Lehtinen, E., S. Aaltonen, M. Koskela, E. Nevasaari & M. Skog-Södersved (toim.) 2011. AFinLA-e Soveltavan kielitieteen tutkimuksia 2011 / n:o 3. 48–61.

Jarmo Harri Jantunen
Oulun yliopisto

Avainsana-analyysi annotoidun oppijankieliaineiston tutkimisessa: Alustavia havaintoja

This paper documents the preliminary findings from a survey in which corpus-driven keyword analysis is employed to investigate a lemmatised and annotated learner language corpus. Keyword analysis is seldom used to analyse grammatically annotated data, and to my knowledge, never in analyses of tagged learner data. This article illustrates the kinds of over- and underused items that can be found in learner corpus data using keyword analysis. These include grammatical tags, content keywords, and tentative learner language keywords. The analysis reveals that annotated data yield a more complete picture of the nature of the atypical frequencies of linguistic items in learner language. The article also discusses the role of other methodological choices, such as the criteria for defining the level of proficiency (learning hours vs. CEFR).

Keywords: corpus-driven analysis, keywords, annotation, learner corpora, learner Finnish

1 Johdanto

Artikkelissani[1] raportoin kokeilusta, jossa testataan korpusvetoisen tutkimusmenetelmän soveltumista annotoidun oppijankieliaineiston analyysiin. Fokusoin erityisesti menetelmään ja esitän alustavia havaintoja ja pohdintoja menetelmän ja aineiston yhdistämisestä saatujen tulosten avulla. Artikkelissa oletetaan, että kieliopillisesti annotoidun aineiston käyttäminen mahdollistaa kieliopillisten luokkien yli- ja alikäytön tutkimisen leksikaalisten piirteiden ohessa. Menetelmän avulla saadaan lisätietoa kielenaineisten epätyypillisistä frekvensseistä, jotka eivät riko kohdekielen systeemiä mutta jotka poikkeavat sen normista tilastollisin perustein. Kielenainesten epätyypillisistä frekvensseistä on tehty runsaasti havaintoja sekä leksikon ja leksikaalisten suhteiden (mm. Hasselgren 1994; Granger 1998; Nesselhauf 2005) että kieliopin (mm. Granger 1997, 1999) osalta, mutta laajaa kieliopillisten ja leksikaalisten piirteiden analyysia ei ole olemassa.

2 Korpusvetoisuus tutkimusmenetelmänä

Sähköisiin aineistoihin perustuva tutkimus voidaan jakaa karkeasti kahteen menetelmällisesti eroavaan päälinjaan: korpuspohjaiseen (*corpus-based*) ja korpusvetoiseen (*corpus-driven*) tutkimukseen (Tognini-Bonelli 2001). Korpuspohjaiselle tutkimukselle on tyypillistä, että tutkittavat kielelliset ilmiöt tai yksiköt on päätetty etukäteen ennen aineiston käsittelyä. Tällöin on tavallista, että esimerkiksi aiemmin manuaalisilla menetelmillä saatuja tuloksia testataan uudestaan, ehkä hieman modifioituina, aiempaa tutkimusta huomattavasti laajemmilla aineistoilla, jolloin tavoitteena on falsifioida tai verifioida aiemmat tulokset tai saada niistä muuten entistä tarkempaa tietoa. Tutkittavat ilmiöt voivat määräytyä ennalta myös tutkijan oman intuition ja aiemman kokemuksen perusteella. (Tognini-Bonelli 2001; ks. myös Hardie & McEnery 2010.) Oppijankielen[2] – ja yleensäkin kielenoppimisen – tutkimuksen näkökulmasta intuition perustana tai sen rinnalla voi olla opettaja-tutkijan havainto ilmiön ongelmallisuudesta kielenoppimisessa ja siten tutkimuksen tarpeellisuudesta. Näin ollen tutkimushypoteesit voivat nousta jostakin muusta kuin käytössä olevasta aineistosta. (Esimerkkejä oppijankielen tutkimuksesta ks. Barlow 2005.) Suurin osa tähän mennessä tehdystä korpustutkimuksesta

1 Artikkeli perustuu AFinLAn 2010 Metodit L2-korpustutkimuksessa -työpajassa pidettyyn esitelmään. Tutkimusta ovat rahoittaneet Pohjoismaiden ministerineuvosto (2934/2007), Ruotsin valtionpankin juhlarahasto (F10-0421:1) ja Oulun yliopisto.

2 Oppijankielellä tarkoitan kielivarianttia, joka on kielenoppijoiden tuottamaa ja joka on kohdekielen normistoa ja ominaispiirteitä jossain määrin noudattavaa, mutta joka oletettavasti pitää sisällään myös piirteitä oppijan äidinkielestä, ja mahdollisesti myös piirteitä, jotka ovat tyypillisiä erityisesti juuri tälle kielivariantille (vrt. välikieli, ks. Latomaa 1996: 98–99; Ellis & Barkhuizen 2005: 4–6; Jantunen 2008).

on nimenomaan korpuspohjaista. Menetelmällä on saatu runsaasti tietoa kielestä, ja nyttemmin myös oppijankielen luonteesta – erityisesti tämä koskee oppijanenglantia.

Korpusvetoinen analyysi puolestaan nojaa aineistoon jo heti tutkimuksen alkuvaiheessa: tutkijalla on mielessään laaja tutkimuskehys, esimerkiksi oppijankielen fraseologisten epätyypillisyyksien selvittäminen, mutta varsinaiset tutkimusyksiköt, joilla ilmiötä tutkitaan ja kuvataan, valitaan ja hypoteesi muokataan aineiston analyysin perusteella (ks. esim. Kallioranta 2009). Menetelmässä olennaisessa asemassa ovat frekvenssit, toistuvuus ja usein myös tilastollinen merkitsevyys. Tognini-Bonellin (2001: 86) mukaan korpusvetoisuuden etuna on se, että sen avulla tutkittavaksi voi nousta sellaisia ilmiöitä, joita tutkija ei pelkän intuitionsa perusteella havaitse mielenkiintoisiksi tai tärkeiksi. Korpusvetoisen menetelmän lähtökohtana voidaan käyttää esimerkiksi avainsana-analyysia, sanalistoja tai konkordansseja (Barlow 2005; Scott & Tribble 2006). Korpusvetoisuuden käyttö on harvinaista oppijankielen tutkimuksessa, ja sen soveltaminen lingvistisesti annotoidun oppijankieliaineiston tutkimukseen on olematonta.

