doi: 10.5128/LV23.04

Korpuspohjaista oppijansanakirjaa tekemässä: esimerkkinä *ConLexis*

JARMO HARRI JANTUNEN
Jyväskylän yliopisto

MARJO KUMPULAINEN
Oulun Aikuiskoulutuskeskus

TANJA TAMMIMIES
Oulun Aikuiskoulutuskeskus

TEEMU TOKOLA
Oulun yliopisto

Tiivistelmä. Artikkelimme käsittelee korpuspohjaisia oppijansanakirjoja. Sähköisiin aineistoihin ja kielenkäytöstä hankittuun tutkimukselliseen tietoon perustuvat sanakirjat ovat edelleen harvinaisia, ja kielenoppijoille niitä on tarjolla hyvin vähän. Suomesta tällaisia sanakirjoja ei ole ollut lainkaan. Esittelemme yleisesti oppijoille suunnattujen sähköisten sanakirjojen vaatimuksia ja tarkemmin uuden tekeillä olevan ConLexis-verkkosanakirjan suunnittelutyötä, tavoitteita ja sisältöä. ConLexis on suunnattu B1-tason ja sitä edistyneemmille kielenoppijoille, mutta sitä voivat opettajat käyttää opetuksensa tukena alemmillakin taitotasoilla. Sana-artikkeleissa esitetty tieto perustuu laajoihin korpusanalyyseihin, ja sanakirjan näkökulma kielenkäyttöön on vahvasti fraseologinen. Sen rinnalle on tekeillä ConPraxis-harjoituskokonaisuus, joka muodostaa yhdessä sanakirjan kanssa itseopiskelupaketin.

Avainsanat: oppijansanakirjat; verkko-oppimateriaalit; korpussanakirjat; fraseologia; kollokaatiot; synonyymit; antonyymit

1. Johdanto

Sähköisiin tekstimassoihin perustuvat sanakirjat yleistyvät jatkuvasti maailmalla. Tyypillistä näille korpuspohjaisille sanakirjoille on, että sana-artikkelien kirjoittaja nojautuu aineiston esiin nostamaan tietoon sanojen käytöstä. Perinteisempi, intuitiiviseen näkemykseen perustuva kuvaus saa tällaisissa sanakirjoissa aiempaa huomattavasti pienemmän roolin. Luonnollisestikaan aineistopohjaisuus ei vähennä sanakirjan tekijän merkitystä sana-artikkelien luojana, mutta nyt tekijä tarvitsee hieman erilaista tietoa ja taitoa työssään. Verrattuna ns. perinteisen sanakirjan kirjoittamiseen korpuspohjaisessa sanakirjatyössä korostuvat erityisesti korpus- ja tilastollisten menetelmien ja kieliteknologian hallinta, lisäksi vaaditaan usein uudenlaista näkemystä sanojen käyttötavoista ja kielenkäytöstä yleensäkin.

Marja Järventausta nostaa esiin tarkkanäköisessä artikkelissaan "Kakkossuomen perussanakirja" (2009) sen, että kasvavasta opiskelijamäärästä huolimatta suomen kielen oppijoille tarkoitettua sanakirjaa ei suomen kielestä ole juuri tehty. Artikkelin kirjoittamisen jälkeen on kuitenkin ilmestynyt *Suomen kielen sanakirja maahanmuuttajille* (Saarikalle & Vilkuna 2010). Sanakirjassa on noin 3 500 sanan verran suomen kielen keskeistä sanastoa. Sanat selitetään yksinkertaisesti mutta kokonaisin virkkein. Sanan taivutusta tai käyttöön liittyviä muotteja ei eksplisiittisesti näytetä, mutta esimerkit kuvastavat sanan tyypillisiä käyttöyhteyksiä. Lisäksi sanan yhteydessä on tietoa idiomaattisesta käytöstä, synonyymeista ja antonyymeista. Käytännönläheisyydessään sanakirja palvelee esimerkiksi Suomeen muuttaneita heidän arjessaan jo kieliopintojen alkuvaiheessa.

Tätä aikaisemmin on julkaistu *Suomen sanakirja opiskelijoille ja ulkomaalaisille, Finnish Dictionary for Students and Foreign Learners* (Nurmi 2009), joka on alkuaan tarkoitettu hakuteokseksi vieraskielisille suomen opiskelijoille (1. painos 1999 nimellä *Suomen kielen sanakirja ulkomaalaisille, Finnish Dictionary for Foreigners*, 2. painos 2004 nimellä *Nykysuomen sanasto*), mutta kirjan tekijät ovat myöhemmin katsoneet

sen soveltuvan myös laajempaan käyttöön, muun muassa koululaisille ja opiskelijoille. Sanakirjassa on yli 17 000 sana-artikkelia. Taivutusmuodot ja tavutus ovat keskeisellä sijalla sanojen esittelyssä. Käyttöyhteys esitellään kokonaisin esimerkkivirkkein. Kirja sisältää myös idiomaattista kieltä. Vaikka sanojen taivutus esitellään perusteellisesti, Bessonoffin (2000: 318) arvion mukaan ulkomaalaisnäkökulma ei juurikaan näy sanaselityksissä tai lause-esimerkeissä.

Järventaustan mukaan sanakirjojen tarve on suuri, ja hankkeeseen pitäisi tarttua S2-opettajien yhdessä leksikografien ja korpuslingvistien kanssa. Saarikallen ja Vilkunan (2010) sanakirja vastaa tarpeeseen osittain. Sanakirja sopii kielenoppijalle ensimmäiseksi yksikieliseksi sanakirjaksi, mutta kielitaidon kehittyessä sanakirja käy auttamatta suppeaksi sanemäärältään ja fraseologiselta informaatioltaan. Artikkelissamme käsittelemme aluksi yleisesti korpuspohjaisia sanakirjoja ja verkkosanakirjan mahdollistamaa esittämistapaa. Tämän jälkeen esittelemme yhtä suomen oppijoille tarkoitettua verkkosanakirjaa *ConLexista*[1], jota on työstetty tähän mennessä neljän vuoden ajan. Esittelemme *ConLexis*-sanakirjan taustaperiaatteet, tekotavat ja tekniset ratkaisut sekä kerromme sana-artikkelien rakenteesta. Lopuksi esittelemme vielä *ConPraxis*-harjoituksia, jotka on tehty sanakirjan rinnalle sanakirjan käyttöä ja opiskelua tukemaan.

2. Oppijansanakirjat

Niin sanottujen pienten kielten opiskelijoiden kohtalona on joutua turvautumaan syntyperäisille kielenkäyttäjille tarkoitettuihin sanakirjoihin, sillä harvoista kielistä on saatavilla laajoja, nimenomaan kielenoppijoille tarkoitettuja yksikielisiä sanakirjoja; englanti ja muutamat muut kielet ovat poikkeuksia. Suomen kielen opiskelijoiden osalta tilanne on kutakuinkin surkea.

Sanakirjojen käyttöä pidetään kielenoppimisessa pääasiassa hyödyllisenä. Eastin (2008: 2–3, 7) mukaan sanakirjojen avulla saavutetaan

[1] http://wiki.virtues.fi/conlexis (10.7.2013).

autenttisuutta ja varmuutta kielenkäyttöön, toisaalta kuitenkin pelätään, että sanakirjat sitovat liikaa kielenkäyttäjää, johtavat virhetulkintoihin ja vievät yksinkertaisesti liikaa aikaa tuotosprosessissa.[2] Joka tapauksessa sanakirjat ovat kielenoppijoiden ja -opettajienkin tärkeimpiä referenssityökaluja. On tavallista, että sanakirjojen käyttö muuttuu kieliopintojen edetessä: alkeistason opiskelijat etsivät taajaan vastineita äidinkielensä ilmauksille kaksikielisistä sanakirjoista, yksikielisten sanakirjojen käyttö puolestaan lisääntyy kielitaidon karttuessa (Järventausta 2009: 90). Vaikka yksikielisten sanakirjojen käyttöä suositellaankin yleensä kielenoppijalle (ks. Thompson 1987), varsinkaan opintojen alkuvaiheessa kielitaito ei välttämättä riitä ymmärtämään yksikielisen sanakirjan tarjoamaa informaatiota, sillä sanaselitykset ovat liian haasteellisia (Järventausta 2009: 89). Tällöin opiskelija tyytyy helposti yksinkertaistettuihin vastaavuussuhteisiin, joita kaksikieliset sanakirjat tarjoavat. (Sanakirjojen eroista ja eduista ks. esim. Thompson 1987; Eats 2008; Järventausta 2009).

Suomen kielen yleissanakirjoista ei yksikään sovellu sellaisenaan oppijan tarpeisiin. Ne on laadittu lähinnä syntyperäiselle suomenpuhujalle (ks. kuitenkin Saarikalle & Vilkuna 2010), eivätkä myöskään pienet kaksikieliset, lähinnä matkailun tueksi tarkoitetut taskusanakirjat palvele kovin hyvin suomenoppijaa (ks. Järventausta 2009). Millainen kielenoppijoille tarkoitettu sanakirja sitten yleensä on? Rundellin (1998: 316–318) ja Järventaustan (2009) mukaan oppijansanakirjalle on tyypillistä syntyperäisille tarkoitettuihin sanakirjoihin verrattuna

1) pienempi sana-artikkelimäärä,
2) laajempi kieliopillinen kuvaus,
3) laajempi ja monipuolisempi esimerkistö,
4) tarkempi fraseologinen kuvaus.

Lisäksi Atkins (2008) ja Järventausta (2009) painottavat, että oppijansanakirjan olisi

[2] Havainnot koskevat sanakirjojen käyttöä testitilanteissa, mutta ovat myös yleistettävissä muuhunkin sanakirjojen käyttöön.

5) annettava yksityiskohtaista tietoa sanojen käyttöalasta (mm. tyyli, tekstilaji ja fraseologia),
6) katettava kielen keskeinen sanasto,
7) kerrottava sanojen suhteesta muihin ilmauksiin (merkitys- ja syntagmaattiset suhteet),
8) esitettävä sanojen tärkeimmät taivutusmuodot,
9) pyrittävä yksinkertaisiin ja pelkistettyihin selitteisiin,
10) pohjauduttava laajoihin sähköisiin tekstimassoihin.

