hyväksymispäivä arvosana

arvostelija

Suomenkielisten tekstien morfologinen analysointi

Pirjo Suominen

Helsinki 10.10.2010

HELSINGIN YLIOPISTO

Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet –Faculty	Laitos – Institution – Department
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen laitos

Tekijä – Författare – Author

Pirjo Suominen

Työn nimi – Arbetets titel – Title

Suomenkielisten tekstien morfologinen analysointi

Oppiaine – Läroämne – Subject
Tietojenkäsittelytiede

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
	10.11.2010	12 sivua

Tiivistelmä – Referat – Abstract

Tässä kirjoituksessa käsitellään suomenkielisten tekstien morfologista analysointia. Ensin käsitellään sanaluokkia kielitieteellisessä analyysissa, sitten käydään läpi analysoinnin vaiheet ja vähän välineitä. Lopuksi selvitetään mihin morfologisesti analysoitua aineistoa voidaan käyttää ja mainitaan muutama olemassa oleva korpus.

Avainsanat – Nyckelord – Keywords
kielitieteelliset aineistot, morfologinen analysointi

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Sisältö

1 Johdanto

Tässä kirjoituksessa käsitellään suomenkielisten tekstien morfologista analysointia. Morfologia on kielitieteen osa, joka tutkii sanojen muotoja. Pienin kielellinen muoto, jolla on merkitys, on nimeltään morfeemi [HVK04]. Esimerkiksi sanassa *lastu* on yksi morfeemi, kun sana *lastusta* sen sijaan koostuu vartalomorfeemista *lastu* ja siihen liittyneestä päätemorfeemista *-sta*.

Kielitieteellisten aineistojen (korpus) tutkimuksissa vertaillaan usein erilaisia tekstilajeja, tekstityyppejä tai rekistereitä. Korpus on tiettyä tarkoitusta varten tietyillä periaatteilla koottu tekstikokoelma. Sana ja sanaluokka ovat tällöin keskeisiä käsitteitä, ja tutkittavana ovat sanaluokkien esiintymistiheys tai suhteelliset taajuudet.

Morfologisessa analyysissa tutkitaan tekstin muoto-opillisia piirteitä. Ennen varsinaisten morfo-syntaattisen tiedon lisäämistä aineisto muokataan niin, että se voidaan tallettaa tekstimuodossa, siitä poistetaan esimerkiksi tekstinkäsittelyohjelman esitysmuotoon liittyvät ominaisuudet.

Varsinaisessa morfologisessa analyysissa tekstin sanojen perusmuodot liitetään sanoihin. Samalla myös morfologisia ominaisuuksia, kuten sanaluokkia, aikamuotoja, tapaluokkia ja taivutusmuotoja, kuvaavia merkitsimiä (”tageja”) liitetään sanoihin. Analyysin perusteella saadaan tietoja mm. sanojen yleisyydestä ja sanaluokista sekä verbien perusmuodoista, pää- ja tapaluokista ja aikamuodoista.

Artikkelissa käsitellään ensin sanaluokkia koulukieliopin ja kielitieteellisen analyysin näkökulmista. Sitten käydään läpi analysoinnin vaiheet ja välineitä analyysin tekemiseen. Lopuksi selvitetään mihin morfologisesti analysoituja aineistoja käytetään ja mainitaan muutama olemassa oleva korpus.

2 Sanaluokista

2.1 Sanaluokat koulukieliopissa

Suomen kieli on morfologisesti rikas kieli, sillä sen substantiiveilla, verbeillä ja adjektiiveilla on teoreettisesti ajatellen tuhansia erilaisia tai ainakin erilaiselta vaikuttavia sanamuotoja. Esimerkiksi sanasta *lastu* on mm. muotoja *lastu, lastusta, lastuista*. Arppe esittää artikkelissaan, että Suomen kielessä on arvioitu olevan nomineja yli 1850, adjektiiveja 6000 ja verbejä arviolta 20000 [Arp06].