Edellä kuvatut kaksi menetelmää eivät välttämättä esiinny tutkimuksessa kuitenkaan ”puhtaina”, vaan niitä on voitu yhdistellä yhdessäkin tutkimuksessa. Ns. aineistoesimerkein tuetussa tutkimuksessa korpusta käytetään puolestaan vain eräänlaisena esimerkkivarantona, jolloin ei kuitenkaan ole enää kyse korpuksen käytöstä metodologisena keinona (Jantunen 2009; ks. myös Tummers, Heylen & Geeraerts 2005).

Korpusvetoinen avainsana-analyysi perustuu esiintymistaajuuksiin ja tilastollisuuteen. Sen avulla pystytään selvittämään, mitkä sanat yli- ja aliedustuvat tutkimusaineistossa vertailuaineistoon verrattuna. Avainsanalla (*keyword*) tarkoitetaan korpustutkimuksessa kielenainesta, joka esiintyy aineistossa tilastollisesti merkitsevästi useammin kuin vertailuaineistossa (Scott ja Tribble 2006: 58–59). Menetelmä pohjautuu tutkimus- ja vertailuaineistosta tehtyjen sana- tai sananmuotolistojen vertailuun; tyypillisesti avainsana-analyysia onkin käytetty juuri sanojen tai sananmuotojen tunnistamiseen aineistosta, mutta menetelmä lienee sovellettavissa myös lingvistisen metatiedon tarkasteluun.

Avainsanat voidaan jakaa positiivisiin eli vertailuaineiston suhteen yliedustuviin ja negatiivisiin eli aliedustuviin sanoihin. Lisäksi avainsanat voidaan luokitella sisältöavainsanoihin ja sellaisiin avainsanoihin, jotka kuvaavat pikemminkin tekstilajia tai tyyliä (Jantunen 2009, tulossa; ks. myös Berber-Sardinha 1999). Sisältöavainsanat kertovat, mistä tekstissä puhutaan. Esimerkiksi S2-alan artikkelien korpuksen avainsanoja voisivat olla *suomenoppija*, *oppijansuomi*, *transfer* ja *kohdekieli*, sillä nämä esiintynevät tällaisessa korpuksessa useammin kuin keskimäärin erilaissa tieteellisissä teksteissä.

Genreavainsanat puolestaan kuvaavat tekstiaineistoa nimenomaan tekstilajille tyypillisten piirteiden näkökulmasta. Tieteellisten tekstien genreavainsanoja ovat mm. numeraalit ja lyhenteet (*mt*., *ks*., *esim*.), kaunokirjallisten tekstien avainsanoja puoles-

taan persoonapronominit (*minä*, *mä*, *hän*) ja kommunikaatioverbit (*sanoa*, *kysyä*). (Jantunen tulossa.) On todennäköistä, että myös erilaisista kielivarianteista on löydettävissä samaan tapaan "kielivarianttiavainsanoja" eli ilmauksia, jotka ovat ominaisia juuri tuolle kielivariantille riippumatta tekstien aihepiiristä tai tuottajasta. Olisiko siten löydettävissä oppijankielen ilmauksia, jotka tilastollisesti yli- tai aliedustuvat riippumatta tekstin aihepiiristä tai oppijan kielitaustasta ja jotka antaisivat teksteille eräänlaisen oppijankielimäisyyden tunnun? Tätä selvitetään alustavasti tässä artikkelissa avainsana-analyysin avulla, jossa leksikaalisten piirteiden lisäksi analysoidaan myös kieliopillisia kategorioita.

3 Oppijankieliaineiston annotointi

Annotoinnilla tarkoitetaan lingvistisen metatiedon eli tietoa kuvailevan tiedon tuottamista digitaaliseen aineistoon (Leech 2005). Tyypillisimmillään se on erilaisten kieliopillisten tunnisteiden eli tagien (puoli)automaattista lisäämistä tekstitiedostoon tai -korpukseen, ja tuloksena voi olla esimerkiksi morfologinen tai syntaktinen annotaatio. Oppijankielen korpuksissa tavallisin annotaatio on virheannotaatio (Barlow 2005: 341; ks. myös Granger 2004). Analysaattorien avulla tehty automaattinen annotaatio ei ole koskaan kuitenkaan täysin virheetön. Annotointi on silti kannattavaa, sillä aineiston käyttömahdollisuudet moninkertaistuvat annotoinnin myötä ns. raakatekstiin verrattuna (ks. Meunier 1998; Leech 2005), vaikkakin toisaalta annotointi on jo yksi (jäsentimen taustalla olevaan kielikäsitykseen perustuva) analyysi aineistosta ja siten tutkija saa käyttöönsä jo kerran tehdyn tulkinnan teksteistä (Sinclair 2004: 191; Hunston 2002: 93).

Tässä analyysissa käytetty aineisto on jäsennetty Connexorin Fi-fdg-dependenssijäsentimellä, joka on suomen kielen morfologinen ja syntaktinen analysaattori. Jäsennin pohjautuu Tesnièren (1959) verbikeskeiseen dependenssikielioppiin. Analyysin tuloksena saadaan mm. sanan perusmuoto eli lemma, morfologista tietoa taivutuksesta sekä syntaktista tietoa lauseenjäsennyksen muodossa. (Ks. Heikkinen, Lounela & Voutilainen tulossa; Fi-fdg). Tässä artikkelissa käytetystä oppijankielen aineistosta otettu tekstijakso näyttää annotoituna ja lemmatisoituna karkeasti riveittäin seuraavalta:

rivi	saneet	lemma	syntaktinen suhde ja morfologinen merkintä
1	Lapsena	LAPSI	NH N SG ESS
2	minä	MINÄ	NH PRON SG P1 NOM
3	oli	OLLA	MAIN V ACT IND PAST SG P3
4	tyynisti	TYYNESTI	ADVL ADV
5	mutta	MUTTA	CC CC
6	hyvin	HYVIN	PREMOD ADV
7	itsepäinen	ITSE#PÄINEN	NH A SG NOM
8	.	.	

