Suomen kielen hallinta ja sen kehitys

Peruskoululaiset ja nuoret aikuiset

Toimittaneet

ANNELI PAJUNEN JA MARI HONKO

SUOMALAISEN KIRJALLISUUDEN SEURA • HELSINKI • 2021

SUOMALAISEN KIRJALLISUUDEN SEURAN TOIMITUKSIA 1472

Teos on Suomalaisen Kirjallisuuden Seuran nimeämien asiantuntijoiden tarkastama.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Kannen suunnittelu: Timo Numminen
Taitto: Maija Räisänen
EPUB: Tero Salmén

ISBN 978-951-858-407-3 (nid.)
ISBN 978-951-858-408-0 (EPUB)
ISBN 978-951-858-409-7 (PDF)

ISSN 0355-1768 (nid.)
ISSN 2670-2401 (verkkojulkaisut)

DOI https://doi.org/10.21435/skst.1472

Teos on avoimesti saatavissa osoitteessa https://doi.org/10.21435/skst.1472 tai lukemalla tämä QR-koodi mobiililaitteella.

TEHTY SUOMESSA
PRODUCERAD I FINLAND

Hansaprint Oy, Turenki 2021

https://doi.org/10.21435/skst.1472

Unelmieni päivä -sanasto kehityksellisestä näkökulmasta

Anneli Pajunen
https://orcid.org/0000-0001-8597-7491

Mari Honko
https://orcid.org/0000-0002-3240-622X

Johdanto

Sanaston hyvä hallinta on lukuisissa tutkimuksissa osoitettu hyvän luku- ja kirjoitustaidon keskeiseksi taustatekijäksi (ks. esimerkiksi Carlisle 2000; Qian 2002; Nippold 2006, 25–29, 89; Wagner, Muse & Tannenbaum 2007; Milton 2009, 170–192; Schmitt 2010, 3–12; suomen osalta Saarela 1997; Honko 2013; Kusnetsoff 2017). Hart ja Risley (1995, 2003) ovat osoittaneet, että Yhdysvalloissa heikko sanatieto päiväkoti-ikäisenä selittää heikosta lukutaidosta yhdeksänvuotiaana jopa 90 prosenttia. Nuorimpien osalta on myös tuloksia, joiden mukaan sanaston monimuotoisuus ennustaa taitoa kirjoittaa kertomuksia (Olinghouse & Wilson 2013). Varhaislapsuudessa syntyneet erot sanastonhallinnassa voivat siis olla kauaskantoisia ja joko haitata tai edistää toiminnallisen kielitaidon kehittymistä. Suomesta vastaavat tutkimukset puuttuvat,

https://doi.org/10.21435/skst.1472

tosin äidinkielen arviointitutkimuksissa on joitain viitteitä heikon osaamisen ja sanaston yhteydestä (ks. esimerkiksi Kuusela 2011).[10]

Aikaisemmin on havaittu, että kielen, sanaston ja peruskäsitteiden osaaminen kehittyy ikäluokkatasolla yläkoulun aikana lukemisen ja kirjoittamisen taitoa vähemmän ja joskus jopa taantuu (Harjunen & Rautopuro 2015, 24–25). Nykyisissä opetussuunnitelman perusteissa tarkoitus on kuitenkin tukea oppilaita tarkastelemaan kieltä vuorovaikutuksessa ja myös sanastoa sen keskeisenä osana. Tekstien tulkinnan ja tuottamisen kuvauksen yhteydessä oppilaita kannustetaan erityisesti laajentamaan sana- ja käsitevarastoaan. Sanasto mainitaan erikseen kunkin ikätason äidinkielen ja kirjallisuuden oppimäärien kuvauksissa, ja sanasto on nimettynä sisältönä mukana myös opetussuunnitelman perusteiden erityisen tuen osiossa osana kommunikaatiotaitoja, joihin sanojen tunnistaminen ja käyttökin kuuluvat (OPH 2014, 72).

