# VENÄJÄN KIELEN MORFOLOGISET ONGELMAT TIEDONHAUSSA: RIITTÄÄKÖ SIJAMUOTOJEN RAJOITETTU TUOTTAMINEN RATKAISUKSI?

Marja Holstila

Tampereen yliopisto
Informaatiotieteiden yksikkö
Informaatiotutkimus ja
interaktiivinen media
Pro gradu -tutkielma
Elokuu 2011

TAMPEREEN YLIOPISTO, Informaatiotieteiden yksikkö
Informaatiotutkimus ja interaktiivinen media
HOLSTILA, MARJA: Venäjän kielen morfologiset ongelmat tiedonhaussa: Riittääkö sijamuotojen rajoitettu tuottaminen ratkaisuksi?
Pro gradu -tutkielma, 73 s., 1 liites.
Elokuu 2011

---

Tämän tutkielman tarkoituksena on tarkastella venäjän kielen morfologiaa tiedonhaun näkökulmasta sekä selvittää Frequent Case Generation (FCG) -menetelmän toimivuutta venäjänkielisessä tiedonhaussa. Kimmo Kettusen ja kumppaneiden FCG-hakumenetelmän on todettu toimivan hyvin morfologialtaan venäjän tavoin mutkikkailla kielillä. Voimakkaasti taipuvien kielten substantiivien ja adjektiivien lukuisat sijamuodot ovat tiedonhaun kannalta haastavia. Rajoitetusti sijamuotoja hakuavaimille tuottava FCG perustuu siihen, että kielten potentiaalisista sanamuodoista vain harvat esiintyvät teksteissä usein.

Venäjän kielen morfologian pääpiirteitä käydään tutkielmassa läpi yhtäältä kielioppikirjallisuuden ja toisaalta tiedonhaun kielitypologian ja tiedonhaun kieliteknologian lingvististen menetelmien näkökulmasta. Tutkielman empiirisen osan tiedonhaun laboratoriokokeessa kolmea FCG-versiota vertaillaan hakuun morfologisesti käsittelemättömillä avaimilla seuraten venäjän sijamuotojen tuottamisessa Kettusen ja kumppaneiden (2007) esimerkkiä. Testikokoelmana kokeessa on erittäin suuri KM.ru-tietokanta.

Tiedonhaun kannalta olennaiseen taivutusmorfologiaan liittyy venäjässä piirteitä, joiden perusteella FCG:n voidaan olettaa soveltuvan venäjänkieliseen tiedonhakuun hyvin ja jopa paremmin kuin yleisesti käytössä olevien menetelmien. Tämän tutkielman FCG-kokeen tulokset ovat samansuuntaisia kuin Kettusen ja kumppaneiden (2007), ja em. oletukset saavat lisäoikeutusta. Venäjän adjektiivien sukukategoria saattaa kuitenkin aiheuttaa ongelmia, jos hakuavaimet perusmuotoistetaan FCG-käsittelyä varten automaattisesti.

Tutkielman rajoitus on se, että FCG-menetelmää ei kokeessa verrata muihin hyviksi havaittuihin tiedonhakumenetelmiin, vaan verrokkina on ainoastaan haku taivutusmuotoisilla avaimilla. Samalla testikokoelmalla tulisi kokeilla myös reduktiivisia menetelmiä ja verrata niiden tuloksellisuutta FCG-tuloksiin.

Tutkielma tarjoaa yleiskatsauksen venäjän morfologiaan tiedonhaussa sekä lisänäyttöä siitä, että sijamuotojen rajoitettu tuottaminen on venäjänkielisessä tiedonhaussa toimiva ratkaisu. Frequent Case Generation -menetelmää ei ole aiemmin kokeiltu suurella venäjänkielisellä aineistolla. Suuri kokoelma lisää koetulosten luotettavuutta.

Avainsanat: tiedonhaku, kieliteknologia, morfologia, venäjän kieli, FCG

# 1 JOHDANTO

Tämän tutkielman tarkoituksena on analysoida venäjän kieltä tiedonhaun kyselyjen ja dokumenttien kielenä. Käsittely on rajattu kielen morfologian tiedonhaulle aiheuttamiin ongelmiin sekä eräisiin niiden ratkaisemiseksi ehdotettuihin tiedonhaun kieliteknologisiin menetelmiin. Venäjää puhuu äidinkielenään noin 165 miljoonaa ihmistä, ja toisena kielenä se on noin 110 miljoonalle (Dolamic & Savoy 2009, 2540), sillä esimerkiksi ukrainalaisista ja valkovenäläisistä monet ovat kaksikielisiä. Lisäksi kielen ydinalueen ulkopuolella on laaja joukko ihmisiä, joilla on ainakin passiivinen venäjän kielen taito. Internet World Stats[1] -sivuston mukaan vuonna 2010 venäjänpuhujista 42.8 prosenttia käytti Internetiä, ja heitä arvioitiin olevan yhdeksänneksi eniten kaikista käyttäjistä. Tekstitiedonhaku on keskeisessä asemassa internetin käytössä. Venäjänkielisen tiedonhaun menetelmien kehittämiselle näyttää siis olevan tarvetta, mutta Venäjän federaation ulkopuolella aiheen tutkimus on ollut verrattain vähäistä. Venäjä kuuluu slaavilaisiin kieliin, ja on mahdollista, että sitä koskevien tiedonhaun ongelmien ratkonnasta hyötyy myös muiden ryhmän kielten tiedonhakua koskeva tutkimus.

Vuonna 2001 Hedlund ja kumppanit julkaisivat tutkimuksen ruotsin kielen ominaisuuksista yksikielisen ja kieltenvälisen tiedonhaun kannalta. He huomauttivat, ettei kunnollisia tiedonhakujärjestelmiä eri kielille voi kehittää tutkimatta kunkin kielen erityisiä ominaisuuksia. Tuolloin tiedonhaun kieliteknologisten menetelmien kehittäminen oli pitkälti keskittynyt englanninkieliseen tiedonhakuun. Siinä hyväksi havaitut menetelmät eivät kuitenkaan välttämättä toimi muuntyyppisillä kielillä. Tiedonhaun lingvistisiä ongelmia Hedlund ja kumppanit (2001) kuvaavat toteamalla, että hakujen tuloksellisuutta heikentävät nimenomaan kielelliset ilmiöt, joista yksi on morfologinen vaihtelu. Haku ei onnistu, jos hakusanat ja tietokannan hakemiston indeksitermit eivät kohtaa niiden muotojen ollessa erilaiset. Tämä koskee taivutusmuotoja ja johdoksia sekä yhdyssanojen osia. Kielen morfologinen

[1] http://www.internetworldstats.com/

kompleksisuus vaikuttaa hakumenetelmien tehokkuuteen ja tarpeellisuuteen sekä seikkoihin, jotka on otettava huomioon niitä suunniteltaessa.

Tässä työssä arvioidaan venäjän kielen morfologisia ominaisuuksia tiedonhaun näkökulmasta. Tutkielman johdantoluvussa määritellään aiheen kannalta keskeiset käsitteet ja esitetään tutkimuskysymykset. Toisessa luvussa kuvataan tutkimusasetelma. Kolmannessa luvussa käydään läpi lingvistisiä ratkaisumalleja, joita tiedonhaun tutkimuksessa on morfologisiin ongelmiin esitetty. Neljäs luku keskittyy venäjän kielen morfologian tiedonhaulle aiheuttamiin haasteisiin. Tutkielman empiirinen osa, sanamuotojen rajoitettua tuottamista koskeva tiedonhakukoe esitellään luvussa 5. Koko tutkielman yhteenveto on luvussa 6.

Tutkielmassa käytettävät venäjänkieliset esimerkit on translitteroitu Vahroksen ja Kahlan vuonna 1967 esittämän säännöstön mukaan. Taulukossa 1 ovat venäjän kielen aakkoset ja niitä vastaavat translitteroinnissa käytetyt merkit. Lisäksi aakkosiin kuuluu kirjain ё, joka esimerkeissä esiintyessään on translitteroitu käytetyn säännöstön mukaisesti merkillä ë. Kirjain esiintyy kuitenkin harvoin venäjänkielisissä teksteissä muualla kuin lastenkirjoissa tai opetusaineistoissa. Sen tilalla käytetään kirjainta е.

**Taulukko 1** Venäjän aakkoset ja niiden translitterointi

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| а (a) | б (b) | в (v) | г (g) | д (d) | е (e) | ж (ž) | з (z) |
| и (i) | й (j) | к (k) | л (l) | м (m) | н (n) | о (o) | п (p) |
| р (r) | с (s) | т (t) | у (u) | ф (f) | х (h) | ц (c) | ч (č) |
| ш (š) | щ (šč) | ъ (") | ы (y) | ь (') | э (ė) | ю (ju) | я (ja) |

Luonnollista kieltä voidaan tarkastella viiden eri osajärjestelmän tasolla: fonologisella, morfologisella, leksikaalisella, syntaktisella ja semanttisella tasolla. Kielen morfologisen tason ilmiöt vaikuttavat osaltaan tekstitiedonhaun tuloksellisuuteen. Erityisen tärkeitä morfologian kysymykset ovat suomen tai venäjän kaltaisten morfologialtaan kompleksisten kielten tiedonhakua koskevassa tutkimuksessa.

Morfologia eli muoto-oppi tutkii sanojen sisäistä rakennetta ja sanojen muodostamista. Se jaetaan taivutus- ja johtomorfologiaan. Taivutuksessa sanoista muodostetaan morfologisin keinoin taivutusmuotoja, jotka ilmaisevat sanojen perusmerkitysten lisäksi niiden välisiä kieliopillisia suhteita (Karlsson 2006; Pirkola 2001). Johtamalla sanoista

muodostetaan taivutusmuotojen sijaan uusia sanoja. Morfologian keskeisiä käsitteitä ovat morfeemi, vartalo ja juuri. Morfeemi on pienin kielitieteellinen yksikkö, jolla on itsenäinen merkitys tai kieliopillinen funktio (Karlsson 2006; Airio 2009). Morfeemi on abstraktio, ilmiöön viittaava yleiskäsite, jonka konkreettisista esiintymistä käytetään nimitystä morfi. Morfeemit jaotellaan vapaisiin ja sidonnaisiin. Sidonnaisia morfeemeja ovat ne, joilla ei ole itsenäistä merkitystä, kuten affiksit. Venäjän kielen affikseja ovat esimerkiksi suffiksit eli jälkiliitteet ja prefiksit eli etuliitteet, jotka yhdistettyinä toisiin morfeemeihin toimivat taivutuspäätteinä tai johtimina. Vapaita morfeemeja ovat tyypillisesti sanat, joilla on itsenäinen merkitys, kuten ”kissa” tai ”ja”. Yhdyssanat koostuvat kahdesta tai useammasta vapaasta morfeemista. Morfien segmentoinnilla tarkoitetaan niiden tunnistamista ja erottelua sanoissa. Vartalo on sanan perusosa, josta affiksit on karsittu pois. Juuri on vartalon pienin osa, jota ei voida jakaa osiksi, ja joka edustaa sanan semanttista merkitystä. Vartalon merkitys on taivutettaessa muuttumaton, affiksien merkitys vaihtelee. Vartalo ilmaisee konkreettisempaa merkitystä kuin affiksit. (Karlsson 2006.)

Tiedonhaun tutkimus käsittelee tiedonhakua ilmiönä, mutta on suuntautunut kehittämään menetelmiä käytännön sovelluksiin eli tiedonhakujärjestelmiin. Tiedonhakujärjestelmien tarkoitus on mahdollistaa halutun, dokumentteihin tallennetun informaation löytyminen. Tekstitiedonhaussa hakijan tiedontarve samoin kuin dokumenttien tietosisältö ilmaistaan sanoin. Siksi sanat, niiden merkitykset, keskinäiset suhteet ja sisäinen rakenne ovat tärkeitä paitsi kielitieteessä, myös tiedonhaussa. Tiedonhaun tutkimuksen kysymyksenasettelu on kuitenkin toisenlainen. Huomion kohteena ei ole kielen rakenne sinänsä, vaan sen vaikutukset dokumenttien löytymiseen. Morfologian osalta tiedonhaun tutkimus keskittyy sanamuotojen vaihtelun automaattiseen hallintaan. Ratkaisevaa on hallintamenetelmien vaikutus tiedonhaun tulokseen, ei morfeemien erottelun oikeaoppisuus. Järvelinin (2007, 978) mukaan tiedonhaun laboratoriotutkimuksessa keskeinen kysymys koskee sitä, mitkä asiat voidaan tehokkaasti automatisoida. Tässä tiedonhaun tutkimus on osa tietojenkäsittelytiedettä.

Tiedonhaun kieliteknologisten menetelmien tutkimus on usein järjestelmäsuuntautunutta laboratoriotutkimusta. Alalla kehitetään kuitenkin menetelmiä, joissa perinteiseksi muodostuneesta järjestelmäsuuntautuneisuudesta käännytään kohti käyttäjiä (esim. Järvelin 2007). Tämän tutkielman empiirisessä osassa

tehtävä tiedonhakukoe on tyypillinen järjestelmäsuuntautunut laboratoriokoe, jossa todellisia käyttäjiä ei ole mukana. Tutkielman luvussa 5 vertaillaan kahta tiedonhakumenetelmää sen suhteen, kuinka tehokkaasti niillä löydetään relevantteja dokumentteja. Haut ovat niin sanottuja ad hoc -hakuja, ja relevanssi tarkoittaa tässä tapauksessa aiherelevanssia. Relevantteja dokumentteja ovat ne, joiden aiheen katsotaan vastaavan hakutehtävän aihetta (Croft et al. 2010, 238).

Vertailtavat menetelmät ovat haku morfologisesti käsittelemättömillä hakuavaimilla ja toisaalta hakuavaimille rajoitetusti sijamuotoja tuottavat FCG (Frequent Case Generation) -menetelmät. Kimmo Kettusen ja kumppaneiden (Kettunen & Airio 2006; Kettunen et al. 2007; Kettunen 2009) kehittämässä FCG:ssä hakulausekkeisiin lisätään kyseessä olevassa kielessä useimmin esiintyviä hakuavainten sijamuotoja. Kettunen ja kumppanit (2007) ovat tutkineet menetelmän soveltuvuutta myös venäjänkieliseen tiedonhakuun, mutta erot vertailtaviin menetelmiin eivät olleet tilastollisesti merkitseviä. Tässä tutkielmassa pyritään osittain toistamaan edellä mainittujen tekijöiden venäjänkielinen FCG-koe käyttäen suurempaa aineistoa. Tarkoitus on selvittää, parantaako FCG tiedonhaun tuloksellisuutta venäjänkielisessä tiedonhaussa merkittävästi verrattuna hakuun taivutusmuotoisilla avaimilla. Toiseksi halutaan tietää, mikä testattavista kolmesta FCG-menetelmästä on tuloksellisin. Kokeiltavat FCG-versiot eroavat toisistaan tuotettavien sijamuotojen määrässä.

Tutkielmassa pyritään vastaamaan seuraaviin kysymyksiin:

1) Mitkä ovat venäjän kielen morfologian pääpiirteet ja miten ne vaikuttavat tiedonhakuun?

2) Parantaako FCG venäjänkielisen tiedonhaun tuloksia verrattuna käsittelemättömillä avaimilla tehtyyn hakuun ja millaiset ovat FCG-versioiden väliset erot?

# 2 TUTKIMUSASETELMA

Tiedon tallennus ja -hakujärjestelmissä ovat morfologisten haasteiden kannata olennaisia ne tavat, joilla järjestelmä käsittelee haettavien dokumenttien ja toisaalta hakutehtävien sanamuotoja. Tässä luvussa kerrotaan tutkielman luvussa 5 raportoitavan tiedonhakukokeen tutkimusasetelmasta ja siihen liittyvistä tiedonhaun ja -tallennuksen menetelmistä sekä niiden arvioinnista eli evaluoinnista.

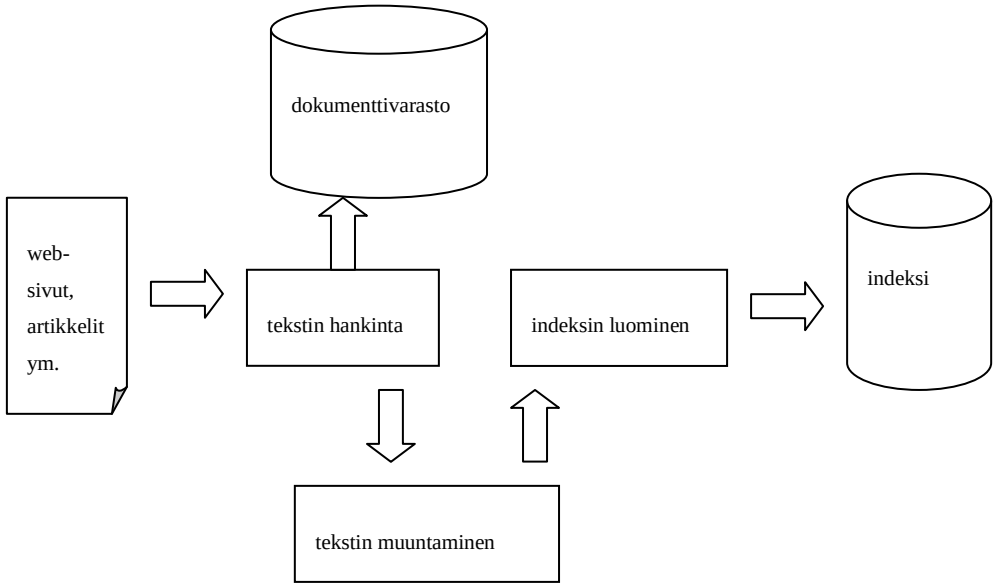
## 2.1 Indeksointi

Dokumenttien teksteissä mahdollisesti tuhansissa eri muodoissa esiintyvien sanojen ja sanoiksi puetun tiedontarpeen vertailun mahdollistaa tiedonhakujärjestelmän hakemistorakenne. Kokotekstitiedonhaussa, josta tässä tutkielmassa puhutaan, tietokannan hakemiston luominen eli indeksointi on ilman muuta automaattista. Nykyisten tiedonhakujärjestelmien, erityisesti web-hakukoneiden, kokoelmat sisältävät miljoonia dokumentteja ja ovat yleensä kasvavia. Myös dokumenttien sisältö muuttuu päivittämisen myötä. Niinpä indeksoinnin on oltava nopeaa. Dokumenttien teksteissä esiintyviä sanamuotoja edustavat tietokannan indeksissä eli hakemistossa automaattisesti muodostetut hakemisto- eli indeksitermit, jotka on voitu pelkistää eri tavoin sääntöjen ja sanakirjan avulla. Jokaista indeksitermiä kohden on lista dokumenteista**,** joissa termin lähtökohtina olevat sanamuodot esiintyvät. Indeksiä sanotaan myös käänteistiedostoksi, sillä se on käänteinen dokumenteille, joissa puolestaan listataan kaikki niiden sisältämät indeksitermit (Croft et al. 2010, 15). Haku tietokannasta tapahtuu siten, että käyttäjän syöttämiä hakusanoja eli hakuavaimia tai hakutermejä verrataan indeksissä oleviin termeihin. Jos termit täsmäävät ja muut mahdolliset ehdot täyttyvät, haetaan käyttäjän tarkasteltaviksi ne dokumentit, joihin termit viittaavat. Hakulausekkeen sanoja voidaan tarvittaessa käsitellä hakujärjestelmän indeksointitapaan sopiviksi joko automaattisesti tai käyttäjän toimesta.

Kuvio 1 esittää hakukoneen indeksointiprosessia. Croft ja kumppanit (2010, 14 – 15) kuvaavat indeksointia erilaisista komponenteista koostuvana prosessina, jonka luomat rakenteet mahdollistavat tiedonhaun. Tekstin hankinta -komponentti voi tarkoittaa olemassa olevan kokoelman käyttöä tai uuden kokoelman rakentamista. Komponentti luo dokumenttien teksteistä ja metadatasta tietokannan dokumenttivaraston, ja siirtää

dokumentit tekstin muuntamiskomponenttiin. Tämä muuntaa dokumentit indeksitermeiksi tai -piirteiksi. Indeksitermit ovat niitä dokumenttien osia, joita käytetään haussa. Indeksin luomiskomponentti luo nimensä mukaisesti indeksitermeistä indeksin, johon haku kohdistuu. (Croft et al. 2010: 14, 15.)

**Kuvio 1** Indeksointiprosessi (Croft et al. 2010: 15)




## 2.2 Evaluointi ja testikokoelmat

Tiedonhakutulosten evaluointiin ovat vakiintuneet saannin ja tarkkuuden mitat. Saanti on haettujen relevanttien dokumenttien määrä jaettuna tietokannan dokumenttivaraston (Kuva 1) kaikkien relevanttien dokumenttien määrällä. Tarkkuus on haettujen relevanttien dokumenttien määrä jaettuna kaikkien haettujen dokumenttien määrällä. Tässä tutkielmassa evaluointimittarina käytetään keskitarkkuuksien keskiarvoa (Mean Average Precision, MAP). Interpoloimaton keskitarkkuus (Average Precision AP) -mittari laskee kyselyittäin hakutuloksen tarkkuuden tuloslistan kunkin relevantin dokumentin kohdalla ja ottaa näiden keskiarvon. Kun kaikkien kyselyjen keskitarkkuudelle lasketaan keskiarvo, saadaan tiedonhaun tutkimuksessa yleisesti käytetty mittari MAP. (Sanderson 2010, 280.)

Tutkielmassa tehtävä tiedonhaun laboratoriokoe noudattelee Cranfield-mallia, jossa tiedonhaun evaluointi perustuu testikokoelmiin. Tässä tapauksessa evaluoinnin

kohteena ovat kielen morfologiaan eri tavoin suhtautuvat hakustrategiat, jotka kohdistuvat yhden tiedonhakujärjestelmän taivutusmuotoindeksiin. Tarkoitus on verrata eri menetelmien tuloksellisuutta. Tuloksellisuus (*effectiveness*) merkitsee tässä tiedonhakumenetelmällä löydettyjen dokumenttien relevanssia suhteessa tiedontarpeeseen. Tiedontarpeeseen suhtaudutaan siten, että se on kuvattu tyhjentävästi testikokoelman hakuaiheissa eli topiikeissa (topics), eikä todellisen maailman tiedontarvitsijoille eli käyttäjille suoda ajatustakaan. Tällaisen järjestelmäsuuntautuneisuuden etuna on koeasetelman toistettavuus ja hallittavissa olevat muuttujat. Käyttäjäsuuntautuneemmissa kokeissa inhimilliset tekijät hankaloittavat muuttujien määrittelyä ja tekevät koetilanteista ainutkertaisia. Järjestelmäsuuntautunutta tutkimusta voidaan arvostella siitä, että sen tulokset eivät välttämättä pidä paikkaansa tosielämän tilanteissa, joissa inhimilliset käyttäjät joka tapauksessa astuvat kuvaan. Vaikka käyttäjä- ja järjestelmäsuuntautuneisuus ovat lähentyneet toisiaan kokeellisessa tiedonhaun tutkimuksessa, on puhtaan järjestelmäsuuntautunut laboratoriotutkimus pitkälti säilyttänyt paradigman asemansa (Järvelin 2007, 972).

Haku taivutusmuotoisilla eli morfologisesti käsittelemättömillä avaimilla ei tunnetusti tuota hyviä tuloksia venäjän kaltaisten morfologialtaan kompleksisten kielten tiedonhaussa. Kyseinen menetelmä on koeasetelmassa vertailtava perustaso eli baseline. Voidaan olettaa, että FCG on baselinea tuloksellisempi, sillä Kettunen ja kumppanit (2007) ovat saaneet rohkaisevia tuloksia kokeillessaan menetelmää venäjän kielelle. Muilla morfologialtaan samantyyppisillä kielillä FCG on toiminut hyvin. FCG pohjautuu havaintoon siitä, että sijamuotojen jakauma kielessä on vino, ja vain kaikista yleisimpien sijamuotojen tuottaminen hakuavaimille voi siten olla riittävän tuloksellista.

Sen selvittäminen, mikä on riittävä sijamuotojen määrä, eli eri FCG-menetelmien tuloksellisuuden vertailu liittyy tehokkuuteen (*efficiency*). Tehokas menetelmä on nopea ja taloudellinen. Taivutusmuotoindeksi on indeksityypeistä kaikkein edullisin ja nopein rakentaa ja päivittää. Haku taivutusmuodoilla on hakumenetelmistä vähätöisin ja nopein: käyttäjän syöttämät hakuavaimet ajetaan indeksiä vasten sellaisinaan. Baseline-menetelmä on edullinen, mutta venäjän kielellä sen tulokset ovat mitä todennäköisimmin heikkoja. Tehokkuuden ja tuloksellisuuden väliltä pyritään löytämään tyydyttävä keskitie. Tässä tutkielmassa kiinnitetään huomiota menetelmien tehokkuuteen, mutta sitä ei varsinaisessa kokeessa mitata.

Korkea tuloksellisuus edellyttää usein suurta työpanosta tai muita resursseja, mikä vähentää menetelmän taloudellisuutta. Oletettavasti kaikkien sijamuotojen tuottaminen hakuavaimille parantaisi tiedonhaun tuloksellisuutta, mutta hakulausekkeet venyisivät pitkiksi ja prosessointi hidastuisi. Kettusen ja kumppaneiden (Kettunen & Airio 2006; Kettunen et al. 2007) ajatuksen mukaan vain muutaman sijamuodon tuottaminen kuitenkin riittää tuloksellisuuden huomattavaksi parantamiseksi suhteessa baseline-menetelmään, sillä tekstitilastojen valossa vain muutama yleisin sanamuoto kattaa suurimman osan sijamuotojen esiintymistä venäjänkielisissä teksteissä. Tässä tutkielmassa pyrkimys on siis määritellä, kuinka monen yleisimmän sijamuodon tuottaminen venäjänkielisille hakuavaimille riittää parantamaan tiedonhaun tuloksellisuutta niin, että tehokkuus pysyy silti riittävän hyvänä. Sitä, mikä on riittävän tehokasta, ei tässä tutkielmassa kuitenkaan määritellä, eikä esimerkiksi hakujen prosessointinopeutta mitata. Kokeessa keskitytään hakujen tuloksellisuuden evaluointiin.