KUVIO 1. Esimerkki lemmatisoidusta ja annotoidusta virkkeestä.

Oppijankielen korpusaineistojen automaattisessa annotoinnissa on luonnollisesti natiiviaineistoja suurempi virhemahdollisuus kirjoitusvirheiden ja väärien tai väärinmuodostettujen taivutusmuotojen vuoksi (de Haan 2000: 71), joskaan edistyneiden kielenoppijoiden ja natiivien tuottamien tekstien annotointitulosten välillä ei enää Grangerin (2002: 58) mukaan ole juurikaan eroja. Van Rooy ja Schäfer (2003: 836) raportoivat esimerkiksi virheellisesti merkittyjen sanaluokka-annotaatioiden (*POS annotation*) määrän olevan edistyneiden englanninoppijoiden teksteissä 3,7–13,7 % analysaattorista riippuen; de Haan (2000: 69) on päätynyt puolestaan viiden prosentin virhemäärän. Ongelmia onkin luonnollisesti enemmän annotoitaessa alemmalla tasolla olevien oppijoiden tuotoksia, joissa korostuvat kirjoitus- tai taivutusvirheet – toisaalta annotointia voivat kuitenkin helpottaa lyhyet ja yksinkertaiset virkerakenteet.

Oppijankieltä sisältävä aineisto voidaan aluksi korjata kirjoitusvirheiden ja väärinmuodostettujen muotojen osalta (kuviossa 1 *tyynisti* > *tyynesti*), minkä jälkeen aineisto annotoidaan ja lopuksi kielenoppijan kirjoittamat virheet palautetaan tiedostoon. Niiden palauttaminen on erittäin tärkeää, jotta tutkijan käytössä on teksti sellaisena kuin se on alun perin kirjoitettu. Näin toimien vältetään se Grangerin (2004: 128) mainitsema ongelma, että automaattisessa analyysissa ei yleensä tavoiteta virheellisiä muotoja. Käsillä olevan tutkimuksen aineisto on annotoitu vaiheittain siis seuraavasti: tekstin virhetarkistus ja virheiden korjaaminen > automaattinen annotointi Fi-fdg-jäsentimellä > virheiden palauttaminen aineistoon. Annotaation tarkistaminen ja disambiguointi eli vaihtoehtoisista annotaatioista oikean valitseminen on vielä käynnissä tätä kirjoitettaessa.

4 Aineistot

Tutkimusaineistona on *Kansainvälisen oppijansuomen korpuksen* (ICLFI, Jantunen & Piltonen 2009) annotoitu osakorpus, joka sisältää äidinkieleltään vironkielisten tallinnalaisten ja tarttolaisten suomenoppijoiden tekstejä. Alkeistason (myöhemmin VICLFI-A) aineiston koko on 51 000 sanetta ja edistyneiden (VICLFI-E) 37 000 sanetta.[3] Vertailuaineistona käytän n. 4 miljoonan saneen *Käännössuomen korpuksen* (Mauranen 2000) alkuperäissuomen osakorpusta, joka on annotoitu tätä tutkimusta varten ja joka sisältää samanlaisia tekstityyppejä (fiktiivisiä ja ei-fiktiivisiä tekstejä) kuin ICLFIkin. Tutkimustyökaluna käytän *WordSmith Tools* -korpustyökalupakettia (Scott 2008).

5 Tulokset

Seuraava analyysi perustuu kolmeen listaan, jotka sisältävät kunkin aineiston (VICLFI-A, VICLFI-E ja natiiviaineisto) saneet, lemmat ja annotoinnin tuloksena saadut tagit eli tunnisteet. Olen verrannut aluksi alkeistason teksteistä tehtyä listaa natiivitekstien listaan ja myöhemmin analysoinut edistyneiden oppijoiden tuottamia tekstejä. Analyysi perustuu tilastolliseen merkitsevyyteen, joka on saatu log likelihood -testauksella (p = 0.000001, minimifrekvenssi 20). Testi huomioi sanan tai tunnisteen määrän tutkittavassa ja vertailuaineistossa ja vertaa määriä aineistojen kokoihin.

5.1 Alkeistason tekstien avainelementit

Taulukossa 1 on esitetty lajittelemattomasti alkeistason tekstien 30 merkitsevintä avainelementtiä. Tulos sisältää kieliopillisia tunnisteita (P1, PRES), numeroita (jotka usein kertovat annotaation rivin, ks. kuvio 1), lemmoja (MINÄ, HUONE) ja sananmuotoja (*on*, *menen*, *minun*). Taulukossa luetellut elementit esiintyvät virolaisilla alkeistason suomenoppijoilla siis useammin kuin verrannollisessa natiiviaineistossa. Luettelosta voi havaita, että se sisältää myös ilmauksia, jotka kuvaavat tekstien sisältöä (mm. *huone*, *syön*, *opiskella*, *Tartto*). Ne voidaan lukea tekstien sisältöavainsanoiksi.

3 Tasomääritelmät perustuvat tässä tutkimuksessa opiskelijoiden saaman yliopisto-opetuksen määrään: alkeistason tekstien kirjoittajat ovat saaneet opetusta alle 200 tuntia ja edistyneet yli 400 tuntia. ICLFIn tekstejä on alettu arvioida myös eurooppalaisen viitekehyksen mukaan, mutta tähän tutkimukseen taitotasoarviot eivät ole vielä käytössä.