Kieli läpäisee kaikki oppiaineet, mutta erityisesti sanasto huomioidaan opetussuunnitelman perusteissa kielten oppimäärien kuvauksissa sekä suomen kielen ja kirjallisuuden kuvauksissa. Alkuopetuksessa korostetaan sanojen havainnointia, sanojen ja sanontojen merkitysten sekä tekstien sanavalintojen pohtimista ja oman käsitevarannon laajentamista. Vuosiluokilla 3.–6. rohkaistaan sana- ja käsitevarannon laajentamiseen sekä analyyttiseen tarkasteluun kuten ”selittämään, vertailemaan ja pohtimaan sanojen, niiden synonyymien, kielikuvien, sanontojen ja käsitteiden merkityksiä ja niiden hierarkioita” (mts. 162–164). Oman tuottamisen yhteydessä mainitaan tekstien elävöittäminen ja sanavalintojen tarkastelu osana tekstin merkityksiä. Vuosiluokilla 7.–9. oppilaita kannustetaan vakiinnuttamaan ja laajentamaan sana- ja käsitevarantoaan sekä pohtimaan kielellisten valintojen merkityksiä ja seurauksia. Erikseen kannustetaan sanastoon liittyvien rekisteri- ja tyylipiirteiden tutkimiseen sekä eri kielten sanaston vertailuun ja kuhunkin tekstiin sopivien ilmaisu-

10 Hartin & Risleyn (1995) tutkimuksen mukaan lapsen leksikon koko on Yhdysvalloissa riippuvainen vanhempien sosioekonomisesta asemasta. Michaels (2013) kritisoi tätä väitettä jyrkästi. Hänen mukaansa tutkimuksen asetelma on rakennettu tulosta ennustavaksi eikä sana-aineistoa ole koodattu esimerkiksi kehitykselliseltä kannalta kriittistä abstraktisanastoa erotellen. Kolmas ja yhteiskunnalliselta kannalta vakavin kritiikin aihe on Michaelsin mukaan se, että huono koulusuoriutuminen langetetaan tutkimuksessa lasten ja heidän vanhempiensa syyksi.

https://doi.org/10.21435/skst.1472

tapojen valikointiin. (Mts. 291–291.) Valtakunnalliset opetussuunnitelman perusteet eivät kuitenkaan anna konkreettisia neuvoja siihen, millä tavalla sanatietoa tulisi opettaa tai tuottamisen ja tulkinnan taitoja harjoitella. Paikallisella opetussuunnitelmatyöllä, opettajien omalla valveutuneisuudella ja aktiivisuudella sekä oppikirjantekijöillä onkin merkittävä rooli siinä, miten esimerkiksi sanamerkitysten havainnointiin ohjataan.

Oma kokemuksemme opettajina ja opettajankouluttajina sekä aikaisempi tutkimus antavat viitteitä siitä, että sanastoa käsitellään opetuksessa usein yhä melko yksipuolisesti ja mekaanisesti niin, että painopiste on irrallisten sanojen nimeämisessä sekä luokittelussa. Kauppisen, Tarnasen ja Aallon (2014) tutkimuksen mukaan luokanopettajaopiskelijat eivät myöskään hahmota sanastoa osaksi äidinkielen opetuksen sekä reaaliaineiden opetuksen oppiaineintegraatiota, vaan yhteyksiä nähdään enemmän tekstilajituntemukseen ja jopa kielitiedon opettamiseen. Silloin, kun sanastoa ja sanatietoa opetuksessa tarkastellaan, huomio tuntuu kiinnittyvän kykyyn nimetä ja luokitella sanoja annettujen kategorioiden mukaan, ei itse prosessiin tai luokitteluperusteisiin kuten sanojen käytön, merkityksen ja muodon analyyttiseen havainnointiin. Kognitiiviseen päättelyyn ja esimerkiksi luokitteluperusteiden kielentämiseen ohjaamisen kulttuuri tarvitsisi vahvistusta, jotta oppiminen – ajattelun ja kielellisten taitojen tukeminen – nousisi keskiöön kategorioiden läpi käymisen sijaan. Oletus tuntuu olevan, että sanasto opitaan riittävän hyvin käyttöyhteyksistä ja erityisesti omaehtoisen lukemisharrastuksen avulla. Tämä näkyy paitsi oppimateriaaleista ja opetussuunnitelmista myös reaktioista: kun lukutaidon PISA-tulokset ovat laskeneet, ratkaisuksi on kehitelty tuettuja lukukampanjoita.[11] Tutkimuksen valossa tämä on tärkeää mutta tuntuu riittämättömältä tilanteessa, jossa suomalaislasten ja nuorten lukemisaktiivisuus vähenee ja luku- ja kirjoitustaito heikkenee.