Tiedonhakukoe ja evaluointi perustuvat tässä tutkielmassa testikokoelmaan ja tiettyjen evaluointimittareiden käyttöön. Tiedonhaun tutkimuksessa testikokoelmat (*test collections*) ovat tärkeällä sijalla, ja niitä kehitetään jatkuvasti erilaisilla tutkimusfoorumeilla, kuten kansainvälisissä CLEF (Cross Language Evaluation Forum) ja TREC (Text REtrieval Conference) -konferensseissa (Sanderson 2010, 250). Tässä tutkielmassa käytetty testikokoelma on venäläisen ROMIP eli Российский семинар по Оценке Методов Информационного Поиска (Rossijskij seminar po Ocenke Metodov Informacionnogo Poiska - Russian Information Retrieval Evaluation Seminar) -foorumin KM.ru-kokoelma. ROMIP:in periaatteet ja tavoitteet ovat samantapaiset kuin TREC-konferenssin. TREC kokoontui ensikerran vuonna 1992 tavoitteinaan luoda testikokoelmia erilaisille tiedonhakutehtäville, edistää näiden hakutehtävien mahdollisimman laajaa tutkimusta ja järjestää kokoontumisia, joissa tutkijat voisivat jakaa TREC-kokoelmilla tekemäänsä tutkimusta (Sanderson 2010, 275). ROMIP aloitettiin vuonna 2003. Se keskittyy venäjänkielisen tiedonhaun menetelmien riippumattomaan evaluointiin ja on tarkoitettu ensisijaisesti venäläisten tutkijoiden keskinäiseen tiedonjakoon.

Testikokoelmat sisältävät yleensä dokumenttikokoelman, joukon topiikkeja (topics) ja joukon relevanssiarvioita. Kullakin dokumentilla on oma tunnistenumeronsa (*docid*). Topiikkeja sanotaan myös kyselyiksi (*queries*), ja niillä on omat tunnisteensa (*qid*).

Tässä tutkielmassa *topic* on suomennettu hakuaiheeksi. Relevanssiarvioiden joukko, johon tässä työssä viitataan termillä saantikanta, on lista hakuaiheiden ja dokumenttien tunnistepareista. Siinä eritellään kunkin dokumentin relevanssi suhteessa hakuaiheisiin. Englanniksi saantikannasta puhutaan myös termillä *query relevance set* eli *qrels*. (Sanderson 2010, 250)

TREC-hakuaiheet (*topics*) koostuvat yleensä tunnisteesta, otsikosta, kuvauksesta ja narratiivista. Hakuaiheen otsikko koostuu muutamasta sanasta, joista voidaan muodostaa lyhyt kysely. Kuvaus luonnehtii tiedontarvetta useammalla lauseella. Narratiivissa selitetään tarkemmin, millaisia dokumentteja pidetään hakuaiheen suhteen relevantteina. ROMIP:n ad hoc -trackin web-hakuaiheet (*tasks*) sisältävät vain yhdestä muutamaan sanaa, ja ne ovat verrattavissa lähinnä TRECin hakuaiheen otsikkoon. Termille *track* (venäjäksi *дорожка)* ei ole vakiintunutta suomennosta. Tässä tutkielmassa siitä käytetään nimitystä tutkimuslinja. Se merkitsee tiedonhaun evaluointiseminaarin tai -konferenssin osaa, jossa kehitetään ja evaluoidaan menetelmiä tietyntyyppisten hakutehtävien ratkaisuun. Tässä tutkielmassa käytetyn testikokoelman relevanssiarvioiden joukot eli saantikannat on luotu vuoden 2009 ja sitä aikaisempien vuosien ROMIP-seminaarin ad hoc -tutkimuslinjan web-hakuosiossa. Tulosten evaluoinnissa käytetään TREC:in TrecEval-ohjelmaa, joka laskee kokeen ajotuloksista erilaisia mittareita.

Kokeiltavissa FCG-hauissa tuotettavien sijamuotojen valinnassa ja määrässä on yksinkertaisesti seurattu Kettusen ja kumppaneiden (2007) esimerkkiä. Heidän tutkimuksestaan tässä tehtävä koe eroaa siinä, että testikokoelma ja sen myötä hakuaiheet eli topiikit ovat toiset. Kettusen ja kumppaneiden (2007) käyttämä venäjänkielinen kokoelma oli pieni, ja kokeessa havaitut eri menetelmien väliset erot eivät olleet tilastollisesti merkitseviä.

Pidemmällä tähtäimellä FCG-menetelmän tutkimisen tarkoitus on, että sitä alettaisiin hyödyntää käytännön sovelluksissa eli tosielämän hakukoneissa niillä kielillä, joilla sen kiistatta todetaan toimivan hyvin. Laajemmin ajatellen tavoitteena on parantaa ihmisten mahdollisuuksia löytää tarvitsemansa informaatio. Vaikka tämän tutkielman antina voi parhaimmillaankin olla vain lisäselvitys sijamuotojen tuottamisen toimivuudesta venäjänkielisessä tiedonhaussa, on se kuitenkin osa tiedonhaun kieliteknologian tutkimuksen kokonaisuutta, jonka on pyrkimyksenä sekä ymmärtää tiedonhaun ilmiöitä

että vaikuttaa parempien tiedonhakujärjestelmien kehittämiseen. Tutkielman teoreettisen osan tarkoitus on keskustella venäjän kielen morfologisista ominaisuuksista tiedonhaussa yleisemmällä tasolla, empiirisessä osassa keskitytään yhteen venäjän morfologian hallitsemiseksi ehdotettuun ratkaisuun – Frequent Case Generation -menetelmään.

# 3 MORFOLOGIA TIEDONHAUN TUTKIMUKSESSA

Tässä luvussa käsitellään sanamuotojen morfologisen vaihtelun hallintaan kehitettyjä tiedonhaun menetelmiä. Alaluvussa 3.1 kerrotaan lyhyesti kielitieteen morfologisesta typologiasta taustaksi tutkielmassa hyödynnettävälle Pirkolan (2001) tiedonhaun kielitypologialle. Myös sanamuotojen rakennetta käsitellään, sillä tiedonhaun lingvistiset menetelmät pitkälti perustuvat sen lainalaisuuksille. Alaluvuissa 3.2. ja 3.3 sekä luvussa 4 esitellään ja vertaillaan näitä menetelmiä.

## 3.1 Morfologinen typologia ja sanojen morfologinen rakenne

Kielitieteen morfologinen typologia luokittelee kieliä niiden synteesin ja fuusion asteen mukaan. Kielillä, joissa morfeemien määrä sanaa kohden on suuri, on korkea synteesin aste. Näiden vastakohtana ovat niin sanotut isoloivat kielet, joissa sanoja ei taivuteta eikä johdeta lainkaan, ja kieliopillisia suhteita osoitetaan esimerkiksi prepositioilla. Fuusio tarkoittaa Pirkolan (2001) mukaan sitä, että morfien yhdistelmä, sanamuoto, ei ole osiensa summa. Ilmiötä kutsutaan myös morfofonologiseksi vaihteluksi. Kielessä, jonka fuusion aste on korkea, päätteet, kannat ja muut morfeemit eivät yhdisty toisiinsa sellaisinaan, vaan yhdistyminen saa aikaan morfologisia muutoksia, jotka vaikeuttavat morfeemien tunnistamista ja erottelua. Myös morfeemien merkitykset voivat olla fuusioituneita siten, että esimerkiksi tietty pääte ilmaisee sekä sijaa että lukua. Asteikon toisessa ääripäässä ovat agglutinatiiviset kielet, joissa morfeemit ovat yksimerkityksisiä ja ne liittyvät toisiinsa muuttumattomina. Useimmat kielet sijoittuvat jonnekin synteettisen ja isoloivan sekä fuusioivan ja agglutinoivan välimaastoon. Puhtaan isoloivassa kielessä kaikki sanat koostuisivat vain yhdestä morfeemista, joten morfeemien yhdistämisen fuusioivuus tai agglutinatiivisuus ei koskisi sitä. (Pirkola 2001; Karlsson 2006; Nikunlassi 2002.)

Venäjän kielen synteesin astetta voi luonnehtia melko korkeaksi. Kielessä on paljon morfeemeja sanaa kohti johtuen siitä, että kieliopillisia suhteita ilmaistaan ja uusia sanoja muodostetaan suurelta osin affikseilla. Taulukko 2 kuvaa eri kielten synteesin astetta Pirkolaa (2001) mukaillen. Taulukon arvot on laskettu sadan sanan tekstinäytteistä. Venäjä on taulukon esimerkkikielistä toiseksi synteettisin.

| **Taulukko 2** Synteesin aste (Pirkola 2001, 337) | |
|---|---|
| kieli | synteesin aste |
| vietnam | 1.06 |
| joruba | 1.09 |
| englanti | 1.68 |
| muinaisenglanti | 2.12 |
| swahili | 2.55 |
| turkki | 2.86 |
| venäjä | 3.33 |
| inuiitti | 3.72 |

Venäjän sanamuodoissa affiksit ja muut morfeemit eivät aina ole helposti eroteltavissa. Sanojen vartaloissa voi niihin affikseja liitettäessä tapahtua äännevaihtelua, kuten vokaalin väistymistä tai konsonantin vaihtumista toiseksi. Morfeemeihin sisältyy tavallisesti useita kieliopillisia merkityksiä. Yksi taivutuspääte voi siis ilmaista yhtä aikaa esimerkiksi maskuliinisukua, yksikköä ja genetiivisijaa, eikä näitä merkityksiä voi erotella omiin morfeemeihinsa. Venäjän kielen morfologiset prosessit ovat siis jossain määrin fuusioivia.

Kielen morfologiseen analyysiin perustuvien tiedonhaun kieliteknologisten menetelmien kannalta sanamuotojen rakenneosien, kuten sanavartaloiden ja affiksien tunnistaminen on tärkeää. Venäjän affiksiryhmät ovat prefiksit eli etuliitteet, suffiksit eli loppuliitteet, taivutuspäätteet eli fleksiot, postfiksit eli jälkiliitteet ja interfiksit eli sisäliitteet. Affikseja esiintyy taipuvissa ja johdetuissa sanoissa ilmaisemassa kieliopillisia ja sananmuodostuksellisia merkityksiä. Suffiksit ja prefiksit ovat sananmuodostusaffikseja, ja niitä voi yhdessä sanassa olla useita. Suffiksit liittyvät vartalon perään ennen fleksiota. Prefiksit sijoittuvat vartalon edelle. Niiden merkitykset voivat olla sekä kieliopillisia että leksikaalisia. Erityisesti venäjän verbeissä etuliitteet ovat usein niin sanottuja salkkumorfeja, jotka sekä muuttavat verbin leksikaalista merkitystä että määrittelevät sen aspektin. Substantiiveissa ovat kieliopillisten suffiksien lisäksi yleisiä leksikaaliset suffiksit, jotka liittävät vartaloon esimerkiksi arvottavia tai kokoa luonnehtivia lisämerkityksiä (Nikunlassi 2002). Tiedonhaun tutkimuksessa kaikista sanan vartalon perään sijoittuvista affikseista puhutaan yleensä suffikseina.

Sanojen kieliopillisilla muodoilla eli sanamuodoilla tarkoitetaan saman sanan eri muotoja, joita se saa erilaisissa kieliopillisissa asemissa konkreettisissa

kielenkäyttötilanteissa. Eri sanojen kokemat ulkoiset muutokset ovat vastaavissa kieliopillisissa merkityksissä usein samankaltaisia, mikä mahdollistaa yleistysten tekemisen. Päätteen tai suffiksin kieliopillinen merkitys ei ole sidottu tietyn vartalon merkitykseen, vaan ne säilyttävät merkityksensä kokonaisten sanaluokkien yhteydessä. Kuten suomessa, myös venäjässä eri sanaluokille ovat ominaisia erilaiset päätteet ja muut affiksit. Tämä helpottaa niin sanaluokkien ja lauseenjäsenten kuin morfeemienkin tunnistamista. (Rahmanova & Suzdal'ceva 2003.)

Sanamuodon vartalo on se sanan osa, joka jää jäljelle, kun affiksit irrotetaan, tai kokonainen taipumaton sana. Vartalo on sanan keskus, morfeemi, joka sisältää sanan ydinmerkityksen. Muut morfeemit liitetään vartalon ympärille. Esimerkiksi sanassa рассмеялØся (rassmejalØsja) sanamuodon vartalo on -смея- (-smeja-), joka sisältää nauramisen leksikaalisen merkityksen, prefiksi -рас- (-ras-) tuo sanaan nauruun purskahtamisen merkityksen sekä perfektiivisen aspektin merkityksen. Tässä tapauksessa perfektiivinen aspekti merkitsee sitä, että toiminta on yhtäkkistä ja kertakaikkista, sillä on tietty alkupiste eikä se jatku prosessinomaisesti. Taivutuspääte eli fleksio -л- (-l-) merkitsee preteritiä eli mennyttä aikaa, Ø-merkillä merkitty nollapääte on tässä maskuliinin tunnus, -ся- (-sja-) refleksiiviverbiä merkitsevä postfiksi. Alla esimerkkisanojen morfit on erotettu pystyviivoilla. Kunkin sanan alla morfit on nimetty siinä järjestyksessä, kun ne ovat kyseisessä sanassa.

рас|смея|л|Ø|ся (ras|smeja|l|Ø|sja)

prefiksi|vartalo|fleksio|fleksio|postfiksi

велик|ий (velik|ij)

vartalo|fleksio

велич|айш|ий (velič|ajš|ij)

vartalo|suffiksi|fleksio

Sana великий (velikij) on suurta, mahtavaa merkitsevän, maskuliinista substantiivia määrittelevän pitkän adjektiivin positiivin yksikön nominatiivi, величайший (veličajšij) sen superlatiivi. Tämän adjektiivin vartalossa tapahtuu vertailuasteita muodostettaessa konsonanttivaihtelua, joka on yksi tiedonhakua vaikeuttavista morfologisen vaihtelun muodoista.

## 3.2 Morfologisen vaihtelun hallinnan lingvistisiä menetelmiä tiedonhaussa

Synteesin ja fuusion asteet, joilla kielitieteessä arvioidaan kielten morfologista kompleksisuutta, kuvaavat yleisesti sekä taivutus- että johtomorfologiaan liittyviä ilmiöitä. Ari Pirkola (2001) on kehittänyt tätä typologiaa edelleen tiedonhaun tarpeisiin. Pirkolan (2001) tiedonhaun kielitypologia kiinnittää huomion morfologian eri osa-alueisiin. Synteesin ja fuusion asteet on jaettu kuvaamaan erikseen taivutus- ja johtomorfologiaa sekä yhdyssanoja. Synteesin asteesta erotetaan synteesin taivutusaste (*inflectional index of synthesis*), synteesin johtamisaste (*derivational index of synthesis*) ja synteesin yhdyssana-aste (*compound index of synthesis*). Ensin mainittu on kielen taivutusmorfeemien määrä jaettuna sanojen kokonaismäärällä. Synteesin johtamisaste taas on johdosmorfeemien määrä jaettuna kaikkien kielen sanojen määrällä. Synteesin yhdyssana-aste on yhdyssanojen osien määrä jaettuna sanojen kokonaismäärällä.

Fuusiota voi ilmetä niin semanttisella kuin morfologisella tasolla. Siksi Pirkola (2001) esittää erilliset jaottelut morfologisen ja semanttisen fuusion asteelle. Tässä tutkielmassa näistä kiinnitetään huomiota lähinnä morfologiseen fuusioon eli siihen, tapahtuuko morfeissa äännevaihtelua niiden yhdistyessä. Pirkola (2001) jakaa morfologisessa fuusion asteen kolmeen osaan: fuusion taivutusaste (*inflectional index of fusion*), fuusion johtamisaste (*derivational index of fusion*) ja fuusion yhdyssana-aste (*compound index of fusion*). Fuusion taivutusaste lasketaan jakamalla fuusioituneiden taivutettujen sanojen määrä kaikkien sanojen määrällä. Fuusion johtamisaste on fuusiojohdosten määrä jaettuna sanojen kokonaismäärällä ja fuusion yhdyssana-aste fuusioituneiden yhdyssanojen määrä jaettuna kaikkien sanojen määrällä. Kielten vertailua varten tiedonhaun kielitypologian asteet lasketaan erikielisistä rinnakkais- tai verrannollisista korpuksista, kuten myös yleisemmät synteesin ja fuusion asteet.

Suunniteltaessa kielen morfologista prosessointia Pirkolan (2001) typologia auttaa ratkaisemaan, käsitelläänkö johdoksia vai vain taivutusta, ja onko yhdyssanat syytä osittaa. Mikäli esimerkiksi pitäydytään vain taivutusmorfologiassa, tärkeitä ovat synteesin taivutusaste (*inflectional index of synthesis*) ja fuusion taivutusaste (*inflectional index of fusion*). Niiden perusteella voidaan arvioida aiotun morfologisen käsittelyn kustannuksia ja hyötyjä kyseisen kielen tiedonhaussa.

Morfologisia haasteita lähestytään tiedonhaussa kieliteknologisilla menetelmillä, jotka voidaan jakaa lingvistisiin ja kieliriippumattomiin. Kieliriippumattomia menetelmiä ovat esimerkiksi n-grammit, tilastolliset stemmerit ja lemmatoijat, muut tilastolliset menetelmät sekä tekstikorpuksesta morfologiaa itsenäisesti oppivat ohjelmat. Näitä ei käsitellä tässä tutkielmassa lainkaan. Lingvistiset menetelmät jaetaan edelleen reduktiivisiin ja generatiivisiin. Reduktiivisia ovat lemmatointi ja stemmaus, joissa useita saman lekseemin eri sanamuotoja edustavat ja yhdistävät tavalla tai toisella pelkistetyt hakemisto- ja hakuavaimet. Generatiivisiksi sanotaan sanamuotojen ja -vartaloiden tuottamista hakuavaimille (Kettunen & Airio 2006; Kettunen et al. 2007; Kettunen 2009).

Tekstitietokatojen indeksit ovat kolmen tyyppisiä: taivutusmuotoindeksejä, stemmattuja ja perusmuoto- eli lemmattuja indeksejä. Taivutusmuotoindeksissä kaikki tietokannan dokumenteissa esiintyvät sanamuodot mahdollisia sulkusanoja lukuun ottamatta on indeksoitu sellaisinaan. Stemmatussa indeksissä on sanavartaloita (*stem*), jotka edustavat useampia dokumenteissa esiintyviä sanamuotoja. Stemmatun indeksin indeksitermit on muodostettu poistamalla sanamuodoista affikseja karsinta-algoritmilla eli stemmerillä. Stemmattu sanavartalo ei välttämättä vastaa sanan kieliopillista vartaloa. Karsinta voidaan kohdistaa myös hakuavaimiin. Lemmatussa indeksissä eri sanamuotoja edustavat perus- eli sanakirjamuodossa olevat indeksitermit eli hakemistoavaimet. Perusmuoto-ohjelman eli lemmatoijan analyysi perustuu yleensä sääntöihin ja laajaan sanakirjaan, jossa voi olla kymmeniä tuhansia sanoja. Stemmauksessa semanttisessa suhteissa olevat eri sanamuodot yhdistetään yksiin vartaloihin käyttämällä joko vain affiksisääntöjä tai sääntöjä ja sanakirjaa. Lemmatointia pidetään yleisesti tuloksiltaan parhaana lähestymistapana. (Kettunen et al. 2007).

Stemmaus on tiedonhaun morfologisista lähestymistavoista käytetyin (Kettunen 2009, 270). Stemmatut sanavartalot ovat yhteisiä joko sanojen kaikille eri muodoille tai vain osalle niistä. Samaa merkitystä edustavalle sanalle eli lekseemille voi siis stemmatussa indeksissä olla useita taivutusvartaloita. Näin silloin, jos sanan taivutus on erityisen mutkikas, niin että sanamuotojen yhteinen vartalo on hyvin lyhyt tai sitä ei ole lainkaan. Ideaalitapauksessa stemmauksen lopputuloksena on sanavartalo, joka on yhteinen kaikille semanttisesti samaa tarkoittaville sanamuodoille ja vain niille.

Tiedonhaun tutkimuksessa ollaan melko yksimielisiä siitä, että automaattinen morfologinen käsittely parantaa tiedonhaun tuloksia morfologialtaan kompleksisilla kielillä. Redusoivista lähestymistavoista liberaalimpi ja halvempi, sääntöperustainen stemmaus ei yksiselitteisesti paranna hakutuloksia englanninkielisessä tiedonhaussa, mutta esimerkiksi saksan-, ruotsin- ja suomenkielisessä tiedonhaussa sen on osoitettu toimivan hyvin. Stemmaus parantaa hakutuloksia myös venäjänkielisessä tiedonhaussa (esim. Dolamic & Savoy 2009; Kettunen et al. 2007).

Lemmatointi ja stemmaus voivat kohdistua sekä dokumenttien sanoihin että hakuavaimiin. Tietokannan dokumentit indeksoidaan käyttäen redusointia, ja hakulausekkeiden sanoille tehdään vastaava käsittely ennen niiden vertaamista indeksitermeihin. Generatiiviset menetelmät sen sijaan kohdistuvat hakuavaimiin. Haettaessa taivutusmuotoindeksistä eli tietokannan indeksistä, jossa sanat ovat niissä muodoissa kun ne dokumenteissa esiintyvät, voidaan hakua kohentaa tuottamalla hakuavaimille automaattisesti eri taivutusmuotoja tai taivutusvartaloita. Generatiivisten menetelmien yhteydessä tiedonhakujärjestelmän indeksi on siis normalisoimaton eli taivutusmuotoindeksi. Perusmuodossa tai katkaistuina syötetyille hakuavaimille tuotetaan joukko taivutusvartaloita tai kokonaisia taivutusmuotoja, joita verrataan indeksissä oleviin taivutusmuotoisiin indeksitermeihin (esim. Kettunen 2009).

Taivutusvartaloiden tuottamisen idea on siinä, että tuottamalla perusmuotoiselle hakuavaimelle useita vartaloita saadaan esimerkiksi voimakkaasti taipuvalla suomen kielellä tarkempi tulos kuin redusoimalla kaikki sanan taivutusmuodot yhdeksi vartaloksi. Kaikille taivutusmuodoille yhteinen vartalo voi olla niin lyhyt, että siihen täsmäävät myös sanat, joilla ei ole semanttista yhteyttä haluttuun sanaan. Toisin kuin kevyesti stemmatun indeksin saman lekseemin eri taivutusvartaloilla, tuottamalla hakuavaimille muutama hieman pidempi taivutusvartalo saadaan katettua kaikki halutut sanamuodot ja vältetään samalla ei-toivottujen termien hakeminen. Taivutusvartaloilla jatkettujen kyselyjen ajaminen taivutusmuotoindeksiin on kuitenkin Kettusen (2009, 273) mukaan kokotekstitiedonhaun tarpeisiin liian hidasta.

Kokonaisten taivutusmuotojen tuottaminen hakuavaimille on toinen generatiivinen lähestymistapa. Kimmo Kettunen on tutkinut erityisesti sijamuotojen rajoitettua tuottamista hakuavaimille ja kehittänyt Frequent Case Generation (FCG) menetelmän (Kettunen & Airio 2006; Kettunen et al. 2007). Kettunen ja kumppanit (2007) panevat

merkille, että lemmausta ja stemmaustakin huomattavasti vähätöisempiä, taivutusmuotojen esiintyvyyteen perustuvia menetelmiä on tutkittu melko vähän. Generatiiviset menetelmät eivät ole käytössä laajalti, vaan lemmatointi ja stemmaus ovat edelleen vallitsevia lähestymistapoja.

Kettunen ja kumppanit (2007) mainitsevat lähtöajatuksekseen, ettei tiedonhakuun yksikielisestä kokoelmasta, edes voimakkaasti taipuvan kielen kyseessä ollessa, tarvita laajaa, sanakirjapohjaista lemmausta, kuten usein on oletettu. Kettunen ja Airio (2006) sekä Kettunen, Airio ja Järvelin (2007) ovat osoittaneet, että vaikka taivutusmorfologialtaan monimutkaisissa kielissä, kuten suomessa, sanalla voi teoriassa olla tuhansia eri muotoja, käytännössä dokumenttikokoelmissa esiintyvät sanamuodot ovat huomattavasti vähäisemmät, ja usein esiintyviä muotoja on sitäkin vähemmän. FCG tuottaa sääntöjen pohjalta perusmuodossa annetuista hakuavaimista taivutusmuotoja rajoittuen yleisimmin esiintyviin. Lähestymistapa on sovellettavissa useimmille eurooppalaisille kielille. Kettunen ja Airio (2006) testasivat ratkaisua suomen kieleen hyvin tuloksin. Menetelmä tuotti rohkaisevia tuloksia myös testattaessa (Kettunen et al. 2007) sitä ruotsin, saksan ja venäjän kielille. Venäjän kielen osalta tulokset tosin olivat epäluotettavia, sillä käytettävissä olleet tekstikokoelmat olivat riittämättömän kokoisia. Silti FCG näytti olevan vähintään yhtä hyvä ratkaisu kuin stemmaus. Sanamuotojen esiintyvyyttä koskevat tilastot saatiin Venäjän kansalliskorpuksesta (Russian National Corpus).