Perinteisissä koulukieliopeissa sanaluokat määriteltiin yleensä semanttisesti eli sanojen merkityksen perusteella. Tällöin substantiivit merkitsivät esineitä, asioita ja olioita, verbit ilmaisivat tekemistä ja adjektiivit kuvasivat, millainen jokin on. Muita vastaavia koulukieliopin sanaluokkia ovat numeraalit, pronominit ja partikkelit [HeL04].

Sanaluokat voidaan määritellä myös morfologisesti, jolloin samalla tavalla taipuvat sanat kuuluvat samaan luokkaan. Tällöin esimerkiksi verbit muodostavat oman luokkansa, koska ne taipuvat persoonassa ja adjektiivit omansa vertailuasteineen [HVK04].

Käytännössä kumpaakaan tapaa ei käytetä ainoana luokitteluperusteena, vaan luokittelu perustuu osaksi edellä mainittuihin, osaksi lauseopillisiin perusteisiin. Suomen kieliopeissa nykyään käytetty luokittelu perustuu osaksi taivutukseen, osaksi syntaktiseen käyttäytymiseen ja merkitykseen. Taivutuksen perusteella sanat voidaan luokittaa kolmeen pääryhmään: nomineihin, joilla on sija- ja lukutaivutus, verbeihin, joilla on persoona-, tempus- ja modustaivutus sekä sanoihin, joilla on korkeintaan osittaista taivutusta. Syntaktisen käyttäytymisen perusteella nominit jakautuvat substantiiveihin, adjektiiveihin, pronomineihin ja numeraaleihin ja taipumattomat tai vajaasti taipuvat sanat adpositioihin, adverbeihin ja partikkeleihin. Verbien infiniittiset muodot ovat sanaluokaltaan epäselvempiä: ne ovat lausekkeen muodostuksen kannalta verbimäisiä, mutta niillä ei ole finiittiverbien persoona-, tempus- eikä modustaivutusta vaan nominien piirteitä.

2.2 Sanaluokat kielitieteessä

Suuria aineistoja analysoitaessa sanaluokan määritelmä tulee yleensä automaattisen analysaattorin jäsennyskieliopin mukana. Niitä ei kuitenkaan ole yleensä dokumentoitu kunnolla, jolloin analysaattorin käyttäjälle ei yksiselitteisesti kerrota, mitä kyseisen analysaattorin jäsennyskieliopissa tarkoitetaan verbillä tai adverbilla tai millaisia sanaluokkia kielioppi sisältää. Analysaattorin verbi saattaa tarkoittaa jotain muuta kuin analysaattorin käyttäjä on koulukieliopissa oppinut [HeL04].

Heikkinen ja Lounela ovat selvittäneet tutkimuksessaan kolmen suomen kielen analysaattorin sanaluokkakäsityksiä. Analysaattorit olivat Connexor Oy:n jäsennin Fi-lite, Kielikone Oy:n Texmorfo -jäsennin ja Lingsoft Oy:n morfologinen analysaattori Fintwol. Taulukossa 1 on kuvattu eri analysaattoreiden sanaluokkatulkinnat.

	Textmorfo	Filite	Fintwol
Substantiivi	Noun	N	N
Erisnimi	Proper		
Verbi	Verb	V	V
Adjektiivi	Adjective	A	A
Averbi	Adverb	ADV	ADV
Ad-adjektiivi			AD-A
Pronomini	Pronoun	PRON	PRON
Konjunktio	Conjunction	CC (rinnastus) / CS (alistus)	C
Numeraali	Numeral	NUM	NUM
Kvantifioija			Q
Prepositio	Preposition	PRE	PP (suom) / PREP (ulkom)
Postpositio		PSP	PSP / PP
Artikkeli	Article		ART (ulkom)
Lyhenne	Abbreviation		ABBR
Interjektio	Interjection	INTERJ	INTJ
Erotinmerkki	Delimiter		PUNCT (välimerkki) / CHAR (muu merkki
Koodi	Code		
Yhdysosa	CompPart		

Taulukko 1: Sanaluokkatulkinnat [HeL04]

Eri analysaattoreiden sanaluokkakäsitykset poikkeavat toisistaan. Textmorfo antaa jokaiselle tekstin sanalla yhden sanaluokan, Fi-lite vähintään yhden ja Fintwol kaikki mahdolliset sanaluokat. Taulukossa 2 on eri analysaattoreiden antamat tulkinnat tekstille: ”Tässä on kuollut mies. Mies on kuollut eilen.” Taulukosta käy ilmi, että tulkinnat eroavat useassa kohtaa toisistaan [HeL04].