Koulutuksen pitäisi tasata ja sen oletetaan tasaavan osaamistasoa (ks. esimerkiksi Hadley & Rispoli 2012), mutta oman tutkimuksemme

11 Esimerkiksi Opetushallituksen Lukuliike ja siihen nidotut kampanjat koettavat lisätä lukemisharrastusta (ks. https://lukuliike.fi). Tavoite on hyvä, mutta ei välttämättä kohdistu siihen kasvavaan joukkoon, jonka ongelma on kielenhallinnassa tai ainakin kirjoitetun kielimuodon vieraudessa.

https://doi.org/10.21435/skst.1472

mukaan sanamerkityksen hallinnan taso, kyky analyyttisesti määritellä sanojen merkityksiä ja taito ymmärtää tai tuottaa johdoksia vaihtelee paljon myös ylioppilastutkinnon suorittaneilla nuorilla aikuisilla, joiden ikäluokastaan koulutetuimpina pitäisi omata melko yhtenevä taitotaso (ks. Pajunen, Itkonen & Vainio 2015; 2016; Vainio, Pajunen & Häikiö 2019; Laasanen, Pajunen & Häikiö, arvioitavana). Sanaston hallinnan tutkimus on Suomessa ollut vähäistä, mutta kansainvälinen tutkimus on kasvavaa ja tiedetään, että vahvat sanastotaidot ovat yhteydessä paitsi hyviin tekstitaitoihin ja niiden edelleen kehittymiseen – myös hyvään koulumenestykseen ja opintopolulla etenemiseen (ks. esimerkiksi Hart & Risley 2003; Treffers-Daller & Milton 2013; Nation & Coxhead 2021). Sanaston hyvä hallinta on myös itseisarvo, joka säilyttää niin kulttuuria kuin edistää yksilön kykyä ilmaista itseään täsmällisesti ja osuvasti.

Kartoitamme tässä yhteydessä koululaisten sanastotaitojen kehittymistä tutkimalla, miten eri-ikäisten kirjoittajien sanaston käyttö muuttuu samasta aiheesta eli unelmien päivästä laadituissa Unelmakirjoitelmissa (alakouluikäisten unelmasanastosta tarkemmin ks. Laine-Leinonen 2013). Ikäluokat ovat alakoululaiset ja yläkoululaiset, joita tarkastellaan myös koululaisten ryhmänä, ja osaksi ryhminä ovat kolmas- ja kuudesluokkalaiset sekä yhdeksäsluokkalaiset. Vertailukohtana ovat nuoret aikuiset. Aineiston sanaesiintymien määrä on yli 200 000 ja ryhmittäin korkeintaan 10 000 eri sanaa (lekseemiä). Oletus aiemman tutkimuksen perusteella on, että käytetyn sanaston yleisyys laskee samalla, kun sanaston diversiteetti eli eri lekseemien määrä ja myös semanttinen monipuolisuus lisääntyy. Käsittelemme aineistoa ryhmätasolla, ja tavoitteena on paitsi osoittaa sanaston monipuolistuminen myös luoda tietoa ikäryhmittäisestä sanaston hallinnan keskimääräisestä tasosta. Jotta yksilöiden väliset erot eivät peittyisi, kuvaamme analyysien yhteydessä myös yksilöllistä vaihtelua. Käsittelemme ensin tutkimuksen teoreettista taustaa ja keskeisiä käsitteitä sekä aineistoa, ja keskitymme sitten kirjoitelmasanaston yleisyyteen, diversiteettiin ja lopuksi semanttiseen luonteeseen. Loppulukuun kootaan tulokset.