TAULUKKO 1. Raakatulos VICLFI-A-aineiston avainelementeistä.

1	P1	11	TAVALLISESTI	21	3
2	KELLO	12	OPISKELLA	22	OPISKELEN
3	2	13	TARTTO	23	LUENTO
4	ON	14	1	24	SYÖDÄ
5	PRES	15	CC	25	0
6	MINÄ	16	KAKSI	26	KOTI
7	HUONE	17	ISO	27	SISKO
8	MENEN	18	ASUA	28	SÄNKY
9	SYÖN	19	MINULLA	29	TARTOSSA
10	MINUN	20	PERHE	30	OLEN

Jotta avainelementeistä saadaan täsmällisempi kuva, on analyysin tuloksena syntyneen luettelon elementtejä syytä luokitella ryhmiin. Taulukkoon 2 on koottu samasta aineistosta 20 merkitsevintä avaintunnistetta (eli tagia), sisältöavainsanaa ja mahdollista oppijankieltä luonnehtivaa avainsanaa.

TAULUKKO 2. VICLFI-A-aineiston 20 yleisintä avainelementtiä luokittain yleisyysjärjestyksessä. [1]Digitoinnissa tekstin kirjoittajan nimen korvaava tunniste.

Avaintunnisteet
< 7, P1, PRES, CC, CARD, NUM, QN, LOC, ADE, NOM, IND, NAME[1], SG, INE, DUR, NH, TMP, ORD, AD, COM

Sisältöavainsanat
HUONE, SYÖN, OPISKELLA, TARTTO, ASUA, PERHE, OPISKELEN, SYÖDÄ, LUENTO, TARTOSSA, SISKO, SÄNKY, KOTI, ASUN, VELI, KAAPPI, PERHEENI, KEITTIÖ, KOTINI, PÄIVÄNI

oppijankielen avainsanat (?)
KELLO, ON, MINÄ, MENEN, TAVALLISESTI, MINUN, ISO, KAKSI, MINULLA, OLEN, KÄYN, PIDÄN, OLLA, MENNÄ, KÄYDÄ, KAHDEKSAN, PALJON, KOTOISIN, PIENI, YKSI

Sisältöavainsanojen ryhmästä pystyy jälleen helposti päättelemään, mistä alkeistason opiskelijat ovat teksteissään kirjoittaneet: kodistaan, perheestään, opiskelustaan ja päivän tapahtumista. Kaikki aiheet ovat tyypillisiä teemoja alkeistason opiskelussa, eivätkä niitä kuvaavat avainsanat ole sinänsä oppijankielen analyysin kannalta tärkeitä. Kiinnostavampia sen sijaan ovat avaintunnisteet: ne kuvaavat kieliopillisia kategorioita, jotka esiintyvät tuotoksissa taajaan. Näitä ovat yksikköä (SG), 1. persoonaa (P1), verbimuotoja (PRES, IND), rinnastuskonjunktiota (CC), sijamuodoista adessiivia, nominatiivia ja inessiiviä (ADE, NOM, INE) sekä numeraaleja (NUM, CARD, ORD, QN) kuvaavat tunnisteet. Samalla kun nämä tunnisteet kertovat alkeisoppijoiden kielestä, ne saattavat ehkä osittain juontaa juurensa myös omaan elämänpiiriin liittyvistä kirjoitustehtävistä, joissa esimer-

kiksi yksikön 1. persoonan käyttö ovat tavallisia (ks. kuvio 2). Adessiivi näyttäisi ainakin osittain selittyvän omistusrakenteen runsaudella (*minulla on isä, äiti ja vain yksi veli*); tätä tukee myös taulukon 2 *minulla*-muoto.

SG NOM 20 **nautin** nauttia @MAIN V ACT IND **PRES SG P1** 21 usein usein tmp>20 @ADVL ADV
CC 8 **meniin** mennä cc>3 @MAIN V ACT IND **PAST SG P1** 9 takasiin takaisin goa>8 @ADVL AD
SG GEN 6 . . 7 1 **Minun** minä attr>2 @PREMOD **PRON SG P1** GEN 2 äitini äiti subj>3 @NH N SG

KUVIO 2. Esimerkki P1-tunnisteesta aineistossa.

Tunnisteiden avulla selviää myös, että rinnastus (CC) on erittäin tavallista tässä tekstimassassa. Numerot (#<7) puolestaan kertovat jossain määrin siitä, että oppijoiden tuottamissa teksteissä virkkeet ovat lyhyempiä kuin natiiviaineistossa, koska juuri pienet saneiden riveistä kertovat numerot yliedustuvat. Tämä näkyy myös tilastollisissa suureissa: alkeistason keskivirkepituus on 7 sanetta (*s*=3,67), edistyneiden 13 (*s*=7,80) ja natiivitekstien 11 (*s*=7,59). Keskihajonta (*s*) on pieni alkeistason teksteissä, mikä kertoo aineiston virkkeiden ja tekstien homogeenisuudesta pituuden suhteen.

Luettelo antaa myös viitteitä siitä, että paikan, ajan, keston ja suhteen adverbiaalit (LOC, DUR, TMP, COM) ovat alkeistasolla ylikäytettyjä samoin kuin ovat likimääräistä määrää ilmaisevat tai määrää intensifioivat adverbiaalit (AD). Aineistosta selviää, että näitä ovat mm. *yli*, *noin, hyvin* ja *erittäin*. Fi-fdg-jäsennintä ei liene kuitenkaan muokattu täysin suomen lauseenjäsennykseen sopivaksi, joten näitä tuloksia on pidettävä alustavina (ks. myöhemmin myös negatiivisten avaintunnisteiden käsittelyä).