Mitkään edellä mainituista ominaisuuksista tai vaatimuksista eivät ole uusia, mutta etenkin vaatimukset (5–10) toteutuvat sanakirjoissa vaihtelevasti. Englanninopiskelijoille tarkoitetut sanakirjat ovat kuitenkin olleet tienraivaajia ja suunnannäyttäjiä monessa suhteessa: esimerkiksi Cobuild (1987) ja MEDAL (*MacMillan English Dictionary for Advanced Learners*, 2007) ovat molemmat frekvenssitietoon perustuvia korpuspohjaisia sanakirjoja, joista jälkimmäinen sisältää myös erittäin mielenkiintoisen oppijankorpukseen perustuvan katsauksen siihen, miten englanninoppijoiden sanasto ja sen fraseologisuus poikkeaa natiivinkaltaisesta käytöstä.

3. Verkko- ja korpussanakirjat sekä fraseologisuuden esittäminen

Nykyisin ilmestyvistä sanakirjoista yhä suurempi osa on sähköisiä sanakirjoja, ja niiden käyttö kasvaa jatkuvasti nopeuden ja helppouden vuoksi sekä saatavuuden parannuttua. Edellä mainituista englanninoppijoille tarkoitetuista korpuspohjaisista sanakirjoista (Cobuild ja MEDAL) on olemassa myös sähköinen versio; suomen kielestä sen sijaan ei luonnollisestikaan ole sähköistäkään oppijansanakirjaa vielä olemassa. Järventausta (2009) toivookin, että jos oppijansanakirja suomen kielestä joskus tehdään, se ilmestyisi myös sähköisenä – ja perustuisi korpusmateriaaliin.

Monet tällä hetkellä verkossa tarjolla olevista sanakirjoista ovat ilmaisia kaksikielisiä sanakirjoja. Ne esittelevät yleensä yksinkertaistaen

vastaavuussuhteita ja ehkä myös ääntämis- ja taivutusohjeita, mikä riittääkin monelle sanakirjankäyttäjälle. Jos kielenoppijan (tai esimerkiksi kääntäjän) tavoitteena on kuitenkin tuottaa autenttiselta tuntuvaa natiivinkaltaista kieltä, osoittautuvat sanavastaavuuksiin tyytyvät sanakirjat yleensä nopeasti riittämättömiksi: niissä annetaan harvoin tietoa sanojen ja ilmausten todellisesta käytöstä. Sanojen "käyttöohjeisiin" kuuluu yhtä hyvin tieto rekisteri- ja tekstilajipreferensseistä, tyyliväristä, konnotatiivisista merkityksistä kuin myös tieto sanojen tyypillisistä leksikaalis-kieliopillis-semanttisista esiintymisympäristöistä. Verkkosanakirjat antavat mahdollisuuden hyvin monipuolisille sana-artikkeleille, sillä hyperlinkkeihin voidaan ilman painetun sanakirjan tilarajoituksia kuvata monenlaista tietoa sanojen käytöstä sekä lisätä erilaisia taulukoita ja tietoruutuja; vain sanakirjantekijöiden mielikuvitus on tässä rajana.

Oma ryhmänsä sanakirjojen joukossa ovat niin sanotut korpussanakirjat. Ne perustuvat todelliseen, laajaan sähköiseen tekstimateriaaliin, jota on käytetty sanakirjatyössä hyväksi nimenomaan systemaattisesti ja analyyttisesti. Korpuksen käyttö ei tällaisessa sanakirjassa voi rajoittua siis pelkästään sopivien käyttöesimerkkien poimintaan korpusaineistosta, vaan systemaattisuus tarkoittaa koko sanakirjatyön läpäisevää analyyttistä sanaston käytön kartoitusta. Käytännössä tämä tarkoittaa ensiksikin, että mukaan tulevat hakusanat valitaan taajuuden perusteella korpuksesta tehdyistä (tai korpuksiin pohjautuvista valmiista) sanalistoista. Vaihtoehtoisesti sen mukaan, mikä on sanakirjan kohderyhmä, hakusanat voidaan valita kohderyhmän tarpeiden mukaan taajuuden perusteella suoraan korpuksesta.[3] Seuraavina vaiheina korpustyössä ovat sanojen käytön tutkiminen, merkitysten ja funktioiden selvittäminen ja viimein käyttöä parhaiten kuvaavien esimerkkien poimiminen. Suurin ja aikaa vievin työvaihe on sanojen käytön analyysi: mitä tarkempaan kuvaukseen pyritään, sitä enemmän yhden sana-artikkelin

[3] Tämä vaatii yleensä aineiston lemmatisointia, jos käytössä on lemmatisoimaton raakatekstikorpus. Sanakirjoissahan hakusanana tarjotaan yleensä kaikki taivutusmuodot sisältävä abstraktio, ellei kyseessä ole tiettyyn taivutusmuotoon leksikaalistunut sana. (Ks. Rundell & Kilgarriff 2011: 264.)

rakentaminen vie luonnollisesti aikaa. Rundell ja Kilgarriff (2011: 261) ovat kuvanneet korpussanakirjan tekemisen monipolvisuutta ja työläyttä esitellessään korpussanakirjojen historiallista kehitystä. Olemme seuraavassa täydentäneet heidän luetteloaan oman kokemuksemme perusteella (luettelo ei pyri olemaan täydellinen prosessin kuvaus eikä se kuvaa työtä kronologisesti):

- sanakirjan tavoitteiden ja rakenteen suunnittelu lähtien kohderyhmän tarpeista
- korpuksen kokoaminen (ellei sopivaa korpusta ole valmiina, mikä on usein asiaintila)
- sana-artikkelien rakenteen suunnittelu
- hakusanalistojen tekeminen ja hakusanojen valinta
- korpusmateriaalin analyysi
    - fraseologisten (leksikaalisten) yksiköiden ja niiden merkitysten etsiminen ja selittäminen
    - fraseologisten yksiköiden tärkeimpien tunnuspiirteiden analysoiminen
        - paradigmaattiset suhteet
            - taivutusparadigmat (koskee etenkin taivutusjärjestelmältään rikkaita kieliä)
            - synonyymisyys ja antonyymisyys sekä assosiatiiviset merkitykset (kuten konnotaatiot)
        - syntaktiset funktiot
        - syntagmaattiset suhteet
            - kollokationaaliset (ja myös semanttiset kotekstuaaliset suhteet)
            - kolligationaaliset eli kotekstin kieliopillisiin rakenteisiin liittyvät suhteet (kuten suomen kielen rektiosäännöt)
            - tilannekontekstiin liittyvät suhteet (kuten rekisteri- ja tekstilajitieto)
- määritelmien muokkaaminen

- käyttötapojen havainnollistaminen edustavien korpusesimerkkien avulla
- sana-artikkelien muokkaaminen yhtenäiseksi kokonaisuudeksi ja toimituksellisten periaatteiden mukaisiksi.

Koska laajoihin korpuksiin perustuva kielentutkimus on nostanut fraseologisuuden myös kielenopetuksessa periferiasta keskiöön (Skiepmann 2008: 185), etenkin englantia koskevat oppijansanakirjat ottavat nykyisin huomioon kielen fraseologialuonteen ainakin jossain määrin. Skiepmann näkee tässä kuitenkin edelleen suuria puutteita, sillä etenkään kaksikieliset oppijansanakirjat eivät juurikaan kuvaa fraseologisuutta. Syitä on hänen mukaansa kolme: 1) lingvistiikassa lisääntyvä nykyinen fraseologiatutkimus ei ole tavoittanut leksikografeja, 2) sanakirjoja tekevät yleensä vain natiivit kielenpuhujat, jotka eivät intuitionsa perusteella havaitse äidinkielensä tyypillisimpiä ja oppijan kannalta tärkeitä rakenteita ja 3) käytössä on varsin vähän ja pieniä puhekielen korpuksia (mts. 193). Puhekielen fraseologisuuden voi katsoa olevan oppijan kannalta ensisijaista kirjoitettuun verrattuna, lisäksi puhuttu kieli sisältää runsaasti kiteytynyttä fraseologiaa. Fraseologisuudella tarkoitamme tässä yhteydessä kielenainesten syntagmaattisia suhteita laaja-alaisesti (ks. esim. Sinclair 1996; Jantunen 2009), emme siis pelkästään keskusteluruutiinien, idiomien, vertausten ja sen kaltaisten kuvaannollisten ilmausten (suhteellisen jähmeitä) kokonaisuuksia. Sanakirjoissa fraseologisuus on esitetty tyypillisimmin kollokationaalisia eli sanojen leksikaalisia syntagmaattisia suhteita kuvaavina rakenteina tai idiomaattisina ilmauksina. Esimerkiksi MEDAL:issa *ground*-sanan eri merkitysten kollokaatiosuhteita on kuvattu seuraavasti (MEDAL s.v. *ground*):

ground noun
1 'surface of Earth':
above/below ground *They were working 250 metres <u>below</u> ground.*
2 'an area of land':
open ground *She had to cross <u>open</u> ground to get to the sea.*
waste ground *A piece of <u>waste</u> ground about 60 feet square.*

3 'a reason for what you say or do…':
reasonable grounds *He believes he has reasonable grounds for making the demand.*
on medical/legal/financial etc grounds *The army turned him down on medical grounds.*
4 'the subject, idea, or information being talked or written about':
cover ground *We'll be covering a lot of new ground in today's lecture.*
5 'an environment in which ideas can develop':
fertile ground *Germany in the 1920s and 30s was fertile ground for such ideas.*

Toisinaan kollokaatit on esitetty vielä eksplisiittisemmin kollokaattilaatikoina:

listen verb
1 'to pay attention to a sound, or to try to hear a sound'

Collocation
Adverbs frequently used with listen 1 ▪ attentively, carefully, closely, hard, intently, politely

Edellä esitetyn kaltaisen kollokaatiotiedon kuvaaminen vaatii korpusten avulla tehtyjä yksityiskohtaisia ja aikaa vieviä analyyseja sanojen syntagmaattisista suhteista (ks. Rundell & Kilgarriff 2011), perustuupa tieto kollokaateista sitten puhtaisiin frekvensseihin tai tilastollisiin analyyseihin.[4] Siksi käytännön syistä usein tyydytään pelkästään kollokaattien esittämiseen eikä esimerkiksi semanttista käyttöympäristöä (mm. semanttista preferenssiä tai semanttista prosodiaa, ks. Sinclair 1996; Jantunen 2009) pystytä kuvaamaan. Verkkosanakirjan rakenne kuitenkin mahdollistaisi tällaisenkin tiedon esittämisen, sillä kontekstia koskevat tiedot voitaisiin esittää muun muassa erilaisina taulukoina ja hyperlinkkien avulla.