https://doi.org/10.21435/skst.1472

Sanaston frekvenssitekijät, rikkausluvut sekä semanttinen luokitus

Sanaston hallinnan tutkimuksessa tavoitteena on yleensä saada tietoa käytössä olevan sanaston yleisyydestä, sen moninaisuudesta ja semanttisesta luonteesta. Yleisyyttä voidaan tarkastella lekseemi- ja esiintymätasolla sekä hyödyntämällä kattavuuslukuja, yleisyystasoja ja niin sanottuja perhefrekvenssejä. Näin voidaan tehdä ennusteita yleisimmin hallitun ja käytetyn sanaston luonteesta, sanojen oppimisiästä, prosessoinnin luonteesta ja niin edelleen. Puhutaan niin sanotuista frekvenssitekijöistä (ks. Bybee & Hopper 2001; Baayen 2014): yleisyys vaikuttaa suuresti siihen, minkälaiselle kielelle kielenkäyttäjät altistuvat ja on siksi yhteydessä myös sanaston tyypilliseen oppimisjärjestykseen (ks. esimerkiksi Treffers-Daller & Milton 2013). Sanaston moninaisuutta teksteissä arvioidaan erilaisilla variaatio- ja rikkausluvuilla (ks. esimerkiksi Jarvis 2013), ja semanttinen analyysi puolestaan edellyttää aineiston alaluokitusta ensin sanaluokkakohtaisesti ja edelleen merkityksen mukaan (ks. esimerkiksi Miller & Fellbaum 1991).

Sanojen yleisyyden arviointiin tarvitaan massakorpus, jonka koon olisi luotettavuuden vuoksi hyvä olla vähintään kymmenissä miljoonissa. Massakorpuksessa koko on yleensä tärkeämpi kuin sen koostumus, koska mikään korpus ei ole harvinaisten sanojen osalta edustava ja koska yleisimmät sanat selviävät luotettavasti pienemmästäkin massakorpuksesta. Itse käytämme HS2000-korpusta (n. 24 milj. sanaesiintymää),[12] jota täydennämme Tieteellisen laskennan antamin yleisyyksin (CSC korpus, 44 milj. sanaesiintymää). Korpusfrekvenssit annetaan yleisyytenä miljoonassa sanassa.

Yleisyystasoja lasketaan jakamalla lekseemit yleisyyden mukaan tuhannen siivuihin: Ensimmäinen tuhat sisältää kielen yleisimmät sanat. Yleensä riittää, jos erotellaan kymmenen yleisintä tasoa (ks. Laufer & Nation 1999, 39; Milton 2009, 22–43), ja tällöin harvinaiset sanat voidaan määritellä niiksi, jotka kuuluvat yleisyystasoille 9 tai sen yli ja

12 HS2000-korpuksen kokonaissanamäärä on noin 31 miljoonaa; laskelmat perustuvat sana-ainesmerkkimäärään.

https://doi.org/10.21435/skst.1472

joiden esiintymiä miljoonassa sanassa on yksi tai vähemmän. Mitä harvinaisempi sana on, sen epätodennäköisempää myös on, että se osuu laskelmiin käytettävän korpuksen sanastoon. Siten harvinaiset sanat ovat myös niitä, joita ei korpuksista löydy. Koska harvinaisista sanoista on vaikea saada luotettavaa yleisyystietoa (vrt. esimerkiksi Keuleers, Stevens, Mandera & Brysbaert 2015), niiden tuttuus tietyssä ikäluokassa on parempi testata erikseen käyttämällä subjektiivisia testejä (vrt. Pajunen & Itkonen 2019). Kehityksellisen tutkimuksen tarpeisiin riittää usein karkea yleisyystasojaottelu hyvin yleisestä tavalliseen ja harvinaiseen, koska aikuisenkin puhujan sanatieto alkaa nopeasti heiketä jo yleisyystasoilla 9–10.