Missä määrin taulukon 2 alimman rivistön avainsanat ovat tosiasiassa oppijankielen piirteitä eivätkä tehtävästä johtuvia, tarvitsee tähän artikkeliin mahtuvaa tarkastelua tarkemman analyysin. Johtuuko MINÄ-sanan ylikäyttö tehtävänannosta vai onko kysymys eksplisiittisestä rakenteesta, jossa pronomini on kirjoitettu näkyviin vaikka persoona tuleekin ilmi verbistä, tai äidinkielen aiheuttamasta transferista? Ovatko MENNÄ-sanat perua virolaisten suomenoppijoiden vaikeudesta erottaa *mennä*- ja *lähteä*-verbit toisistaan tai mahdollisesta *mennä*-verbien suosimisesta astevaihtelullisen *lähteä*-verbin sijaan?[4] Entä johtuvatko OLLA-verbin muodot rajallisesta verbivalikoimasta vai osittain esimerkiksi yllä mainitusta omistusrakenteesta? Aiemmat tutkimukset sen sijaan antavat jo viitteitä siitä, että PALJON (Kallioranta 2009) ja KELLO (Jantunen 2009) esiintyvät oppijantuotoksissa taajaan ja ovat vahvoja kandidaatteja oppijankielen avainsanoiksi.

VICLFI-A-aineiston negatiiviset eli aliedustuneet avainelementit on vuorostaan lueteltu taulukossa 3. Verbitunnisteista merkitsevimmin aliedustuneita ovat imperfekti (PAST), partisiippi (PCP) ja passiivi (PASS). Myös infinitiivin (INF) muodot (F3, F1, F4

4 Sähköpostikeskustelu Kristi Pällinin kanssa 25.1.2011.

eli MA- ja A-infinitiivit ja *on tekeminen* -nesessiivirakenne) ovat harvinaisia, samoin ovat konditionaali (CND) ja imperatiivi (IMP). Tunnisteet AUX ja NEG viittaavat vähäisiin kieltomuotoihin (AUX= kieltoverbi apuverbinä, NEG= pääverbin kieltomuoto). Myös runsas joukko sijamuodoista on aliedustunut alkeistason teksteissä: Merkittävimmin on aliedustunut genetiivi, ja myös toinen kieliopillinen sija, partitiivi, esiintyy harvoin vironkielisten alkeistason teksteissä. Myös paikallissijat illatiivi, elatiivi, allatiivi ja essiivi, sekä translatiivi ja harvinaiset abessiivi ja instruktiivi ovat aliedustuneita.

TAULUKKO 3. VICLFI-A-aineiston negatiiviset avainelementit yleisyysjärjestyksessä.

Avaintunnisteet
(>10), PAST, PL, PCP, PASS, GEN, ATTR, PREMOD, ILL, INF, P2, A, CS, OBJ, ESS, AUX, CND, NEG, ALL, F3, GOA, TRA, PTV, INS, CLI, P3, F1, PREMARK, IMP, MAN, ELA, PROP, V, MOD, F4, ABBR, PRON, SOU, PM, ABE, RSN
Avainsanat
SE, ETTÄ, OLI, EI, -KIN, KUIN, SANOA, JOKA, VOIDA, KUN, NIIN, TÄMÄ , VAIKKA , SITÄ, OLLUT, SAADA, MUU, NAINEN, TULLA, NÄHDÄ, ASIA, MIKÄ, OLIVAT, PÄÄ, TIETÄÄ, TAAS, MITÄ, VAIN, ENÄÄ, JO, KOKO, SIITÄ, IHMINEN, KÄSI, NYT, KAIKKI, SILLÄ, JOKIN, VASTA

Sanaluokista alkeistason oppijat käyttävät vähän adjektiiveja (A), verbejä (V) ja pronomineja (PRON), ja erityisesti pronomineja ja verbejä on sekä lemmoina että taivutusmuotoina myös negatiivisten avainsanojen joukossa. Huomionarvoista on, että kieltoverbi *ei* (eli AUX-tunnuksen reaalistuma) on myös luettelossa. Alistuskonjunktioiden[5] aliedustus näkyy konjunktioiden *että*, *kun*, *vaikka* vähäisyytenä sekä CS-tunnisteen (alistuskonjunktio) että PREMARK-tunnisteen aliedustumisena: paitsi että jälkimmäinen on adpositiolausekkeen määritteen[6] syntaktinen tunniste (eli tunnisteen aliedustus voi kertoa myös adpositiolausekkeiden vähäisestä määrästä), jäsennin merkitsee sillä myös sivulauseen aloittavan alistuskonjunktion. Sivulauseiden osuus lienee siis pieni, mikä puolestaan liittyy edellä käsiteltyyn virkepituuteen. Tätä ja virkkeiden lyhyyttä tukee myös se, että negatiivisten avaintunnisteiden listassa on paljon kymmentä suurempia numeraaleja (#>10), eli oppijoiden teksteissä ole ollut kovin paljon yli kymmensanaisia virkkeitä analysoitavana.

Avaintunnistelista sisältää myös jonkin verran syntaktisiin suhteisiin liittyvää tietoa: Tavan (MAN), syyn (RSN) ja lokatiivisten (GOA, SOU) adverbiaalien esiintyminen listalla viitannee siihen, että ne olisivat teksteissä aliedustuneita. PREMOD- ja ATTR-tunnisteet viittaavat puolestaan sekä nominien että adverbien määritteiden vähyyteen

5 Jäsennin käyttää käsitettä alistuskonjunktio, VISK (§ 816) puolestaan pääsääntöisesti käsitteitä yleis- ja adverbiaalikonjunktiot.

6 Jäsennin analysoi adpositiolausekkeen pääsanaksi nominin ja määritteeksi adposition, VISK:n (§ 687) tuoreempi tulkinta pitää pääsanana adpositiota.

aineistossa. Lyhyt analyysi kuitenkin paljastaa, että automaattinen analyysi ei ole aukoton: esimerkiksi elatiivimuotoiset sanat ovat usein saaneet lokatiivisen lähdettä merkitsevän tunnisteen (SOU), vaikka kyse ei tällaisesta olisikaan (*pidän serkuistani* vs. *kotoisin Virosta*). Lisäksi näitä syntaktisiin suhteisiin viittaavia tunnisteita ei ole disambiguoinnin yhteydessä tarkistettu, joten lauseenjäsennyksen osalta analyysi on suuntaa antava. Oman ryhmänsä avainsanojen joukossa muodostavat myös adverbit, joskaan ne eivät olekaan luokkana aliedustettu; erityisesti aikaa merkitsevät adverbit *taas*, *enää*, *jo*, *nyt* ja *vasta* ovat aliedustettuja. Tosin sanoja olisi lähestyttävä monifunktioisina ja -merkityksisinä ja niiden käyttötapojen eroja olisi tutkittava tarkemmin.