Fraseologisuus on moniaalle haarova ongelma kielenoppimisessa: muun muassa opiskelijoiden ja opettajienkin tietämättömyys ilmiöstä, todellisiin aineistoihin ja todelliseen kielenkäyttöön perustuvien

[4] MEDAL:in esipuheesta ei käy ilmi, millä perusteella esitetyt kollokaatit on valittu.

oppimateriaalien ja apuneuvojen puute, aineistoihin perustuvan fraseologiatutkimuksen puute ja olemassa olevan tutkimustiedon hyödyntämättömyys ovat johtaneet siihen, että kielenoppiminen on näiltä osin hankalaa (ks. Granger & Meunier 2008). Oppijat käyttävät vähemmän ja epätyypillisiä kohdekielen mukaisia fraseologisia ilmauksia; he suosivat omia, mutta usein kohdekielen usuksesta poikkeavia rakenteita, jotka saattavat olla perua oppijan äidinkielestä; he muodostavat tekstejä sana sanalta, sen sijaan että hyödyntäisivät monisanaisia yksiköitä eli he mieltävät merkitysyksikön yhdeksi sanaksi useampisanaisen sijaan; heidän vastaanottokykynsä hidastuu, koska kielikompetenssiin ei sisälly frekventtejä monisanaisia yksiköitä, jotka nopeuttavat ymmärtämistä ja vastaanottamista (ks. Siepmannin (2008) koontia aiemmasta tutkimuksesta; myös Nesselhauf 2005; Jantunen 2009). Ilmiö koskettaa siis laajasti kaikkia, jotka jollakin tavalla osallistuvat kielenoppimisen prosesseihin, myös sanakirjantekijöitä.

4. Uusi ConLexis-verkkosanakirja

4.1. Sanakirjan taustaa

ConLexis-verkkosanakirjaa on tehty pääasiallisesti vuosina 2008–2010 Opetushallituksen rahoittaman hankkeen tuotteena. Hanke on toteutettu Oulun yliopiston ja Oulun Aikuiskoulutuskeskuksen yhteistyönä. Yksikielinen sanakirja on suunnattu suomea toisena tai vieraana kielenä opiskeleville oppimateriaaliksi sekä heidän opettajilleen työkaluksi. Siinä on tässä vaiheessa 302 sana-artikkelia, jotka esittelevät sanojen merkityksen, taivutuksen sekä sanan käytön kontekstissaan. Vuosina 2011–2012 ConPraxis-hankkeessa sanakirjaan on tuotettu harjoituksia sanakirjan sanoista.

Sanakirjan sana-artikkelien määrä on vielä hyvin pieni, mutta kahden hankekauden aikana on luotu pohja sana-artikkelille sekä kehitetty sitä vastaamaan paremmin oppijoiden tarpeita. Sanakirja on suunnattu oppijoille, joiden kielitaidon taso on vähintään Eurooppalaisen viitekehyksen taitotaso B1 (kielitaitotasoista ks. EVK). Käyttäjäpalautteen

perusteella sanakirja soveltuu oppijoille parhaiten B2-taitotasolta alkaen, jolloin sana-artikkelien esimerkkejä pystyy ymmärtämään kohtuullisen vaivattomasti ja myös hyödyntämään niiden tarjoamaa tietoa. Sanakirjaa voi kokemustemme perusteella käyttää ohjatusti jo A2-taitotasolta lähtien.

Oppijoiden lisäksi sanakirja palvelee myös opettajia, jotka voivat käyttää sanakirjaa pohjamateriaalina omassa opetuksessaan ja oppimateriaaleissaan. Oppimateriaalit tehdään usein natiivin kielenkäyttäjän intuition pohjalta, ja myös luokkahuoneessa opettaja on intuitionsa varassa, kun hänelle esitetään kysymyksiä sanojen käytöstä. Intuitioon tukeutuminen on monin tavoin epäluotettavaa: Se ei aina anna oikeaa kuvaa sanojen todellisesta käytöstä eikä riitä kertomaan kielen piirteiden yleisyydestä tai niiden variaatiosta kielessä. Usein intuitiivisesti tunnistetaan herkemmin epätyypillisiä kuin tyypillisiä ilmiöitä. Opettajienkin on siis helpompaa tunnistaa oppijankielen epätyypillisyydet kuin ohjata oppijaa tyypilliseen sanojen käyttöön. Luonnollisesti vielä vaikeampaa sanaston fraseologisten piirteiden opettaminen pelkästään intuitioon tukeutuen on niillä opettajilla, jotka eivät äidinkielenään puhu kohdekieltä. (Biber ym. 1998: 1–5; Mauranen 2004: 96.) Koska havainnot kielen käytöstä eivät riitä kuvaamaan kielen todellista käyttöä, tarvitaan suuria aineistoja, joiden avulla kielen tarkastelu muuttuu tarkemmaksi ja mahdollistaa kielen rakenteen ja käytön uudenlaisen kuvaamisen (Tsui 2004: 41–42). Suomen kielestä ei ole saatavilla helppokäyttöistä ja avointa korpusta, jota voisi käyttää opetuksen tukena. ConLexis-sanakirja tarjoaa korpuspohjaista ja tutkimukseen perustuvaa tietoa sanojen käytöstä, jota opettaja voi varsinaisen korpuksen puuttuessa käyttää.

Sanaston käyttöesimerkkien aineisto on Tieteen tietotekniikan keskuksen hallinnoimasta Suomen kielen tekstipankista. Esimerkit ovat suurimmaksi osaksi tekstipankin sanomalehtiteksteistä, mutta esimerkkejä on täydennetty myös uudemmalla Internet-aineistolla. Sanat on valittu Suomen kielen tekstipankin frekvenssiluettelosta[5]; lähtökohdaksi

[5] http://www.csc.fi/tutkimus/alat/kielitiede/taajuussanasto-B9996/view (10.7.2013).

otettiin 100 yleisintä sanaa, joista suurin osa on nyt mukana sanastossa. Sanoja ei siis ole valittu esimerkiksi eri taitotasoilla, oppimis- tai käyttötilanteissa tarvittavan sanaston perusteella, koska tilanteinen tai tasoittainen variaatio on kuitenkin hyvin laajaa. Näiden sanojen ympärille on muodostettu sanapesyeitä yleisyyden mukaan valitun pääsanan synonyymeistä ja antonyymeistä. Sanan valinta siis perustuu ensi kädessä sen yleisyyteen, mutta sanapesyeiden myötä sanastoon on tullut mukaan myös harvinaisia sanoja.

Jokaista sana-artikkelia varten on tehty tutkimusta käyttämällä Suomen kielen tekstipankin tekstikorpusta. Lähtökohtana sana-artikkelien suunnittelussa ovat olleet käytön tyypillisyys, autenttisuus sekä fraseologisuuden esittäminen kollokaatioiden ja useampisanaisten ilmausten (klustereiden) avulla. Leksikaalisen tiedon lisäksi esitetään kieliopillista kontekstuaalista tietoa, kuten morfosyntaksin osalta rektioita.

Rundellin (2008: 222–223) mukaan oppijan sanakirjoille on tyypillistä, että niissä pyritään esittämään sanan käyttöä mahdollisimman monipuolisesti. ConLexis-sanakirjassa esitetään sanojen tavallisimmat käyttöyhteydet, eli sana-artikkeleista jää puuttumaan paljon täysin mahdollisia käyttöyhteyksiä. Tarkoituksena ei ole kuitenkaan ohjata oppijaa olemaan käyttämättä harvinaisempia käyttöyhteyksiä; sanojen tyypilliseen käyttöön on tässä keskitytty, jotta sana-artikkelit pysyisivät luettavina, ja myös siksi, että itse sana-artikkelit ovat niin työläitä laatia. Jos sana-artikkeleihin otettaisiin mukaan kaikki materiaali, mitä korpus tutkimuksen aikana tarjoaa, olisi sana-artikkelien määrä nykyistäkin pienempi ja toisaalta raja tyypillisen ja frekventin käytön sekä harvinaisemman käytön välillä hämärtyisi.

Pedagogisesta näkökulmasta katsottuna ConLexiksen tarkoituksena on antaa materiaalia kielen oppimiseen ja opettamiseen mahdollisimman tehokkaasti. Nesselhaufin (2005) tutkimuksen tulokset tukevat sitä, että kollokaatioiden opettamisen pitäisi olla tietoisempaa ja keskittyä itse kollokaatioiden lisäksi erilaisten kollokaatioiden oppimiseen liittyvien taitojen opettamiseen. Tutkimuksessaan hän nimittäin havaitsi, että

1) kohdekielestä poikkeavien kollokaatioiden osuus on oppijankielessä suuri,
2) opiskeluvuosien lisääntymisellä on vain vähäinen merkitys kollokaatioiden oppimisessa,
3) pitempi altistuminen kielelle parantaa kollokaatioiden osaamista vain vähän.

Tietoisuuden lisääminen ja havaitsemisen opettaminen ovat ensiaskelia kollokaatioilmiön opiskelussa. Opetuksen pitäisi myös olla systemaattista ja keskittyä sellaiseen ainekseen, jota opiskelijat eivät hallitse. Konkreettisesti kollokaatiota voi opettaa esimerkiksi siten, että sanojen käyttötapoja esiteltäessä tarkastellaan sanojen kollokationaalista variaatiota. Seuraavassa vaiheessa opiskeltavan sanan lähisynonyymejä voi tarkastella suhteessa opiskeltavaan sanaan. Lisäksi huomiota kannattaa kiinnittää usein esiintyviin rakenteisiin ja erityisesti semanttiseen prosodiaan[6]. (Nesselhauf 2005: 253, 265–269.) Opettajat tunnistavat kollokaation ilmiönä, mutteivät aina pysty selittämään sitä opiskelijoille, koska ei ole olemassa materiaalia eikä riittävästi tietoa, johon voisi nojautua (Tsui 2004: 43). Etenkin suomen kielestä on hyvin vähän materiaalia kollokaatioiden opettamiseen. ConLexis-sanakirjan tarjoaman tiedon pohjalta opettajat voivat myös laatia lisää materiaalia luokkahuoneeseen tukemaan sanaston oppimista.