Perhefrekvenssejä voidaan laskea suomen tapaisessa kielessä muun muassa tutkimalla sanueitten jäsenkokoa (kantasana ja sen johdokset ja yhdyssanat): mitä suurempi sanue, sen todennäköisemmin sana tunnetaan hyvin ja prosessoidaan nopeasti (ks. Pajunen, Vainio & Itkonen 2015; Vainio, Pajunen & Häikiö 2019). Suomen kielessä substantiiviperheet ovat paljon suurempia kuin verbiperheet, koska substantiivilekseemien ja verbilekseemien suhde on korkeintaan 8:2 ja koska substantiiveja muodostetaan myös yhdistämällä, verbejä vain harvoin. Substantiiviperhe on suuri, jos jäseniä on vähintään parisenkymmentä, mutta verbiperhe on suuri, jos niitä on vain muutamiakin. Perhefrekvensseihin kuuluu myös niin sanottu taivutusperhefrekvenssi, joka saadaan yksinkertaisesti laskemalla massakorpuksesta sanan kaikki taivutusmuototyypit ja esiintymät tyypeittäin. Kaikki korpuksessa esiintyvät muodot, kuten kunkin substantiivisanan yksikkö ja monikko eri sijoissa tai liitepartikkeleissa taivutettuina, ovat laskennassa eri tyyppejä. Esimerkiksi *hykerrellä*-verbin taivutusperhekoko on HS2000-korpuksessa 8 ja yleisyys 21 (ks. taulukko 3). Nyt verbien taivutusperheet ovat paljon suurempia kuin substantiivien. Verbilekseemien toistuvuus on korkeampi kuin substantiivilekseemien, ja se nostaa eri konteksteissa ja siten useammissa eri taivutusmuodoissa esiintymistä.

Taivutusperheluvut korreloivat melko hyvin esiintymäyleisyyteen korkeimpien lukujen osalta, mutta pienet taivutusperheet viittaavat sanan käytön vähäisyyteen tai rajoituksiin ja merkityksen vähittäiseen idiomaattisuuteen. Taivutusperhekoko vaihtelee yhdestä kahteen kol-

https://doi.org/10.21435/skst.1472

meen sataan. Koko osoittaa, kuinka monipuolinen malli kielenkäyttäjällä sanasta on. Iso taivutusperhe tukee esimerkiksi vartalonvaihteluiden hallintaa, pieni vastaavasti heikentää.[13] Kehityksellisestä näkökulmasta katsoen yleisen sanan vähäinen taivutus voisi viitata rajoittuneeseen käyttökontekstiin, mutta melko yleisen sanan maksimaalinen taivutus viittaisi monipuoliseen käyttökontekstiin. Jälkimmäinen tilanne edistää sanan opittavuutta.

Taulukko 3. Esimerkki taivutusmuotoperheestä.

Lekseemi	Yleisyys	Taivutusperhe	Sanamuodon esiintymä	Sanamuodon yleisyys
hykerrellä	21	8	*hykerrellen*	1
			hykerrellä	2
			hykertelee	7
			hykertelemään	1
			hykertelevät	1
			hykertelevät	3
			hykerteli	4
			hykertelivät	2

Englannissa samaan perhefrekvenssiin lasketaan sekä sanue- eli derivaatioperhe että taivutus(muoto)perhe (ks. Bauer & Nation 1993). Vastaava laskelma sisältäisi suomessa esimerkiksi paitsi kaikki *käsi*-sanan muodot myös sen johdokset ja kaikkien näiden johdosten kaikki eri taivutusmuototyypit. Tällainen perhefrekvenssi on suomesta mahdoton laskea, koska mikään ohjelma ei yhdistä kannan kaikkia johdoksia yhteen – yhdyssanoista puhumattakaan – eikä laske lisäksi kaikkia niiden taivutusmuototyyppejä. Koska suomessa johtamattomien kantasanojen määrä on pieni ja johtaminen keskeinen sananmuodostuskeino, voi myös epäillä, olisiko tällaisista perheluvuista hyötyä. Nuorimmat kouluikäiset eivät myöskään vielä hallitse suomen sananmuodostuksen pe-