5.2 Edistyneen tason avainelementit

Seuraavaksi tarkastelen edistyneiden suomenoppijoiden tekstejä samaan tapaan kuin edellä alkeisopiskelijoiden tekstejä. Tulokset on esitetty tiivistetysti jälleen 20 merkitsevimmän – sekä positiivisen että negatiivisen – avainelementin osalta taulukossa 4. Positiiviset avaintunnisteet ovat osittain samoja kuin alkeistasolla (PRES, NUM, CARD, ORD, INE, AD), eli tässä on viitteitä siitä, että tietyt piirteet saattaisivat yliedustua oppijankielessä tasosta riippumatta. Huomion arvoista on, että monet edistyneiden aineiston yliedustuneista tunnisteista on sellaisia, jotka olivat alkeistasolla aliedustuneita (vrt. taulukko 3). Koska näitä ovat mm. partitiivi (PTV), lauseiden alistussuhteiden ilmaiseminen (CS) ja A-infinitiivin käyttö (F1), ei voitane väittää, että kyse olisi esimerkiksi edistyneellä tasolla opitun asian yliedustumisesta, koska ainakin nämä asiat ovat esillä jo melko varhaisessa vaiheessa opintoja. Yli kymmentä suuremmat numerot (#>10) johtunevat osittain pitkistä virkkeistä. Ne voivat olla perua siitä, että aineisto sisältää tieteellistyyppisiä asiatekstejä (referaatteja, esseitä ja tutkielmia; ks. positiivisia sisältöavainsanoja) suhteessa enemmän kuin vertailuaineisto, jolloin pitkät virkkeet korostunevat aineistossa (vrt. aiemmin esitettyihin virkepituuksiin). Myös aliedustuneet avaintunnisteet ovat osittain samoja kuin edellä. Muun muassa verbitunnisteita PAST, PCP ja NEG on tässäkin aineistossa vähemmän kuin natiiviteksteissä, samoin liitepartikkeleita (CLI), mikä näkyy negatiivisten avainsanojenkin listassa (*-kin*, *-kO*). Listassa on kuitenkin myös tunnisteita, kuten PROP (erisnimi), POSS (omistusliite) ja IMP (imperatiivi), joita ei esiintynyt alkeistekstien negatiivisessa avaintunnistelistassa.

TAULUKKO 4. VICLFI-E-aineiston positiiviset ja negatiiviset avainelementit (maks. 20).

positiiviset avaintunnisteet
#>10, PRES, NUM, F1, INE, AD, CARD, ADV, CS, ORD, ATTR, PM, ADVL, MOD, RSN, PTV
positiiviset sisältöavainsanat
KIELI, KIELEN, OPPIMINEN, OPPIJA, SUOMI, OPPIA, VERBI, VIRO, KIELTÄ, SUOMEN, OPETUS, VIRON, OPPIJAN, ÄIDIN, OMAKSUA, KIELENÄ, OPISKELU, SANA, KONTRASTIIVINEN, OPISKELLA
positiiviset oppijankielen avainsanat (?)
KOSKA, ON, JOKIN, TAKIA, MYÖS, PALJON, ETTÄ, JNE, ELI, JOHONKIN, MIELESTÄNI, MITEN, ERI, RIIPPUU, RIIPPUA, JOKU, ENEMMÄN, AIKA, KIINNOSTAVA, HYVIN
negatiiviset avaintunnisteet
PAST, PROP, P2, CLI, POSS, P3, PCP, IMP, PL, SUBJ, GOA, NOM, ALL, V, ACT, A, NEG, IND, ILL, SG
negatiiviset oppijankielen avainsanat (?)
HÄN, OLI, -KIN, KUIN, -KO, ME, OLIVAT

Positiiviset oppijankielen avainsanaehdokkaat ovat varsin erilaisia kuin alkeistason teksteissä: luettelossa on paljon kieliopillisia sanoja, kuten konjunktioita, pronomineja ja postpositioita. Kun esimerkiksi viimeksi mainittuihin kuuluvan *takia*-postposition esiintymismäärää verrataan sen synonyymien *vuoksi*, *tähden*, *johdosta* ja *ansiosta* käyttöön, havaitaan, että *takia* on aineistossa lähes yksinomainen; verrannollisessa natiiviaineistossa käyttö jakautuu tasaisemmin ja *vuoksi* on taajakäyttöisempi kuin *takia*. Oman mielenkiintoisen ryhmänsä muodostavat kvantifiointia osoittavat ilmaukset *jokin*, *johonkin*, *joku*, *paljon*, *enemmän*, *aika* ja *hyvin*. Sanojen osuus positiivisten avainsanojen joukossa on suuri (mainittakoon listan ulkopuolelta vielä sanat *erittäin* ja *eniten*), mikä viittaisi kielenoppijoiden viehtymykseen kvantifioida ilmiöitä joko indefiniittisillä pronomineilla, kvanttoriadverbeilla tai vahventavilla astemääritteillä (aiemmista kvantifiointiin liittyvistä havainnoista ks. Hasselgren 1994; Granger 1998; Jantunen 2008; Kallioranta 2009). Negatiivisten avainsanojen määrä on pieni; *hän-* ja *me*-pronominien aliedustus voi johtua fiktiivisten tekstien vähyydestä VICLFI-E-aineistossa.

