Riitta Alkula

Merkkijonoista suomen kielen sanoiksi*

Alkula, Riitta: Merkkijonoista suomen kielen sanoiksi: Suomen kielen morfologisten tulkintaohjelmien liittäminen tekstitiedonhakujärjestelmään ja liittämisen vaikutukset tekstin tallennukseen ja hakuun. Informaatiotutkimuksen laitos, Tampereen yliopisto. Acta Universitatis Tamperensis 763.

Väitöskirjatutkimuksessa selvitettiin, miten suomen kielen morfologisten tulkintaohjelmien avulla voidaan ratkaista sellaisia tiedon tallennuksen ja haun ongelmia, jotka johtuvat suomen kielen erityispiirteistä.

Tutkimusta varten rakennettiin oma testausympäristönsä, jossa samasta tekstiaineistosta (23244 sanomalehtiartikkelia) tuotettiin joukko erilaisia tietokantoja. Eri tietokantoja luotaessa sovellettiin usealla eri tavalla suomen kielen morfogisia tulkintaohjelmia, ja näitä tietokantoja sekä niistä tehtyjen tiedonhakujen tuloksia vertailtiin toisiinsa saanti- ja tarkkuusarvojen perusteella. Projekti oli siis luonteeltaan laboratorioympäristössä toteutettu evaluointitutkimus. Testauksissa käytetty hakujärjestelmä oli käänteishakemistoon perustuva, Boolen operaattoreita käyttävä BASIS.

Vertaillut tutkimusympäristöt olivat seuraavat:

T1) **Perinteinen hakeminen**: hakijan katkaisemat hakusanat (kysely kohdistui taivutusmuotohakemistoon, joka sisälsi dokumenttien sanat sellaisinaan, taivutusmuodoissaan)

T2) **Automaattinen katkaisu**: perusmuotoisten hakusanojen syöttäminen taivutusvartaloita tuottaville ohjelmille, kysely vartaloilla (kysely kohdistui taivutusmuotohakemistoon)

T3) **Seulonta**: automaattinen taivutusvartaloiden tuottaminen, kysely vartaloilla sekä tulosten seulonta perusmuotoon palauttavalla ohjelmalla (kysely kohdistui taivutusmuotohakemistoon)

T4) **Perusmuotojen** ja yhdyssanojen alkuosien hakeminen (kysely kohdistui perusmuotohakemistoon, jossa morfologisen tulkintaohjelman tunnistamat sanat olivat perusmuodossa, tunnistamatta jääneet sanat taivutusmuotoisina)

T5) **Perusmuotojen ja** yhdyssanan kaikkien **osien** hakeminen (kysely kohdistui ositettuun perusmuotohakemistoon, jonne perusmuotojen lisäksi on tallennettu yhdyssanoista kaikki niiden osat sekä näiden osien yhdistelmät)

Lisäksi tutkimuksessa rakennettiin yksi tutkimusympäristö (T6), jossa tutkittiin tarkemmin kyselyjä, jotka eivät perusmuotohakemistossa tuottaneet oikeaa tulosta. Perusmuotoistamisen riskinä nimittäin on, että tulkintaohjelmille tuntemattomat sanat tulkitaan väärin, jolloin hakemistoon päätyy vääriä sanoja. Tutkimuksessa kokeiltiin muutamia yksinkertaisia korjausmenetelmiä, joilla tällaiset väärät tulkinnat voidaan hakuvaiheessa kiertää ja siten varmistaa dokumenttien löytyminen tai parantaa hakutulosten tarkkuutta.

Tutkimusta varten koottiin 26 kyselyn perusjoukko, jonka lisäksi johdoksia ja yhdyssanoja tarkasteltiin tarkemmin omissa osajoukoissaan. Kyselyistä muodostettiin kahdeksan eri tyyppiä. Ensimmäisissä neljässä tyypissä kyselyä laajennettiin hakusanojen perusmuodoista lähtien:

A) **Peruskysely,** joka sisälsi alkuperäiset hakupyynnössä esiintyneet sanat perusmuodossaan

AB) **Johdoskysely,** joka sisälsi peruskyselyn hakusanat sekä näiden johdokset perusmuodossaan

AC) **Yhdyssanakysely,** joka sisälsi peruskyselyn hakusanat sekä yhdyssanat, jonka osana hakusanat esiintyivät

ABC) **Yhdistelmäkysely,** joka sisälsi peruskyselyn hakusanat ja näiden johdokset sekä yhdyssanat, joiden osana hakusanat tai niiden johdokset esiintyivät