13 Sanat *keriminen*, *keritseminen* ja *kerkeäminen* voivat muodosta riippuen saada kantasanakseen joko *keriä* tai *keritä*-muodon ja toisinaan jokin taivutusmuoto laukaisee myös *kerjetä*-verbin. HS2000-korpuksessa taivutusmuotoja on *keritä*-verbistä 10 (joista 1 liittyy keritsemiseen, 9 kerkeämiseen) ja *keriä*-verbillä 14 (joista esimerkiksi passiivimuoto voi liittyä sekä keritsemiseen että kerimiseen). Lampaan keritsemiskuvan kielennystehtävässä suurin osa koehenkilöinä toimineista nuorista aikuisista nimesi kuvan kerimistapahtumaksi (Pajunen 2017). Toisin sanoen *keritsemisen* pieni taivutusperhe korpuksessa ennusti virhetulosta, vaikuttipa siihen mikä tahansa tekijä.

https://doi.org/10.21435/skst.1472

riaatteita vaan joko tuntevat tai eivät tunne yksittäisiä johdoksia näkemättä samaan perheeseen kuuluvien johdosten välisiä yhteyksiä (ks. Kusnetsoff 2017).

Laskemiseen liittyy synteettisessä kielessä monia periaatteellisia ongelmia, joita toistaiseksi ei ole ratkaistu, vaikka hyviä laskemistyökaluja löytyisikin. Miten esimerkiksi määritellään sanueen lähtökohtana oleva (kanta)sana, tai miten otetaan produktiivisuus huomioon.[14] Paras tuntemamme väline taivutusmuotojen laskemiseen on WordMill (ks. Laine & Virtanen 1999), joka antaa tiedot suoraan hakutuloksena mutta määrittelee kantasanan (tai lemman) paikoin epätoivottavasti (esimerkiksi *sammakoiden* < *sammakoida). Ohjelma myös sisällyttää taivutusmuotoihin muutamia johdoksia, jotka halutessaan voi erotella vain manuaalisesti. Sanueiden kokoa joutuu toistaiseksi laskemaan sanakirjojen avulla. Synteettisestä kielestä saadaan myös sanarakennetta ja sanapituuksia arvioimalla kuva frekvenssitekijöistä ja sanaperheistä: johtamattomat sanat ovat yleensä johdoksia ja yhdyssanoja yleisempiä ja lyhyet sanat pitkiä yleisempiä.

Käytämme sanaston kehittymisen tutkimuksen tukena yleisyyslukuja (/miljoona sanaesiintymää), yleisyystasotietoja ja taivutusperhefrekvenssejä. Sanaperhelukuja emme käytä mutta niiden vaikutus peruskouluikäisten kielenhallintaan tuli esiin hankkeen derivaatiotestissä (ks. Vainio ja muut 2019). Lisäksi tarkastelemme sanaston monimuotoisuutta (diversiteettitunnusluku *sum of probabilities*) sekä sen jakautumista sisältö- ja kieliopillisiin sanoihin sekä eri sisältösanaluokkien sanoihin. Nämä tarkastelutavat mahdollistavat hyvin erilaajuisten sekä kokonaisrakenteeltaan että tyyliltään erilaisten tekstien vertailun, ja ne ovat perusteltuja myös aikaisemman oppijankielen tutkimuksen perusteella. Varsinaisia oppimisikätietoja ei ole suomen kielestä kerätty.

Analysoimme kirjoitelmien sanaston myös semanttisesti. Semanttinen luokitus voi perustua yksin käsiteanalyysiin, jolloin ei tarvitse tehdä

14 Hakulinen (1979, 478) toteaa sekä yksilön sanaston että murresanastojen kokoa arvioidessaan, että suomessa lukuihin on laskettava mukaan paitsi johtamaton sana myös jokainen johdos. Sanakirjoissahan hyvin produktiiviset johtimet, kuten teettokausatiivijohtimet ja *minen*-johdin/muoto jätetään pääasiassa huomiotta. Kummankin oppiminen on kehityksellisesti mielenkiintoista. Lopputulema kuitenkin on, että derivaatioperheen laskeminen on suomessa erityisen vaikeaa.