Konjunktiot jaetaan Fi-litessä alistus- ja rinnastuskonjunktioihin, muut luokittelevat ne toisin. Vain Textmorfo käsittelee erisnimet omana sanaluokkana. Textmorfo laskee välimerkit sanaluokkamerkitsinten joukkoon, Fi-litessa taas välimerkeillä ja muilla merkeillä on eri sanaluokkakoodit. Fintwol ei mainitse välimerkkejä sanaluokkana, mutta ei-aakkosnumeeriset merkit saavat PUNCT-merkitsimen.

	Textmorfo	Fi-lite	Fintwol
Tässä	Adverb	PRON	DEM PRON
on	Verb	V	COP V
kuollut	Verb	A	V / PCP2 / PCP2 A
mies	Noun	N	N
.	Delimiter	PUNCT	PUNCT FULLSTOP
Mies	Noun	N	N
on	Verb	V	COP V
kuollut	Verb	V	V / PCP2 / PCP2 A
eilen	Adverb	ADV	ADV
.	Delimiter	PUNCT	

Taulukko 2: Laboratoriotekstin sanaluokka-analyysit [HeL04]

Partisiippien käsittelyssä on myös eroja. Textmorfo tulkitsee ne verbeiksi, Fi-lite pyrkii erottamaan adjektiivi- ja verbitulkinnat ympäristön perusteella. Fintwol pitää partisiippia sanaluokkana, mutta tarjoaa kaikki tulkinnat. Se tarjoaa kuitenkin partisiippimuodoille myös tulkinnan, jossa on sekä adjektiivi- että partisiippimerkitsin, mikä hämärtää sanaluokan käsitettä.

3 Aineistojen analysointi

3.1 Analysoinnin vaiheet

Kieliteknologiaa hyödyntävä tutkimusprosessi voidaan jakaa neljään osaan: aineiston valinta, aineiston valmistelu koneellisesti ymmärrettävään muotoon, automaattisen tunnuslukujen laskenta ja tunnuslukujen ja tekstien tulkinta. Tässä käsitellään lähinnä tutkimusprosessin toista vaihetta, aineiston valmistelua.

Ennen analysointia teksti on talletettava tietokoneelle tekstimuodossa, eli se ei saa olla paperilla eikä tekstinkäsittelyohjelman talletusmuodossa. Eri analyysiohjelmat suhtautuvat eri tavoin isoihin kirjaimiin, sanoissa kiinni oleviin välimerkkeihin ja

muihin muotoseikkoihin [LeL04]. Siksi teksti normalisoidaan esikäsittelemällä se ennen varsinaista analyysia, josta esimerkki kuvassa 1.

lasketaanko vastuu vähittäiskaupan harteille , teollisuuden ,
valtiovallan vai julkisen sanan kannettavaksi ? sitä emme
valitettavasti itsekään vielä tiedä .

Kuva 1: Esikäsitelty teksti [HLL00]

Kun teksti on normalisoitu analyysiohjelman ymmärtämään muotoon, se voidaan analysoida morfologisesti, disambiguoida, jäsentää pintatasolla tai sille voidaan tehdä täysi syntaktinen jäsennys. Disambiguoinniksi kutsutaan prosessia, jonka avulla sanan tulkinta tehdään yksiselitteiseksi.