---

*Alkulan väitöstilaisuus pidettiin 25.8.2000 Tampereen yliopistossa. Väitöskirja on luettavissa kokotekstinä osoitteessa http://acta.uta.fi/pdf/951-44-4886-3.pdf

Neljässä muussa kyselytyypissä jaettiin osiinsa sellaiset hakusanat, jotka olivat yhdyssanoja. Tämän jälkeen kyselyihin lisättiin nämä yhdyssanojen osat seuraavasti:

Aa) **Osien peruskysely,** joka sisälsi alkuperäiset hakupyynnössä esiintyneet sanat sekä yhdyssanojen osat perusmuodossaan

ABab) **Osien johdoskysely,** osien peruskysely laajennettuna hakusanojen ja näiden osien johdoksilla perusmuodossaan

ACac) **Osien yhdyssanakysely,** osien peruskysely laajennettuna sellaisilla yhdyssanoilla, joiden osana jokin hakusanan osa oli

ABCabc) **Osien yhdistelmäkysely,** edellisten kyselytyyppien yhdistelmä

Kun hakemistoon tallennettavat sanat perusmuotoistettiin, perusmuotohakemisto vei vähemmän muistitilaa kuin taivutusmuotohakemisto. Tämä päti myös ositetussa perusmuotohakemistossa eli kun hakemistoon tallennettiin perusmuotojen lisäksi myös yhdyssanojen osat ja niiden yhdistelmät.

Kun kyselyjen tulosjoukkoja vertailtiin saannin keskiarvojen perusteella, paras tulos saatiin ositetusta perusmuotohakemistosta (T5) ja toiseksi paras perusmuotohakemistosta (T4). Erot toisiin tutkimusympäristöihin olivat systemaattisia, mutta yleensä eivät tilastollisesti merkitseviä. Kolmanneksi paras oli taivutusmuotohakemisto (T1), jonne tekstien sanat oli tallennettu taivutusmuodossaan ja hakija käytti katkaistuja hakusanoja. Tosin automaattisella katkaisulla (T2) päästiin lähes samoihin tuloksiin. Selvästi huonoimman tuloksen tuotti seulonta (T3) - ero toisiin tutkimusympäristöihin oli myös tilastollisesti merkitsevä.

Tarkkuuden keskiarvojen vertailussa parhaat tarkkuusarvot sai seulonta (T3). Seuraavaksi parhaat tarkkuusarvot saatiin perusmuotohakemistosta (T4), mutta lähes yhtä hyvät tarkkuusarvot saatiin ositetusta perusmuotohakemistosta (T5). Tarkkuudeltaan huonoimpia olivat hakijan katkaisemilla hakusanoilla taivutusmuotohakemistosta (T1) saadut tulosjoukot. Hakusanojen katkaiseminen automaattisesti (T2) paransi tarkkuutta, mutta varsin vähän. Eri tutkimusympäristöjen tarkkuusarvojen väliset erot eivät olleet tilastollisesti merkitsevät.

Perusmuotohakemisto oli yleisesti ottaen tarkempi kuin taivutusmuotohakemisto: kun esimerkiksi molemmissa käytettiin täsmälleen samoja, taivutusvartaloohjelman katkaisemia hakusanoja, perusmuotohakemistosta saatujen tulosjoukkojen tarkkuus oli parempi kuin samanlaisella kyselyllä taivutusmuotohakemistosta saatujen tulosjoukkojen tarkkuus.

Toisaalta perusmuotohakemistoon tehdyissä kyselyissä ei kannata käyttää pelkkiä hakusanan perusmuotoja. Kun perusmuotohakemistosta haettiin antamalla muuten samat hakusanat kuin taivutusmuotohakemistosta haettaessa, mutta jättämällä ne katkaisematta, saanti romahti. Kun kyselyyn lisättiin hakusanan perusmuotojen lisäksi sen johdokset tai hakusanan sisältävät yhdyssanat, tulosjoukon saanti nousi useampia prosenttiyksikköjä kuin tarkkuus samalla laski.

Perusmuotohakemistoista haettaessa hakijan on siis muistettava ottaa myös johdokset ja yhdyssanat huomioon. Toisaalta hakija pystyy perinteistä hakutapaa (kysely katkaistuilla hakusanoilla taivutusmuotohakemistosta) paremmin valitsemaan, haluaako painottaa saantia vai tarkkuutta.