4.2. ConLexis-sanakirjan kehystekstit

ConLexis-verkkosanakirjan kehystekstejä ovat oppijalle suunnattu ohjeistus sanakirjan ja harjoitusten käyttöön, “Käyttö ja rakenne” sekä “Kuvaus ja aineisto”. “Käyttö ja rakenne” -osuus sisältää kuvauksen sanaartikkelin rakenteesta sekä tekstinä että esimerkkiartikkelina. Lisäksi siinä esitellään sanakirjan keskeiset toiminnot. Osio on suunnattu sekä oppijoille että sanakirjaa käyttäville suomen kielen opettajille. Oppijalle

[6] Semanttisella prosodialla tarkoitetaan yleensä ilmauksen esiintymistä joko positiivisessa, negatiivisessa tai neutraalissa käyttöympäristössä (ks. esim. Jantunen 2009).

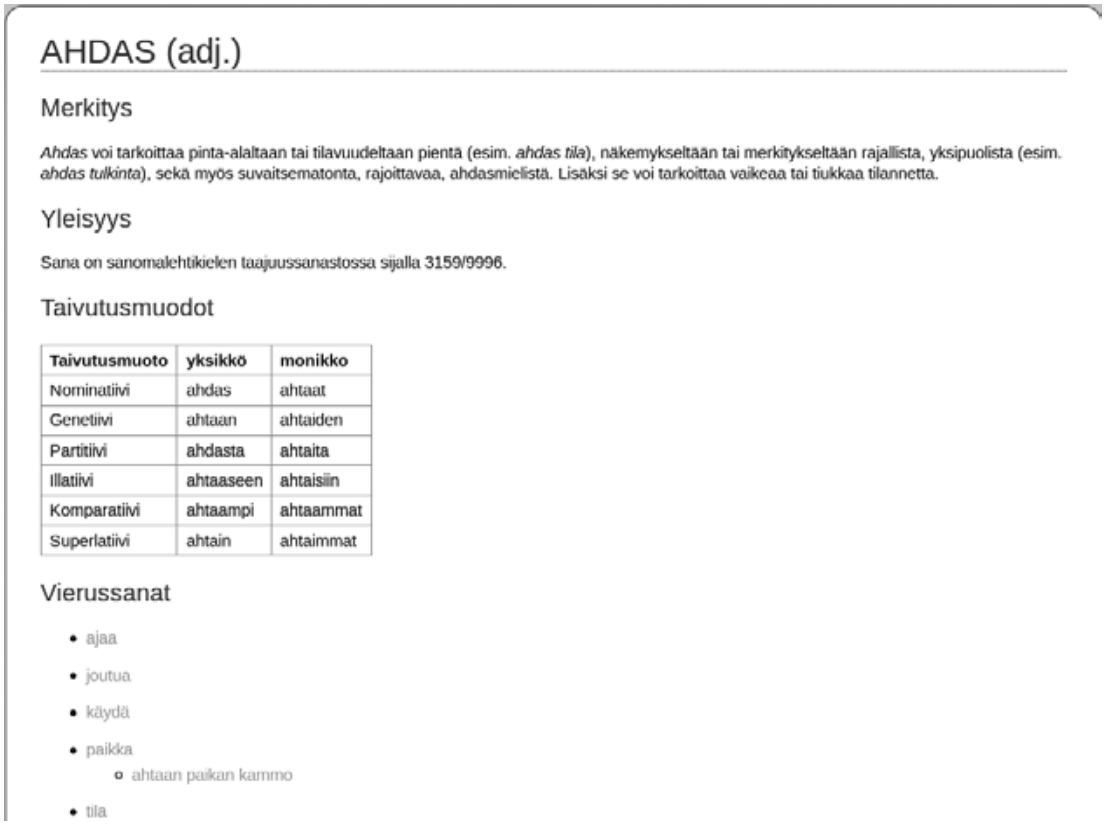
suunnattu ohjeistus painottuu harjoitusten tekemisen opastamiseen. Se on kirjoitettu selkeästi ja yksinkertaisesti, ja siinä ohjataan opiskelijaa harjoitusten tekemisessä. Lisäksi sivulla on selitetty sanakirjan peruskäsitteet, jotta oppija pystyy ymmärtämään sana-artikkelien rakenteen.

"Käyttö ja rakenne" -sivu sisältää esimerkkiartikkelin ja kuvauksen sana-artikkelin rakenteesta sekä sanakirjan käytöstä. Sivun tekstit on kirjoitettu pitäen mielessä sekä oppija että opettaja käyttäjinä. Kieli ei ole yhtä yksinkertaista kuin oppijalle suunnatulla sivulla, mutta esimerkkiartikkeli on havainnollinen tapa esittää sanakirjan käyttöä, ja se avautuu varmasti myös oppijalle. "Kuvaus ja aineisto" -sivu sisältää teoreettista taustatietoa ja sanakirjan käyttämiseen liittyvät olennaiset käsitteet. Lisäksi sivustolle on tulossa vinkkejä opettajalle sanakirjan hyödyntämisestä kielen opetuksessa.

4.3. Sana-artikkelin rakenne

Sanakirjat määrittelevät sanoja perinteisesti synonyymien kautta. Synonyymeinä pidettyjen sanojen käyttö voi kuitenkin olla keskenään hyvinkin erilaista. (Biber ym. 1998: 24.) Synonyymien luetteleminen ei välttämättä anna tarkkaa kuvaa sanan merkityksestä, ja ilman lisätietoja se voi jopa johtaa oppijaa harhaan. ConLexis-sanakirjassa sanan merkitys on selitetty kokonaisilla virkkeillä ja synonyymien listausta on pyritty välttämään. Merkitysten lisäksi artikkelissa on taivutusmuototaulukko, johon on valittu sellaiset muodot, joissa näkyvät sanojen erilaiset vartalot tai jotka ovat muuten olennaisia, esimerkiksi adjektiiveilla komparaatio. Esimerkki sana-artikkelin alusta on kuvassa 1.

Kollokaatit ja niihin liittyvät klusterit sijaitsevat artikkelin keskellä (kuva 1). Kollokaatteja on sanan yleisyydestä ja käytöstä riippuen 0–10. Kollokaattilistassa on linkki jokaisesta sanasta erilliselle kollokaattisivulle tai klusterisivulle, jossa ovat esimerkit kollokaatiosta.



AHDAS (adj.)

Merkitys

Ahdas voi tarkoittaa pinta-alaltaan tai tilavuudeltaan pientä (esim. *ahdas tila*), näkemykseltään tai merkitykseltään rajallista, yksipuolista (esim. *ahdas tulkinta*), sekä myös suvaitsematonta, rajoittavaa, ahdasmielistä. Lisäksi se voi tarkoittaa vaikeaa tai tiukkaa tilannetta.

Yleisyys

Sana on sanomalehtikielen taajuussanastossa sijalla 3159/9996.

Taivutusmuodot

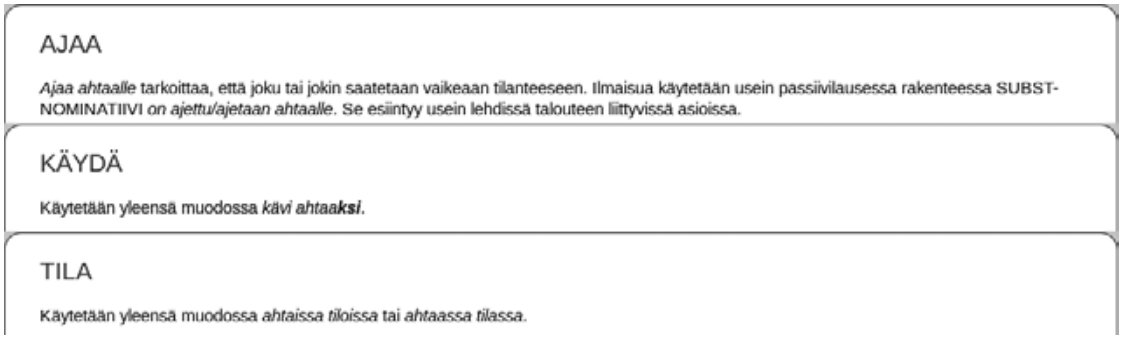
Taivutusmuoto	yksikkö	monikko
Nominatiivi	ahdas	ahtaat
Genetiivi	ahtaan	ahtaiden
Partitiivi	ahdasta	ahtaita
Illatiivi	ahtaaseen	ahtaisiin
Komparatiivi	ahtaampi	ahtaammat
Superlatiivi	ahtain	ahtaimmat

Vierussanat

- ajaa
- joutua
- käydä
- paikka
  - ahtaan paikan kammo
- tila

Kuva 1. *Sana-artikkelin alku sanalle* ahdas

Kollokaatti- tai klusterisivulla on myös tietoa rakenteista, joita käytetään sanan yhteydessä, jos siitä on jotain erikseen mainittavaa. Esimerkiksi *ahdas*-sanan kollokaattisivuilla (kuva 2) on erilaisia tapauksia siitä, millaisia kieliopillisia rakenteita sanan kollokaattien yhteydessä esiintyy ja miten ne on selitetty.



AJAA

Ajaa ahtaalle tarkoittaa, että joku tai jokin saatetaan vaikeaan tilanteeseen. Ilmaisua käytetään usein passiivilausessa rakenteessa SUBST-NOMINATIIVI *on ajettu/ajetaan ahtaalle*. Se esiintyy usein lehdissä talouteen liittyvissä asioissa.

KÄYDÄ

Käytetään yleensä muodossa *kävi ahtaa**ksi***.

TILA

Käytetään yleensä muodossa *ahtaissa tiloissa* tai *ahtaassa tilassa*.

Kuva 2. *Esimerkkejä* ahdas*-sanan kollokaattisivujen alussa olevista rakennetiedoista*

Esimerkit on esitetty konkordanssien muodossa taulukossa, koska se havainnollistaa hyvin sanojen yhteydessä esiintyviä tyypillisiä kieliopillisia rakenteita ja toiston kautta vahvistaa sanan ymmärtämistä. Esimerkkejä kustakin kollokaatiosta tai klusterista on kymmenen. Jos sanan käyttöön liittyy erilaisia merkityksiä, ne on erotettu omiksi taulukoikseen, esimerkiksi *ahdas paikka* (kuva 3) voi tarkoittaa pientä tilaa tai vaikeaa tilannetta. Esimerkkilauseet on valittu korpusaineistosta, mutta osaa niistä on editoitu. Lauseita on lyhennetty, niistä on poistettu tai korvattu sanoja, jotka viittaavat toisaalle tekstiin ja jotka tekisivät lauseen kontekstistaan irrotettuna hankalan ymmärtää. Myös ei-julkisuuden henkilöiden nimet on korvattu kuvauksilla tai persoonapronomineilla.