Teksteihin voidaan tässä vaiheessa merkitä myös kappale- ja virkerajat ja otsikot, mikäli teksti halutaan esimerkiksi XML-muotoon [LeL04]. Morfologinen analyysi voidaan suorittaa automaattisesti käyttämällä esimerkiksi professori Kimmo Koskenniemen kehittämää kaksitasomallia [Kos84]. Analysoitaessa teksti morfologisesti sanoille etsitään niiden perusmuodot ja ne liitetään sanoihin. Lisäksi sanoihin liitetään niiden morfologisia ominaisuuksia kuten sanaluokkia, aikamuotoja ja taivutusmuotoja kuvaavia merkitsimiä. Näitä tehtäviä varten on olemassa automaattisia työkaluja, mutta osa työstä on tehtävä itse tai ainakin jokaisen vaiheen lopputulos on tarkistettava huolellisesti [Lou07b].

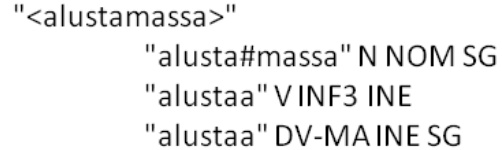
Kaksitasomallin ongelmaksi muodostuvat kuitenkin sananmuodot, jotka voidaan tulkita eri tavoin, esimerkiksi *teillä* on sekä sanan *tie* että sanan *te* muoto. Siksi morfologisen analyysin jälkeen tekstille voidaan tehdä disambiguoitu analyysi, jolla sanan tulkinta tehdään yksiselitteiseksi. Ihminen pystyy yleensä helposti tulkitsemaan oikein monitulkintaiset muodot. Koneellisessa analyysissä ne ovat kuitenkin suuri ongelma. Ratkaisuna käytetään muun muassa professori Fred Karlssonin kehittämää rajoitekielioppia. Siinä kontekstista löytyvää tietoa käytetään väärien tulkintojen poistamiseen.
Disambiguointia varten tarvitaan jonkin verran tietoa sanojen määrite-pääsana -suhteista, jolloin samalla kun tehdään disambiguoitua analyysia, voidaan lisätä myös pintasyntaktisen analyysin merkitsimet. Ne voivat olla esimerkiksi määreisiin liitettäviä, jolloin ne kertovat missä suunnassa pääsana on ja nimeävät määritesuhteen [HLL00].

Joillakin analysaattoreilla voidaan tuottaa tällaista analyysia, mutta usein aineiston valmistelija (ihminen) poistaa analyysista kontekstissaan väärät tulkinnat. Valmiit analysaattorit saattavat tehdä tässä vaiheessa paljon virheitä, joiden etsiminen ja korjaaminen on lähes yhtä työlästä kuin väärien tulkintojen poistaminen käsityönä, jolloin valmistelija voi valita mieluummin käsityön [Lou07b].
Mikäli aineistolle tehdään syntaktinen analyysi tarkoittaa se sitä, että sanojen määritesuhteet tuodaan eksplisiittisesti esiin. Tällöin merkitsimistä voi lukea myös riippuvuuden tyypin ja pääsanan järjestysnumeron virkkeessä [HLL00].

Näiden vaiheiden jälkeen teksti on analysoitu ja se on valmis automaattiseen tunnuslukujen ja listojen tuottamiseen. Analysoinnin löytämiä tekstin piirteitä voi tämän jälkeen käyttää tekstin tutkimiseen, esimerkiksi määrälliseen analysointiin [HLL00]. Tutkimusprosessin viimeinen vaihe, tunnuslukujen ja tekstien tulkinta, suoritetaan täysin ihmisvoimin. Automaattisen analyysin tuottamia lukuja ja listoja tutkitaan ja verrataan. Luvut ja listat saattavat osoittaa teksteistä kiinnostavia ominaisuuksia, joita voidaan tutkia edelleen syventymällä kvalitatiivisesti itse teksteihin [Lou07b].