Testikokoelmina Kettunen ja kumppanit (2007) käyttivät CLEF-aineistoja. Ruotsin ja saksan kielille CLEF 2003 -aineistoja ja venäjän kielelle CLEF 2004 aineistoa. Venäjänkielisessä kokoelmassa oli alle 20 000 dokumenttia, kun ruotsin- ja saksankieliset aineistot sisälsivät satoja tuhansia dokumentteja. Relevantteja dokumentteja venäjänkielisessä aineistossa oli myös verrattain vähän. Tietokantojen indeksien ja kyselyjen normalisoimiseen käytettiin SWETWOL-, FINTWOL- ja GERTWOL-lemmatoijia, jotka kaikki perustuvat Koskenniemen (1996) TWOL-menetelmään, sekä Snowball-stemmereitä. Venäjän kielelle lemmatoijaa ei ollut saatavilla, ja reduktiivisia menetelmiä venäjänkielisessä kokeessa edusti yksin stemmaus.

Pirkolan (2001) typologiaa voi hyödyntää yksittäisen kielen kohdalla mm. arvioitaessa ennalta morfologisen käsittelyn tehoa ja esimerkiksi tehokkaan stemmerin tai

lemmatoijan luomisen edellyttämää työmäärää. Vaikka tiedonhaun tutkija tai sovellusten kehittäjä ei hallitsisi käsittelemäänsä kieltä, voi sen tiedonhakuun liittyvät morfologian ongelma-alueet paikantaa kielitypologian avulla. Koska sanamuotojen vaihtelu esimerkiksi venäjässä on samantyyppistä kuin muissa morfologisesti kompleksissa kielissä, voidaan näille kielille sopivien morfologisten käsittelytapojen olettaa olevan kokeilemisen arvoista myös venäjässä.

## 3.2 Lingvististen menetelmien vertailua

Generatiiviset menetelmät ovat yksinkertaisia, lähes yhtä tehokkaita kuin reduktiiviset ja morfologialtaan kompleksisten kielten tiedonhaussa jopa tehokkaampia kuin yksinkertaiset stemmerit. Stemmauksella voi olla hakua laajentava vaikutus. Erityisesti jos samankantaiset johdokset redusoidaan yksiksi vartaloiksi, laajentaminen saattaa heikentää tarkkuutta merkittävästi. Johdosten merkitykset saattavat olla kaukana pelkän alkuperäisen kannan ja siitä tuotettujen muiden johdosten merkityksistä. Taivutuspäätteidenkin katkaiseminen voi johtaa siihen, että jäljelle jäävä vartalo on liian lyhyt, eikä kaikilla siihen täsmäävillä dokumenttien sanamuodoilla ole semanttista yhteyttä keskenään. Tällaista ilmiötä kutsutaan ylistemmaukseksi. (Esim. Kettunen 2009.)

Muita mahdollisia ongelmia stemmaustuloksessa ovat väärinstemmaus ja alistemmaus. Alistemmauksessa pääteainesten karsinta on niin varovaista, että se tuottaa liian pitkiä vartaloita. Tällöin esimerkiksi kaikki saman sanan yleisetkään taivutusmuodot eivät täsmää stemmattuun vartaloon. Puutteelliset affiksin esiintymisympäristöä koskevat stemmaussäännöt voivat aiheuttaa väärinstemmausta, kuten sellaisten merkkien karsintaa, jotka tietyssä sanayhteydessä eivät olekaan affikseja, vaan osa sanan kantaa.

Aluksi stemmaus perustui ainoastaan sääntöihin, mutta nykyään myös sanakirjoja käyttävät stemmerit ovat melko yleisiä. Ilman sanakirjaa toimivat lemmatoijat ovat mahdollisia, mutta harvinaisia. Sanakirjojen käytöllä on hyvät ja huonot puolensa. Ne lisäävät täsmällisyyttä, mutta aiheuttavat ongelmia, mikäli sanaa ei löydykään sanakirjasta. (Kettunen 2009.)

Sanakirjan ulkopuolisia sanoja on teksteissä ja hakulausekkeissa käytännössä aina jo yksin kirjoitusvirheiden vuoksi. Internet-hakukoneiden kaltaiset järjestelmät, joiden

kokoelma on dynaaminen, ovat sanastoltaan jatkuvasti muuttuvia ja kasvavia. Tällaisissa tapauksissa sanakirjaperusteinen morfologinen analyysi voi olla riittämätöntä ja liian joustamatonta, sillä sanakirjan jatkuva päivittäminen on kallista.

Kettunen (2009, 281) vetää yhteen kriteerejä, joiden perusteella morfologisia analyysivälineitä yleensä arvotetaan. Etuna pidetään muun muassa sitä, ettei käyttäjän tarvitse huolehtia morfologisesta vaihtelusta, vaan järjestelmä hoitaa asian. Näin on yleensä lemmatoinnissa ja stemmauksessa, joissa sama automaattinen redusointi voidaan kohdistaa sekä dokumenttien että kyselyjen sanoihin, ja käyttäjä voi syöttää hakusanat missä muodossa hyvänsä. Generatiiviset menetelmät edellyttävät yleensä perusmuotoisia hakuavaimia morfologisen käsittelyn lähtökohdaksi. Tilansäästö on niin ikään reduktiivisten menetelmien etu. Stemmatun tai lemmatun indeksin koko on pienempi kuin taivutusmuotoisen, sillä yksi indeksitermi viittaa useisiin kohteisiin dokumenteissa. Toisaalta taivutusmuotoisen indeksin rakentaminen ja ylläpito vaatii vähemmän ihmistyötä ja aikaa. Tämä on huomattava etu etenkin laajoja web-kokoelmia indeksoitaessa.

Hakutulosten paraneminen on tietysti kriteereistä ensisijainen, ja se suhteutetaan muihin kriteereihin. Eniten hakutulosten saantia ja tarkkuutta voimakkaasti taipuvien kielten tiedonhaussa näyttää parantavan lemmatointi. Menetelmä on kuitenkin työläs ja sanakirjaan perustuvana myös tehoton ainakin web-tiedonhaussa. Stemmaus parantaa hakutuloksia vastaavilla kielillä kohtuullisesti siihen nähden, että menetelmä on varsinkin sääntöperustaisena melko keveä ja halpa. (Kettunen 2009.)

Myös generatiivisilla menetelmillä on päästy hyviin hakutuloksiin. Erityisesti rajoitettu täysien taivutusmuotojen tuottaminen (FCG) on edullista ja tehokasta, mutta sitä ei ole toistaiseksi tutkittu kovinkaan paljon eikä ohjelmia todelliseen käyttöön kehitetty. Kettunen (2009) toteaa, että stemmauksesta on tullut tiedonhaun tutkimuksessa suorastaan standardi morfologisen vaihtelun hallintamenetelmä. Stemmereitä on saatavilla helposti yli 20 kielelle, ja Snowballin stemmeristä on olemassa versio 15 kielelle.

Snowball on ohjelmointikieli, jolla stemmausalgoritmi voidaan kirjoittaa erilaisia luonnollisia kieliä varten. Se on kehitetty M. Porterin vuonna 1980 julkaiseman englannin kielen suffikseja karsivan algoritmin pohjalta. Englanninkielisessä tiedonhaussa stemmaus onnistuu melko yksinkertaisten karsittavia päätteitä koskevien

sääntöjen ja niiden rajoitusten avulla. Venäjä on morfologialtaan huomattavasti englantia monimutkaisempi. Dolamicin ja Savoyn (2009) mukaan morfologisesti kompleksisten kielten tiedonhaussa leksikaalinen, sanakirjaan ja kielioppisääntöihin perustuva stemmeri olisi tulosten kannalta hyvä ratkaisu. Laajan sanakirjan ja säännöstön luominen ja ylläpito vaativat kuitenkin paitsi työvoimaa, mahdollisesti myös prosessointiaikaa. Siksi kohtuulliset hakutulokset tuottava algoritminen stemmeri voisi olla paras ratkaisu, mutta esim. venäjän tai suomen kielille sellaisen suunnittelussa on otettava huomioon useampia seikkoja kuin englannin kohdalla (Dolamic & Savoy 2009). Alempana hieman lähemmin tarkasteltavat stemmerit ovat sääntöpohjaisia eli algoritmisia stemmereitä.

Dolamicin ja Savoyn tutkimuksen (2009) tulokset osoittavat stemmauksen toimivan tilastollisesti katsoen aina morfologisen analyysin poisjättämistä paremmin. Stemmauksen edut morfologialtaan kompleksisille kielille onkin havaittu kiistatta useissa tutkimuksissa. Pirkola (2001) huomauttaa, että yksittäisissä hauissa reduktiivisista menetelmistä saattaa olla enemmän haittaa kuin hyötyä. Hänen esimerkkinsä liittyy homonymiaan. Suomen sana *kuusi* voi tarkoittaa puuta tai lukua, mutta sen taivutusmuodot eivät ole homonyymisia. Moniselitteisyys siis ratkeaa esimerkiksi sanamuodoissa *kuusen* ja *kuuden*, joiden merkityksestä ei ole samanlaista epävarmuutta kuin näiden sanojen nominatiivimuotojen merkityksestä. Lemmatointi tai ahne stemmaus voi siis lisätä monimerkityksisyyttä. Lemmatoija palauttaisi sanat sanakirjamuotoon, joka on yksikön nominatiivi *kuusi*. Stemmatussa indeksissä lukusanaa merkitsevän sanan *kuusi* eri sijamuotoihin luultavasti viittaisi kaksi eri indeksitermiä. Nämä olisivat taivutusvartalot *kuud* ja *kuus*. Jos sanan kaikkia muotoja edustaisi indeksissä yksi stemmattu vartalo, täytyisi sen olla niinkin lyhyt kuin *kuu*, joka viittaisi paitsi kuusipuuta tarkoittaviin sanamuotoihin, myös esimerkiksi kuuta, kuumuutta ja kuusamaa tarkoittaviin. Tällaiselta epätarkkuudelta vältytään generatiivisilla menetelmillä, etenkin kokonaisten sijamuotojen rajoitetulla tuottamisella, FCG:llä.

Dolamicin stemmerin määrittelemät karsittavat päätteet ovat substantiivien ja adjektiivien sijapäätteitä ja niihin liittyviä suku- ja lukusuffikseja, mutta niiden kanssa voi osua yksiin myös muiden sanaluokkien sanojen pääteaineksia. Vaikka stemmerin tarkoitus ei ole laajentaa hakua yli sanaluokkarajojen, täsmäävät substantiivien ja adjektiivien karsitut vartalot usein myös esimerkiksi verbimuotoihin. Stemmauksella

onkin taipumus parantaa saantia, mutta samalla tarkkuus ei välttämättä parane samassa suhteessa. FCG-menetelmässä morfologinen käsittely kohdistuu vain haluttuihin sanaluokkiin, eivätkä hakuavaimet myöskään täsmää muihin kuin oman sanaluokkansa sanoihin. FCG:ssä sijamuotoja tuotetaan hakuaiheiden substantiiveille ja adjektiiveille. Mahdolliset muiden sanaluokkien sanat jätetään siihen muotoon, missä ne sattuvat olemaan, sillä niiden vaikutus haun tuloksellisuuteen on vähäinen (Baeza-Yates & Ribeiro Neto 1999; Kettunen 2006; Kettunen et al. 2007; Loponen & Järvelin 2010). FCG:n tarkoituksena on tuottaa vain kaikkein keskeisimmät taivutusmuodot, jotka voidaan rajata muutamiin. Näin halulausekkeet voidaan pitää lyhyinä ja prosessointi nopeana.

FCG-hakumenetelmän on arveltu soveltuvan erityisen hyvin web-tiedonhakuun (Kettunen et al. 2007), jossa hakukoneiden kokoelmat ovat erittäin suuria ja dynaamisia, ja morfologisen analyysin tekeminen indeksointivaiheessa mahdollisesti hidasta. Erityisesti sanakirjapohjaisten indeksointimenetelmien käyttö web-hakukoneessa on paitsi hidasta, myös tehotonta sikäli, etteivät sanakirjat pysy jatkuvasti muuttuvan kokoelman dokumenttien sanaston perässä ainakaan ilman jatkuvaa päivitystä, mikä puolestaan on kallista. FCG:n tekemä morfologinen käsittely, sijamuotojen rajoitettu tuottaminen, kohdistuu ainoastaan hakulausekkeisiin, ja tietokannan indeksi voi siten olla helpoiten tehtävä ja ylläpidettävä taivutusmuotoindeksi (Kettunen et al. 2007).

# 4 VENÄJÄN KIELEN MORFOLOGIA TIEDONHAUSSA

Tässä luvussa tarkastellaan venäjän kielen morfologiaa tiedonhaun näkökulmasta käyttäen apuna Pirkolan (2001) tiedonhaun kielitypologiaa. Alaluvuissa 4.1 – 4.5 venäjän kielen morfologian keskeisiä osa-alueita käydään läpi keskittyen seikkoihin, jotka tiedonhaun näkökulmasta vaikuttavat olennaisilta. Näitä venäjän morfologian piirteitä vedetään yhteen suhteessa tiedonhaun kielitypologinaan alaluvussa 4.6. Tiedonhaun lingvistisiä menetelmiä venäjän morfologian käsittelyssä pohditaan vielä alaluvussa 4.7.

Konkreettisten esimerkkien saamiseksi tarkasteluun on otettu mukaan kaksi venäjän kielen stemmaukseen tarkoitettua karsinta-algoritmia, Ljiljana Dolamicin (Dolamic & Savoy 2009) kevyt stemmeri ja Snowball-stemmeri. Martin Porterin (2001) erikielisten stemmausalgoritmien kirjoittamiseen luodun Snowball-kielen verkkosivu tarjoaa venäjän kielestä 49 673 sanan näytteen ja stemmatun version siitä. Stemmatusta versiosta sekä Dolamicin stemmerillä käsitellystä samasta sananäytteestä on tähän otettu muutamia esimerkkejä. Lisäksi tämän tutkielman empiiriseen osaan (Luku 5) liittyviä kyselyjä muotoiltaessa ilmenneitä seikkoja nostetaan paikka paikoin esiin jo tässä luvussa.

Venäjän ulkopuolella tehtyä venäjän kielen käsittelyä koskevaa tiedonhaun tutkimusta on rajoittanut saatavilla olevien testikokoelmien vähyys. Vuoden 2003 CLEF-kampanjassa tuotettiin pieni venäjänkielinen kokoelma, jonka jälkeen siihen on kohdistettu muutamia tutkimuksia (Kettunen et al. 2007). Dolamicin ja Savoyn (2009) venäjän kielen morfologian hallintamenetelmien vertailussa käyttämä testikokoelma luotiin vuosien 2005–2008 CLEF-kampanjoissa. Kokoelma koostuu Russian Social Science -korpuksesta (RSSC), jossa on 94,581 dokumenttia, ja INION-korpuksesta, joka sisältää bibliografisia tietoja venäjänkielisistä yhteiskunta- ja taloustieteellisistä teksteistä. INION koostuu 145 802 dokumentista. Kokoelmien dokumentit ovat lyhyitä.

Dolamic ja Savoy (2009) ovat vetäneet yhteen venäjän kielen morfologisia piirteitä. He muistuttavat Kettusen ja Airion (2006) huomiosta, että konkreettisissa käyttötilanteissa kielen teoreettinen morfologinen monimutkaisuus harvoin ilmenee täydessä mitassaan. Dolamic ja Savoy (2009) sivuavat myös korpustilastoja ja panevat merkille esimerkiksi sijamuotojen todellisen käyttöfrekvenssin tuntemisesta mahdollisesti koituvat hyödyt.

Käydessään läpi venäjän kielen morfologiaa koskevaa tiedonhaun tutkimusta he mainitsevat venäläisestä tutkimuksesta ROMIP- kampanjat, joissa on koottu testikokoelmia pääosin webin venäjänkielisistä dokumenteista. Kokoelmat eivät olleet vapaasti saatavilla ja tiedot korpuksista, evaluointimenetelmistä sekä hakuun ja tallennukseen käytetyistä lingvistisistä keinoista olivat ROMIP:in verkkosivuilla ainoastaan venäjäksi. Poisluettuaan näin venäläisen tutkimuksen Dolamic ja Savoy (2009) päätyvät siihen, että siihenastisista venäjän kielen käsittelyn käytännön sovelluksista tiedonhaussa parhaat tulokset tuotti Snowballin stemmeri. Tässä luvussa on käytetty jonkin verran myös Venäjällä tehtyä tutkimusta, mutta pääasiallisesti venäjän kielen morfologiasta tehtäviä huomioita peilataan Dolamicin ja Savoyn (2009) sekä Kettusen ja kumppaneiden (2007) havaintoihin.

## 4.1 Sanaluokat ja sulkusanat

Kielen morfologiaa voidaan jäsentää lähtien sanaluokista, joihin sanoja jaetaan niiden muotojen ja funktioiden perusteella. Venäjän kieliopit luokittelevat kielen sanat perinteisesti itsenäisiin ja apusanaluokkiin sen mukaan, voivatko ne toimia lauseenjäseninä. Itsenäisiä sanaluokkia ovat verbit, substantiivit, adjektiivit, pronominit, lukusanat ja adverbit. Apusanaluokkia ovat prepositiot, konjunktiot, interjektiot ja partikkelit. Lauseenjäsennykseen perustuva luokitus liittyy ennen muuta sanojen leksikaaliseen merkitykseen eli siihen, mitä kielen ulkopuolista olentoa sana tarkoittaa. Morfologian ja kieliopillisten merkitysten kannalta luokittelu ei ole aukoton, koska esimerkiksi lukusanoista osa taipuu adjektiivien tavoin, osa taas muistuttaa substantiiveja sekä taivutukseltaan että sukuominaisuutensa perusteella. (Nikunlassi 2002, 123.)

Taivutuksen pohjalta venäjän kielen sanat voitaisiin jakaa luokkiin samaan tapaan kuin suomenkin sanat. Nomineilla on sijataivutus, verbeillä verbitaivutus, ja partikkelit eivät taivu lainkaan. Tämän jaon mukaan adverbit, konjunktiot, prepositiot ja interjektiot kuuluvat partikkelien luokkaan, substantiivit, adjektiivit, pronominit ja lukusanat nominien. Verbejä ovat sanat, joilla on persoona- tai aikakategoria. Sanaluokat eivät kummallakaan tavalla jaotellen ole täysin suljettuja, sillä esimerkiksi partisiipeilla on sekä verbi- että nominitaivutuksen piirteitä. Kontekstista riippuen tietty sana voi olla esimerkiksi adjektiivi tai partisiippiverbi. Sana voi myös kokonaan siirtyä sanaluokasta

toiseen. Esimerkiksi adjektivoitunut partisiippiverbi voi ajan myötä poistua käytöstä verbinä tai substantivoitunut adjektiivi vakiinnuttaa substantiivimerkityksensä. Täysin samanmuotoisten sanojen esiintyminen eri sanaluokissa on venäjässä kuitenkin huomattavasti harvinaisempaa kuin vaikkapa englannissa.

Sanojen kieliopilliset muodot voidaan jakaa kieliopillisiin kategorioihin. Esimerkiksi verbin aikamuodot, joita venäjän kielessä on kolme, kuuluvat ajan kieliopilliseen kategoriaan. Venäjän kielen nominien muodoissa vaikuttavat luvun, sijan, suvun, vertailuasteen ja elollisuuden sekä elottomuuden kategoriat. Verbien kieliopilliset kategoriat ovat aspekti, persoona, tempus, tapaluokka, pääluokka ja suku. Nominien kieliopillisista kategorioista substantiivien taivutukseen vaikuttavat vain sija ja luku. Elollisuus tai elottomuus sekä suku ovat substantiivien muuttumattomia ominaisuuksia, jotka vaikuttavat muihin nomineihin kategorioina. Adjektiivien muodoissa vaikuttavia kategorioita ovat sija, suku, luku, vertailuasteet sekä elottomuus tai elollisuus.

Tässä tutkielmassa tarkoitetaan Dolamicin stemmerillä substantiivien ja adjektiivien sijapäätteitä karsivaa ”kevyttä” karsinta-algoritmia, jonka kehittämisestä kertovat Dolamic ja Savoy (2009). Itse stemmerin kuvauksessa sen tekijäksi mainitaan Ljiljana Dolamic. Kehittäessään stemmereitä venäjänkieliseen tiedonhakuun Dolamic ja Savoy (2009) lähtevät siitä, että verbien taivutusmuodot on syytä jättää käsittelyn ulkopuolelle, sillä ne runsaudessaan aiheuttaisivat liian paljon virheitä. Savoy (2006) on päätynyt siihen, että stemmereitä suuniteltaessa huomio kannattaa kohdistaa lähinnä substantiiveihin ja adjektiiveihin. Tiedonhaun kieliteknologian tutkimuksessa pidetäänkin jo yleisesti (Baeza-Yates & Ribeiro Neto 1999; Kettunen & Airio 2006; Kettunen et al. 2007; Loponen & Järvelin 2010) selvänä, että verbit voidaan jättää morfologisessa käsittelyssä huomiotta. Venäjän kielen Snowball-stemmeri karsii kuitenkin myös verbipäätteitä. Sekä Dolamicin että Snowballin nomineista karsittavien suffiksien listat on tehty silmällä pitäen ainoastaan substantiivien ja adjektiivien sijapäätteitä, mutta käsitellyiksi tulevat myös vastaavasti taipuvat pronominit ja lukusanat.

Tekstitiedonhakujärjestelmään liittyykin yleensä sulkusanalista, jolla dokumenttien informaatiosisällön kannalta merkityksettömät sanat suljetaan pois indeksointi- ja hakuprosesseista. Tällaisia ovat prepositiot, kuten englannin *in* ja venäjän *в (v)*, konjunktiot, kuten suomen *ja*, interjektiot eli huudahdussanat sekä yleisimmät adverbit,

esimerkiksi suomen kielen sana *paljon*. Nomineista pronominit ja yleisimmät lukusanat laitetaan usein sulkusanalistalle, jottei niitä turhaan käsiteltäisi. Myös useimmin esiintyvien substantiivien, adjektiivien ja verbien, esimerkiksi olla-verbin, katsotaan olevan sellaisia, että niiden ottaminen mukaan indeksiin ja hakuun kuormittaisi järjestelmää. Etenkin suuria dokumenttikokoelmia käsiteltäessä sulkusanalista voi nopeuttaa indeksointia sekä myös hakujen suorittamista merkittävästi. Dolamicin ja Savoyn (2009) stemmereidensä oheen suunnittelema 420 sanan lista on saatavilla Neuchâtelin yliopiston monikielisen tiedonhaun resurssisivulla. Snowball-stemmerin yhteydessä voi halutessaan käyttää Snowballin sulkusanalistaa.

Tämän tutkielman empiirisessä osassa käytetyn KM.ru-testikokoelman taivutusmuotoindeksiä rakennettaessa on hyödynnetty hieman muokattua versiota Dolamicin ja Savoyn (2009) sulkusanalistasta. Sulkusanalistoille kuuluvista sanoista on erilaisia näkemyksiä. KM.ru-kokoelmaa indeksoitaessa mahdollisimman monen sanan saaminen sulkusanalistalle oli sikäli tärkeää, että kokoelman suuri koko hankaloitti prosessia. Sen keventämiseksi ja nopeuttamiseksi olivat kaikki keinot tervetulleita. Liian laaja sulkusanalista voi kuitenkin huonontaa hakutulosta. Eräs testikokoelman hakuaiheista on ’лагеря второй пятилетки’ suomeksi ’toisen viisivuotiskauden leirit’. Sen valossa järjestyslukusanojen jättäminen sulkusanalistan ulkopuolelle näyttäisi perustellulta, sillä hakijan voi kuvitella haluavan tietoa juuri kyseisen viisivuotiskauden leireistä, eikä esimerkiksi neljännen viisivuotiskauden. Muita esimerkkejä hakutuloksen kannalta olennaisilta vaikuttavista järjestyslukusanoista ovat ’ensimmäinen maailmansota’ ja ’Toinen Internationaali’.

Dolamicin ja Savoyn (2009) lista ei ole järjestyslukujen suhteen kattava. Siinä on sana второй (vtoroj, toinen), joka on oletettavasti yksikön nominatiivin maskuliinimuoto. Esimerkkihakuaiheen yksikön genetiivin feminiinimuoto on sen kanssa homografinen (taivutusmuotohomografiasta esim. Alkula 2000). Ilmeisesti siksi, että sulkusanalistalla ovat vain kaikkein yleisimmät lukusanat, siinä ei ole sanan второй (vtoroj) muita muotoja. Sama koskee muita järjestyslukuja: sulkusanalistalla ovat vain niiden maskuliinipäätteiset sanakirjamuodot. Ne ja niiden homografit eivät siis ole indeksissä, mutta muut muodot ovat, esimerkiksi feminiinipäätteinen nominatiivi вторая (vtoraja – toinen). Yksittäisissä hauissa sulkusanalista saattaa vaikuttaa hakutuloksen laatuun. Kokonaisvaikutus selviäisi vain kokeilemalla erilaisia sulkusanalistoja. Tutkiessaan indeksointi- ja hakustrategioiden soveltuvuutta venäjänkieliseen tiedonhakuun Dolamic

ja Savoy (2009, 2545) havaitsivat, ettei sulkusanalistan käytöllä tai poisjättämisellä ollut mainittavaa vaikutusta hakutulosten keskitarkkuuksien keskiarvoihin. Pyrittäessä keventämään indeksointiprosessia sulkusanalista kuitenkin ansaitsee huomiota.

Snowball:in tarjoama venäjän kielen sulkusanalista sisältää peruslukuja, mutta ei järjestyslukuja. Se on muodostettu sanojen esiintyvyystiheyden perusteella. Listaan kuuluvat vain kaikista yleisimmät niistä sanoista, joiden indeksointi katsotaan turhaksi. Siinä ei esimerkiksi ole muita possessiivipronomineja kuin yksikön ensimmäisen persoonan maskuliinin nominatiivi мой (moj), sekä persoonapronominien genetiivimuotojen kanssa homografiset, taipumattomat kolmannen persoonan possessiivipronominit ей (ej) его (ego) ja их (ih). Ensimmäisen ja toisen persoonan possessiivipronomineilla on suku- luku- ja sijataivutus, ja yksin niistä saataisiin 50 eri sanamuotoa, jotka informaatiosisältönsä puolesta sopisivat hyvin sulkusanalistalle. Kooltaan KM.ru-kokoelmaa vastaavia testikokoelmia indeksoitaessa voisi kaikki pronominimuodot kattavasta listasta olla hyötyä.