PAIKKA

Ahdas paikka tarkoittaa pientä paikkaa (konkordanssi 1) tai vaikeaa tilannetta (konkordanssi 2).

takaisin sanasivulle

- paikka
  - ahtaan paikan kammo

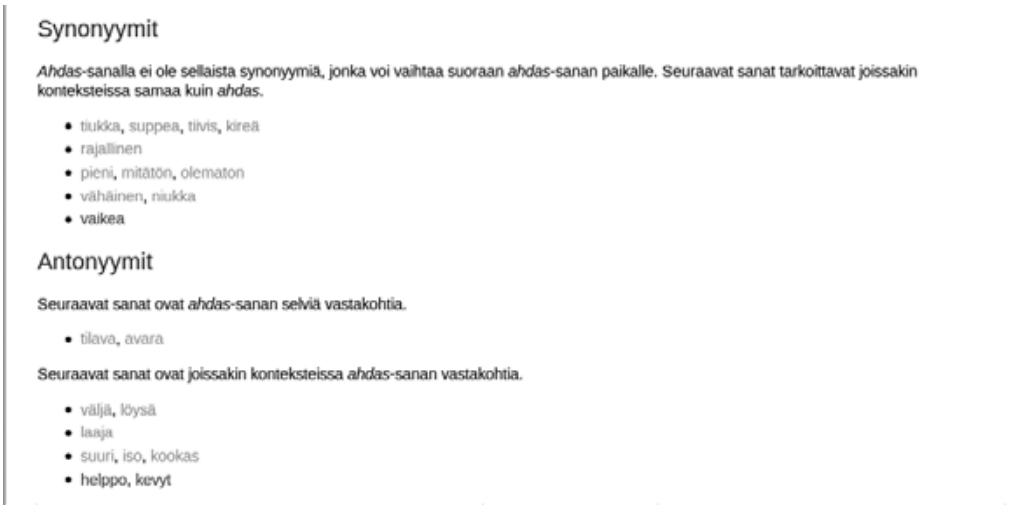
	Ahtaat paikat	aiheuttavat suuria turvallisuusongelmia.
Äiti arveli, että pienokaiselle alkoi tulla vatsassa liian	**ahtaat paikat.**	
	Ahtaisiin paikkoihin	pysäköitäessä peilejä saa käännettyä alas 60 astetta.
Ikean muotoilija kehitteli naulakon, joka sopii	**ahtaisiin paikkoihin.**	
Lattian pesussa nykyään suosittuja ovat nivelöidyt pesimet, joilla pääsee	**ahtaisiinkin paikkoihin.**	
Uusi tie rakennettiin	**ahtaaseen paikkaan**	nykyisen tien ja rautatien väliin.
Kiitosta saa erityisesti kevyt käsiteltävyys, joka saa auton	**ahtaissa paikoissa**	tuntumaan pienemmältä.
Ilmavoimien isot helikopterit eivät sovi	**ahtaiden paikkojen**	pelastustoimiin.
Ottelu ratkesi toisen joukkueen parempaan pallon hallintaan	**ahtaissa paikoissa.**	

Kun Hongkongin yllä liehuu Kiinan lippu tulee Taiwanille	**ahtaat paikat.**	
Onneksi asunnon voi aina myydä, kun tulee	**ahtaat paikat**	lainan kanssa.
Yritys siirtyi ulkomaille kun Suomessa tuli	**ahtaat paikat.**	

Kuva 3. Ahdas-*sanan kollokaattisivu vierussanalle* paikka

Hyvän esimerkin pitäisi esittää kollokaation tyypillistä käyttöä ja olla merkityksen kannalta havainnollinen ja ymmärrettävä oppijalle, eli se ei saa sisältää liian vaikeaa sanastoa tai kielioppia (Kilgarriff ym. 2008: 426). Vaikka käytössä on laaja sanomalehtiaineisto, satojen esimerkkien joukosta voi toisinaan olla vaikea löytää sopivia ja riittävän pelkistettyjä

esimerkkejä havainnollistamaan sanan käyttöä halutulla tavalla. Toisaalta monipuolinen ja monentasoista kielenainesta sisältävä esimerkistö mahdollistaa sanaston soveltuvuuden kielenoppijoille eri taitotasoilla. Kaikkea ei tarvitse ymmärtää saadakseen käsityksen sanan käytöstä.
Kollokaatioiden jälkeen sana-artikkelissa seuraa luettelo sanan synonyymeistä ja antonyymeistä, jotka on jaettu merkityskenttien mukaan erilaisiin ryhmiin. Niiden yhteydessä on myös pyritty antamaan tietoa siitä, kuinka selvästi sanat ovat synonyymisiä tai antonyymisiä sana-artikkelin pääsanan kanssa (kuva 4).



Kuva 4. Ahdas-*sanan synonyymi- ja antonyymilistat sanasivulla*

Viimeisenä osana sana-artikkelissa on Lisätietoa-osio, jossa sana-artikkelin kirjoittajan on mahdollista kuvata esimerkiksi sanan käytön kannalta olennaisia kieliopillisia rakenteita tai semanttista profiilia.

4.4. Korpusteknologiasta ja ConLexis-sanakirjasta

Tietotekniikan sovelluksissa on yhtenä merkittävänä trendinä ollut tietojen ja alustojen avautuminen: laitteisto- ja ohjelmistovalmistajat ovat avanneet omia alustojaan ulkopuolisille sovelluskehittäjille, ja tämä on johtanut toimivien ohjelmistoekosysteemien syntyyn. Englanninkielisten korpusten suuren tarjonnan ansiosta on syntynyt vastaavanlainen

ekosysteemi, jossa kieliteknologian sovelluksia kehittäville on tarjolla paljon vapaasti käytettävissä olevia korpuksia. Suomenkielisten korpusten osalta tilanne ei ole yhtä hyvä, etenkään annotoitujen korpusten osalta.

ConLexiksen tämänhetkinen sanasto on koottu asiantuntijatyönä käsin, eli yhtä sana-artikkelia kohden tutkijat ovat tehneet useita hakuja korpustietokantaan. Kirjoitusprosessia voitaisiin kuitenkin tukea antamalla tiettyjä työvaiheita tietokoneiden tehtäväksi. Näitä työvaiheita voivat olla mm.

- sana-artikkelien toteutusjärjestyksen valinta,
- sana-artikkelien pohjatietojen haku ja artikkelipohjan luonnostelu,
- esimerkkien, kollokaattien ja muiden suhteiden etsiminen korpuksesta suoraan sana-artikkeliin sanakirjantekijän hyväksyttäväksi,
- esimerkkien luokittelu tekstilajeihin.

Sanakirja itsessään on toteutettu käyttäen MoinMoin-wikiteknologiaa, joka mahdollistaa sana-artikkelien muokkaamisen suoraan sivustolla ja niiden luontevan linkittämisen toisiinsa. Käytetty ohjelmisto hoitaa automaattisesti myös esimerkiksi sanastojen päivittämisen sekä artikkelien ja harjoitusten linkittämisen toisiinsa.

5. ConPraxis-oppimateriaali

5.1. Edistyneen kielenoppijan sanasto-oppimateriaali: nykytilanne ja tarpeet

Suomen kielen varhaisimmat oppimateriaalit on tehty nimenomaan vieraana kielenä oppimisen näkökulmasta, sillä suomea opiskeltiin pitkään lähinnä ulkomaisissa yliopistoissa. Koska ulkomailla suomea opiskelivat pääasiassa lingvistit, myös oppimateriaali painottui kielen rakenteisiin ja kielen yhteen standardimalliin. Suomessa kielenopetuksen kysyntä oli pitkään aika vähäistä, joten käytännössä samat oppimateriaalit olivat

käytössä myös suomalaisissa kansalais- ja työväenopistoissa ja yliopistoissa, joissa kielikursseja tarjottiin. Traditio vaikuttaa vieläkin suomi toisena kielenä -opetuksen ja oppimateriaalien taustalla, ja osa lähtökohtaisesti vieraan kielen opiskeluun tarkoitetuista materiaaleista on yhä aktiivisessa käytössä. (Suni 2008: 30–31; Tanner 2012: 31.)

Erilaiset yleiskorpukset ja oppijankielen korpukset ovat mahdollistaneet kielentutkimuksen painopisteen siirtymisen kielen käytön näkökulmaan, nimenomaan kielen elementti- ja valmisrakenteisuuteen, kuten kollokaatioihin, rutiineihin ja erilaisiin diskurssimuotteihin (esim. Granger 2005). Suomen kielen sanakirjoissa ja oppimateriaaleissa kielen fraseologisia piirteitä kyllä jo jonkin verran esitellään (Jönsson-Korhola & White 1999; Muikku-Werner ym. 2008), mutta kielen rutiineja, idiomeja ja rektioita vapaammin varioivat tapaukset ovat toistaiseksi jääneet oppimateriaalissa vähemmälle huomiolle.

S2-oppimateriaalia on nykyään runsaasti saatavilla niin painettuna kuin verkossakin, mutta suurin osa materiaalista on suunnattu lähinnä alkeistasolta toimivan peruskielitaidon tasolle (ks. Jokinen ym. 2011). Toimivan peruskielitaidon tai itsenäisen kielitaidon saavuttamiseksi oppijoille on verrattain vähän sanaston oppimiseen liittyvää materiaalia, mutta esimerkiksi Saunelan (2008a) *Harjoitus tekee mestarin* -kirjasarjan kolmannessa osassa käsitellään muun muassa sanojen johtamista, yhdyssanojen muodostamista, synonyymeja ja antonyymeja sekä toisiinsa sekoittuvia sanoja. Sarjan neljäs osa (Saunela 2008b) sisältää monipuolisia kirjallisia tehtäviä keskitason yleiseen kielitutkintoon valmistautuville. Kirjassa on aiemman YKI-testikäytännön mukaisesti erikseen myös sanasto- ja rakennetehtäviä. *Suomen kielen sanastoharjoituksia* (Kangasniemi 2003) sisältää monentasoisia, myös hieman edistyneemmälle oppijalle soveltuvia sanastotehtäviä, joissa sanastoa käsitellään hyvin monipuolisesta näkökulmasta. Tehtävissä etsitään muun muassa synonyymeja, antonyymeja ja homonymiaa. Myös puhekielelle tyypillisiä piirteitä on esitelty. Tehtävät sisältävät esimerkiksi slangijohtimia, lapsenkieltä ja idiomeja. Edellä kuvatun tyyppiset sanastoharjoitukset ovat omiaan laajentamaan oppijan sanavarastoa,

mutta ne eivät aina anna täsmällistä tietoa siitä, millainen sanan koteksti on, miten sanaa tulisi käyttää erilaisissa konteksteissa tai millainen sana on tyylisävyltään. Esimerkiksi Kangasniemen (2003) synonymiaa esittelevät tehtävät eivät anna oppijalle informaatiota sanojen semanttisista eroista tai syntaktisesta käytöstä, vaikka synonyymiset sanat ovat ani harvoin merkitykseltään ja käytöltään toisiaan vastaavia.