3.2 Ohjelmistoja analysointiin

Lingsoft Oy:n Fintwol analysaattori perustuu Kimmo Koskenniemen kehittämään kaksitasomalliin ja se tuottaa jokaiselle tekstin sanalle morfologisen tiedon [Kos84]. Mikäli sanalla voi olla useita sanamuotoja, Fintwol listaa kaikki vaihtoehdot. Esimerkiksi sana ”*alustamassa*” voidaan tulkita tarkoittavan yhdyssanaa ”*alusta*” ja ”*massa*” tai kolmatta infinitiivimuotoa verbistä ”*alustaa*” tai deverbalisoitua johdannaista verbistä ”*alustaa*” [Lou05]. Fintwol analysaattori tekee kuitenkin vain morfologisen analyysin, josta esimerkki kuvassa 2.



Kuva2: Esimerkki Fintwolin morfologisesta analyysista [Lou05]

Fi-litellä voidaan tehdä morfologisen analyysin lisäksi disambiguoitu analyysi, jolloin poistetaan sanan väärät tulkinnat. Samalla voidaan lisätä myös pintasyntaktisen analyysin merkitsimet. Fi-lite poikkeaa Fintwolista analysointitavan lisäksi myös siinä, että se antaa jokaiselle tekstin sanalla vähintään yhden sanaluokan, ei välttämättä kaikkia.

FDG on suomenkielisten tekstien morfologinen ja syntaktinen jäsennysohjelma, joka jäsentää tekstiä virkkeittäin dependenssisyntaksiin perustuen, jolloin myös sanojen määritesuhteet näkyvät selkeästi analyysin tuloksessa.

Kielikoneen Textmorfo-ohjelma jäsentää ja yksiselitteistää suomenkielistä tekstiä. Ohjelma sisältää sekä Kielikone Oy:n morfologisen jäsentimen että dependenssijäsentimen, joka yksiselitteistää morfologisen jäsentimen mahdolliset moniselitteiset analyysit [YKL10]. Textmorfon morfologinen analyysi eroaa Fintwolista ja Fi-litestä ainakin siinä, että se antaa jokaiselle tekstin sanalla vain yhden sanaluokan.

4 Morfologisesti analysoidun tekstin käyttö

Korpusten avulla voidaan tutkia kielen rakennetta ja käyttöä eri aikoina ja siten myös kielen muuttumista. Morfologisesti analysoituja aineistoja on käytetty tutkittaessa sanojen kieliopillisia suhteita.

Sanaluokkajakaumia tutkittaessa on huomattu, että uutisissa ja virkakielessä sanaluokkien käyttö on erilaista. Virkakielessä yleisin sanaluokka substantiivi on yleisempi virkakielessä kuin uutisissa. Verbeissä tämä suhde on juuri päinvastainen. [Lou07a]. Sekä verbien finiittimuotojen määrä että sanaluokkajakauma kertovat eräiden tutkijoiden mukaan siitä, kuinka staattinen tai dynaaminen on tekstien kuvaama todellisuus. On osoitettu, että taiteellisessa tyylissä finiittiverbien osuus kaikista tekstisanoista on suurin (n.18 %), tieteellisessä tyylissä pienin (n.12 %). Oletus on, että mitä finiittiverbivoittoisempaa teksti on, sitä dynaamisempaa se on. Substantiiveista on osoitettu, että niitä on eniten lakiteksteissä ja vähiten saarnoissa [HLL00].

Morfologisesti analysoimattomista aineistoista voi saada kiinni vain melko yleisluontoisia määrällisiä piirteitä. Mikäli aineistot ovat pienehköjä, nykykielisiä

tekstijoukkoja, joiden morfologinen ja rakenteellinen analyysi on ollut mahdollista, niistä on voitu laskea eri tavoin tekstien ominaisuuksista kertovia vertailutietoja kuten perusmuoto-, sanaluokka-, aikamuoto- ja sijamuotolistat. Automaattisesti tuotetut luvut ja listat ovat hyvä renki, mutta huono isäntä. Aineiston valmistelussa mahdollisesti tapahtuneet virheet tai muut aineiston puutteet voivat aiheuttaa yllättäviä tuloksia varsinkin pienistä tekstimääristä tehdyissä laskelmissa. Tutkimuksissa on otettava huomioon myös aineiston koko: mitä pienempi aineisto on, sitä enemmän virheet tai yhden ison tekstin ominaisuudet voivat vaikuttaa koko joukon tuloksiin. Kustakin tekstijoukosta lasketut tulokset pätevät siihen itseensä, ja niistä voi ehkä tehdä varovaisia oletuksia sen tekstityypin, tekstilajin tai genren ominaisuuksista, jota tekstijoukko edustaa [Lou07a].