Koska verbien on ylipäätään todettu olevan tiedonhaussa vähämerkityksisiä, ne voidaan jättää morfologisen käsittelyn ulkopuolelle, kuten Dolamic ja Savoy (2009) ovat tehneet. Käsiteltäviksi jäävät substantiivit ja adjektiivit, joilla venäjässä on sijataivutus. Verbit ja muut sanat, jotka eivät sisälly sulkusanalistalle, mutta jotka eivät täytä prosessoitaviksi valittujen sanaryhmien määritelmää, ovat indeksissä ja hauissa alkuperäisissä muodoissaan.

## 4.2 Yhdyssanat

Yhdyssanat eivät ole venäjässä samanlainen ehtymätön sananmuodostuskeino kuin suomessa, mutta niitä kuitenkin esiintyy, eikä niiden muodostaminen ole agglutinatiivista. Yhdyssanojen ensimmäisinä osina toimiessaan substantiivit ja adjektiivit esiintyvät usein -o tai -e -loppuisina. Toisin sanoen yhdyssanan määriteosan vartaloa yhdistää perusosaan liitosmorfeemi (*fogemorpheme*) (Hedlund 2002). Esimerkiksi substantiivi вода (voda), *vesi*, saa asun водо (vodo) sanassa водопад (vodopad – vesiputous*)*. Adjektiivi железный (železnyj – rautainen) muuttuu samantyyppisesti yhdyssana-adjektiivissa железнодорожный (železnodorožnyj – rautatieläinen, rautatie-). Liudentuneen konsonantin jälkeen vastaavassa asemassa oleva kirjain on -e.

Kun morfi loppuu samaan äänteeseen kuin seuraava morfi alkaa, limittyvät morfit siten, että kyseinen äänne esiintyy sanassa vain kerran. Myös yhdyssanojen osat limittyvät: morfofonologia on venäjäksi морфонология (morfonologija). Venäjässä on huomattavan paljon leksikaalistuneita lyhenteitä ja vakiintuneita tapoja lyhentää varsinkin usein käytettyjä yhdyssanojen alkuosia, kuten государственный (gosudarstvennyj – valtiollinen), joka lyhenee muotoon гос (gos), esim. господдержка (gospodderžka – valtion tuki). Tällaisten piirteiden takia yhdyssanojen osittaminen lemmatisoinnin yhteydessä ei vaikuta hyödylliseltä. Venäjän kielessä yhdyssanat eivät myöskään ole yhtä yleisiä kuin suomessa, ruotsissa tai saksassa. Ruotsinkielisessä tiedonhaussa yhdyssanojen osittamisesta on todettu olevan hyötyä (Ahlgren & Kekäläinen 2006), samoin esimerkiksi suomenkielisessä (Alkula 2000), mutta Dolamic ja Savoy (2009, 2545) havaitsivat, että venäjänkielisessä haussa osittamisesta on haittaa.

Monissa tapauksissa, joissa suomen kielessä käytettäisiin määrittely- ja perusosasta koostuvaa yhdyssanaa, korvaa venäjässä määrittelyosan erillinen adjektiivi. Esimerkiksi suomen informaatioteknologia on venäjäksi инфомационная технология (informacionnaja tehnologija). Yhdyssanojen vähäisyys eli matala synteesin yhdyssana-aste (*compound index of synhesis*) on stemmauksen toimivuuden kannalta suotuisa piirre. Pääteainesten karsinnassa yhdyssanarikkaalla kielellä vaarana on huono saanti, sillä relevanteissa dokumenteissa yhdyssanojen osat saattavat esiintyä erikseen. Venäjässä tämä ongelma on luultavasti vähäinen. Luvun 5 tiedonhakukokeessa käytetyt 47 ROMIP-hakuaihetta sisältävät vain muutamia yhdyssanoja. Näistä esimerkiksi keskiaikaista tarkoittava ilmaus, joka, kuten suomessa, on venäjässä yhdyssana, on niin vakiintunut, että sen jakaminen osiinsa olisi haun tuloksen kannalta luultavasti pikemminkin vahingollista kuin hyödyllistä.

## 4.3 Johdokset

Englanninkielisessä tiedonhaussa stemmauksen kohteeksi otetaan yleensä taivutuspäätteiden lisäksi myös johdosaffiksit, joten eri sanaluokkia edustavat, samankantaiset sanat palautuvat samaan vartaloon. Dolamicin ja Savoyn (2009) sekä Segalovichin (2003) mukaan venäjän kielessä johdosaffiksien karsiminen aiheuttaisi tiedonhaun tuloksen heikkenemistä, sillä johdokset voivat olla lähtökohdastaan semanttisesti jo hyvin kaukana. Toisaalta merkitysero voi olla myös erittäin vähäinen,

kuten verbien aspektipareilla tai diminutiivisilla johdoksilla. Diminutiivinen johdossuffiksi liittää sanaan pienen koon tai muun vähäisyyden merkityksen.

Johdoksia muodostetaan venäjässä prefikseillä ja suffikseilla. Stemmauksessa prefiksit eli etuliitteet on syytä jättää koskematta, sillä niiden avulla muodostetut sanat ovat usein hyvin kaukana johtamisen lähtökohtana olevan sanan merkityksestä. Lisäksi prefiksien automaattiseksi tunnistamiseksi on hankalampi luoda sääntöjä kuin suffiksien tunnistamiseksi, sillä useiden sanojen vartaloon kuuluvat ensimmäiset kirjaimet ovat identtisiä joidenkin prefiksien kanssa (Segalovich 2003). Snowball-stemmeri asettaakin karsinnan ehdoksi, että jäljelle jäävässä vartalossa on oltava ainakin yksi vokaali. Muuten vaarana olisi joidenkin sanojen karsiutuminen olemattomiin. Dolamicin ja Savoyn (2009) suunnittelema aggressiivisempi stemmeri karsii myös johdosaffikseja keskittyen kuitenkin vain suffikseihin. Karsimalla adjektiiviset ainekset se pyrkii yhdistämään substantiivin ja siitä johdetun adjektiivin samaan vartaloon. Taivutusaffiksein karsintaan kevyellä stemmerillä riitti 57 sääntöä. Johdosaffiksien karsimiseksi aggressiiviseen stemmeriin lisättiin vielä 40 sääntöä.

Tätä lukua varten on testattu Dolamicin stemmereistä tekijän suosittelemaa kevyempää versiota, joka karsii sanoista ainoastaan taivutusaffikseja. Snowball-stemmerin karsittavissa suffikseissa on yksi johdossuffiksi, -ост- (-ost-) ja sen versio -ость (-ost'). Sana надменность (nadmennost' - ylimielisyys) karsiutuu muotoon надмен (nadmen). Sanan morfologinen rakenne on seuraava:

prefiksi на- (na-), kanta -дме- (-dme-), suffiksit -ен- (-en-) -н- (-n-) -ость (-ost')

Jos stemmeri poistaisi sanasta prefiksin ja kaikki johdossuffiksit, ja sen kanta jäisi stemmattuun indeksiin karsituksi vartaloksi, olisi kyse selvästä ylistemmauksesta, jonka seurauksena termiin täsmäisivät merkitykseltään hyvin kaukaiset sanat, esimerkiksi предмет (predmet – asia, esine) ja подмести (podmesti – lakaista).

Pirkolan (2001) mukaan lähes kaikissa kielissä morfeemit järjestyvät siten, että taivutuspäätteet ovat sanojen lopussa viimeisinä, ja johdossuffiksit sijoittuvat niiden ja sanan kannan väliin. Johdosaffiksien tahaton poistaminen on siis epätodennäköistä, koska stemmerit yleensä karsivat affikseja sanan lopusta lukien. Johdosten määrän ja eroteltavuuden mittaamisen voi venäjänkielisen tiedonhaun tutkimuksessa jättää joka tapauksessa sivummalle, koska niiden käsittelyyn aiempi tutkimus ei kannusta. Kun

kielen synteesin yhdyssana-aste on matala, keskeisiksi nousevat taivutusmorfologia sekä Pirkolan (2001) muuttujista indeksin taivutusaste ja fuusion taivutusaste.

Snowball-stemmerin harjoittaman verbien päätteiden poiskarsimisen lopputulos on kuitenkin samantyyppinen kuin johdossuffiksien karsimisen: verbien vartalot käyvät yksiin substantiivien ja adjektiivien vartaloiden kanssa. Käytännössä Snowball siis usein yhdistää johdokset yksiin vartaloihin. Näin käy joissain tapauksissa myös Dolamicin stemmerillä, kun esimerkiksi adjektiivin karsinnan tuloksena oleva vartalo täsmää myös adverbeihin, substantiiveihin ja verbeihin. Verbien morfologista käsittelyä vastaan puhuu niiden vähäisen informaatioarvon lisäksi runsas johtomorfologia, jota käsitellään seuraavassa alaluvussa yhdessä verbitaivutuksen kanssa.

## 4.4 Verbien taivutus- ja johtomorfologiasta

Venäjän verbit esiintyvät personaisina eli finiittisinä sekä toisaalta infinitiiveinä, partisiippeina ja gerundeina. Finiittimuodot ilmaisevat tapaluokkaa eli modusta. Kuten suomessa, moduskategoria on kolmijäseninen, ja siihen kuuluvat indikatiivi, imperatiivi ja konditionaali. Verbit jakautuvat aspektipareihin, eli perfektiivisen ja imperfektiivisen aspektin verbeihin. Aikamuotoja venäjän kielessä on yleisen tulkinnan mukaan kolme: preteriti, preesens ja futuuri. Preteritin eli menneen ajan merkitykset kattavat suomen imperfektin, perfektin ja pluskvamperfektin. Aikamuodot liittyvät kiinteästi aspekteihin, sillä esimerkiksi preesensiä eli puhehetkellä tapahtuvaa toimintaa voi ilmaista ainoastaan imperfektiivisen aspektin verbillä. Mennyttä aikaa ilmaisee suffiksi -л-, mutta perfektin, pluskvamperfektin ja futuurin merkitykset on tulkittava aspektin perusteella, milloin kyseessä ei ole liittofutuuri, eli esimerkiksi suomen *tulen kirjoittamaan* -tyyppistä ilmaisua vastaava, mutta huomattavasti yleisempi muoto буду писать (budu pisat'). Kuten useissa kielissä, venäjän verbien persoonakategoriassa on kolme muotoa: ensimmäinen, toinen, ja kolmas. Verbin pääluokkakategorian ryhmät ovat passiivi ja aktiivi.

Tiedonhaun kannalta mielenkiintoinen verbimuoto on partisiippi. Se esittää verbin kuvaaman toiminnan substantiivin tarkoitteen ominaisuutena samalla tavalla kuin adjektiivi. Partisiipissa on erityistä se, että siihen liittyy sekä verbin että adjektiivin piirteitä. Sillä on verbin aikakategoria ja pääluokka, mutta ei ole modusta eli tapaluokkaa. Adjektiivin tavoin sillä on sija-, suku- ja lukukategoriat, ja sen rooli

lauseessa on samanlainen kuin adjektiivin. Se esiintyy attribuuttina kuten adjektiivi ja lyhyessä muodossa myös predikaattina kuten adjektiivin lyhyet muodot (Rahmanova & Suzdal'ceva 2003, 418). Suomessa partisiipin rooli on samantyyppisesti kuvaileva, esimerkiksi partisiipissa *kuvaileva*. Venäjässä partisiipin päätteet ovat lisäksi samanlaisia kuin adjektiivin, joten sanaryhmiä on muodon perusteella vaikea erottaa toisistaan.

Snowballin partisiippien tunnusten karsintaan liittyy ehto, jonka mukaan passiivin partisiipin preteritin tunnuksen -нн- (-nn-) edellä on oltava kirjain a, jotta tunnus karsittaan. Kyseisessä verbimuodossa tunnuksen edellä voi kuitenkin olla myös kirjain e (tai ë), esimerkiksi sanassa наполненные (napolnennye – täytetyt). Siitä sekä Snowball että Dolamicin stemmeri karsivat adjektiiviset sijapäätteet, joihin kuuluu muiden muassa ые (ye), ja soveltavat kahden н(n)-kirjaimen typistämissääntöä. Molempien stemmaustulos on наполнен (napolnen). Vartaloon täsmäävät substantiivit наполненность (napolnennost' – täyteys) ja наполнение (napolnenie – täyte), mutta ei verbi наполнить (napolnit' – täyttää). Verbinkin tavoittaminen olisi kuitenkin oletettavasti Snowballin karsinnan tarkoitus.

Toisaalta sana напряженный (naprjažennyj) luokitellaan venäjänkielisessä wikisanakirjassa adjektiiviksi eikä partisiipiksi. Sen ensisijainen merkitys on *kireä*, vaikka se sanamuodon puolesta voisi yhtälailla olla partisiippi *jännitetty*. Molemmat ohjelmat redusoivat sanan vartaloksi напряжен (naprjažen), jolla tavoitetaan adjektiivin eri muodot. Verbi напрягать (naprjagat' – jännittää) on morfologisesti etäämmällä, sillä siinä on konsonanttivaihtelun seurauksena kirjaimen ж (ž) tilalla г (g). Kyseessä on tyypillinen verbejä koskeva morfologisen fuusion ilmiö, joka nähtävästi on Snowballin partisiippien karsintaa rajoittavan ehdon osasyy. Ilmeisesti ainoastaan ehdon täyttävät sanamuodot ovat yksiselitteisesti partisiippiverbejä, eivätkä esimerkiksi adjektiiveja, eikä niiden taustalla olevissa verbeissä esiinny konsonanttivaihtelua. Tässä tarkasteltuun otokseen sisältyvät sanamuodot невинный (nevinnyj – syytön) ja неожиданною (neožidannoju – äkillistä) ovat myös adjektiiveja, vaikka niissä on partisiippiverbeihinkin kuuluva suffiksi -нн- (-nn-). Dolamicin stemmeri jättää tietysti kaikki partisiippiverbien tunnukset karsimatta, ja kohtelee näitä sanamuotoja adjektiiveina, sillä adjektiivien ja partisiippien päätteet ovat samanlaiset. Verbien koko problematiikan voi ohittaa Dolamicin ratkaisulla, joka ilmeisesti on hakutulostenkin

puolesta hyvä. Dolamicin stemmeri tietysti käytännössä karsii myös verbien päätteitä niiltä osin, kun ne käyvät yksiin nominien sijapäätteiden kanssa.

Venäjän verbien morfologiaa mutkistavat aspektit. Aspektien muodostamiseen käytetään sekä suffikseja, prefikseillä että vartalon äännevaihtelua. Prefikseillä voidaan perfektiivisen aspektin muodostamisen lisäksi muuttaa sanan merkitystä esimerkiksi päinvastaiseksi влюбить (vljubit' – rakastua) расслюбить (rassljubit' – lakata rakastamasta). Imperfektiivinen aspekti merkitsee puhehetkellä jatkuvaa tai toistuvaa toimintaa, perfektiivinen kertaluonteista, yleensä loppuun saatettua tai saatettavaa toimintaa. Lähes kaikki venäjän verbit ovat parillisia. Niillä on siis perfektiivinen ja imperfektiivinen muoto. Kyse ei kuitenkaan ole saman sanan eri muodoista vaan eri sanoista, mutta niiden leksikaalinen merkitys on ytimeltään sama. Aspektiero koskee tekemisen laatua, mutta ei itse tekemisen merkitystä. Useimpia suomenkielisiä verbejä vastaavat sanakirjassa aspektiparit. Seuraavassa esimerkissä kenoviivojen oikealla puolella olevat sanat ovat perfektiivisen aspektin verbejä.

lukea – читать / прочитать (čitat' / pročitat')

kerätä – собирать / собрать (sobirat' /sobrat')

ostaa – покупать / купить (pokupat' / kupit')

antaa – давать / дать (davat'/ dat')

Viimeisessä esimerkissä imperfektiivinen aspekti on muodostettu suffiksilla -ва- (-va-). Ostamista merkitsevässä sanassa aspektiero luodaan etuliitteellä ja vartalon äännevaihtelulla. Prefikseistä on mahdoton erotella niitä, joiden vaikutus verbeihin olisi vain aspektia eikä koko merkitystä muuttava. Lisäksi monet etuliitteet sekä muuttavat verbin aspektin perfektiiviseksi että muuttavat sen merkitystä. Jos stemmauksella haluttaisiin saada kaikki perusmerkitykseltään samat verbit yhtenäistettyä, olisi myös aspektiparit voitava yhdistää toisiinsa. Se ei kuitenkaan vaikuta järkevältä tavoitteelta. Etuliitteet onkin jätetty Snowball-stemmerissä karsinan ulkopuolelle. Lemmatisoijien tuottama sanakirjamuoto edustaa niin ikään aina vain yhtä verbin aspekteista.

Hankaluutta tosiaan semanttisesti lähellä olevien johdosten yhtenäistämisessä aiheuttavat myös liikeverbit. Liikeverbit jakautuvat iteratiivisiin ja duratiivisiin eli yhteen päämäärään suuntautuvaa liikettä kuvaaviin ja edestakaista tai päämäärätöntä liikettä kuvaaviin. Liikeverbejä on venäjässä paljon, sillä eri kulkuneuvoilla

tapahtuvalle liikkumiselle tai kuljettamiselle on venäjässä eri verbit. Suomessakin toki liikkumisen tapaa voidaan määritellä esimerkiksi verbeillä *kävellä*, *lentää* tai *uida*. Näille perusmerkityksille on venäjässä kullekin iteratiivinen ja duratiivinen vastineensa, joilla molemmilla on lisäksi eri aspekteja edustavat versionsa. Venäjän synteesin astetta nostaa se, että monet toiminnan tapaan liittyvät merkitykset, joita suomessa ilmaistaan verbillä ja sitä määrittelevillä adjektiiveilla tai adverbeillä, on venäjässä ilmaistu affiksein.

Tämä venäjän kielen piirre nostaa nimenomaan sen synteesin johtamisastetta. Prefikseillä muodostetaan uusia sanoja, ei saman verbin eri muotoja. Myös suffiksein ja prefiksein tai kokonaan eri kannoista muodostetut verbien aspektiparit sekä iteratiiviset ja duratiiviset parit ovat erillisiä sanoja, joilla on oma taivutuksensa. Tiedonhaun näkökulmasta verbien erilaisten johdosten merkitykset ovat kuitenkin useissa tapauksissa lähes samat. Kuviteltu esimerkki monikielisestä tiedonhausta voisi olla sellainen, että venäjäksi käännettävässä suomenkielisessä haussa olisi yhtenä avaimena sana *lentää*. Jos käännökseen haluttaisiin kaikki merkitystä *lentää* ilmaisevat venäjän kielen sanat, siihen olisi sisällytettävä iteratiivinen ja duratiivinen liikeverbipari, aspektiparit sekä tavallisimmat etuliittein ilmaistut liikkumisen suunnan merkityksiä sisältävät sanat.

Snowball:in karsinnan tuloksena eri aspektien verbeille sekä iteratiivisilla ja duratiivisille liikeverbeille on erilaiset karsintavartalot eli stemit. Verbien persoona-, suku- ja lukupäätteitä sekä imperfektin tunnuksia ohjelma karsii muutamin varauksin. Gerundin ja partisiipin tunnuksista se karsii vain osan. Rajoituksiin on ilmeisesti tarkkuuden vaalimiseen liittyviä morfologisia syitä. Käytännössä esimerkiksi lentämiseen liittyviä dokumentteja haettaisiin todennäköisemmin toimintaa kuvaavilla substantiiveilla. Kuten sanottu, verbien on todettu olevan tiedonhaun kannalta vähämerkityksisiä. Dolamicin stemmaustavalla vältetään verbien käsittelyn edellyttämät monimutkaisemmat säännöt ja prosessointiaika luultavasti lyhenee. Toisaalta Snowball-menetelmällä tietokannan indeksi pienenee, kun erimuotoisten verbien määrä vähenee. Verbien karsinnasta voi olla haittaa, jos dokumenttien informaatiosisältöä huonosti kuvaavien verbien karsitut vartalot täsmäävät hauissa esiintyviin nomineihin.

## 4.5 Sijataivutus

Nominit määritellään sanaluokiksi, joilla on sijataivutus. Sija ilmaisee nominin suhteita lauseessa, sen asemaa subjektina, objektina, muita sanoja määrittelevänä attribuuttina ja niin edelleen. Kielen morfologista kompleksisuutta arvioidaan usein sen perusteella, montako sijamuotoa siinä on (Pirkola 2001). Taulukossa 3 on esitetty kahdeksan kielen sijamuotojen eli kieliopillisen sijakategorian morfosyntaktisten piirteiden määrä.

**Taulukko 3** Sijamuotojen määrä kahdeksassa kielessä (Pirkola 2001, 340)

| kieli | sijamuotojen määrä |
|---|---|
| englanti | 2 |
| saksa | 4 |
| venäjä | 6 |
| liettua | 7 |
| serbia | 7 |
| sanskrit | 8 |
| suomi | 14 |
| unkari | 21 |

Esimerkiksi suomen kielessä 14 sijamuodon erilaisten päätteiden suuri määrä yksikössä ja monikossa sekä lukuisat liitepartikkelit nostavat kielen synteesin astetta ja tekevät siitä morfologisesti hyvin kompleksisen (Pirkola 2001, 340). Venäjässä sijakategoria on kuusijakoinen, lukukategoria useiden kielten tavoin kaksijakoinen, ja monikon sijapäätteet ovat erilaiset kuin yksikön. Venäjän kielen substantiivilla ei erilaisia sijapäätteitä silti ole 12 vaan enintään 10, sillä muodoissa on päällekkäisyyttä. Tämän toteavat myös Kettunen ja kumppanit (2007). Adjektiiveilla mahdollisia sijamuotoja on 11, sillä ne voivat saada joko maskuliini-, feminiini tai neutritaivutuksen sen mukaan, mitä näistä kolmesta kieliopillisesta suvusta adjektiivin määrittelemä substantiivi edustaa.

Sijamuotojen määrällä mitattuna venäjän morfologinen kompleksisuus on keskitasoa. Taulukon 3 kielistä venäjää vähemmän kompleksisia ovat vain englanti ja saksa, mutta indoeurooppalaisista kielistä muidenkin germaanisten kielten morfologia on venäjän morfologiaa yksinkertaisempi, samoin romaanisten kielten, kuten espanjan, italian ja ranskan.

### 4.5.1 Substantiivit ja sijataivutus

Nominien vartaloiden ja sijapäätteiden välisten suhteiden kuvaamisessa on tavallista puhua deklinaatioista eli taivutusluokista. Venäjän substantiivit jaetaan yleensä yksikön nominatiivin päätteen mukaan kolmeen deklinaatioon (mm. Andrews 2001; Kratkaja russkaja grammatika 1989; Rahmanova & Suzdal'ceva 2003; Šeljakin 2006), joiden avulla sekä päätteiden valitseminen että tunnistaminen helpottuvat. Vartalo, jolla on tietty deklinaatiopiirre, yhdistyy tietynlaiseen päätteeseen. Ensimmäiseen deklinaatioon (Taulukko 4) kuuluvat konsonanttiin – tai oikeammin nollafleksioon – päättyvät maskuliinit sekä -o, -e, ja -ë -loppuiset neutrit. Toiseen deklinaatioon (Taulukko 5) kuuluvat -a ja -я -loppuiset maskuliinit ja feminiinit. Kolmas deklinaatio (Taulukko 6) kattaa nollapäätteiset eli pehmeämerkkiin päättyvät feminiinit ja maskuliininen sana путь (put').

Venäjän kuusi sijamuotoa ovat nominatiivi, genetiivi, datiivi, akkusatiivi, instrumentaali ja prepositionaali. Dolamic ja Savoy (2009) huomauttavat, ettei jokainen sija vaadi omanlaistaan päätettä, ja ottavat esimerkiksi ensimmäisen deklinaation akkusatiivin, joka on joko nominatiivin tai genetiivin kaltainen. Toisessa deklinaatiossa datiivi ja prepositionaali ovat samankaltaisia. Kolmannen deklinaation sijataivutuksessa on vielä enemmän päällekkäisyyttä. Taulukoissa 4, 5 ja 6 näkyvät deklinaatioittain substantiivien kussakin sijassa saamat päätteet yksikössä ja monikossa.

Taulukoiden selkeinä pitämiseksi päätteitä ei ole niissä translitteroitu. Päätteisiin vaikuttaa substantiivien jakautuminen kova- ja pehmeävartaloisiin. Esimerkiksi ensimmäisen deklinaation genetiivin pääte on kovavartaloisilla substantiiveilla -a, pehmeävartaloisilla -я. Kyse on fuusiosta, koska päätteet ovat saman morfeemin – päätteen -a – eri allomorfeja eli ilmenemismuotoja, eivätkä eri päätteitä (esim. Nikunlassi 2002). Vartalo saa siis aikaan muutoksia affiksissa. Morfologisesta vaihtelusta on kyse myös päätteissä -ы/-и (-y/-i), -ов/-ев (-ov/-ev) -ами/-ями (-ami/-jami) -ом/-ем (-om/-em) -ах/-ях (-ah/jah) ja ой/-ей (-oj/ej). Kova- ja pehmeävartaloisuuden lisäksi päätteiden äännevaihtelua aiheuttavat eräiden suhuäänteiden yhteydessä esiintyviä vokaaleja rajoittavat säännöt. Dolamic ja Savoy (2009) eivät mainitse päätteissä tapahtuvaa äännevaihtelua. Ilmiö nostaa venäjän fuusion taivutusastetta (*inflectional index of fusion*), mutta sen vaikutus stemmaukseen

rajoittuu siihen, että karsittavien suffiksien lista ja sen myötä prosessointiaika pitenee. Tiedonhaun tulosta päätteiden vaihtelu ei ainakaan stemmauksessa heikennä.