Luokkahuoneessa kielen fraseologisia piirteitä opetetaan kaiketi paljon enemmän kuin oppimateriaalien tarkastelu antaa ymmärtää, mutta fraseologisuutta systemaattisemmin käsittelevillä oppimateriaaleilla voisi olla oma roolinsa ilmiön vakiinnuttamisessa opetukseen. Kuten on jo todettu, ilman opiskelua edistyneenkin kielenoppijan on vaikea tunnistaa esimerkiksi sanojen tyypillisiä käyttöympäristöjä. Siksi kielen fraseologisia ominaispiirteitä tulisi opettaa systemaattisesti. Koska opettajat huomaavat intuitionsa varassa parhaiten kielen epätyypillisyyksiä, olisi kielitaidon kehittymisen kannalta tärkeää, että edistyneille kielenoppijoille olisi tarjolla myös todelliseen kielenkäyttöön perustuvaa oppimateriaalia.

Autenttisuudella voidaan tarkoittaa paitsi materiaalin autenttisuutta, myös toiminnan autenttisuutta. Tiukimman tulkinnan mukaan mikään oppimateriaali ei ole enää autenttista, sillä todellisuus ei siirry tekstin mukana alkuperäisestä ympäristöstään oppikirjaan (Widdowson 1998: 711–712). Tällaisella tulkinnalla päädytään kuitenkin siihen, että autenttisuus on kielenopetuksen ulottumattomissa. Äärimmäisen autenttisuuden sijaan keskeisempää olisi miettiä, miten oppimistilanne tai -materiaali suhteutuu merkitykselliseen toimintaan ja kuinka relevanttia esimerkiksi opetustekstin sisältö on oppijalle. Tällä hetkellä autenttisuus näkyy oppimateriaaleissa lähinnä teksteissä, joita on joko sellaisenaan tai toisinaan helpotettuina otettu mukaan oppimateriaaleihin ja jotka toimivatkin sanaston opettamisessa yhtenä hyvänä välineenä. (Esim. Kangasniemi 2006.) Myös keskusteludialogeihin otetaan natiivikielestä mallia, mutta oppikirjoihin päätyvät simulaatiot ovat lopulta melko kaukana todellisesta kielenkäytöstä (ks. esim. Tanner 2012: 32). Joitakin poikkeuksiakin mahtuu toki joukkoon.

5.2. ConPraxis-sanastoharjoitukset edistyneille kielenoppijoille

ConLexis-sanakirjaa on vuosina 2011–2012 täydennetty sanastoharjoituksilla, joiden tavoitteena on ollut paitsi parantaa ConLexis-sanakirjan käytettävyyttä kielenopiskelussa, myös tarjota uudenlaista oppimateriaalia edistyneemmille kielenoppijoille. Harjoitukset on suunnattu toimivan ja sujuvan peruskielitaidon tasolta aina taitavan kielitaidon tasolle asti (EVK). Tehtävät sopivat myös yleisten kielitutkintojen keski- ja ylimmän tason testiin valmistautujille. Harjoituksissa on hyödynnetty ConLexis-sana-artikkeleiden tutkimustietoa sanojen käytöstä. Koska harjoitusten on sanakirjan tavoin tarkoitus pohjautua autenttiseen aineistoon ja esitellä tyypillistä kielenkäyttöä, on harjoitusten oikeisiin vastauksiin otettu mukaan ainoastaan sellaiset vaihtoehdot, jotka on esitelty sana-artikkelissa. Tämä ei tarkoita sitä, etteivätkö muut vaihtoehdot olisi mahdollisia. Oikeiden vastausten rajaamista pelkästään sana-artikkeleissa esitettyihin vaihtoehtoihin voi perustella kuitenkin sillä, että materiaalia ei ole haluttu rakentaa tekijöidensä intuitiivisen kielitiedon varaan. Kielenoppijakin selviää vähemmällä: hänellä on pelkän sana-artikkelin opiskeltuaan riittävästi tietoa harjoituksen tekemiseen.

Harjoitusten aineistona on käytetty Internetissä julkaistua tekstimateriaalia, jolloin valittujen esimerkkien kirjosta on tullut laajempi ja vaikeustasoltaan vaihtelevampi kuin Tekstipankin kokoelmissa. Esimerkkejä on etsitty Internetin hakukoneilla muun muassa verkkolehdistä, verkkojulkaisuista, blogeista ja jopa keskustelupalstoilta. Valintaa on ohjannut paitsi kielellinen ilmaisu itsessään, myös ConLexis-sana-artikkeleissa kerrottu kuvaus ja esimerkit käyttökontekstista ja tyypillisyyksistä. Vaikka oppimateriaalin taustalla onkin autenttinen aineisto, on sekä aineistopohjaisuudesta että autenttisuudesta joustettu, jotta materiaali palvelisi paremmin itse tarkoitusta, kielenoppimista. Aineistoista poimittuja esimerkkejä on muokattu samoin periaattein kuin sana-artikkeleissa ja lisäksi täydennetty tarpeen mukaan, jotta vastausvaihtoehtojen määrä olisi rajatumpi ja jotta asiayhteydestä irrotettunakin ne olisivat ymmärrettäviä. Vastausvaihtoehtoja on alkuperäisen

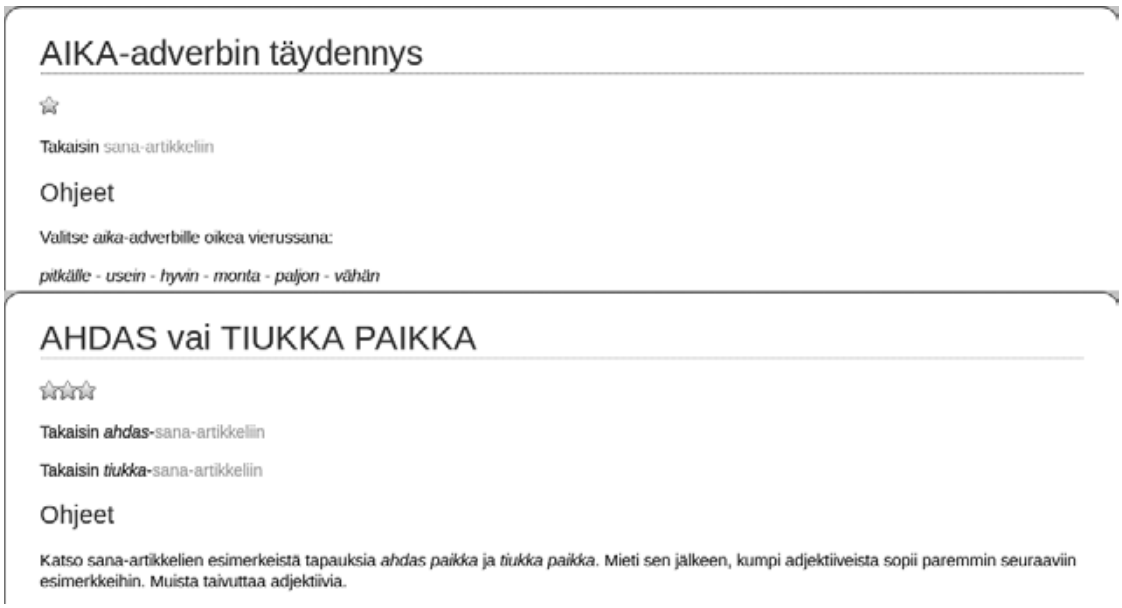
vastauksen lisäksi täydennetty muilla kontekstiin sopivilla, sana-artikkeleissa esitetyillä vierussanoilla. *aika*-adverbin vierussanatäydennyksestä poimituissa esimerkeissä (1 ja 2) alleviivatut sanat ovat oikeita vastauksia, ja harjoituksen tekijän lisäämät täydennykset on osoitettu lihavoinnilla:

(1) **Presidentti Sauli** Niinistö kertoo jaksamisestaan **positiiviseen sävyyn**: "Peruskuntoni on aika hyvä." (**)
(2) Kaikkea ei tarvitse ostaa itse. Varailen ja tilailen aika usein / **paljon** materiaalia eri kirjastoista. (*)

Harjoitukset on luokiteltu kolmeen eri tasoon siten, että yhden tähden harjoitukset on suunnattu B1-tasolle, kahden tähden harjoitukset B2-tasolle ja kolmen tähden harjoitukset C1-tasolle (ks. esimerkkejä 1 ja 2). Tehtävät ovat samantyyppisiä kaikilla tasoilla, ja tällä hetkellä harjoituksissa on synonyymin valinta- ja vierussanatäydennystehtäviä. Helpoimmalla tasolla esimerkit ovat lyhyitä ja niin aihepiiriltään kuin sanastoltaankin mahdollisimman yleisiä. Vastausvaihtoehtojen määrä on rajoitetumpi, ja oppijalle annetaan ohjeistuksessa enemmän vihjeitä oikeiden vastausten löytämiseksi. Esimerkiksi yhden tähden tehtävän ohjeistuksessa rajataan valmiiksi vierussana- ja taivutusvaihtoehdot, jotka kyseisessä tehtävässä ovat mahdollisia. Kolmen tähden tehtävän haasteena on *ahdas*- ja *tiukka*-adjektiivien käytön hienojakoinen ero *paikan* vierussanana. Ohjeistuksessa muistutetaan sana-artikkeleihin tutustumisesta ja sanantaivutuksesta, mutta muita vihjeitä ei anneta (kuva 5, s. 111).

Sanastoharjoituksissa keskeisessä asemassa on kollokoivien sanaparien oppiminen (esimerkiksi *sää* + *sallia* ja *heittää* + *ilma* muodostavat kollokaatit), mutta kollokaatteja harjoitellessaan kielenoppija tulee samalla tutustuneeksi kielen elementtirakenteisiin, esimerkiksi kolligaatioihin tai muotteihin (3–4).