https://doi.org/10.21435/skst.1472

jyrkkää sanaluokkaeroa. Käytännössä on helpompi lähteä leksikaalisista luokista, jotka perustuvat ontologis-semanttiseen kolmijakoon 'oliot', 'tapahtumat' ja 'ominaisuudet'. Nämä luokat kattavat leksikaalisesta massasta valtaosan (91 % omassa aineistossamme) ja niiden jäsenet vaikuttavat siihen, millä täsmällisyydellä ja abstraktiuden tasolla maailmassa esiintyviä todellisia ja kuviteltuja olioita ja tapahtumia pystytään nimeämään. Ne myös ilmaisevat kulttuuris-yhteiskunnalliset sävyt ja tyylilliset erot, ja muun muassa frekvenssitekijöiden suhteen ne eroavat toisistaan selvästi. Peruskolmijakoa tarkastellaan alaluokittelemalla ne useihin merkityksen alaluokkiin.

Myös syntaktiset, kielikohtaiset sanaluokat, kuten konjunktiot ja postpositiot, on koodattu aineistoon, mutta ne sivuutetaan tässä yhteydessä. Sanaluokkajako on luonnollisesti jatkumonluonteinen, on selvät substantiivit, adjektiivit ja verbit ja sitten luokkien väliin jäävät erikoisryhmät, kuten itsenäisesti käytettävät adjektiivit tai substantiivien luokasta irronneet adverbit, joten luokat jossain määrin risteävät. Verbi- ja adjektiiviluokituksissa voi tukeutua melko runsaaseen kirjallisuuteen, sen sijaan toimivia substantiiviluokituksia ei ole olemassa. Paras malli on ehkä Millerin (1998) käsitteellinen luokitus, joka listaa niin johtamattomat kuin johdetut substantiiviluokat. Sama malli on lukemattomissa eri kielten WordNet-sovelluksissa käytössä. Semanttinen analyysi on tehtävä manuaalisesti, mutta sen avulla saadaan tietoa sekä teemakeskeisestä sanastosta että kirjoitelmien sanaston abstraktiusasteesta. Yleisyystekijät, diversiteetti ja sanaston abstraktistuminen ovat sanatasolla keskeisiä kehityksellisiä muuttujia.

Aineistona Unelmasanasto

Unelmasanasto on koostettu Koulu- ja Unelmakorpuksen aineistoista, joissa on yhteensä lähes 1 900 kirjoitelmaa ja noin 230 000 sanaesiintymää (ks. tarkemmin johdantoluvusta). Sana-analyysiä varten aineisto valikoitiin siten, että kaikki erikoisaines poistettiin, koska tavoitteena oli saada näkyviin leksikaalinen erilaisuus lekseemimäärän sijaan. Sanaesiintymät muutettiin lekseemeiksi noudattamalla seuraavia periaatteita:

https://doi.org/10.21435/skst.1472

1) Aineistosta poistettiin:
   a) proprit (henkilön- ja paikannimet sekä tavaranimet), sanaliitot, numerot ja merkit,
   b) epäselvyydet (ei analysoitavissa, rauniosanat tms.),
   c) vieraskieliset sanat sekä
   d) interjektiot ja numeroalkuiset sanat (*9-v*).

2) Osa johdetuista sanaesiintymistä on laskettu edellisen derivaatiovaiheen sanaan:
   a) Esimerkiksi *minen*-johdoksen esiintymät on laskettu vastaavan verbin esiintymiin, mikäli se löytyy aineistosta. Jos verbiä ei ole aineistossa, *minen*-johdoksen esiintymä on luettu omaksi lekseemikseen.
   b) Kaikki partisiippien (aktiivin ja passiivin partisiipit sekä *ma*-partisiipit) esiintymät on laskettu vastaaviin verbilekseemeihin, mutta jos kyseinen verbilekseemi puuttuu aineistosta, partisiippi on oma lekseeminsä.