**Taulukko 4** Ensimmäisen deklinaation päätteet

| Sija | Yksikkö | Monikko |
|---|---|---|
| Nominatiivi | -ø, -о/-е | -ы, -и , -а, -я |
| Genetiivi | -а, -я | -ов, -ев, -ей, -ø |
| Datiivi | -у, -ю | -ам, -ям |
| Akkusatiivi | -а, -я (elolliset maskuliinit) | -ов, -ев, -ей (elolliset) |
| | -ø, -о/-е (elottomat) | -ы, -и , -а, -я (elottomat) |
| Instrumentaali | -о(-ё)м, -ем | -ами, -ями |
| Prepositionaali | -е, -и | -ах, -ях |

**Taulukko 5** Toisen deklinaation päätteet

| Sija | Yksikkö | Monikko |
|---|---|---|
| Nominatiivi | -а, -я | -ы, -и |
| Genetiivi | -ы, -и | -ø, -ей |
| Datiivi | -е, -и | -ам, -ям |
| Akkusatiivi | -у, -ю | -ø, -ей (elolliset)<br>-ы, -и (elottomat) |
| Instrumentaali | -ой, -ей | -ами, -ями |
| Prepositionaali | -е, -(и)и | -ах, -ях |

**Taulukko 6** Kolmannen deklinaation päätteet

| Sija | Yksikkö | Monikko |
|---|---|---|
| Nominatiivi | -ø | -и |
| Genetiivi | -и | -ей |
| Datiivi | -и | -ам, -ям |
| Akkusatiivi | -ø | -ей (elolliset)<br>-и (elottomat) |
| Instrumentaali | -ью | -ами, -ями |
| Prepositionaali | -и | -ах, -ях |

Kuten Dolamic ja Savoy (2009, 2542) panevat merkille, kaikissa tapauksissa sijamuodot eivät ilmene suffikseina, vaan päinvastoin yksikön nominatiivissa esiintyvä pääte voi pudota pois. He antavat esimerkin kirjaa merkitsevästä sanasta, jonka yksikön nominatiivi on книга (kniga), mutta monikon genetiivin книг (knig). Tätä kutsutaan venäjän kielioppia koskevassa kirjallisuudessa (esim. Nikunlassi 2002; Rahmanova & Suzdal'ceva 2003) nollapäätteeksi. Yksikön nominatiivissa eli perusmuodossa

konsonanttiin sekä pehmeämerkkiin päättyvät substantiivit tulkitaan myös nollapäätteisiksi, sillä näiden sanojen viimeinen konsonantti kuuluu sanan kantaan, ja pehmeämerkin katsotaan ainoastaan ilmaisevan oikeinkirjoituksessa sanan viimeisen konsonantin liudentuneisuutta. Substantiivien deklinaatioita kuvaavissa taulukoissa (Taulukot 4, 5 ja 6) nollapäätettä osoittaa merkki ø.

Monikossa monet yksikölle tyypilliset deklinaatioiden väliset erot häviävät. Kaikissa deklinaatioissa kaikki monikon substantiivit muutamia harvoja poikkeuksia lukuun ottamatta saavat datiivissa päätteen -ам/-ям (-am/-jam), instrumentaalissa -ами/-ями (-ami/-jami) ja prepositionaalissa -ах/-ях (-ah/-jah). Myös eri deklinaatioiden genetiivin ja nominatiivin päätteet ovat monessa tapauksessa samanlaisia.

Substantiivien deklinaatioiden ulkopuolelle jäävät adjektiivien tavoin taipuvat sekä taipumattomat substantiivit, kuten пальто (pal'to – päällystakki), шимпанзе (šimpanze – simpanssi), jotka ovat samanmuotoisia kaikissa sijoissa ja kummassakin luvussa. Taipumattomien sanojen sijamerkitykset ilmaistaan prepositioilla tai ne käyvät ilmi esimerkiksi niiden verbien semantiikasta, joihin substantiivit ovat sisällöllisessä suhteessa. Singularia tantum, eli kirjaimellisesti ne substantiivit, joilla on vain yksikkö, esim. еда (eda - ruoka), гордость (gordost' – ylpeys), горение (gorenie – palaminen), voivat potentiaalisesti saada myös monikkomuodon. Vaikka se on harvinaista, muodostaminen on mahdollista ja tulos ymmärrettävä.

Erityistapauksen muodostavat ns. pluralia tantum, ne, joilla kirjaimellisesti on ainoastaan monikko, esim. сутки (sutki – vuorokausi), часы (časy – kello), ножницы (nožnicy – sakset). Pluralia tantum -substantiivit eivät kuulu mihinkään sukuun. Monikossa sukukategoria ei muutoinkaan vaikuta substantiivia määrittelevien sanojen taivutukseen (Rahmanova & Suzdal'ceva 2003, 284.) Substantivoituneet adjektiivit ja erisnimistä -ов/-ев (-ov/-ev) ja -ин/-ы (-in/y) -loppuiset sukunimet noudattavat adjektiivien sijataivutusta (Taulukko 9). Sekä substantiivien että adjektiivien taivutustyyppien ulkopuolelle jäävät taipumattomat sekä -о tai -ых/-их (-yh/-ih) -loppuiset sukunimet. Erisnimiin liittyvät erityispiirteet luultavasti heikentävät venäjänkielisen tiedonhaun tehokkuutta esimerkiksi henkilöihin ja paikkoihin liittyviin faktoihin kohdistuvassa web-hauissa.

Karlssonin (2006) mukaan suomen kielessä paikannimet esiintyvät usein paikallissijoissa. Venäjän kielessä suomen paikallissijoja vastaavat merkitykset

ilmaistaan prepositioiden ja sijapäätteiden avulla, jolloin esimerkiksi merkitys 'Moskovasta' ilmaistaan prepositiolla из (iz) ja kaupungin nimen genetiivimuodolla Москвы (Moskvy). Prepositioiden käyttö vaikuttaa osaltaan siihen, että venäjässä tarvitaan vähemmän sijamuotoja kuin suomessa. Merkitys 'Moskovassa' ilmaistaan prepositiolla в (v) ja prepositionaalimuodolla Москве (Moskve). Prepositionaali on Kettusen ja kumppaneiden (2007) mukaan venäjän substantiivien kuudesta sijamuodosta yksikössä vasta viidenneksi yleisin, joten sitä ei tuoteta yhdessäkään tässä tutkielmassa kokeiltavista yleisimpiä sijamuotoja tuottavan menetelmän versioista. Monikossa venäjän substantiivien yleisin sijamuoto on genetiivi (Kettunen et al. 2007). Paikannimet harvoin esiintyvät monikon genetiivissä. Tällaiset seikat saattavat vaikuttaa FCG-menetelmän tuloksellisuuteen yksittäisissä hauissa. Vaikka erisnimet eivät ole tilastollisesti tekstissä esiintyvien sanamuotojen kärkipäässä, on luultavaa, että webin tavallisissa hakulausekkeissa ne esiintyvät usein.

Venäjän kieliopissa erisnimien sijataivutussäännöt ovat jokseenkin moninaiset. Erityyppisiä etu- ja sukunimiä koskevien lainalaisuuksien merkitystä lisää se, että etu-, suku ja isännimet taipuvat kaikki. Näin tapahtuu samassakin lauseyhteydessä, toisin kuin suomessa. Sekä titteli että etu- ja sukunimi taipuvat samassa sijassa: с президент*ом* Дмитри*ем* Медведев*ым* (s prezident*om* Dmitri*em* Medvedev*ym*) = president*in* Dmitri*n* Medvedev*in* kanssa.

Venäläisillä etu- ja sukunimillä, slaavilaisten maiden asukkaiden sekä entisen Neuvostoliiton alueella asuvien kansojen etu- ja sukunimillä on sijataivutus. Tästä yleisestä säännöstä on joukko poikkeuksia. Nykykirjakielen normin mukaan esimerkiksi seuraavia morfologisia tyyppejä edustavat erisnimet eivät taivu: sukunimet, jotka päättyvät kirjaimiin -o, -e, -и, -y, -ых/-их (-o, -e, -i, -u, -yh/-ih) konsonanttiin päättyvät naisten sukunimet, kaksiosaisten sukunimien ensimmäiset osat, mikäli ne eivät ole perinteisiä venäläisiä sukunimiä. Yleisnimien kaltaisista sukunimistä, kuten Жук (Žuk - kovakuoriainen), ei ole sääntöä, vaan niitä milloin taivutetaan, milloin ei. (Ramanova & Suzdal'ceva 2003, 314–315.)

Tyypillisten itäslaavilaisten -ок, -ек, ец, -яц, -ень, ель (-ok, -ek, -ec, -jac, -en', -el') -loppuisten miesten sukunimien taivutuksessa tapahtuu vokaalin väistymistä vaihtelevasti, kun nimet ovat homonyymisiä yleisnimien kanssa, tai niillä on samanlaiset päätteet kuin yleisnimillä, joissa on väistyvä vokaali. Teksteissä

tämäntyyppiset sukunimet saattavat siis esiintyä eri tavoin taivutettuina. Esimerkiksi jänistä tarkoittavan sanan заяц (zajac) tavoin taivutettuna genetiivimuoto Зайца (Zajca) voi olla muodostettu joko sukunimestä Заяц (Zajac), Заец (Zaec) tai Зайц (Zajc). Kun sukunimessä esiintyvä vokaali tulkitaan väistyväksi, on nimen perusmuodon päättely sijamuodosta siis mahdotonta. Teksteissä tulkinnasta esiintyy variaatiota, eikä asiasta ole myöskään yksiselitteisiä suosituksia. Toinen esimerkki on genetiivi muoto Журавля (Žuravlja), jonka lähtökohtana voi olla joko sukunimi Журавль (Žuravl') tai Журавель (Žuravel'). Tällainen morfologisen vaihtelun variaatio voi heikentää tiedonhaun tulosta. Sukunimi Заяц (Zajac) esiintyy teksteissä genetiivissä sekä muodossa Зайца (Zajca) että Заяца (Zajaca), sukunimi Заец (Zaec) muodoissa Зайца (Zajca) ja Заеца (Zaeca), Журавель (Žuravel') muodoissa Журавля (Žuravlja) ja Жураравеля (Žuravelja). (Rahmanova & Suzdal'ceva 2003, 315.)

Sukunimien, jotka päättyvät -онок/-енок (-onok/-enok), viimeinen vokaali väistyy taivutettaessa aina. Tällaisissa nimissä ei siis esiinny kirjakielen normeihin mahtuvaa variaatiota, toisin kuin yllä mainituissa -ок, -ек (-ok, -ek) -loppuisissa sukunimissä. Puolalaiset, tšekkiläiset ja slovakialaiset -ек (-ek) ja -ел (-el) -loppuiset nimet taipuvat nykyvenäläisissä teksteissä ja käännöksissä useimmiten niin, että vartalo säilyttää kaikki vokaalinsa. Sen sijaan samantaustaisten -ец (-ec) -loppuisten sukunimien tulkitaan pääsääntöisesti sisältävän väistyvän vokaalin. (Emt.)

Huomattakoon, että venäjän kirjain й on puolivokaali, jota Suomessa tutkijoiden ja kirjastotyöntekijöiden käytössä olevan, Vahroksen ja Kahlan (1967) esittämän tarkan tavan mukaan translitteroidaan j:ksi. Kirjaimet я, е, ё ja ю ilmaisevat foneettisesti katsoen vokaaleja a, e, o ja y (suomen u:ta vastaava) sekä niitä edeltävän konsonantin liudentuneisuutta tai, edellisen kirjaimen ollessa vokaali, j:n ääntymistä niiden edellä. Näin ollen sanan заяц (zajac) genetiivimuodossa зайца (zajca) kirjaimen я (ja) vokaaliaines a on väistynyt, mutta j:tä on oikeinkirjoituksessa jäänyt edustamaan й. Kyse on siis väistyvästä vokaalista, joka aiheuttaa merkkijonotasolla muunkinlaista fuusiota. Useimmat tiedonhakujärjestelmät eivät tee eroa isojen ja pienten kirjainten välillä, joten sukunimi Жук (Žuk) ja kovakuoriaista tarkoittava sana жук (žuk) ovat saavat tiedonhaussa useimmiten samanlaisen kohtelun. Tällä on hakua laajentava ja mahdollisesti tarkkuutta huonontava vaikutus.

Tiedonhaun kannalta pluralia tantum ja singularia tantum sekä erisnimet voivat olla hankalia. Tässä tutkielmassa tehtävän tiedonhakukokeen hakuaiheissa on melko paljon sanoja, jotka tuskin koskaan esiintyvät monikossa, vaikka sellaisen muodostaminen on mahdollista. Esimerkiksi hakuaiheessa, jonka avaimet taivutusmuotoisissa hauissa ovat muodossa химия пищеварения (himija piščevarenija - ruuanlaiton kemia), пищеварение (piščevarenie -ruuanlaitto) -sanalle tuotettavat monikolliset sijamuodot tuskin parantaisivat haun tehokkuutta. Hakuaiheiden 99 substantiivista ja adjektiivista 20 oli erisnimiä tai niistä muodostettuja adjektiiveja. Erityisesti maantieteellisiä erisnimiä käytetään harvoin monikossa. Monikkomuotojen tuottaminen saattaa siis turhaan kuormittaa prosessia. Se ei kuitenkaan heikennä hakutulosta.

Elollisuus tai elottomuus on substantiivin ominaisuus, joka vaikuttaa taivutuskategoriana adjektiiveihin, lukusanoihin ja partisiippiverbeihin. Kategorialla on merkitystä kuitenkin vain akkusatiivissa, jossa määrittelevien sanojen sijapääte määräytyy substantiivin elollisuuden mukaan. Yksikön akkusatiivissa kategoria liittyy vain osaan substantiiveista, monikossa kaikkiin. Paitsi elollisilla ensimmäisen deklinaation substantiiveilla yksikössä, myös kaikilla elollisiksi luokitelluilla substantiiveilla monikossa akkusatiivin pääte on genetiivin kaltainen. Toisen deklinaation substantiiveilla on yksikön akkusatiiville oma -y/-ю (-u/-ju) -pääte. Ensimmäisen deklinaation elottomien sekä kolmannen deklinaation kaikkien substantiivien yksikön akkusatiivipääte lankeaa yhteen nominatiivin päätteen kanssa. Erilaisia sanamuotoja ei muun muassa juuri akkusatiivin luonteeseen liittyen ole venäjässä niin paljon kuin sijamuotojen määrästä voisi päätellä.

Kettunen ja kumppanit (2007) ovat selvittäneet venäjän kielen substantiivien ja adjektiivien sijamuotojen esiintymistaajuutta Venäjän kansalliskorpuksesta lasketuilla tekstitilastoilla. Taulukoissa 7 ja 8 esitetyt tilastot perustuvat 5 miljoonan sanan osakorpukseen (Kettunen et al. 2007, 426). Yleisimmät sijamuodot sekä substantiiveilla että adjektiiveilla ovat nominatiivi, genetiivi ja akkusatiivi. Kimmo Kettusen (Kettunen & Airio 2006; Kettunen et al. 2007) kehittämälle yleisimpiä sijamuotoja tuottavalle menetelmälle adjektiivin akkusatiivimuodon riippuvuus substantiivin elollisuudesta ei siis aiheuta ongelmia venäjänkielisten kyselyjen muotoilussa, sillä sekä akkusatiivi että sen kanssa vaihtelevasti identtiset nominatiivi ja genetiivi kuuluvat tuotettaviin sijoihin. Ellei näin olisi, adjektiivien sijamuotojen generoimisessa ilman kontekstiaan saattaisi sattua virheitä.

**Taulukko 7** Venäjän substantiivien sijamuotojen frekvenssit (Kettunen et al. 2007)

| Sija | Yksikkö | % | Monikko | % |
|---|---|---|---|---|
| Nominatiivi | 327,637 | 32.7 | 76,500 | 25.6 |
| Genetiivi | 236,917 | 23.7 | 97,737 | 32.7 |
| Datiivi | 53,021 | 5.3 | 14,812 | 4.9 |
| Akkusatiivi | 195,340 | 19.5 | 56,929 | 19.0 |
| Instrumentaali | 98,789 | 9.9 | 24,132 | 8.1 |
| Prepositionaali | 89,253 | 8.9 | 29,132 | 9.7 |
| Yhteensä | 1,000,957 | 100 | 299,246 | 100 |

**Taulukko 8** Venäjän adjektiivien sijamuotojen frekvenssit (Kettunen et al. 2007)

| Sija | Yksikkö | % | Monikko | % |
|---|---|---|---|---|
| Nominatiivi | 76,059 | 31.8 | 26,371 | 26.8 |
| Genetiivi | 53,482 | 22.4 | 29,989 | 40.5 |
| Datiivi | 9,051 | 3.8 | 3,983 | 4.1 |
| Akkusatiivi | 44,079 | 18.5 | 17,843 | 18.1 |
| Instrumentaali | 24,360 | 10.2 | 7,890 | 8.0 |
| Prepositionaali | 31,799 | 13.3 | 12,347 | 12.5 |
| Yhteensä | 238,830 | 100 | 98,423 | 100 |

### 4.5.2 Sukukategoria ja adjektiivien sijataivutus

Substantiivien suku määräytyy ensinnäkin niiden tarkoitteiden niin sanotun luonnollisen suvun eli sukupuolen mukaan. Suurimmassa osassa tapauksista, kuten esineitä tarkoitettaessa, sukupuoleen ei tietenkään voi turvautua kieliopillista sukua pääteltäessä. Suku liittyykin pääsääntöisesti siihen, minkä päätteen sanat saavat yksikön nominatiivissa. Kaikki konsonanttiin eli nollapäätteeseen ja й(j)-kirjaimeen päättyvät substantiivit ovat maskuliineja, -о, -е/-ё ja -мя (-mja) -päätteiset puolestaan neutreja. Suurin osa -а ja -я (-ja) -päätteisistä substantiiveista on feminiinejä ja loput luonnollisen suvun perusteella maskuliineja. Osa substantiiveista päättyy pehmeämerkkiin eli ь-merkkiin, jota translitteroinnissa tarkoitetaan '-merkillä. Myös pehmeämerkkiin päättyvät substantiivit ovat kieliopin näkökulmasta nollapäätteisiä. Näiden sanojen sukua, joka on joko feminiini tai maskuliini, ei voi päätellä niiden merkityksestä tai muodosta. Vierasperäisten, taipumattomien substantiivien suku puolestaan määräytyy vaihtelevin perustein.

Sukukategoria vaikuttaa adjektiivien sijapäätteisiin siten, että kaikille kolmelle suvulle on oma taivutuksensa. Adjektiivin päätteet siis määräytyvät sen substantiivin suvun perusteella, jonka attribuuttina adjektiivi on. Taulukossa 9 (Zaliznjak 1977, 27) ovat adjektiivien positiivien pitkien muotojen sijapäätteet eri suvuissa. Kuten substantiivit, adjektiivit jakautuvat kova- ja pehmeävartaloisiin, mikä vaikuttaa päätteiden muotoon. Vinoviivojen (/) oikealla puolella taulukossa 9 olevat päätteet ovat pehmeiden vartaloiden päätteitä. Kovavartaloiset maskuliinit jakautuvat edelleen niihin, joissa paino on sanan vartalolla, ja niihin, joissa se on päätteellä. Taulukossa akkusatiivin kohdalla näkyy substantiivin elollisuusominaisuuden vaikutus sitä määrittelevään adjektiiviin.

**Taulukko 9** Adjektiivien päätteet, pitkät muodot

| | Yksikkö | | | Monikko |
|---|---|---|---|---|
| Sija | Maskuliini | Neutri | Feminiini | Kaikki suvut |
| Nominatiivi | -ый/ -ой -ий | -ое/-ее | -ая/-яя | -ые/-ие |
| Genetiivi | -ого/-его | -ого/-его | -ой/-ей | -ых/их |
| Datiivi | -ому/-ему | -ому/-ему | -ой/-ей | -ым/-им |
| Akkusatiivi eloton | -ый/-ой -ий | | | -ые/-ие |
| Akkusatiivi elollinen | -ого/-его | -ое/-ее | -ую/-юю | -ых/их |
| Instrumentaali | -ым/-им | -ым/-им | -ой/-ей, | -ыми/-ими |
| Prepositionaali | -ом/-ем | -ом/-ем | -ой/-ей | -ых/-их |

Kettusen ja kumppaneiden (2007) mukaan adjektiivit eivät ole tiedonhaussa erityisen tärkeitä. Venäjänkielisessä haussa niiden merkitys saattaa kuitenkin korostua. ROMIP-aineistossa on yhtenä hakuaiheena esimerkiksi pedagogisen psykologian kurssityötä koskeva, hakukoneen lokista poimittu hakulauseke ”курсовая по педагогической психологии” (kursovaja po pedagogičeskoj psihologii), jossa kurssityötä ilmaisevaa fraasia – курсовая работа – edustaa vain adjektiivi курсовая ja substantiivi работа on jätetty kokonaan pois. Etenkin puhekielessä tämä on tavallista. Samanlainen adjektiivin tärkeä asema näkyy ROMIPin Leningradin Jelisejevin tavaratalon johtajaa koskevassa hakuaiheessa ”директор елисеевского ленинград” (direktor eliseevskogo leningrad), jossa Елисеевский магазин (Eliseevskij magazin) on typistynyt pelkkään adjektiivimuotoisen erisnimen genetiiviin елисеевского (eliseevskogo). Venäjässä adjektiivi on varsinkin erisnimien kohdalla usein tehtävässä, jossa suomessa käytettäisiin genetiivimuotoista substantiivia. Esimerkiksi Venäjän federaatio on venäjäksi Российская Федерация (Rossijskaja Federacija) – ”venäläinen federaatio”.

Jos FCG-menetelmää käytettäisiin tiedonhakujärjestelmässä käytännön sovelluksena, tulisi käyttäjien syöttää hakuavaimet perusmuodossa eli yksikön nominatiivissa. Luultavasti he syöttäisivät adjektiivit sen suvun mukaisessa yksikön nominatiivissa, jota hakulausekkeen substantiivi edustaisi. Joissakin tapauksissa substantiivin suvun määrittely voi kuitenkin tuottaa jopa syntyperäiselle puhujalle vaikeuksia. Esimerkiksi kahvia tarkoittava sana on vierasperäisten, taipumattomien, ruokia ja juomia tarkoittavien sanojen tapaan maskuliini. Ehkä siksi, että sanalla on neutreille tyypillinen -e-pääte, ja se ei ääntämyksensä puolesta kuulosta vierasperäiseltä, käytetään sanaa кофе (kofe – kahvi) usein virallisluonteisia tekstejä lukuun ottamatta neutrina (Rahmanova & Suzdal'ceva 2003). Teksteissä sanaa määrittelevät adjektiivit ovat eri sukujen mukaisissa sijamuodoissa, mikä voi heikentää yhden suvun mukaisilla sanamuodoilla tehtävän haun tulosta. Tällaiset tapaukset ovat kuitenkin luultavasti marginaalisia, eivätkä ne vaikuta tiedonhaun tulokseen kokonaisuutena. Tämän tutkielman empiirisessä osassa käytetyissä hakuaiheissa ei ollut yhtään substantiivia, jonka suvusta olisi laajasti eri tulkintoja.

Tiedonhaun tehokkuuteen voi vaikuttaa merkittävämmin se, että vaikka kaikille adjektiiveille on kolme eri sukujen mukaista yksikön nominatiivia, on niistä vain yksi niin sanottu sanakirjamuoto. Niinpä lemmatoija perusmuotoistaa kaikki adjektiivit maskuliinimuotoon. Kuten sanottu, venäjän kielessä suomen yhdyssanojen määrittelyosaa vastaavassa roolissa on usein adjektiivi, kuten luvussa 3.2 käytetyssä informaatioteknologiaa tarkoittavassa esimerkissä. Adjektiivin suku on tärkeä, sillä relevanteissa dokumenteissa ne voivat esiintyä samanlaisessa asemassa kuin hakuaiheessa määrittelemässä esimerkiksi feminiinistä teknologiaa tarkoittavaa sanaa. Jos hakuaiheessa feminiininä tai neutrina esiintyvä adjektiivi perusmuotoistetaan maskuliiniksi, kohtaavat sille tuotetut sijamuodot taivutusmuotoindeksissä vain maskuliinisia adjektiivien sijamuotoja, ja relevantteja dokumentteja voi jäädä löytymättä. Esimerkiksi maskuliinipäätteinen, perusmuotoa edustava adjektiivi информационный (informacionnyj – informaatio-) ja feminiinisukuinen substantiivi технология (tehnologija – teknologia) eivät luultavasti esiinny samassa dokumentissa ainakaan samassa yhteydessä. Tähän palataan tutkielman empiirisessä osuudessa, luvussa 5.4.

RUSTWOL-lemmatoija perusmuotoistaa yksikön nominatiivin maskuliinimuotoon myös lyhyet adjektiivit. Lähes kaikilla venäjän adjektiiveilla on lyhyt ja pitkä muoto.

Pitkiin adjektiiveihin vaikuttavat sija, suku, luku ja vertailuasteet, mutta lyhyisiin vain suku ja luku. Lyhyillä adjektiiveilla ei siis ole sijataivutusta. Käytettäessä indeksointiin ja hakuun lemmatointia, löytyvät sekä pitkät että lyhyet adjektiivit. Samoin stemmauksessa, mikäli stemmatut sanavartalot ovat tarpeeksi lyhyitä. Sen sijaan haettaessa taivutusmuotoindeksistä pitkiksi muodoiksi perusmuotoistetuilla ja niistä edelleen sijamuodoiksi generoiduilla hakusanoilla dokumenttien lyhyet adjektiivit jäävät löytymättä.