(3) Sään salliessa lasten liikuntakerho kokoontuu ulkona. (*)
(4) Arvioita ja mielipiteitä ruoista vaihdettiin innokkaasti, ja arvailuja raaka-aineista ja tekniikoista heitettiin ilmaan. (**)

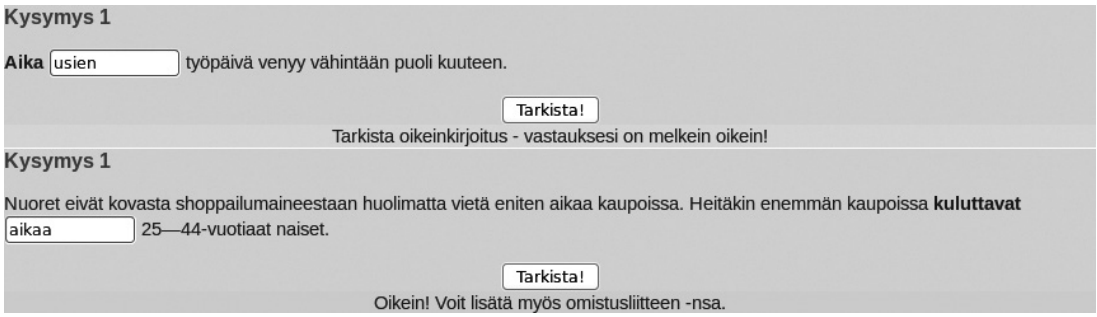


Kuva 5. *Ohjeita harjoituksen tekemiseen*

Iso osa oppimateriaalista on edelleen painettuja oppikirjoja. Kontaktiopetuksessa painettu materiaali onkin yleensä välttämätöntä, sillä oppijoilla ei ole aina käytössään verkko-opiskeluun tarvittavia välineitä. Verkko-opiskelu tarjoaa kuitenkin ajasta ja paikasta riippumattomat mahdollisuudet opiskella kieltä. Sen etuina ovat lisäksi materiaalin päivitettävyys, täydennettävyys ja helppo saatavuus. Verkkoympäristö tarjoaa myös teknisiä ratkaisuja kielenoppimisen tueksi (ks. luku 5.3.). Oppikirjojen avulla itsenäiseen opiskeluun verrattuna verkossa tehdyt harjoitukset antavat oppijalle heti myös palautetta suorituksesta. ConPraxis-harjoituksissa syötetyt oppijanvastaukset tallentuvat sivustolle, jolloin ylläpitäjä voi hyödyntää niitä materiaalin muokkauksessa ja laatimisessa. Tehtävistä voidaan esimerkiksi poistaa epäselviä tai asiasisällöltään vanhentuneita kysymyksiä.

Tekniikan avulla voidaan toteuttaa myös vastauksista annettavaa palautetta. Palautteen tulisi olla oppijalle informatiivista, mutta toteutettavissa automatisoidusti ilman, että jokainen palaute täytyy käsin lisätä. Oppijaa voidaan esimerkiksi kehottaa tarkistamaan sanan oikeinkirjoitus, niin ettei oppijan tarvitse pelkän oikein–väärin-palautteen

ohjaamana lähteä kokeilemaan vastaukseen kokonaan toista sanaa. Palaute voi sisältää myös analyyttisempaa tietoa oppijan antamasta vastauksesta (kuva 6).



Kuva 6. *Esimerkkejä oppijan saamasta palautteesta*

ConPraxis-oppimateriaalia on tähän mennessä työstetty natiivien kielenkäyttäjien tuottaman aineiston pohjalta. Oppimateriaalien laatimisessa olisi jatkossa syytä huomioida paremmin myös saatavilla olevat oppijankielen korpukset. Esimerkiksi sanastomateriaalin suunnittelussa suomen opettaja tai materiaalin tekijä voi hyödyntää oppijankorpusta selvittääkseen, millaisia ongelmia kielenoppijalla on sananvalinnassa, jolloin esimerkiksi monivalintatehtävät voisivat intuitioon tukeutumisen sijaan perustua oppijankielenkorpuksesta saatuun tietoon.

5.3. Sanakirjaharjoituksista yleensä ja ConPraxis-harjoituksista

Oppijansanakirjojen siirtyessä sähköiseen muotoon on täysin luontevaa, että niihin liitetään samalla myös erilaisia harjoituksia, sillä sanakirjan käyttäjähän on jo määritelmänkin mukaan kielenoppija ja siten arvatenkin kiinnostunut opettelemansa kielen harjoittelusta. Korpuspohjaista sanakirjaa luotaessa on verrattain helppoa automatisoida sekä erilaisten korpuspohjaisten tehtävien luonti että niiden tarkastaminen. Sanaartikkeleita luotaessa voidaan tarjota erilaisia automaattisesti luotuja harjoituksia sanakirjantekijälle, ja tämä voi sitten hyväksyä ne sanakirjaharjoituksiin lisättäväksi. ConPraxis-harjoituksissa esimerkit on valittu

käsin, mutta kuten sana-artikkeleiden luonnissakin, voidaan tämä prosessi jatkossa automatisoida.

Automaattisten harjoitusten – kuten puuttuvan sanan täydentämisen – kannalta tehtävien automaattisessa tarkastuksessa on ongelmana vastausten laaja hajonta. Järjestelmän tuleekin osata mukautua mahdollisimman monenlaisiin vastauksiin sen sijaan että se hyväksyisi vain tietyt vastaukset. Järjestelmän antaman välittömän palautteen olisi syytä tukea ainakin seuraavia ominaisuuksia, jotka on toteutettu myös ConPraxiksen tehtävätarkastukseen:

- Kirjoitusvirheiden huomioiminen esimerkiksi tähän usein käytetyllä Levenshteinin (1966) algoritmilla: “Kirjoititko sanan varmasti oikein?”.
- Väärin valittujen taivutusmuotojen huomioiminen: “Oikea sana, mutta nominatiivi on väärä taivutusmuoto.”
- Oikean taivutusmuodon huomioiminen, vaikka sana olisi väärä: “Taivutit sanan oikein monikon genetiivissä, mutta sana on väärä.”

Yllättävien vastausten huomioimisen kannalta olisi tärkeää, että verkkosanakirja pystyisi tekemään korpukseen myös uusia hakuja eli vastaamaan kysymykseen siitä, käytetäänkö kielenoppijan syöttämää sanaa laisinkaan kyseisessä lauseyhteydessä. ConPraxiksessa ei tätä mahdollisuutta ole vielä käytetty. FinnWordNet-hankkeen (Lindén & Carlson 2010) yleiseen käyttöön tarjoama suomennos englanninkielisestä WordNet-hankkeesta (Miller 1995; Fellbaum 1998) puolestaan tarjoaa algoritmeille mahdollisuuden antaa palautetta, joka perustuu sanojen merkitysten välisiin suhteisiin. Kahden edellä olevan perusteella voidaankin listata joukko haastavampia mahdollisuuksia, joita pystytään toteuttamaan korpusten avulla:

- Todellisen käytön huomioiminen korpuksen perusteella: “Tässä yhteydessä ei yleensä käytetä tätä sanaa.”
- Merkityksen huomioiminen sanasuhteiden perusteella: “Etsimme tarkempaa ilmausta.” / “Etsimme tämän sanan synonyymiä.”

- Tekstilajin ja tyylin huomioiminen: “Vastauksesi sopii epäviralliseen kommunikaatioon muttei viralliseen tekstiin.”

Näillä menetelmillä voidaankin syötettyjen lauseiden semanttisen sisällön järkevyys jo varsin pitkälle tarkistaa: siinä missä sellaisen syötteitä tarkistavan tekoälyn luominen, joka ymmärtäisi lauseen merkitykset ihmisen tavoin, on erittäin vaikeaa ja virheille altista, voidaan tämänkaltaisilla tilastollisilla menetelmillä korvata tämä ymmärrys korpukseen pohjautuvalla tilastollisella tiedolla siitä, että ihmiset ovat käyttäneet tai eivät ole käyttäneet kyseistä ilmausta.

6. Lopuksi

Sanakirjatyö on aina haastavaa, olipa kyse minkätyyppisestä sanakirjasta tahansa. Omat erityishaasteensa on kielenoppijoille tarkoitetulla sanakirjalla, etenkin sellaisella, joka perustuu olemassa oleviin tekstimateriaaleihin ja tällaisen materiaalin tarkkaan analyysiin. Sähköisiä sanakirjoja koskevat taas omat vaatimuksensa. (Sanakirjojen laatimisesta ks. esim. Heinonen 2013: luku 4, sähköisistä sanakirjoista mm. Granger & Paquot 2012.) Moonin (2009) mukaan sanakirjatyössä eivät tekijät ja teoreetikot ole välttämättä aina kohdanneet, ja käytännön sanakirjatyö on ollut toisinaan hyvinkin teoriatonta. Käytännön työ ja teoreettinen leksikografia eivät saisi olla kuitenkaan toisistaan irrallisia, sillä samoin kuin missä tahansa tutkimustyössä, teoreettinen tausta antaa perustan menetelmille ja itse työlle – ja jopa auttaa ymmärtämään, miksi sanakirja kannattaa ylipäätään tehdä.

Työmäärän vähentäminen ja sana-artikkelien tuottamisen nopeuttaminen ovat asioita, joihin on tulevaisuudessa pystyttävä luomaan ratkaisuja. Ainakin jossain määrin automaattisesti voidaan jo luoda korpuksia, valita hakusanoja sekä analysoida syntagmaattisia ja paradigmaattisia suhteita. Tulevaisuutta voisi olla se, että tietokoneohjelmat saisivat entistä itsenäisemmin ratkaista, esimerkiksi esiintymien tilastollisuuden mukaan, mikä tieto on kielen kuvauksen ja vastaanottajan

kannalta relevanttia. Tällä hetkellä manuaalisen työn osuus on kuitenkin vielä suuri, ja kielenoppijoille tarkoitetun sanakirjan toimittamisessa sen osuus säilynee merkittävänä mm. siksi, että autenttista esimerkkiaineistoa on muokattava kohderyhmälle sopivaksi. Suuria kysymyksiä ovat myös nykyisten kieliaineistojen riittävyys, ajantasaisuus ja rekisterien kapeus: kun sanakirjatyö on aloitettava todellisten kieliaineistojen kokoamisella, ajantasaistamisella ja monipuolistamisella, on työmäärä loppumaton ja prosessi hidas. Onkin todennäköistä, että (sähköisten) sanakirjojen tekemisen ja yleensäkin leksikografian kehittyminen on sidoksissa kieliteknologisten ratkaisujen kehittymiseen. Myös jo olemassa olevia mahdollisuuksia ja tarpeita on selvitettävä entistä syvällisemmin tiivistämällä eri toimijoiden yhteistyötä.