Tulkintoja tehtäessä on huomattava myös se, että kielen keinot ovat joustavia. Preesensillä tai nominatiivilla voi olla useita eri käyttötarkoituksia samassa tekstijoukossa ja erilaisia tyypillisiä käyttöjä eri tekstijoukoissa. Tämän vuoksi pelkkien lukujen ja listojen perusteella ei voi tehdä kovin pitkälle meneviä päätelmiä tekstijoukkojen viestinnällisistä ominaisuuksista eikä varsinkaan päätellä kovin paljon sellaisen abstraktin olion kuin ”suomen kieli” ominaisuuksista yleensä. [Lou07a].

5 Morfologisia korpuksia

Korpukset ovat muuttuneet vuosien varrella, mm. niiden koostamisperiaatteet, tavat, koot ja käyttömahdollisuudet ovat muuttuneet. Korpukset voivat pitää sisällään tekstiä sellaisenaan ilman mitään lingvististä analyysiä, tai sitten korpuksen sisältö on voitu lingvistisesti annotoida joko automaattisesti, puoliautomaattisesti tai manuaalisesti. Korpukset saattavat sisältää myös metalingvististä tietoa esim. kirjoittajasta, puhujasta, julkaisuajankohdasta, tekstityypistä tai vaikkapa siitä sanomalehden osasta, missä kyseinen teksti on julkaistu.

Korpuksia on kertynyt lukuisia kullekin kielelle, joten niiden kattava esittäminen on mahdoton tehtävä. Verkosta löytyy useita jatkuvasti ylläpidettyjä listauksia eri kielten erityyppisistä korpuksista, joista yksi on Helsingin yliopiston Kielitieteen laitokselta [HLC07]. Tässä mainitaan muutamia Suomessa kerättyjä ja käytettyjä korpuksia.

Valitettavasti osa missä tahansa päin maailmaa kerätyistä korpuksista ei ole käytännössä julkisesti käytettävissä, vaan niihin on erikseen haettava käyttölupa korpusta ylläpitävältä organisaatiolta.

Tunnettuja suomalaisia korpuksia ovat Oulun korpus ja ns. HKV-korpus, kuten myös suomen kielen tekstipankki. Oulun korpusta ryhdyttiin keräämään 1967 ja perusjoukkona siinä on 60-luvun loppupuolen sanoma- ja aikakauslehtiä ja radio-ohjelmia sekä vuosina 1961 - 1966 ilmestynyttä kauno- ja tietokirjallisuutta. Perusjoukko on koottu arpomalla lyhyitä tekstikatkelmia mahdollisimman suuresta määrästä eri tekstejä. Tämän korpuksen pohjalta on tehty Suomen kielen taajuussanasto [HLL00].

HKV-korpus koostuu neutraalista asiaproosasta, jonka alatekstilajeja ovat tietokirjat, ensyklopediat, pääkirjoitukset, kulttuuriartikkelit, pakina ja tiedottavat artikkelit. Erona Oulun korpukseen on se, että otokset koostuvat kokonaisista teksteistä tai ainakin itsenäisistä pitkähköistä tekstikappaleista. Näin siksi, että monet tutkimusten muuttujat ovat sellaisia, että niiden luonne ja yhteisvaikutus on todettavissa vasta kokonaisissa teksteissä eikä yksittäisten lauseiden sisällä [HLL00].