Adjektiivien vertailuasteista yksinkertaisella komparatiivilla ei ole sijataivutusta. Toinen komparatiivityyppi on kahdesta sanasta muodostuva liittomuoto, johon sanamuotoihin kohdistuvat menetelmät eivät vaikuta. Yksinkertaiset komparatiivien -ee-päätteet osuvat yksiin positiivien neutrin nominatiivipäätteen (Taulukko 9) kanssa. Ne ovat sekä Snowball-stemmerin että Dolamicin kevyen stemmerin karsittavien adjektiivin päätteiden listalla. FCG löytäisi taivutusmuotoindeksistä komparatiivit vain neutreille yleisiä sijamuotoja generoitaessa. Toinen komparatiivin pääte on -e, joka myös sisältyy stemmereiden karsittavien päätteiden listoihin. Komparatiivissa sen edellä tapahtuu kuitenkin konsonanttivaihtelua. Komparatiivi muodostetaan lisäksi monella poikkeuksellisella tavalla sekä muutamissa tapauksissa suppletiivisesti eri vartalosta kuin positiivi. Suomen kielessä tällainen tapaus on hyvä – parempi. Dolamic ja Savoy (2009) eivät kiinnitä huomiota komparatiivin muodostamisessa ilmenevään fuusioon, ja Dolamicin stemmeri karsiikin vain sijapäätteitä. Komparatiivit eivät luultavasti vaikutakaan suuremmin tiedonhaun tulokseen.

Myös superlatiivi voidaan venäjässä muodostaa kahdesta sanasta. Yksinkertaisen, eli yhdellä sanamuodolla ilmaistavan superlatiivin muodostamiseen käytetään kolmenlaisia tunnuksia. Superlatiiveilla on sijataivutus, ja tunnukset ovat sijapäätteen edelle asettuvia suffikseja. Näistä Snowball listaa karsittavaksi yhden suffiksin kaksi eri ilmenemismuotoa -ейш- (-ejš-) ja -ейше- (-ejše-), mutta vain tietyissä esiintymisympäristöissä. Tunnuksen -айш- (-ajš-) edellä vanavartalossa tapahtuu konsonanttivaihtelua. Kolmas superlatiivin tunnus on -ш- (-š-). Kun superlatiiveista karsitaan sija- ja suku- ja lukupäätteet, täsmää vartalo vain superlatiiveihin, eikä kaikkiin saman adjektiivin muotoihin. Kuten sanottu, lemmatoinnissa superlatiivit pelkistyvät positiiviin, maskuliinin nominatiiviin. FCG ei tietenkään tuota superlatiivin tunnuksia, koska sekin keskittyy sijamuotoihin. Adjektiiveilla sija-, suku- ja

lukumerkitykset siis sisältyvät samaan morfeemiin, eivätkä ole erotettavissa toisistaan. Substantiiveilla sijapäätteet sisältävät luvun merkitykset.

## 4.6 Venäjän taivutusmorfologian synteettisyydestä ja fuusioivuudesta

Taulukossa 2 (Luku 3.1) näkyvää, verrattain korkealta vaikuttavaa venäjän kielen synteesin astetta nostavat taivutusaffiksien lisäksi johdosaffiksit, joita erityisesti verbeissä on paljon. Myös puhtaasti taivutusmorfologiaan ja siis kieliopillisiin kategorioihin liittyvien affiksien runsautta sekä verbeillä että nomineilla lisää jonkin verran sukukategoria. Nomineilla se vaikuttaa adjektiivien, pronominien ja joidenkin lukusanojen taivutuspäätteiden määrään. Substantiiveilla suku on kunkin sanan pysyvä ominaisuus, joten yhdellä substantiivilla on edellä mainitut enintään 10 mahdollista sijataivutuspäätettä. Tässä ovat mukana kuuden sijan yksikön ja monikon muodot, joissa vähintään kahdessa sijassa on päällekkäisyyttä. Synteesin taivutusaste (*inflectional index of synthesis*, IIS) (Pirkola 2001) kuvaa nimenomaan taivutusaffiksien määrää suhteessa kaikkien sanojen määrään tekstinäytteessä. Tämän tutkielman liitteessä 1 olevista, Tampereen Lenin-museon sivuilta otetuista sadan sanan paralleeleista tekstinäytteistä laskettuna venäjän kielen synteesin taivutusaste on 55/100 eli noin 0.5, suomen 59/100 eli 0.6. Laskutoimitus on tekijän kielitieteellisen asiantuntemattomuuden vuoksi suurpiirteinen, mutta silti suuntaa antavana valaiseva. Synteesin taivutusaste lasketaan jakamalla tekstinäytteen taivutusaffiksien määrä sen sanojen määrällä. Koska venäjän synteesin aste on lähellä suomen tasoa, voidaan olettaa, että samanlaiset taivutusmorfologian hallintatavat, joilla on saavutettu hyviä tuloksia suomenkielisessä tiedonhaussa, toimivat myös venäjän kohdalla.

Dolamicin ja Savoyn (2009) mukaan sanan vartalo ei venäjässä yleensä muutu pääteaffiksin vaikutuksesta, mutta he tuovat esiin väistyvän vokaalin slaavilaisille kielille melko tavallisena affiksin aiheuttamana vartalon äännevaihteluna. Esimerkiksi isää merkitsevässä sanassa yksikön nominatiivin отец (otec) vokaali väistyy genetiivimuodossa отца (otca). Venäjän vokaaleista е (e) ё (ë) ю (ju) ja я (ja) osoittavat niitä edeltävän konsonantin liudentuneisuutta. Suurimalla osalla venäjän konsonanteista on kaksi ääntämistapaa – liudentunut ja liudentumaton. Ääntäminen riippuu konsonantin asemasta sanassa. Tämä näkyy eräässä Dolamicin ja Sanvoyn (2009) antamassa väistyvää vokaalia koskevassa esimerkissä. Sanassa лёд (lëd - jää) vokaali ё

väistyy esimerkiksi datiivissa – льду (l'du) – mutta edellisen konsonantin liudentuneisuutta jää osoittamaan pehmeämerkki ь ('). Toista vokaalivaihtelutyyppiä edustaa Dolamicin ja Savoyn (2009) esimerkeissä sanan *sisko* сестра (sestra) monikon genetiivi on сестёр (sestër). Tässä yksikön nominatiivin päätteen katoamisen lisäksi on sanan lausumista helpottamaan ilmaantunut välivokaali ё. Venäjässä suffiksit saattavat aiheuttaa sanan vartalossa myös konsonanttien vaihtelua tai poisjäämistä. Konsonanttivaihtelu koskee kuitenkin lähinnä verbejä. Äänteellisesti läheiset tai samalla äänteellä alkavat suffiksit aiheuttavat vartalossa yleensä samanlaista vaihtelua, mikä helpottaa morfologista analyysia.

Dolamic ja Savoy (2009) pitävät taivutuksen aiheuttamaa sanojen vartaloissa tapahtuvaa äännevaihtelua venäjän kielessä niin vähäisenä, ettei se heikennä merkittävästi stemmauksen tehoa. Yhteenvetona morfologisesta vaihtelusta he toteavat, ettei se ole yhtä voimakasta kuin eräissä muissa kielissä, esimerkiksi suomessa (Dolamic & Savoy 2009, 2542). He arvioivat venäjän kielen morfologia lähinnä substantiivien ja adjektiivien osalta. Snowball-stemmerin venäjänkielisen version esittelyssä sanavartaloissa tapahtuvien muutosten todetaan liittyvän poikkeuksellisesti taipuviin sanoihin. Snowballin analyysi koskee myös verbien taivutusta.

Tiedonhaun kielitypologian näkökulmasta Dolamic ja Savoy (2009) näyttävät pitävän venäjän kielen fuusion taivutusastetta (*inflectional index of fusion*) matalana. Fuusion asteella tarkoitetaan morfeemien eroteltavuutta myös siinä mielessä, ovatko ne kieliopillisesti yksimerkityksisiä. Venäjässä useat johdos- ja taivutusmorfeemit ovat monimerkityksisiä. Esimerkiksi nominien sijapäätteet ilmaisevat sijan lisäksi useimmiten myös sukua ja lukua. Tästä hyvänä esimerkkinä ovat adjektiivien sijapäätteet, joista sukupäätteet eivät ole eroteltavissa. Kun adjektiiveille tuotetaan sijamuotoja, otetaan väistämättä kantaa myös niiden sukutaivutukseen. Vastaavasti adjektiivien lemmatointi niiden kolmesta erilaisesta yksikön nominatiivimuodosta yhteen – maskuliiniin – sulkee haun ulkopuolelle ne sijamuodot, joiden päätteet sisältävät feminiinin tai neutrin merkitykset. Nikunlassin (2002) mukaan suomi on perusluonteeltaan agglutinoiva kieli, jossa luku ja sija ilmaistaan omilla morfeillaan, kun taas venäjä on fuusioiva. Vaikka päätteet voivat merkitykseltään olla moniselitteisiä eli fuusioituneita, voi ne silti useimmissa tapauksissa erottaa yksiselitteisesti vartalosta.

## 4.7 Pohdintaa venäjän morfologiasta tiedonhaussa

Aihetta koskevassa tutkimuksessa (Loponen & Järvelin 2010) on käynyt ilmi, että verbit voi tiedonhaussa jättää morfologisen käsittelyn ulkopuolelle, ja että etuliitteisiin ja johdosaffikseihin kajoaminen venäjänkielisessä tiedonhaussa on turhaa tai vahingollista. Näin ollen huomio voidaan kohdistaa pelkästään substantiivien ja adjektiivien sijapäätteisiin, jotka kielen fuusioivuuden vuoksi sisältävät myös luku- ja sukukategorian taivutuspiirteet. Vaikka Dolamicin stemmeri määrittelee karsittaviksi vain nominien sijapäätteet, stemmauksessa tarkastelu kohdistuu kaikkiin indeksoitaviin sanoihin ja kaikkiin hakuavaimiin, mikäli ne eivät ole sulkusanoja. Stemmeri karsii myös verbien päätteitä niiltä osin, kun ne käyvät yksiin nominien sijapäätteiden kanssa. Samoin voi käydä taipumattomille adverbeille, jolloin kyse on väärinstemmauksesta. Kaikki adverbit eivät ole Dolamicin ja Savoyn (2009) sulkusanalistalla.

Snowball-stemmerissä mahdollisuuksia väärinstemmaukseen voi olla enemmän, koska myös karsittavia päätteitä ja suffikseja on huomattavasti enemmän. Toisaalta Snowballissa on väärinstemmauksen välttämiseksi paljon affiksien esiintymisympäristöä koskevia ehtoja. Dolamicin stemmerissä ainoa ehto on se, että tarkasteluun otettavan sanan on oltava vähintään neljän merkin pituinen. Esimerkiksi taipumattomien lainasanojen väärinstemmauksen uhka lienee käytännössä vähäinen. Lainasanojen loppuainekset ovat usein epätyypillisiä, erilaisia kuin karsittavat venäläiset päätteet, ja jäävät siten ilman muuta koskemattomiksi. Dolamic ja Savoy (2009) mainitsevat taipumattomat, pääosin vierasperäiset substantiivit venäjän morfologista käsittelyä mahdollisesti hankaloittavana piirteenä, mutta katsovat näiden sanojen määrän olevan vähäinen, eikä niitä ole huomioitu Dolamicin kevyessä stemmausalgoritmissa.

Ahneen ja varovaisen stemmauksen seurauksia on vaikea arvioida muuten kuin tiedonhakukokeella. Esimerkiksi eri sanaluokkien täsmäämisestä samaan vartaloon voi tapauksesta riippuen olla haittaa tai hyötyä. Dolamic ja Savoy (2009) vertasivat omia karsinta-algoritmejaan Snowball-stemmeriin, yhteen n-grammimenetelmään ja stemmaamatta jättämiseen (*baseline*). Kaikki hakumenetelmät, joissa käytettiin kielen morfologista prosessointia, olivat huomattavasti tehokkaampia kuin *baseline.* Dolamicin kevyt stemmeri pärjäsi parhaiten. Snowball karsii sijapäätteiden lisäksi myös verbien päätteitä ja suffikseja, superlatiivin tunnuksia ja yhden johdossuffiksin. Venäjän verbien

morfologia on mutkikkaampaa kuin nominien, ja Snowball-algoritmi onkin huomattavasti Dolamicin stemmeriä monimutkaisempi.

Dolamic ja Savoy (2009) osoittavat, että Snowball-stemmeriä kevyemmällä ja yksinkertaisemmalla menetelmällä voidaan päästä vähintään yhtä hyviin tuloksiin. Sijamuotojen rajoitettu tuottaminen olisi vielä kevyempää. Siinä morfologinen käsittely pystytään rajaamaan tarkasti tiettyihin sanaryhmiin. Yksinomaan hakuavaimiin kohdistuvana FCG sopisi erityisen hyvin hakuun web-kokoelmista. Niissä taivutusmuotoindeksi voi olla nopein ja edullisen ratkaisu.

Stemmaus tekee usein vääryyttä venäjän taivutusmorfologialle, kuten venäjän vokaalivaihtelulle. Dolamic ja Savoy (2009) sekä Martin Porter ja muut venäjän Snowball-algoritmin kirjoittajat eivät pidä kielen fuusion taivutusastetta erityisen korkeana. Sitä, millainen vaikutus stemmauksen suurpiirteisyydellä on hakutuloksen laatuun todellisuudessa, selviäisi sekin vain tiedonhakukokeessa. Olisikin mielenkiintoista testata Dolamicin stemmeriä tämän tutkielman empiirisessä osassa käytettävässä KM.ru-kokoelmassa ja verrata tuloksia FCG:n tuloksiin. Seuraava Dolamicin kevyellä stemmerillä stemmattu esimerkkisana on poimittu KM.ru-testikokoelman hakuaiheista:

| *sanamuoto* | | *stemmattu vartalo* |
|---|---|---|
| египта (egipta) | → | египт (egipt) |

Egyptiä tarkoittavan sanan perusmuoto eli yksikön nominatiivi on Египет (Egipet), jota karsinnan tulos египт (egipt) ei tavoita. Kyseisen erisnimen yksikön nominatiivin kannassa esiintyvä e-vokaali väistyy eri sijamuodoissa. Genetiivimuoto египта (egipta) ei ole perusmuodon eli yksikön nominatiivin египет (egipet) ja genetiivin sijanpäätteen -a summa, vaan taivutusmuoto on fuusioitunut. Automaattisesti sijamuotoja tuottava, webissä vapaasti käytettävissä oleva Morpher.ru-ohjelma tunnistaa venäjän kieleen vakiintuneet erisnimet, ja hallitsee niiden kuten yleisnimienkin fuusion. Sille pitää kuitenkin syöttää nominit perusmuodossa. Todellisessa käytössä FCG-ohjelmaan liitettäisiin luultavasti Morpher.ru-ohjelman tapainen automaattinen generoija. Hakusanat perusmuotoistasi joko hakija, tai perusmuotoistaminenkin voitaisiin automatisoida.

Kettunen ja kumppanit (2007) ounastelevat venäjän olevan ideaali kieli FCG-menetelmälle. Ajatusta tukevat havainnot (Baeza-Yates & Ribeiro-Neto 1999; Dolamic & Savoy 2009; Loponen & Järvelin 2010) siitä, että substantiivien ja adjektiivien sijataivutus on tiedonhaun näkökulmasta morfologiassa keskeisintä. Tiedonhaun kielitypologiassa venäjän synteesin taivutusaste on korkea, mikä kertoo taivutusmorfologian käsittelyn tarpeellisuudesta. Vaikka verbien taivutusmorfologia vaikuttaa synteesin taivutusasteeseen, on venäjä sijamuotojensa puolesta kompleksisuudeltaan keskitasoa. Fuusion taivutusaste näyttää myös olevan melko korkea. Synteettistä, fuusioivaa taivutusmorfologiaa on helpompi käsitellä generatiivisilla kuin reduktiivisilla menetelmillä. Algoritmisessa stemmauksessa sanavartaloissa tapahtuvan äännevaihtelun hallinta voi jää heikoksi. Vaikka venäjän kielen substantiiveilla ja adjektiiveilla taivutukseen liittyvät morfofonologiset muutokset ovat melko säännöllisiä ja suhteellisen vähäisiä, on kieliopillisten merkitysten fuusioituminen samaan affiksiin hyvin tavallista. Sen seurauksena esimerkiksi adjektiivien suku- ja sijapäätteitä ei voi erotella, millä saattaa olla vaikutusta myös FCG:n tehokkuuteen. FCG-menetelmällä taivutusmorfologialtaan synteettinen ja melko fuusioiva suomen kieli on hallittu hyvin (Kettunen & Airio 2006). Seuraavassa luvussa raportoidaan venäjänkielisen tiedonhaun FCG-kokeesta.

# 5 TIEDONHAKUKOE: FREQUENT CASE GENERATION –MENETELMÄ VENÄJÄNKIELISESSÄ TIEDONHAUSSA

Melko yksinkertaisenkin sääntöpohjaisen stemmerin on havaittu parantavan hakutuloksia morfologisesti kompleksisissa kielissä. Vielä kevyempi ja edullisempi morfologisen vaihtelun hallintatapa, generatiivinen Frequent Case Generation eli FCG (Kettunen & Airio 2006; Kettunen et al. 2007) näyttää johtavan yhtä hyviin tai jopa parempiin tuloksiin samoilla kielillä. Tässä luvussa raportoidaan kokeesta, jossa menetelmää käytetään suuressa venäjänkielisessä tekstikokoelmassa. FCG-menetelmää koskeneiden tutkimusten tarkoitus on ollut paitsi verrata lähestymistavan tehokkuutta muihin menetelmiin, myös selvittää, kuinka monen yleisimmän sijamuodon tuottaminen on kullakin kielellä haettaessa riittävää. Tämän kokeen tavoitteet ovat samat.

Tutkielman johdannossa esitetty, empiiristä osaa koskeva kysymys on seuraava:

> 2) Parantaako FCG venäjänkielisen tiedonhaun tuloksia verrattuna käsittelemättömillä avaimilla tehtyyn hakuun ja millaiset ovat FCG-versioiden väliset erot?

Tarkasteltaessa venäjän kielen morfologiaa luvussa 4 kävi ilmi, että adjektiivien sukutaivutus saattaa kaivata lisähuomiota. Tästä ajatuksesta on muotoutunut alakysymys:

> 2a) Vaikuttaako adjektiivin sukutaivutus hakutuloksiin venäjänkielisessä FCG- haussa?

Tässä luvussa käsitellään tiedonhakukokeita, joilla on pyritty etsimään vastauksia näihin kysymyksiin. Alaluvussa 5.1 esitellään kokeissa käytettävä testikokoelma, alaluvussa 5.2 kerrotaan FCG-kyselyiden muotoilusta ja eri hakumenetelmien tulosten evaluoinnista. Tutkimuskysymystä 2 koskevan kokeen keskeiset tulokset ovat alaluvussa 5.4. Kysymystä 2a selvittävästä lisäkokeesta kerrotaan alaluvussa 5.4. Kokeista vedetään johtopäätöksiä alaluvussa 5.5, jossa tuloksia myös verrataan Kettusen ja kumppaneiden (2007) tuloksiin.

## 5.1 Aineisto

Tässä tutkielmassa toistetaan osittain Kettusen ja kumppaneiden (2007) venäjänkielisen tiedonhaun FCG-koe. Testiaineisto on kuitenkin suurempi. Kettunen ja kumppanit (2007) käyttivät venäjänkielisenä aineistona CLEF 2004 testikokoelmia, jossa on 16 716 dokumenttia. Relevantteja dokumentteja kokoelmassa on sen 50 hakuaiheesta (*topic*) vain 34 hakuaiheelle, ja kaiken kaikkiaan kokoelmassa on 123 relevanttia dokumenttia. Tässä tutkielmassa käytettävä ROMIP-foorumin KM.ru-testikokoelma on kopio venäläisestä www.km.ru-yleisportaalista, ja se sisältää noin kolme miljoonaa dokumenttia. Relevantteja dokumentteja testikokoelmassa on löyhemmillä relevanssikriteereillä arvioituna 4855, tiukemmilla 1530. Testikokoelma on indeksoitu Tampereen yliopiston informaatiotieteiden yksikön tiedonhaun laboratorion UNIX-palvelimelle (kastanja.uta.fi) Lemur 4.12 -järjestelmällä, jolla myös haut tehtiin. Taivutusmuotoindeksissä on yli miljoona uniikkia indeksitermiä. Tietokannan indeksointiin on käytetty Neuchâtelin yliopiston monikielisen tiedonhaun resurssisivun englannin ja venäjän sulkusanalistojen yhdistelmää, johon on lisätty termit *km* ja *ru* sekä muutama venäjän kielen interjektio. Englanninkielisiä sulkusanoja on käytetty, koska testikokoelma sisältää myös englanninkielisiä dokumentteja. Testikokoelmassa ja hakulausekkeissa käytetty merkistökoodaus on UTF-8.

Venäläinen tiedonhaun vuotuinen evaluointiseminaari ROMIP aloitettiin vuonna 2003, sillä kansainvälisten foorumien, kuten CLEFin, panostusta venäjän kieleen pidettiin tarpeisiin nähden riittämättömänä. Toiseksi venäläisten tutkijoiden osallistuminen näihin foorumeihin oli vähäistä. ROMIP-seminaarin puitteissa on luotu useita venäjänkielisiä testikokoelmia, joista esimerkiksi BY.web ja KM.ru sisältävät miljoonia dokumentteja. Tämän kokeen haut on tehty käyttäen ROMIP 2009 web-tutkimuslinjan (*Дорожка поиска по Веб коллекции*) hakuaiheita ja saantikantoja. Tutkimuslinjan tarkoitus on web-tiedonhakuun soveltuvien menetelmien evaluointi. Siinä käytetyt kokoelmat ovat KM.ru ja BY.web. Sekä kokoelmien dokumentit että haut ovat aiheeltaan yleisiä, ja ne edustavat tyypillisiä www-dokumentteja ja -hakuja. Hakuaiheet eivät CLEF-testikokoelmien aiheiden tavoin koostu otsikosta, kuvauksesta ja narratiivista, vaan ovat yksinkertaisesti hakukoneiden lokeista poimittuja hakulausekkeita. KM.ru-testikokoelmalle on 2009 web-tutkimuslinjassa kaksi saantikantaa. AND-saantikannassa on käytetty tiukkoja relevanssikriteerejä, OR-

saantikannassa löyhempiä. OR-kannassa on relevantteja dokumentteja 47 hakuaiheelle, AND-kannassa 42 hakuaiheelle. Tässä kokeessa on käytetty vertailun vuoksi molempia kantoja, mutta niiden evaluointitulokset ovat hyvin samansuuntaisia. Saantikannat koostuvat hakuaiheen tunnistenumeroista (*taskid*), niiden dokumenttien tunnistenumeroista (*docid*), joista on tehty relevanssiarviot, sekä kunkin dokumentin hakuaihekohtasiesta relevanssiarviosta. Relevanssiarviot ovat binäärisiä, ja ne on saantikannoissa ilmoitettu relevance=”notrelevant” tai relevance=”vital” -merkinnöillä seuraavaan tapaan:

```
<task id="arw33607">

<document id="1353878" collectionId="KM.RU-2007" relevance="notrelevant"/>

<document id="466469" collectionId="KM.RU-2007" relevance="vital"/>
```

## 5.2 Menetelmä

Kettunen ja kumppanit (2007) muotoilivat hakuaiheista sekä pitkät että lyhyet FCG-kyselyversiot. Pitkiin kyselyihin hakuavaimet poimittiin hakuaiheiden otsikoista ja kuvauksista, lyhyisiin vain otsikoista. Tässä tutkielmassa FCG-kyselyjen pohjana olevat ROMIPin hakuaiheet ovat itsessään lyhyitä kyselyitä. Niistä osa on poimittu KM.ru-portaalin sekä Yandex-hakukoneen lokeista. Kyselyt ovat siis aitoja webin käyttäjien tekemiä hakuja. Eri sanaluokkien rooli tiedonhaussa konkretisoituu tarkasteltaessa näitä KM.ru-testikokoelman hakuaiheita. Yhdessäkään niistä ei esiinny verbejä, vaan hakulausekkeet koostuvat pääosin adjektiiveista ja substantiiveista. Niiden lisäksi lausekkeissa on partikkeleita, prepositiota sekä muutamia pronomineja ja adverbeja.

Tämän tutkielman FCG-kyselyt tehtiin samalla periaatteella kuin Kettusen ja kumppaneiden (2007) kokeessa. Heidän tilastonsa (Taulukot 7 ja 8, luku 4.5.1) venäjän kielen sijamuotojen jakaumista perustuvat Venäjän kansalliskorpuksen viiden miljoonan sanan osakorpukseen, johon morfologiset tiedot on merkattu käsin. Venäjän kielen yleisimmät sijamuodot ovat sekä substantiiveilla että adjektiiveilla yksikössä nominatiivi, genetiivi ja akkusatiivi. Substantiiveilla kolme yleisintä sijaa kattavat 75,7 % kaikista yksikön ja monikon substantiivimuodoista. Monikossa genetiivi on nominatiivia yleisempi. Monikkomuotoja esiintyy venäjänkielisessä tekstissä huomattavasti yksikkömuotoja vähemmän. Neljänneksi yleisin sijamuoto yksikössä on

substantiiveilla instrumentaali. Prepositionaali on instrumentaalia yleisempi monikossa ja adjektiiveilla myös yksikössä. Kettunen ja kumppanit (2007) katsovat, että neljäntenä sijamuotona kannattaa venäjänkielisessä FCG-haussa tuottaa instrumentaali. (Kettunen et al. 2007.)

Tässä kokeessa vertailtiin Kettusta ja kumppaneita (2007) mukaillen neljää menetelmää: hakua taivutusmuotoisilla, hakuaiheista suoraan poimituilla avaimilla sekä kolmea FCG-versiota. Toisin kuin Kettusella ja kumppaneilla (emt.), tässä tutkielmassa FCG-menetelmän verrokkina on ainoastaan haku ilman morfologista käsittelyä, sillä muiden menetelmien käyttö KM.ru-kannassa ei tämän työn puitteissa ollut mahdollista. Koe on rajattu koskemaan vain lyhyitä hakulausekkeita. Tiedonhaun laboratoriokokeissa käytetään yleensä myös pitkiä hakulausekkeita, mutta hyvin lyhyet lausekkeet vastaavat paremmin WWW-ympäristön todellisuutta (Kettunen et al. 2007).