6 Pohdintaa

Artikkelin tavoitteena oli testata korpusvetoisen avainsana-analyysin soveltumista annotoidun oppijankielimateriaalin tutkimiseen. Edellä olevat havainnot ovat alustavia, ja ne herättävät enemmän kysymyksiä kuin antavat vastauksia. Yleisesti voidaan kuitenkin sanoa, että annotoidun aineiston analysoiminen *KeyWords*-työkalulla antaa kielen omi-

naispiirteistä kattavamman kuvan kuin käsittelemättömän datan analysoiminen, jolloin joudutaan tyytymään juuri avain**sanoihin** tai -**sananmuotoihin**. Annotoidusta ja lemmatisoidusta aineistosta työkalu puolestaan nostaa esiin yli- ja aliedustumat niin sananmuoto- kuin lemmatasolla sekä kieliopillisina kategorioina ja syntaktisina suhteina.

Raakateksti- ja virhekoodattua aineistoa käytettäessä joudutaan kielen ilmiöitä lähestymään leksikaalis- tai virhelähtöisesti, ja kieliopillisten piirteiden tarkastelu on haastavaa ja vaatii kekseliäisyyttä. Juuri tähän annotaatio tuo helpotusta. Kun aineisto on annotoitu ja annotaatio korjattu ja disambiguoitu, voi tutkija etsiä esimerkiksi edellä käsitellyt rinnastus- ja alistuskonjunktiot tunnisteiden avulla sen sijaan että hakisi ne aineistosta yksitellen. Kiistattomasti hakuvaihe nopeutuu silloin, kun tutkittavana ovat esimerkiksi verbien ja nominien taivutusmuodot, joiden hakeminen raakatekstistä on hyvin haastavaa – mitä enemmän morfologista variaatiota tutkittavaan luokkaan sisältyy, sitä vaativampaa annotoimattoman materiaalin analysoiminen on.

Etenkin alkeistason oppijoiden tuottamien tekstien automaattinen analyysi on työlästä oikeinkirjoitus- ja taivutusvirheiden vuoksi, eivätkä epätavalliset sanajärjestykset, sijavalinnat tai innovatiiviset sanastolliset ad hoc -ilmaukset tee analyysista mutkattomampaa. Tällaiset karikot voidaan kuitenkin välttää ennen analysaattoriin syöttämistä tapahtuvalla esikorjauksella ja analyysin jälkeen virheiden palauttamisella aineistoon. Työläydestä huolimatta prosessi on kannattavaa, sillä lemmatisointi ja annotointi monipuolistavat oppijankieliaineistojen käytettävyyttä.

Vertailevassa (korpus)tutkimuksessa aineistojen on oltava edustavia ja taustamuuttujiensa osalta luotettavasti dokumentoituja. Tässä tutkimuksessa käytetyn ICLFI-aineiston taitotasot on määritelty yliopistossa annettujen kontaktituntien mukaan. Opetusmäärään perustuva taitotasojaottelu voi kuitenkin antaa tasosta erilaisen tuloksen kuin esimerkiksi eurooppalaisen viitekehyksen mukaan tehty tasoarviointi, ja arviointikriteerien erilaisuus voi myös vaikuttaa tuloksiin (Spoelman 2010). Yllä tehtyä analyysia olisikin syytä tarkentaa eurooppalaisen viitekehyksen taitotasoasteikon mukaan jaotellun materiaalin avulla ja verrata tuloksia edellä saatuihin. Kuusiportainen asteikko antaa myös mahdollisuuden tarkastella tekstejä tasoittain tässä esitettyä tarkemmin. Avainsanamenetelmä voikin paljastaa runsaasti tietoa kehittymisestä tasolta toiselle siirryttäessä sekä leksikaalisista ja kieliopillisista piirteistä taitotasoittain.

Oppijankielen aineistoja verrattiin edellä tasoittain natiivikielen aineistoon, jolloin oppijankielen avainsanaisuus eri taitotasoilla ilmeni suhteessa natiivituotokseen. Toinen vaihtoehto on verrata oppijankielen aineistoja keskenään joko asettamalla vertailuun taitotasot (esim. B1 vs. B2) tai vertailemalla yhtä taitotasoa monitaitotasoiseen aineistoon (esim. B1 vs. ICLFI-A1–C2 tai B2 vs. ICLFI-C1–C2). Taitotasoittainen vertailu natiiviaineistoon noudattaa Grangerin (1996, 2004) kontrastiivisen välikielen analyysin (*Contrastive Interlanguage Analysis*) mallia, jossa vertailtavien aineistojen tärkein muut-

tuja on oppijan äidinkieli. Jälkimmäinen on puolestaan sovellus käännöstutkimuksessa harjoitetusta vertailusta, jossa vertailuaineistona on käytetty monilähdekielisiä aineistoja (ks. esim. Jantunen 2004). Ensimmäinen yksitaitotasovertailu antaa tietoa yhdestä taitotasosta suhteessa toiseen taitotasoon, jälkimmäinen taas yhden taitotason suhteesta koko oppijankieliaineistoon tai johonkin laajempaan taitotasoluokkaan. Menetelmiä yhdistämällä saadaan kielitaidon kehittymisestä tasoittain tarkempi kuva.

Kirjallisuus

Barlow, M. 2005. Computer-based analyses of learner language. Teoksessa R. Ellis & G. Barkhuizen *Analysing Learner Language*. Oxford: Oxford University Press, 335–357.

Berber-Sardinha, T. 1999. Using key words in text analysis: practical aspects. *DIRECT Papers* 42, 1–9 [online, luettu 19.1.2011]. Saatavissa: www2.lael.pucsp.br/direct/DirectPapers42.pdf.

de Haan, P. 2000. Tagging non-native English with the TOSCA-ICLE tagger. Teoksessa C. Mair & M. Hundt (toim.) *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi, 69–79.

Ellis, R. & G. Barkhuizen 2005. *Analysing Learner Language*. Oxford: Oxford University Press.

Fi-fdg: Connexor Machinese Syntax for Finnish [online]. CSC–Tieteen tietotekniikan keskus [luettu 19.1.2011]. Saatavissa: www.csc.fi/english/research/software/fi-fdg.

Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. Teoksessa K. Aijmer, B. Altenberg & M. Johansson (toim.) *Languages in Contrast*: *Text-based Cross-linguistic Studies*. Lund: Lund University Press, 37–51.

Granger, S. 1997. On identifying the syntactic and discourse features of participle clauses in academic English: Native and non-native writers compared. Teoksessa J. Aarts, I. de Mönnink & H. Wekker (toim.) *Studies in English Language and Teaching*. Amsterdam: Rodopi, 185–198.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. Teoksessa A. P. Cowie (toim.) *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 145–160.

Granger S. 1999. Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. Teoksessa H. Hasselgård & S. Oksefjell (toim.) *Out of Corpora: Studies in Honour of Stig Johansson*. Amsterdam: Rodopi, 191–202.

Granger, S. 2002. A bird's-eye view of learner corpus research. Teoksessa S. Granger, J. Hung & S. Petch-Tyson (toim.) *Computer Learner Corpora: Second Language Acquisition and Foreign language Teaching*. Amsterdam: John Benjamins, 3–33.

Granger, S. 2004. Computer learner corpus research: current status and future prospects. Teoksessa U. Connor & T. Upton (toim.) *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi, 123–145.

Hardie, A. & T. McEnery 2010. On two traditions in corpus linguistics, and what they have in common. *International Journal of Corpus Linguistics,* 15 (3), 384–394.

Hasselgren, A. 1994. Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4 (2), 237–260.

Heikkinen, V., M. Lounela & E. Voutilainen (tulossa). Automaattinen analysaattori tekstilajitutkimuksessa. Teoksessa V. Heikkinen, E. Voutilainen, P. Lauerma, M. Lounela & U. Tiililä (toim.) *Tekstilajitutkimuksen käsikirja*. Helsinki: Gaudeamus.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: CUP.

Jantunen, J. H. 2004. Untypical patterns in translations: Issues on corpus methodology and synonymity. Teoksessa A. Mauranen & P. Kujamäki (toim.) *Translation Universals: Do They Exist*? Amsterdam: John Benjamins, 101–126.

Jantunen, J. H. 2008. Haasteita oppijankielen korpusanalyysille: oppijankielen univer-saalit. Teoksessa P. Eslon (toim.) *Õppijakeele analüüs: Võimalused, probleemid, vajadused*. Tallinn: Tallinna ülikool, 67–92.

Jantunen, J. H. 2009. Ei pelkästään mielikuvituksen puutteen vuoksi: kieliaineistojen systemaattinen käyttö kielentutkimuksessa. *Virittäjä* 113, 101–113.

Jantunen, J. H. (tulossa). Korpusvetoinen tekstilajianalyysi: sanalistat ja genreavainsanat. Teoksessa V. Heikkinen, E. Voutilainen, P. Lauerma, M. Lounela & U. Tiililä (toim.) *Tekstilajitutkimuksen käsikirja*. Helsinki: Gaudeamus.

Jantunen, J. H. & S. Piltonen 2009. Oppijansuomen ja -viron sähköiset tutkimusaineistot. *Virittäjä* 113, 449–458.

Kallioranta, O. 2009. Paljon-*adverbin kollokointi oppijansuomessa. Korpusvetoinen tutkimus*. Painamaton pro gradu -tutkielma. Oulun yliopisto. Saatavissa: www.oulu.fi/hutk/sutvi/oppijankieli/tutkimus/

Latomaa, S. 1996. Matkalla uuteen kieleen. Teoksessa H. Ruuska & S.-M. Tuomi (toim.) *Moneja baareja: Tiellä toimivaan kaksikielisyyteen*. Helsinki: Äidinkielen Opettajain Liitto, 97–106.

Leech, G. 2005. Adding linguistic annotation. Teoksessa M. Wynne (toim.) *Developing Linguistic Corpora: A Guide to Good Practice* [online]. Oxford: Oxbow Books [luettu 21.1.2011]. Saatavissa: www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm.

Mauranen, A. 2000. Strange strings in translated language: A study on corpora. Teoksessa M. Olohan (toim.) *Intercultural Faultlines*: *Research Models in Translation Studies I. Textual and Cognitive Aspects*. Manchester: St. Jerome, 119–141.

Meunier, F. 1998. Computer tools for the analysis of learner corpora. Teoksessa S. Granger (toim.) *Learner English on Computer*. London: Longman, 19–37.

Nesselhauf, N. 2005. *Collocations in a learner corpus*. Amsterdam: Benjamins.

Scott, M. 2008. Developing WordSmith. *International Journal of English Studies,* 8 (1), 95–106.

Scott, M. & C. Tribble 2006. *Textual Patterns*: *Key Words and Corpus Analysis in Language Education*. Studies in corpus linguistics 22. Amsterdam: John Benjamins.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Spoelman, M. 2010. The use of the partitive case in Finnish learner language: The operationalization of foreign language proficiency. Esitelmä työpajassa *Metodit L2-korpustutkimuksessa, AFinLAn syyssymposiumi 12.–13.11.2010, Vaasa*.

Tesnière, L. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tummers, J., K. Heylen & D. Geeraerts 2005. Usage-based approaches in cognitive linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory,* 1 (2), 225–261.

van Rooy, B. & L. Schäfer 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. Teoksessa D. Archer, P. Rayson, A, Wilson & T. McEnery (toim.) *Proceedings of the Corpus Linguistics 2003 Conference Lancaster University (UK), 28-31 March 2003*. Vol. 16. Lancaster: UCREL, 835–844.

VISK = Hakulinen, A., M. Vilkuna, R. Korhonen, V. Koivisto, T. Heinonen & I. Alho 2004. *Iso suomen kielioppi*. Verkkoversio [online]. Helsinki: SKS [luettu 21.1.2011]. Saatavissa: http://scripta.kotus.fi/visk URN:ISBN:978-952-5446-35-7