Tomlinson S. 2003. Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003.
http://clef.iei.pi.cnr.it/2003/WN_web/19.pdf. Viitattu 28.4.2004.