ROMIP:in hakuaiheiden substantiiveille ja adjektiiveille on tässä kokeessa tuotettu rajoitetusti yleisimpiä sijamuotoja. FCG_3-versiossa hakulausekkeiden substantiivit ja adjektiivit ovat hakuaiheiden alkuperäisten sanamuotojen lisäksi yksikön nominatiivissa, genetiivissä ja akkusatiivissa. FCG_6-versiossa näille avaimille on lisäksi tuotettu monikon vastaavat sijamuodot. FCG_8-versiossa hakulausekkeisiin on vielä lisätty yksikön ja monikon instrumentaalimuodot. Hakuaiheiksi on otettu kaikki ne, joille ROMIP 2009 saantikannassa on löyhemmillä kriteereillä relevantiksi arvioituja dokumentteja. Tällaisia hakuaiheita on yhteensä 47. Lisäksi haut on tehty erikseen aiheista, joille on vain tiukoilla kriteereillä relevanteiksi arvioituja dokumentteja. Näitä hakuaiheita on 42. Haut on ajettu taivutusmuotoindeksiin ja tulokset evaluoitu TREC_EVAL-ohjelmalla.

Tässä tutkielmassa sijamuodot on tuotettu Morpher.ru-ohjelmalla, josta oli käytettävissä ilmainen versio webissä. Ohjelma tuottaa kaikki sijamuodot yksikön nominatiivissa syötetyille nomineille ja määrittelee annetun sanan suvun. Hakuaiheiden substantiivit ja adjektiivit on ohjelmaa varten perusmuotoistettu käsin. Muut sanaluokat kuin substantiivit ja adjektiivit on tässä kokeessa jätetty siihen muotoon, jossa ne ovat alkuperäisissä hakuaiheissa. Morpher.ru tunnistaa joitakin fraaseja ja sanaliittoja, esimerkiksi hakuaiheisiin sisältyvän средневековая культура (srednevekovaja kul'tura – keskiaikainen kulttuuri) sekä venäläisiä etunimi, isännimi, sukunimi -yhdistelmiä. Hakulausekkeita varten sijamuodot on kuitenkin tuotettu sana kerrallaan, sillä

Morpher.ru-ohjelman käytön tarkoitus on vain nopeuttaa sijamuotojen tuottamista. Automaattisessa tuottamisessa on pyritty kuitenkin johdonmukaisuuteen pitäytymällä Morpher.ru-ohjelman tekemissä ratkaisuissa, vaikka inhimillinen generoija olisi ehkä tulkinnut jotkin sanat toisin. Ohjelman tuottamat sanamuodot ovat oikeita, mutta sen tekemät tulkinnat singularia tantum ja pluralia tantum -sanoista ovat toisinaan vääriä.

Morpher.ru tunnistaa yleiset venäläiset sukunimet, kuten Бабич (Babič), eikä tuota niille lainkaan monikon sijamuotoja. Tämä on hieman kummallista, sillä venäjässä ei ole mitenkään tavatonta käyttää sukunimiä monikossa samaan tapaan kuin suomessa, esimerkiksi ”Mikkosten nuorin poika kirjoitti ylioppilaaksi”. Vierasperäiset sukunimet ohjelma generoi yksikössä ja monikossa. Morpher.ru:n linjan noudattamisesta johtuen esimerkiksi laulaja Mihail Krugin (Михаил Круг) venäläisestä etunimestä generoidaan vain yksikön sijamuotoja, mutta sukunimestä – joka on taiteilijanimi – myös monikon sijamuotoja. Sana круг (krug) on tavallisempi rengasta tai ympyrää tarkoittavana yleisnimenä kuin sukunimenä. Tästä syystä ohjelma ei tunnista sitä erisnimeksi. Maantieteellisille erisnimille, kuten Испания (Ispanija - Espanja) Morpher.ru tuottaa monikkomuodot, vaikka niiden kohdalla monikot ovat käytännössä harvinaisia. Hakulausekkeiden sanoista osa on moniselitteisiä, kuten бутан (butan), joka voi tarkoittaa joko valtiota nimeltä Bhutan tai butaania. Morpher.ru-ohjelmaa käytettäessä tulkinnalla ei ole väliä, sillä erimerkityksisten sanojen taivutus on näissä tapauksissa yhdenmukainen.

Kyselyjen muotoilussa on käytetty Lemurin InQuery-kyselykieltä, joka mahdollistaa hyvin rakenteiset kyselyt, ja jossa vaihtoehtoiset hakuavaimet rinnastetaan #SYN-operaattorilla. Alla olevassa esimerkissä sanalle второй (vtoroj – tässä: toisen) ei ole tuotettu muita sijamuotoja, koska se järjestyslukuna kuuluu sulkusanalistaan.

<query><number>arw33607</number>

<text>#combine(#syn(лагерь лагеря лагерей лагерем лагерями) #syn(второй) #syn(пятилетка пятилетки пятилетку пятилеток пятилеткой пятилетками))</text>

</query>[2]

[2] Käsittelemätön kysely suomeksi: toisen viisivuotiskauden leirit

## 5.3 Tulokset

ROMIP-aineistolla tehdyn kokeen keskeiset tulokset ovat taulukossa 10 ja kuvioissa 2 – 5. Tulokset osoittavat, että FCG on toimiva menetelmä myös venäjän kielelle, kuten Kettusen ja kumppaneiden (2007) tulokset ennakoivat. Kaikissa FCG-versiossa hakutulokset ovat vertailtavan perustason (*baseline*) tuloksia parempia. Parhaat tulokset tuotti FCG_6 mitattuna interpoloimattomalla keskitarkkuuksien keskiarvolla (*mean average precision*, MAP) yli hakuaiheiden. Kyseinen evaluointimittari valittiin, koska sitä käytetään yleisesti vertailtaessa erilaisten hakumenetelmien tehoa relevanttien dokumenttien löytämisessä (Dolamic & Savoy 2009). Eri menetelmien erojen tilastolliseen testaamiseen käytettiin Conoverin (1999) versioon perustuvaa Friedmanin testiä. Myös Kettunen ja kumppanit (2007) käyttivät Friedmanin testiä, sillä se soveltuu useiden menetelmien välisten erojen tarkasteluun. Löyhemmän relevanssiarvion (OR-saantikannan) evaluointitulosten erot olivat Friedmanin testin mukaan tilastollisesti erittäin merkitseviä. Parivertailussa erot käsittelemättömillä avaimilla hakemisen eli vertailtavan perustason (baseline) ja FCG_6-menetelmän välillä olivat tilastollisesti erittäin merkitseviä merkitsevyystasolla p=0,01. Niin myös FCG_8:n ja vertailtavan perustason. Merkitsevyystasolla 0,05 erittäin merkitseviä eroja oli myös menetelmien FCG_3 ja FCG_6 välillä. Ajoissa, jossa kyselyjen määrä oli vain 42, ja joiden evaluoinnissa käytettiin tiukkakriteeristä AND-saantikantaa, tulokset olivat niin ikään tilastollisesti erittäin merkitseviä. Näiden tulosten parivertailussa kuitenkin vain FCG_6-menetelmän ja vertailtavan perustason välinen ero oli merkitsevä, ja sekin merkitsevyystasolla 0,05.

**Taulukko 10** Eri menetelmien MAP-arvot tiukoilla ja löyhillä relevanssikriteereillä

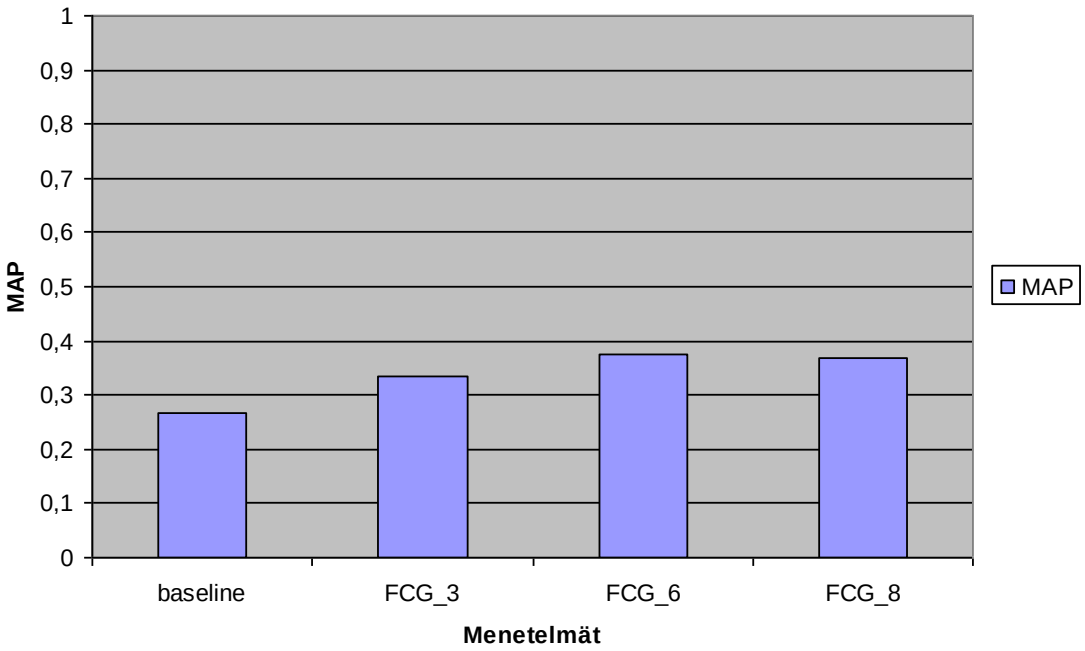
| Menetelmä | Relevanssiarviot | MAP | Relevanssiarviot | MAP |
|---|---|---|---|---|
| baseline | tiukat | 0.2197 | löyhät | 0.2656 |
| FCG_3 | tiukat | 0.2802 | löyhät | 0.3360 |
| FCG_6 | tiukat | 0.3051** | löyhät | 0.3746** |
| FCG_8 | tiukat | 0.2976 | löyhät | 0.3666** |

** tilastollinen merkitsevyys erittäin merkitsevä p<0.01 suhteessa baselineen

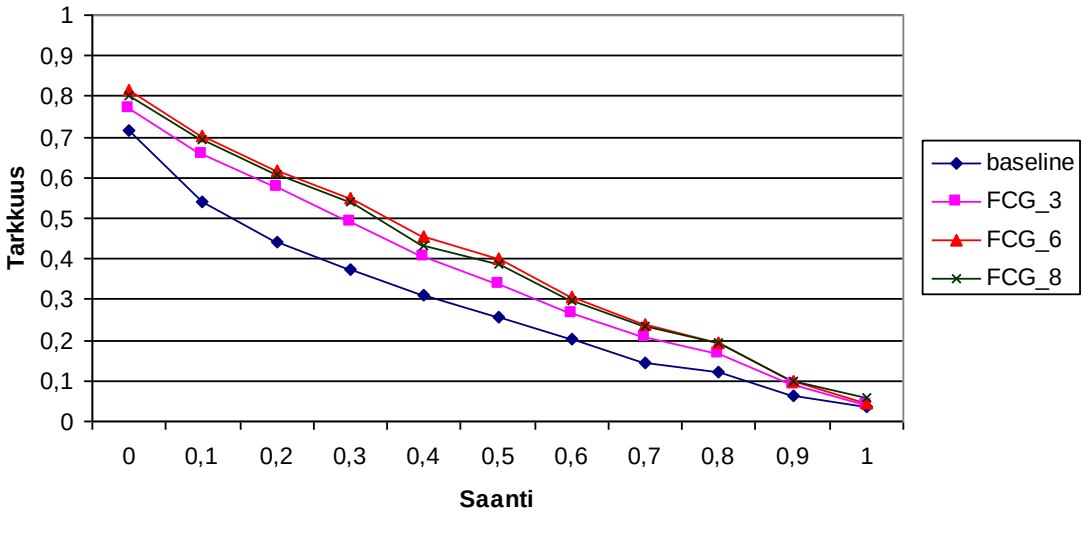
Kuvio 2 esittää keskitarkkuuksien keskiarvon (MAP) yli kyselyjen löyhillä relevanssikriteereillä. Kuviossa 3 ovat samoilla relevanssikriteereillä evaluoidut eri

hakumenetelmien interpoloidut tarkkuudet saantitasoittain. Kuvioissa 4 ja 5 ovat vastaavat arvot tiukoilla relevanssikriteereillä. Kuten Kettusen ja kumppaneiden (2007) kokeessa, FCG_6 menestyi hakumenetelmistä parhaiten. Instrumentaalin yksikön ja monikon lisääminen kyselyihin (FCG_8) heikentää hakutulosten tarkkuutta.

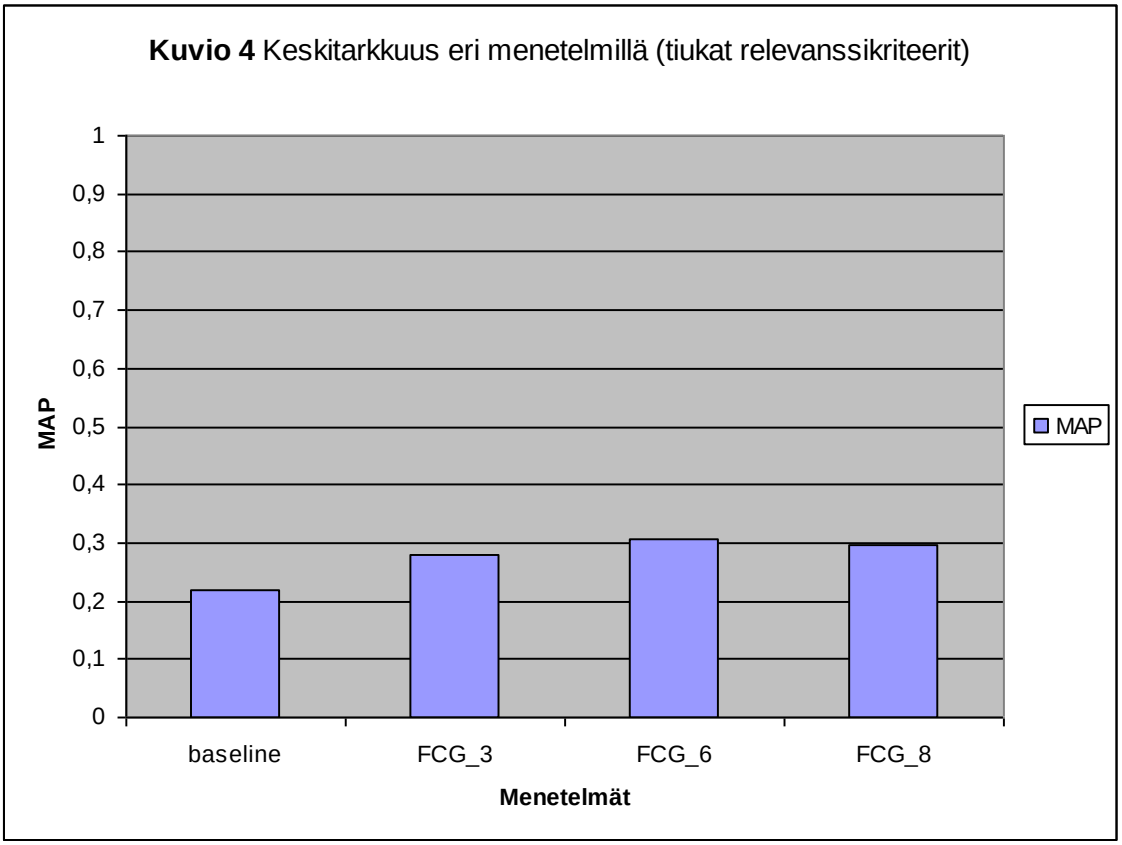
**Kuvio 2** Keskitarkkuus eri menetelmillä (löyhät relevanssikriteerit)



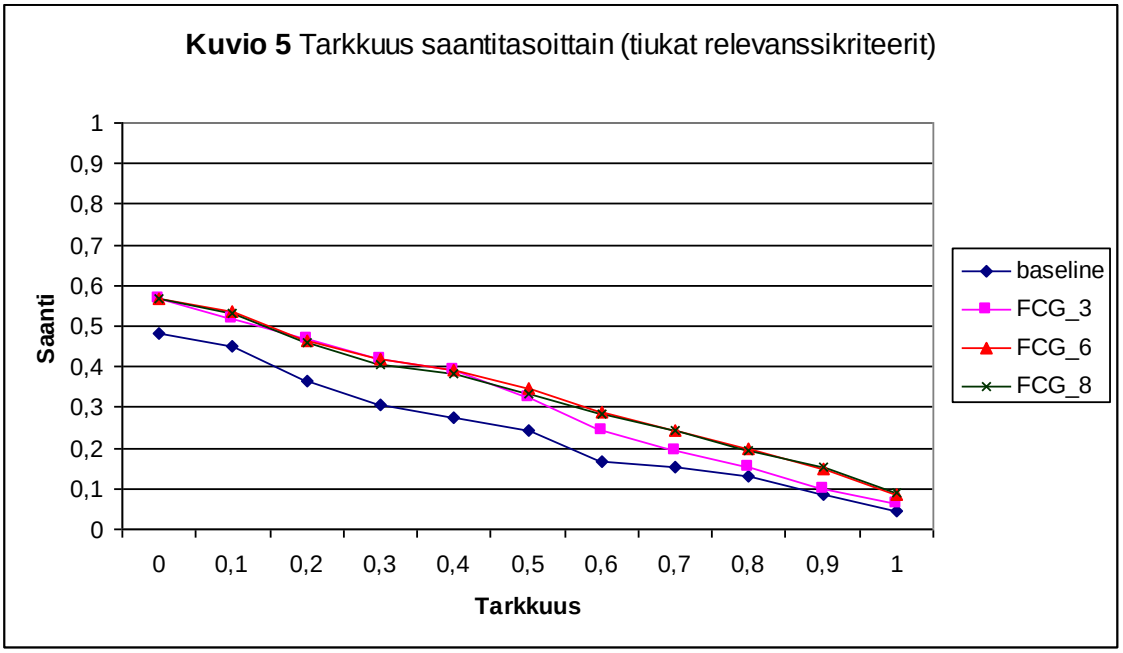

**Kuvio 3** Tarkkuus saantitasoittain (löyhät relevanssikriteerit)



**Kuvio 4** Keskitarkkuus eri menetelmillä (tiukat relevanssikriteerit)




**Kuvio 5** Tarkkuus saantitasoittain (tiukat relevanssikriteerit)




## 5.4 Adjektiivien perusmuotoistaminen

FCG-ratkaisumallissa sanamuotojen tuottaminen tapahtuu siten, että hakuaiheiden taivutusmuotoiset substantiivit ja adjektiivit ensin perusmuotoistetaan. Perusmuodossa –

venäjässä yksikön nominatiivissa – oleville sanoille tuotetaan sitten haluttu määrä sijamuotoja. Tässä työssä FCG-ohjelman toimintaa todellisessa, luultavimmin automaattiseen generointiprosessiin perustuvassa käytössä on simuloitu. Sijamuodot on tuotettu automaattisesti, mutta niitä generoivaa ohjelmaa varten sanat on perusmuotoistettu käsin. Myös tämän vaiheen olisi voinut tehdä automaattisesti. Adjektiivien osalta automaattinen perusmuotoistaminen on kuitenkin ongelmallista.

Venäjässä adjektiivien suku ja luku määräytyvät sen substantiivin mukaan, jonka attribuutteina ne ovat, mutta adjektiiveille on sovittu sanakirja- eli perusmuodoksi yksikön nominatiivin maskuliini. Venäjän kielen lemmatoija eli perusmuotoistaja RUSTWOL, joka muiden TWOL-ohjelmien tavoin perustuu Koskenniemen (1983) luomaan malliin, tarjoaa kaikkien sukujen adjektiivien sijamuotojen, vertailuasteiden ja lyhyiden muotojen perusmuodoksi maskuliinipäätteisen yksikön nominatiivin. Hakuaiheessa ja dokumenteissa adjektiivit ovat kuitenkin maskuliini-, feminiini- tai neutrimuodossa riippuen siitä, mikä on niiden määriteltävänä olevan substantiivin suku. Tästä syystä sanamuodot on edellä päädytty perusmuotoistamaan itse niin, että adjektiivien sijamuotojen generoinnin lähtökohdaksi on otettu maskuliini-, feminiini- tai neutripäätteinen yksikön nominatiivi sen mukaan, mitä sukutaivutusta adjektiivi taivutusmuotoisessa hakulausekkeessa edustaa. Esimerkiksi курсовая (kursovaja) on yksikön nominatiivin feminiini, joka on otettu kyseisen hakulausekkeen sijamuotojen automaattisen generoinnin lähtökohdaksi.

Koska adjektiivit vaikuttavat venäjänkielisessä tiedonhaussa mahdollisesti tärkeämmiltä kuin esimerkiksi suomenkielisessä, on tässä tutkielmassa vertailtu erilaisia adjektiivien käsittelytapoja FCG-menetelmän yhteydessä. Edellä kuvatun, tutkielmassa keskeisemmän kokeen tarkoitus on testata FCG-menetelmän kilpailukykyä verrattuna hakuun ilman morfologista prosessointia. Toisella, seuraavaksi kuvattavalla kokeella haluttiin selvittää, vaikuttaako adjektiivien perusmuotoistamisen tapa hakutuloksiin. Luvussa 4.5.2 käsitellyn adjektiivien sukukategorian feminiini- ja neutrimuotojen kadottamisen aiheuttama saannin heikkeneminen saattaa osoittautua tilastollisesti katsoen vähäiseksi muun muassa siksi, että monikossa sukuominaisuudet häviävät. Vaikka adjektiivien rooli lauseessa vaikuttaa venäjän kielessä merkittävältä, eivät ne esiinny teksteissä läheskään yhtä usein kuin substantiivit (ks. substantiivien ja adjektiivien kokonaismäärät taulukoissa 7 ja 8, luvussa 4.5.1).

Niistä ROMIPin 47 hakuaiheesta, joille testikokoelmassa on löyhempien relevanssikriteerien mukaan relevantteja dokumentteja, 99 sanaa on otettu FCG-käsittelyyn. Näistä 20 on adjektiiveja, joista yksi määrittelee neutrisukuista, yhdeksän feminiini- ja 10 maskuliinisukuista substantiivia. Tässä kokeessa hakutulokset on evaluoitu tiukkojen relevanssikriteerien mukaan, joten kokeeseen on otettu vain ne 42 hakuaihetta, joille AND-saantikannan mukaan on relevantteja dokumentteja. Käytetyissä hakuaiheissa on 19 adjektiivia, joista 10 on maskuliinia, 8 feminiiniä ja yksi neutri. Feminiineistä kaksi on samassa kyselyssä. Adjektiiveja sisältäviä hakuaiheita on siis 18.

Kyselyiden muotoilussa on simuloitu lemmatoinnin käyttöä hakuavainten perusmuotoistamisessa. Hakuaiheiden kaikki adjektiivit on perusmuotoistettu maskuliinisiksi yksikön nominatiiveiksi, ja tämä muoto on otettu muiden sijamuotojen tuottamisen lähtökohdaksi. Morpher.ru-ohjelmalle siis syötettiin maskuliininen perusmuoto kaikista adjektiiveista, ja ohjelma tuotti niille maskuliinisuvun mukaisen sijataivutuksen. Tuotettujen sijamuotojen kokonaismäärä väheni, sillä feminiinin erityinen akkusatiivimuoto jäi pois. Feminiinissä adjektiiveilla ei kokonaisuudessaan ole enempää sijapääteitä kuin maskuliinissa, neutrissa tai monikossa (Luku 4, Taulukko 9). Kaikissa FCG-versioissa tuotetuissa kolmessa yleisimmässä sijassa feminiinipäätteitä kuitenkin on enemmän sekä substantiiveilla että adjektiiveilla, koska akkusatiivi kuuluu tuotettuihin yleisimpiin sijamuotoihin. Toisen deklinaation substantiiveilla ja feminiinimuotoisilla adjektiiveilla akkusatiivin päätteet eivät ole genetiivin tai nominatiivin kaltaisia, vaan sijamuotojen päällekkäisyys koskee datiivia ja prepositionaalia.

Menetelmän verrokkina on edellisessä kokeessa käytetty menetelmä, jossa adjektiiveille tuotettiin maskuliini- feminiini- tai neutrisuvun sijataivutus sen mukaan, minkä sukuinen adjektiivi hakuaiheessa oli. Tässä kokeessa vertailtavana ovat erilaiset tavat tuottaa adjektiiveille sijamuotoja, ei tuotettujen sijamuotojen määrä kuten edellisessä kokeessa. Edellisessä kokeessa ajetuista 42 hakuaiheen FCG_3, FCG_6 ja FCG_8 kyselyistä on tehty uudet versiot, joissa tuotetut adjektiivien sijamuodot siis ovat maskuliinisia. Näitä menetelmäversioita tarkoitetaan tässä nimillä FCG_3_LEM, FCG_6_LEM ja FCG_8_LEM. Ne on ajettu KM.ru-kannan taivutusmuotoindeksiä vasten ja evaluoitu TREC_EVAL-ohjelmalla käyttäen tiukempaa AND-saantikantaa.