Suomen kielen tekstipankki koostettiin vuosina 1996-1998 yhteiseurooppalaisessa PAROLE-projektissa. Suomalaisia osapuolia edustivat Helsingin yliopiston yleisen kielitieteen laitos ja Kotimaisten kielten tutkimuskeskus. Tavoitteena oli kerätä vähintään 20 miljoonan sanan nykykielen korpus kustakin kielestä ja nimenomaan kirjoitetusta kielestä. Tekstit jaettiin neljään luokkaan: kirjoihin, sanomalehtiin, aikakauslehtiin ja muihin teksteihin [HLL00].

6 Yhteenveto

Vaikka analysointiin on kehitetty hyviä ohjelmistoja, ei manuaalista käsityötä voi unohtaa. Analysaattorin lopputulos on tarkistettava, koska suomen kieli on morfologisesti rikas kieli. Eri analysaattorit antavat myös hieman poikkeavia tuloksia ja prosessoivat tietoa hieman eri tavoin, joten niiden käyttäjän on tiedettävä mitä milläkin analysaattorilla saa aikaan ja miten niiden tuloksia luetaan ja tulkitaan.

Tulosten tulkinnassa on oltava varovainen, ettei tee liian yleistäviä ja pitkälle meneviä johtopäätöksiä tekstien viestinnällisistä ominaisuuksista. Niitä on kuitenkin pystytty hyödyntämään mm. tekemällä korpuksen pohjalta Suomen kielen taajuussanasto.

Valmiita suomenkielisiä korpuksiakin on olemassa jo paljon. Verkosta löytyy listauksia erityyppisistä korpuksista, mutta niiden ongelma on se, että niitä ei voi vapaasti käyttää. Lupa on erikseen haettava korpusta ylläpitävältä organisaatiolta, mutta onneksi luvan saa yleensä aika helposti tutkimuskäyttöön.

Lähteet

Arp06 Arppe, A., Frequency Considerations in Morphology, Revisited - Finnish Verbs Differ, Too. *Festschrift in Honour of Fred Karlsson in his 60th Birthday (2006). Special Supplement to SKY Journal of Linguistics, Volume 19/2006, pp. 175-189*. Linguistic Association of Finland, Turku, 2006

HeL04 Heikkinen Vesa, Lounela Mikko, Sanaluokka morfologisen analyysin kategoriana. *Proceedings of the Annual Finnish and Estonian Conference of Linguistics*, Tallinna, 2004, s.87-97

HLC07 Computer corpora of the uralic languages, http://www.ling.helsinki.fi/uhlcs/data/languages.html [10.11.2010]

HLL00 Heikkinen, V., Lehtinen, O., Lounela, M., Ihminen ja kone tekstiä mankeloimassa, kuusikohtauksinen keskustelu. *XXVII Kielitieteen päivät Oulussa*, Oulu, 2000

HVK04 Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T., Alho, I., Iso suomen kielioppi, Suomalaisen Kirjallisuuden Seura, Helsinki, 2004. http://scripta.kotus.fi/visk [10.11.2010]

Kos84 Koskenniemi, K. A general computational model for word-form recognition and production, *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*, s. 178-181, 1984.

LeL04 Lehtinen, O., Lounela, M., A model for composing and (re-)using text materials for linguistic research. *Papers from the 30th Finnish Conference of Linguistics. Studies in Language, University of Joensuu Vol. 39*, s. 73-78, Joensuu, 2004

Lou05 Lounela, M., Exploring Morphologically Analysed Text Material. *Inquiries into Words, Constraints and Contexts Festschrift for Kimmo Koskenniemi on his 60th Birthday*, s. 259-267, 2005

Lou07a Lounela, M., Tekstien kvantitatiivisia piirteitä: teksti ja tekstijoukko määrällisten muuttujien valossa. *Kotimaisten kielten tutkimuskeskus*. 2007

Lou07b Lounela, M., Kieliteknologiasta suomenkielisten tekstien tutkimisessa. *Puhe ja kieli, 27:1* , 2007, s. 41-48.

YKL10 Kieliteknologiset ohjelmat Helsingin yliopiston Yleisen kielitieteen laitoksella http://www.ling.helsinki.fi/atk/sovellusohj/index.shtml [10.11.2010]