Kiitokset

Kiitämme Opetushallitusta aikuisten maahanmuuttajien koulutuksen kehittämiseen (ConLexis- ja ConPraxis-hankkeet) vuosina 2008–2009 ja 2011 myönnetystä rahoituksesta.

Lähteet

Atkins, B. T. Sue 1993. Theoretical lexicography and its relation to dictionary-making. – Dictionaries: Journal of the Dictionary Society of North 14, 4–43. http://dx.doi.org/10.1353/dic.1992.0011

Bessonoff, Salli-Marja 2000. Suomen sanoja muunkielisille. – Virittäjä 104 (2), 314–318.

Biber, Douglas, Susan Condrad, Randi Reppen 1998. Corpus Linguistics. Investigating Language Structure and Use. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511804489

Cobuild 1987 = Collins Cobuild English Language Dictionary. London: Cobuild. Toinen painos nimellä Collins Cobuild English Dictionary 1995. London: HarperCollins.

East, Martin 2008. Dictionary Use in Foreign Language Writing Exams. Impact and Implications. Language Learning & Language Teaching 22. Amsterdam/Philadelphia: John Benjamins.

EVK = Eurooppalainen viitekehys: Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys 2008. WSOY: Helsinki.

Fellbaum, Christiane (Ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Granger, Sylviane 2005. Pushing back the limits of phraseology: How far can we go? http://hdl.handle.net/2078.1/75667 (24.2.2013).

Granger, Sylviane, Fanny Meunier 2008. Phraseology in language learning and teaching. Where to from here? – Fanny Meunier, Sylviane Granger (Eds.). Phraseology in Foreign Language Learning and Teaching. Amsterdam: John Benjamins, 247–252.

Granger, Sylviane, Magali Paquot (Eds.) 2012. Electronic Lexicography. Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199654864.001.0001

Heinonen, Tarja 2013. Idiomien leksikaalinen kuvaus kielenkäytön ja vaihtelun näkökulmasta. Helsingin yliopiston nykykielten laitos.

Jantunen, Jarmo H. 2009. Minulla on aivan paljon rahaa – Fraseologiset yksiköt suomen kielen opetuksessa. – Virittäjä 133 (3), 356–381.

Jokinen, Päivi, Pirjo Immonen-Oikkonen, Leena Nissilä (Toim.) 2011. Kommentoitu luettelo maahanmuuttajataustaisten opetuksen ja koulutusten materiaaleista. Helsinki: Opetushallitus.

Järventausta, Marja 2009. Kakkossuomen perussanakirja. – Virittäjä 113 (1), 89–100.

Jönsson-Korhola, Hannele, Leila White 1999. Tarkista tästä. Suomen sanojen rektioita suomea vieraana kielenä opiskeleville. Toinen, korjattu painos. Helsinki: Finn Lectura.

Kangasniemi, Heikki 2003. Suomen kielen sanastoharjoituksia. Helsinki: Tammi.

Kangasniemi, Heikki 2006. Suomen kielen tekstinymmärtämisharjoituksia. Helsinki: Sanoma Pro.

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell, Pavel Rychlý 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. – Elisenda Bernal, Janet DeCesaris (Eds.). Proceedings of the XIII Euralex International Congress. Barcelona: Institut Universitari de Linguistica Aplicada. Universitat Pompeu Fabra, 425–432.

Levenshtein, Vladimir 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10, 707–710. [Julkaistu alun perin: Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР 163 (4), 845–848.]

Lindén, Krister, Lauri Carlson 2010. FinnWordNet – WordNet på finska via översättning. – LexicoNordica 17, 119–140.

Mauranen, Anna 2004. Spoken Corpora for an Ordinary Learner. – John McH Sinclair (Ed.). How to Use Corpora in Language Teaching. Studies in Corpus Linguistics 12. Philadelphia: John Benjamins, 89–105.

MEDAL = MacMillan English Dictionary for Advanced Learners 2007. 2., uudistettu painos. Oxford: MacMillan Education.

Miller, George A. 1995. WordNet: A Lexical Database for English. – Communications of the ACM 38 (11), 39–41. http://dx.doi.org/10.1145/219717.219748

Moon, Rosamund 2009. Sinclair, lexicography, and the Cobuild Project. The application of theory. – Rosamund Moon (Ed.). Words, Grammar, Text. Revisiting the Work of John Sinclair. Amsterdam: John Benjamins, 1–22.

Muikku-Werner, Pirkko, Jarmo Harri Jantunen, Ossi Kokko 2008. Suurella sydämellä ihan sikana. Suomen kielen kuvaileva fraasisanakirja. Helsinki: Gummerus.

Nesselhauf, Nadja 2005. Collocations in a Learner Corpus. Studies in Corpus Linguistics, 14. Amsterdam: John Benjamins.

Nurmi, Timo 2009. Suomen sanakirja opiskelijoille ja ulkomaalaisille. Finnish Dictionary for Students and Foreign Learners. Helsinki: Gummerus.

Rundell, Michael 1998. Recent trends in English pedagogical lexicography. – International Journal of Lexicography 11 (4), 315–342.

Rundell, Michael, Adam Killgarrif 2011. Automating the creation of dictionaries. Where will it all end? – Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin, Magali Paquot (Eds.). A Taste for Corpora. In Honour of Sylviane Granger. Amsterdam: John Benjamins, 257–281.

Saarikalle, Anne, Johanna Vilkuna 2010. Suomen kielen sanakirja maahanmuuttajille. Helsinki: Gummerus.

Saunela Marja-Liisa 2008a. Harjoitus tekee mestarin 3. Suomen kielen syventäviä harjoituksia maahanmuuttajille. Helsinki: Art House.

Saunela, Marja-Liisa 2008b. Harjoitus tekee mestarin 4. Lisäharjoituksia yleiseen kielitutkintoon valmistautumista varten. Helsinki: Art House.

Sinclair, John 1996. The search for units of meaning. – Textus IX, 75–106.

Skiepmann, Dirk 2008. Phraseology in Learners' Dictionaries. What, where and how? – Fanny Meunier, Sylviane Granger (Eds.). Phraseology in Foreign Language Learning and Teaching. Amsterdam: John Benjamins, 185–202.

Suni, Minna 2008. Toista kieltä vuorovaikutuksessa. Kielellisten resurssien jakaminen toisen kielen omaksumisen alkuvaiheessa. Väitöskirja. Jyväskylä studies in humanities 94. Jyväskylä: Jyväskylän yliopisto.

Tanner, Johanna 2012. Rakenne, tilanne ja kohteliaisuus. Pyynnöt S2-oppikirjoissa ja autenttisissa keskusteluissa. Väitöskirja. Suomen kielen, suomalais-ugrilaisten ja pohjoismaisten kielten ja kirjallisuuksien laitos. Helsinki: Helsingin yliopisto. http://urn.fi/URN:ISBN:978-952-10-7903-0 (25.2.2013).

Thompson, Geoff 1987. Using bilingual dictionaries. – ELT Journal 41 (4), 282–286. http://dx.doi.org/10.1093/elt/41.4.282

Tsui, Amy B. M. 2004. What teachers have always wanted to know – and how corpora can help. – John McH Sinclair (Ed.). How to Use Corpora in Language Teaching. Studies in Corpus Linguistics 12. Philadelphia: John Benjamins, 39–61.

Widdowson, Henry 1998. Context, community and authentic language. – TESOL Quarterly 32 (4), 705–716. http://dx.doi.org/10.2307/3588001

Jarmo Harri Jantunen
Suomen kieli, Kielten laitos
PL 35, 40014 Jyväskylän yliopisto, Finland
jarmo.jantunen@jyu.fi

Marjo Kumpulainen
Oulun Aikuiskoulutuskeskus
Kotkantie 3
90250 Oulu, Finland
marjo.kumpulainen@oakk.fi

Tanja Tammimies
Oulun Aikuiskoulutuskeskus
Kotkantie 3
90250 Oulu, Finland
tanja.tammimies@oakk.fi

Teemu Tokola
Oulun yliopisto, Tietotekniikan osasto
PL 4500, 90014 Oulun yliopisto, Finland
teemu.tokola@oulu.fi

Towards a corpus-based online learner dictionary: ConLexis

JARMO HARRI JANTUNEN
University of Jyväskylä

MARJO KUMPULAINEN
Oulu Adult Education Centre

TANJA TAMMIMIES
Oulu Adult Education Centre

TEEMU TOKOLA
University of Oulu

The theme of our article is corpus-based learners' dictionaries. These dictionaries are based on large amounts of digital text and research results on the use of language. These types of dictionaries are still scarce, and are even more so for language learners – for the Finnish language such dictionaries do not exist at all. Additionally, nowadays more and more dictionaries are published in an electronic format, which makes possible many features not available in paper format dictionaries. Large corpora have also had the effect of making phraseology, study of how words are used in context, become an increasingly studied topic, as already shown by some corpus-based dictionaries in the English language. However, it seems that in this respect there is still a lot to hope for: phraseological emphasis is still missing in a significant part of the dictionaries where such emphasis would benefit the learners.

In this paper we present the general requirements for electronic dictionaries aimed towards language learners and the design, aims and contents of a new ConLexis web dictionary. ConLexis is aimed for language learners of at least B1 level. However, teachers may use it as supporting material also at lower skill levels. Information presented in the ConLexis word articles is based on extensive corpus analysis, and has a phraseological emphasis. Alongside ConLexis, an

exercise material set called ConPraxis is being developed. These together form a corpus-based self-study material for language learners and a teachers' aid for teachers of Finnish language. Each word article in the dictionary presents the word alongside its definition and conjugations, provides links to synonyms and antonyms and a list of most important collocates. For each of the collocates, a list of example phrases derived from the used corpora is presented. Some of these phrases may be shortened or otherwise slightly altered from the form in which they appear in the corpus, to make them more relevant to the exercise, for example by removing content not necessary for demonstrating the use of the word and its collocate. The ConPraxis exercises in turn contain similarly selected phrases, which have been turned into exercises by removing words that the learner needs to fill in. The dictionary provides an instant feedback mechanism that takes into account typing errors and is able to conjugate words to provide meaningful advice in a majority of situations in which the answer is nearly correct.

Keywords: learners' dictionaries; online materials; corpus dictionaries; phraseology; collocations; synonyms; antonyms