Hakutulosten MAP-arvoja verrataan kaksisuuntaista t-testiä käyttäen verrokkina olevien menetelmien MAP_arvoihin. FCG_3_LEM-menetelmää verrataan siis FCG_3-menetelmään, FCG_6_LEM-menetelmää FCG_6-menetelmään ja niin edelleen. Hakumenetelmien MAP-arvojen erot eivät ole tilastollisesti merkitseviä. Esimerkiksi FCG_6-hauissa MAP yli kyselyjen on 0,31, FCG_6_LEM-hauissa 0,29. Taulukossa 11 ovat eri menetelmien keskitarkkuuksien keskiarvot yli kyselyjen (MAP). Vaikka erot eivät adjektiivien vähäisyyden vuoksi ole merkitseviä, vaikuttavat tavallisten eli käsin tuotettujen kyselyjen tulokset systemaattisesti paremmilta kuin automaattisesti tuotettujen eli lemmattujen.

**Taulukko 11** MAP yli kyselyjen (tiukat relevanssikriteerit)

| | Tavallinen | Lemmattu |
|---|---|---|
| FCG_3 | 0,28 | 0,27 |
| FCG_6 | 0,31 | 0,29 |
| FCG_8 | 0,30 | 0,28 |

Kyselykohtaisesti tarkasteltuna haettujen relevanttien dokumenttien määrässä on havaittavissa adjektiivin suvun vaihtumisen vaikutus. Taulukossa 12 näkyvät menetelmäkohtaisesti relevanttien dokumenttien määrät niille hakuaiheille, joiden muotoiluun adjektiivien perusmuotoistamisen tavan muutos on vaikuttanut. Toisin sanoen taulukossa on vain niiden kyselyjen tuloksia, joissa adjektiiveille on tuotettu eri menetelmillä erilaisia sijamuotoja. Hakuaiheessa arw44884, joka koskee matkoja keskiaikaisessa kulttuurissa, tulos heikkenee lemmauksen seurauksena aina huomattavasti. Kolme sijamuotoa tuottavissa FCG-versioissa maskuliinisilla adjektiivimuodoilla on haettu vain seitsemän relevanttia dokumenttia, kun feminiinisiä muotoja käyttäen niitä haettiin 52. Kuusi ja kahdeksan sijamuotoa tuotettaessa erot eivät ole yhtä suuret. Osassa kyselyistä tulosten erot ovat pieniä, ja muutamissa lemmatoinnilla saatu tulos on jopa verrokkiaan parempi.

**Taulukko 12** Adjektiivien lemmaus ennen sijamuotojen tuottamista. Haetut relevantit dokumentit 1000 ensimmäisen joukossa (cut-off value 1000) hakuaiheittain eri FCG-versioissa

| Hakuaihe[3] | FCG_3 | FCG_3_LEM | FCG_6 | FCG_6_LEM | FCG_8 | FCG_8_LEM |
|---|---|---|---|---|---|---|
| arw33012 информационные технологии | 44 | 25 | 77 | 62 | 76 | 64 |
| arw33584 курсовая по педагогической психологии | 15 | 20 | 26 | 26 | 26 | 26 |
| arw38984 демографическая статистика | 36 | 31 | 36 | 33 | 36 | 33 |
| arw40084 информационные технологии в государстве | 2 | 2 | 2 | 2 | 2 | 2 |
| arw41069 компьютерные технологии в спорте | 0 | 4 | 0 | 4 | 0 | 4 |
| arw44884 путешествия в средневековой культуре | 52 | 7 | 55 | 20 | 53 | 21 |
| arw45982 свадебные поздравления | 2 | 2 | 2 | 2 | 2 | 2 |
| arw49617 японские игры | 3 | 4 | 6 | 6 | 6 | 6 |

## 5.5 Johtopäätökset tiedonhakukokeesta

Kokeen tulokset osoittavat, että FCG toimii venäjänkielisessä tiedonhaussa paremmin kuin haku morfologisesti käsittelemättömillä hakuavaimilla. Kettusen ja kumppaneiden (2007) suuntaa antavat tulokset saivat lisävahvistusta. Parhaat hakutulokset tuotti FCG_6-versio, jossa hakulausekkeeseen tuotettiin alkuperäisten hakuavainten lisäksi nominatiivin, genetiivin ja akkusatiivin yksikkö- ja monikkomuodot. Eniten sijamuotoja tuottava FCG_8-menetelmä ei MAP-mittarilla arvioiden ole paras. Tämä näkyi jo Kettusen ja kumppaneiden (2007) tuloksista.

Taulukossa 13 ovat Kettusen ja kumppaneiden (2007) venäjänkielisen tiedonhakukokeen eri menetelmien hakutulosten MAP-arvot. Prosentteina ilmaistujen MAP-arvojen oikealla puolella sulkeissa olevat luvut osoittavat, kuinka monta prosenttia kyseinen menetelmä on parhaiten pärjännyttä Ru-FCG_6-menetelmää heikompi. Inflected-menetelmässä hakuavaimet olivat niissä taivutusmuodoissa, joissa ne esiintyivät hakuaiheiden otsikoissa. Kettusen ja kumppaneiden (2007) Inflected-menetelmä vastaa siis tämän tutkielman baseline-menetelmää.

[3] Hakuaiheet suomeksi: arw33012 informaatioteknologiat, arw33584 pedagogisen psykologian kurssityö, arw38984 demografiset tilastot, arw40084 informaatioteknologiat valtiossa, arw41069 tietokoneteknologiat urheilussa, arw44884 matkat keskiaikaisessa kulttuurissa, arw45982 hääonnittelut, arw49617 japanilaiset pelit

Taulukossa 14 tässä tutkielmassa esitellyn kokeen MAP-arvot on esitetty vastaavalla tavalla. Lyhenne BS tarkoittaa taulukoissa baseline-menetelmää. Tulokset ovat samansuuntaisia kuin Kettusen ja kumppaneiden (2007). Kuusi eri sijamuotoa tuottava FCG_6-menetelmä oli tämän tutkielman kokeessa muihin menetelmiin nähden vielä parempi kuin Kettusen ja kumppaneiden (2007) kokeessa. Kaikki FCG-menetelmät ovat selvästi baselinea eli vertailtavaa perustasoa parempia. Snowball-stemmeri olisi todennäköisesti tässäkin kokeessa toiminut FCG-menetelmää huonommin, kuten Kettusen ja kumppaneiden (2007) kokeessa (Taulukko13).

**Taulukko 13** Kettusen ja kumppaneiden (2007) tulokset (MAP) lyhyillä kyselyillä (title queries)

| Menetelmä | MAP | |
|---|---|---|
| Ru-FCG_6 | 32.0 % | |
| Ru-FCG_8 | 31.7 % | (-0.3) |
| RU-FCG_3 | 31.2 % | (-0.8) |
| Snowball Ru | 27.2 % | (-4.8) |
| Inflected | 25.1 % | (-6.9) |

**Taulukko 14** Tämän tutkielman tuloksia (MAP)

| Löyhemmillä relevanssikriteereillä | | |
|---|---|---|
| Menetelmä | MAP | |
| FCG_6 | 37.5 % | |
| FCG_8 | 36.7 % | (-0.8) |
| FCG_3 | 33.6 % | (-3.9) |
| BS | 26.6 % | (-9.9) |
| **Tiukemmilla relevanssikriteereillä** | | |
| Menetelmä | MAP | |
| FCG_6 | 30.5 % | |
| FCG_8 | 29.8 % | (-0.7) |
| FCG_3 | 28.0 % | (-2.5) |
| BS | 22.0 % | (-8.5) |

Taulukossa 15 ovat Kettusen ja kumppaneiden (2007) venäjänkielisessä tiedonhakukokeessa eri menetelmillä löydettyjen relevanttien dokumenttien määrät. Aineistossa oli kaikkiaan vain 123 relevantiksi arvioitua dokumenttia. Haettujen dokumenttien listaa tarkastellaan ensimmäisten 1000 tuloksen osalta (*cut-off value 1000 documents*). Taulukossa 16 ovat tämän tutkielman kokeessa eri menetelmin haettujen relevanttien dokumenttien määrät. Menetelmäsarakkeessa BS tarkoittaa hakua

morfologisesti käsittelemättömillä avaimilla eli baseline-menetelmää. Löyhemmillä relevanssikriteereillä arvioiden aineistossa on 4855 relevanttia dokumenttia. Tiukemmat relevanssikriteerit täyttää 1530 dokumenttia. Taulukoista näkyy, että eniten sijamuotoja tuottavilla menetelmillä (Ru-FCG_8 ja FCG_8) löydettiin eniten relevantteja dokumentteja. Tämän tutkielman kokeessa, jossa käytettiin tiukoin kriteerein tehtyjä relevanssiarvioita, FCG_8-menetelmällä haettiin jopa 88 prosenttia kaikista relevanteista dokumenteista (Taulukko 16).

Kuusi eri sijamuotoa tuottava menetelmä on sekä Kettusen ja kumppaneiden (2007) että käsillä olevan tutkielman tulosten perusteella saannin ja tarkkuuden kokonaisuuden kannalta kolmesta tutkitusta FCG-menetelmästä paras. Taulukossa 15 tätä menetelmää edustaa Ru-FCG_6, taulukossa 16 FCG_6. Sitä käyttämällä haettujen relevanttien dokumenttien osuudet kaikista relevanteista dokumenteista olivat kummassakin kokeessa lähellä Ru-FCG_8 ja FCG_8 -menetelmillä saatuja osuuksia. Morfologisesti käsittelemättömillä hakuavaimilla tehty haku taivutusmuotoindeksistä tuotti näissä kokeissa aina muita hakumenetelmiä pienemmän osuuden relevanteista dokumenteista.

**Taulukko 15** Kettusen ja kumppaneiden (2007) löydetyt relevantit dokumentit 1000 parhaan haetun joukossa

| Menetelmä | Löydetyt 123 relevantista | % |
|---|---|---|
| Ru-FCG_6 | 84 | 68 |
| Ru-FCG_8 | 86 | 70 |
| Ru-FCG_3 | 78 | 63 |
| Snowball | 81 | 66 |
| Inflected | 67 | 54 |

**Taulukko 16** Tämä tutkielman kokeessa löydettyjen relevanttien dokumenttien määrä 1000 parhaan haetun joukossa

| | Löyhät relevanssikriteerit | | Tiukat relevanssikriteerit | |
|---|---|---|---|---|
| Menetelmä | Löydetyt 4855 relevantista | | Löydetyt 1530 relevantista | |
| FCG_6 | 3159 | 65 % | 1337 | 87 % |
| FCG_8 | 3197 | 66 % | 1347 | 88 % |
| FCG_3 | 2822 | 58 % | 1198 | 78 % |
| BS | 2326 | 48 % | 958 | 63 % |

Tutkielman teoreettisessa osassa herännyt epäilys sijamuotojen tuottamista varten tehtävän automaattisen perusmuotoistamisen hakutulosta heikentävästä vaikutuksesta ei saanut kokeessa kiistatonta vahvistusta. Adjektiivien erilaiset sijapäätteet eri suvuissa näyttävät tällä aineistolla testattuna vaikuttavan tiedonhaun tuloksellisuuteen vain hyvin vähän. Kaikkien sukujen mukaisten adjektiivien yleisimpien sijamuotojen lisääminen hakulausekkeisiin luultavasti parantaisi tulosta, mutta tilastollisesti parannus tuskin olisi merkitsevä.

Tarkastelun kohteena olivat näissä kokeissa vain lyhyet, web-tiedonhaukua simuloivat kyselyt. Tarkkaan ottaen tulokset osoittavatkin FCG-menetelmän toimivan hyvin vain tyypillisissä venäjänkielisissä web-hakulausekkeissa. Menetelmästä on otaksuttu olevan eniten hyötyä juuri web-tiedonhaussa (mm. Kettunen et al. 2007), ja kokeen tulokset vahvistavat aiemman tutkimuksen oikeuttamaa oletusta, että FCG:n hyödyntäminen käytännön sovelluksissa web-hakukoneissa olisi kannattavaa.

Tässä tutkielmassa FCG-menetelmää ei verrattu lemmatointiin eikä stemmaukseen, vaikka se FCG:n toimivuuden toteamiseksi olisi suotavaa. Ljiljana Dolamicin stemmeri, jota käsiteltiin luvussa 4, oli keväällä 2011 käytettävissä Tampereen yliopiston informaatiotieteiden yksikön tiedonhaun laboratoriossa. Niin ikään RUSTWOL-lemmatoija on mahdollista saada tutkimuskäyttöön CSC – Tieteen tietotekniikan keskuksen kautta. Käytetyn testikokoelman suuri koko hidasti kuitenkin indeksointiprosessia niin, ettei stemmatun tai lemmatun indeksin käyttö tutkielman puitteissa ollut mahdollista.

Kettunen ja kumppanit (2007) panevat merkille, että menetelmiä tutkittaessa otetaan liian usein verrokiksi huonoin mahdollinen vaihtoehto – haku taivutusmuotoindeksistä morfologisesti käsittelemättömillä avaimilla. He vertaavat FCG-menetelmiä paitsi käsittelemättömään hakuun, myös parhaina pidettyyn menetelmään, lemmaukseen, ja yleiseen käytäntöön ei stemmaukseen. Lemmatoijaa venäjän kielelle ei ollut saatavilla, mutta stemmausta pidettiin kohtuullisen hyvänä vaihtoehtona. Tässä tutkielmassa raportoidussa kokeessa tyydyttiin vertaamaan FCG-menetelmää hakuun taivutusmuodoilla. Menetelmien erot ovat kuitenkin siksi suuret, että niiden perusteella voi jo sanoa FCG:n toimivan hyvin. Seuraavaksi Dolamicin (2009) tai Snowball-stemmeriä voitaisiin käyttää KM.ru-kokoelmassa ja koetella FCG:n tuloksia verrattuna stemmauksella saatuihin. Koska lemmatointia pidetään parhaana menetelmänä

morfologialtaan monimutkaisten kielten tiedonhaussa, olisi myös RUSTWOL hyvä saada mukaan vertailuun.

# 6 YHTEENVETO

Tässä tutkielmassa tarkasteltiin venäjän kielen morfologiaa ja sen hallitsemiseksi esitettyjä tiedonhaun lingvistisiä menetelmiä. Substantiivien ja adjektiivien sijataivutuksessa tapahtuvan morfologisen vaihtelun hallinta voitiin aikaisemman tutkimuksen pohjalta todeta tiedonhaun näkökulmasta venäjän kielen morfologian olennaisimmaksi piirteeksi. Venäjä on morfologiselta kompleksisuudeltaan kielten keskitasoa, ja sijataivutus on sen keskeinen kieliopillisten suhteiden ilmaisukeino. Tiedonhaun kielitypologian (Pirkola 2001) avulla arvioituna venäjän synteesin taivutusaste on korkea, ja sen morfologian hallinnassa voidaan olettaa olevan hyötyä menetelmistä, jotka kohdistuvat taivutusaffikseihin. Kielen fuusion taivutusastetta (Pirkola 2001) ei tässä tutkielmassa laskettu, mutta kieli todettiin taivutusmorfologian osalta melko fuusioivaksi, koska pääteaffiksit ovat usein kieliopillisesti monimerkityksisiä, eikä eri kategorioita ilmaisevia affikseja voi erotella. Sijataivutus on fuusioivaa myös sikäli, että sanojen vartaloissa tapahtuu paikoin äännevaihtelua, kun niihin liitetään pääteaffikseja. Sanavartaloiden muutokset vaikuttavat stemmauksen tuloksellisuuteen, mutta Dolamic ja Savoy (2009) pitävät venäjän fuusioivuutta tältä osin niin vähäisenä, että eivät katso siitä olevan merkittävää haittaa.

Taivutusmorfologialtaan samalla tavoin kompleksisten kielten tiedonhaun tuloksellisuuden voidaan olettaa paranevan samantyyppisillä morfologian käsittelytavoilla (Kettunen et al. 2007; Dolamic & Savoy 2009). Suomen kielessä sijamuodot ovat keskeisessä asemassa ja synteesin taivutusaste on samaa tasoa kuin venäjän. Kimmo Kettunen ja Eija Airio (2006) havaitsivat, että rajoitettu taivutusmuotojen tuottaminen hakuavaimille toimii suomenkielisessä tiedonhaussa erittäin hyvin. Kettunen ja kumppanit (2007) testasivat menetelmää kompleksisuudeltaan eritasoisille kielille, joiden joukosta venäjä muistutti morfologisilta ominaisuuksiltaan eniten suomea. Tuotettaessa kokonaisia sijamuotoja ei taivutusmorfologian fuusioivuus aiheuta samanlaisia ongelmia kuin sijapäätteitä karsittaessa. Etenkin suomen morfeemeissa voi taivutettaessa tapahtua huomattavaa äännevaihtelua. Myös venäjässä vartalomuutoksista aiheutuvien ongelmien ratkeaminen on tervetullutta.

Venäjän adjektiivien päätteissä ilmenevä suku- ja sijakategorian fuusioituminen olisi saattanut heikentää FCG-menetelmän tuloksellisuutta. Tutkielman alaluvussa 5.4 raportoidussa adjektiivien perusmuotoistamisen tapoja vertailevassa FCG-lisäkokeessa kuitenkin osoittautui, ettei suvun muuttuminen heikennä hakutulosta kokonaisuutena tämän kokoisella aineistolla. Venäjänkielessä tiedonhaussa adjektiivit saattavat olla merkittävämmässä asemassa kuin esimerkiksi suomenkielisessä, mutta asian selvittämiseen tarvittaisiin erilainen aineisto.

Tässä tutkielmassa toistettiin Kettusen ja kumppaneiden (2007) venäjänkielisen tiedonhaun FCG-koe käyttäen suurempaa testikokoelmaa. Verrattuna baseline-menetelmään sijamuotojen rajoitettu tuottaminen osoittautui huomattavan tulokselliseksi hakutavaksi. Kokeen tulokset vahvistavat aiempia suuntaa antavia havaintoja (Kettunen et al. 2007) siitä, että FCG-menetelmä sopii erinomaisesti venäjänkieliseen tiedonhakuun. Tuloksia ei tässä kokeessa verrattu reduktiivisiin menetelmiin, mitä voidaan pitää puutteena. Jatkossa tässä tutkielmassa käytettyyn ROMIP-seminaarin KM.ru-testikokoelmaan voisi kokeilla stemmausta ja lemmausta ja verrata niitä FCG-menetelmään.

# LÄHTEET


Ahlgren P. & Kekäläinen J. (2007). Indexing strategies for Swedish full text retrieval under different user scenarios. Information Processing and Management 43: 81–102.

Airio, E. (2009). Morphological problems in IR and CLIR. Applying linguistic methods and approximate string matching tools. Academic dissertation. Tampere: University of Tampere, Department of Information Studies and Interactive Media. Acta Electronica Universitatis Tamperensis 842.

Alkula, R. (2000). Merkkijonoista suomen kielen sanoiksi. Suomen kielen tulkintajärjestelmien liittäminen tekstitiedonhakujärjestelmiin ja liittämisen vaikutukset tekstin tallennukseen ja hakuun. Väitöskirja. Tampere: Tampereen yliopisto, informaatiotutkimuksen laitos. Acta Universitatis Tampenrensis 763.

Andrews, E. (2001). Russian. SEELRC Reference Grammar. Duke University, Slavic and Eurasian Language Resource Center. Saatavilla: http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=6 Haettu 27.4.2011.

Baeza-Yates, R. & Ribeiro-Neto B. (1999). Modern Information Retrieval. New York: ACM Press.

Conover, W. J. (1999). Practical Nonparametric Statistics, 3rd ed. New York: Wiley.

Croft, W. B., Metzler, D. & Strohman T. (2010). Search Engines. Information retrieval in practice. New York: Pearson.

Dolamic, L. & Savoy J. (2009). Indexing and Searching strategies for the Russian language. Journal of the American society for information science and technology 60 (12): 2540 – 2547.

Dolamcin ja Savoyn sulkusanalista saatavilla: IR Multilingual Resources at UniNE. Université de Neuchâtel, http://members.unine.ch/jacques.savoy/clef/russianST.txt. Haettu 27.4.2011.

Dolamicin stemmeri saatavilla: IR Multilingual Resources at UniNE. Université de Neuchâtel, http://members.unine.ch/jacques.savoy/clef/russianStemmerPerl.txt. Haettu 27.4.2011

Hedlund, T., Pirkola, A. & Järvelin, K. (2001). Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language retrieval. Information Processing & Management, 37(1): 147 – 161.

Hedlund, T. (2002). Compounds in dictionary-based cross-language information retrieval. Information research 7(2). Saatavilla: http://informationr.net/ir/7-2/paper128.html Haettu 12.6.2011.

Internet World Stats (2010). Miniwatts Marketing Group. Saatavilla: http://www.internetworldstats.com/stats7.htm. Haettu 30.6.2011.

Järvelin, K. (2007). An analysis of two approaches in information retrieval: from frameworks to study designs. JASIST 58 (7): 971 – 986.

Karlsson, F. (2006). Yleinen kielitiede. Helsinki: Yliopistopaino.

Kettunen, K. & Airio, E. (2006). Is a morphologically complex language really that complex in full-text retrieval? In: Advances in natural language processing, LNAI 4139: 411 – 422. Berlin Heidelberg: Springer-Verlag.

Kettunen, K., Airio, E. & Järvelin, K. (2007). Restricted inflectional form generation in management of morphological keyword variation. Information retrieval 10: 214 – 244.

Kettunen, K. (2009). Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval. An overview. Journal of Documentation 65 (2): 267 – 290.

Koskenniemi, K. (1983). Two-level morphology: a general computational model for word-form recognition and production. Helsinki: University of Helsinki, Department of General Linguistics, Publications No. 11.

Koskenniemi, K. (1996). Finite state morphology and information retrieval. In: Extended finite state models of language. Proceedings of the ECAI 96 Workshop: 42 – 45.

Kratkaja russkaja grammatika (1989). Moskva: Russkij jazyk.

Loponen, A. & Järvelin, K. (2010). A dictionary- and corpus-independent statistical lemmatiser for information retrieval in low-resource languages. In: Multilingual and multimodal information access evaluation. Proceedings of the International Conference in the Cross-Language Evaluation Forum, CLEF 2010: 3 – 14. Heidelberg: Springer.

Nikunlassi, A. (2002). Johdatus venäjän kieleen ja sen tutkimukseen. Helsinki: Finn Lectura.

Pirkola, A. (2001). Morpholgical typology of languages for IR. Journal of Documentation, 57 (3): 330 – 348.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Saatavilla: http://snowball.tartarus.org/texts/introduction.html. Haettu 1.5.2011.

Rahmanova, L. I. & Suzdal'ceva, V. N. (2003). Sovremennyj ruskij jazyk. Leksika, freazeologija, morfologija. Moskva: Aspekt Press.

Russian Information Retrieval Evaluation seminar, ROMIP. Saatavilla: http://romip.ru/en/index.html Haettu 15.5.2011.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. Foundations and trends in information retrieval 4 (4): 247 – 375.

Savoy, J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. Proceedings of the 21. ACM Symposium on Applied Computing (ACMSAC): 1031–1035. New York: ACM Press.

Segalovich, I. (2003). A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. Yandex. Saatavilla: http://download.yandex.ru/company/iseg-las-vegas.pdf. Haettu 26.4.2011.

Šeljakin, M. A. (2006). Spravočnik po russkoj grammatike. Moskva: Drofa.

Snowball. Russian stemming algorithm. Saatavilla: http://snowball.tartarus.org/algorithms/russian/stemmer.html. Haettu 1.5.2011

Vahros, I. & Kahla, M. (1967). Venäläisten sanojen translitteroinnista. Helsinki: Neuvostoliittoinstituutti.

Zaliznjak, A. A. (1977). Grammatitšeskij slovar russkogo jazyka. Moskva: Russkij Jazyk.

# LIITE 1 Synteesin taivutusaste laskettuna sadan sanan samansisältöisistä suomen- ja venäjänkielisistä tekstinäytteistä

*Syksyllä 1893 V. I. Lenin liittyi Pietarissa opiskelijoiden marxilaiseen opintokerhoon. Kerhon kokouksessa hän arvosteli H. B. Krasinin referaattia ”Markkinakysymys”. Lenin kirjoitti vastauksena artikkelin ”Niin sanotun markkinakysymyksen johdosta”. Hän perusteli artikkelissaan, että Venäjä oli siirtymässä kapitalismiin, mistä oli osoituksena tavarantuottajien jakaantuminen kapitalisteiksi ja proletariaatiksi. Lenin näki saman ilmiön tapahtuvan myös maaseudulla: ”Jos tarkkailemme maanviljelijä-talonpoikia, niin osoittautuu, että toisaalta talonpojat joukoittain hylkäävät maansa, kadottavat taloudellisen itsenäisyyden, muuttuvat proletaareiksi, ja että toisaalta talonpojat laajentavat alituisesti kylvöalaa ja siirtyvät parempiin viljelystapoihin.” Lenin jatkoi perusteluaan ja totesi, että kapitalismi edustajinaan kapitalisti ja kylvöalaa laajentanut talonpoika sekä köyhtyneet joukot eli maalais- sekä työläisproletariaatti eivät sulkeneet pois*

Taivutusmorfeemien määrä jaettuna suomenkielisen tekstinäytteen sanojen määrällä: 59/100

*Осенью 1893 г. В.И. Ленин вступил в Петербурге в марксистский кружок студентов. В том же году, на одном из собрании кружка Ленин подверг критике реферат Х.Б. Красина «Вопрос о рынках» и написал в ответ статью «По поводу так называемого вопроса о рынках». В данной статье Ленин обосновал переход России к капитализму, о чем свидетельствовал раздел производителей на капиталистов и пролетариат. Согласно Ленину, то же самое происходило и в деревне: «Возьмем ли мы крестьян-земледельцев – окажется, что, с одной стороны, крестьяне массами забрасывают землю, теряют хозяйственную самостоятельность, обращаются в пролетариев, с другой стороны, крестьяне расширяют постоянно запашки и переходят к*

Taivutusmorfeemien määrä jaettuna venäjänkielisen tekstinäytteen sanojen määrällä: 55/100

Näytteet on poimittu Tampereen Lenin-museon verkkosivuilta (http://www.lenin.fi/) 5.5.2011.