Näillä kriteereillä esiintymämäärät vähenivät kokonaisesiintymämääristä kaikissa aineiston osissa muutaman prosenttiyksikön (ks. taulukko 4). Nuorimmilla vähennys johtui runsaista nimi- ja numeroesiintymistä, vanhimmilla oli paljon syntaktisia johdoksia. Vaikka esimerkiksi *minen*-johdokset lisääntyvät kehityksellisesti, niiden katsottiin ensi sijassa osoittavan syntaktista taitoa, jolla sanottavaansa saattaa tiivistää. Lekseemimäärät taulukossa 4 ovat ryhmäkohtaisia, eli ne on laskettu ryhmittäin vain kerran, mutta eri ryhmissä samat lekseemit voivat toistua.

Taulukko 4. Unelmasanaston frekvenssit.

Ikäluokka	Esiintymät	Esiintymät karsinnan jälkeen	Ero %	Lekseemit
Alakoulu 1.–5.	64 914	60 860	6,2	4 854
Kuudennet	42 282	39 422	6,8	4 396
Yläkoulu	36 189	34 209	5,5	4 507
Aikuiset	89 192	87 126	2,3	10 069
Yhteensä	232 577	221 617	4,7	

https://doi.org/10.21435/skst.1472

Leksikaalinen sana-aines on jatkoanalysoitu kolmella eri tavalla, jotta saadaan selville iän mukainen muutos ja kehitys sanaston käytössä. Ensinnäkin koko aineistolle laskettiin yleisyystiedot isoista sanomalehtikorpuksista (HS2000-korpus, ks. Pajunen & Virtanen 2002). Tulokset tarkistettiin sekä sanakohtaisesta (ks. Laine & Virtanen 1999) että kontekstia sisältävästä korpuksesta. Yleisyystiedot laskettiin sitten ryhmittäin siten, että alakoulun 6. luokka laskettiin erikseen, koska siitä oli aineistoa paitsi vuodelta 2006 myös vuodelta 2014. Yläkoulun 7.–9. luokkien aineisto laskettiin pienuutensa vuoksi yhtenä ryhmänä. Aikuisryhmät analysoitiin erikseen, mutta erojen pienuuden vuoksi niitä käsitellään yhtenä. Toisekseen eri ikätasojen vertailua varten kolmannen, kuudennen ja yhdeksännen luokan peruskoululaisten sekä ammattikorkeakoululaisten aineistosta tehtiin sanaston diversiteettilaskelmat koehenkilökohtaisesti (ks. taulukko 5). Kaikki tekstit, joissa sanamäärä oli pienempi kuin 42, poistettiin analyysistä, koska lyhyet tekstit vinouttavat tuloksia ja koska käytetty diversiteettitunnusluvun menetelmä (*sum of probabilities*) edellyttää tiettyä vähimmäissanamäärää (McCarthy & Jarvis 2007, 472). Tässä menettelyssä tietyn tekstin sanastollista monimuotoisuutta kuvataan tunnusluvulla, joka lasketaan kunkin tekstissä esiintyvän sanan esiintymistodennäköisyyksien summana. Laskennassa käytetään apuna etukäteen asetettua parametria (tässä tapauksessa 42), joka kuvaa todennäköisyyksien laskennassa käytettävän tekstikatkelman pituutta. Kolmanneksi koko ensimmäisen kohdan lekseemiaineisto analysoitiin myös semanttisesti jakamalla aineisto ensin sanaluokkiin ja sitten luokittelemalla se sanaluokkakohtaisesti.

Unelmasanaston yleisyys

Frekvenssitekijät vaikuttavat sanaston hallintaan ja käyttöön voimakkaasti. Voidaankin ennustaa, että yleisin sanasto hallitsee nuorimpien peruskoululaisten kirjoittamista, mutta iän myötä sanasto harvinaistuu ja monipuolistuu. Tarkastelemme tässä yhteydessä niin sanottuja kattavuuslukuja, lekseemien ja esiintymien yleisyyksiä, käytetyn sanaston

https://doi.org/10.21435/skst.1472

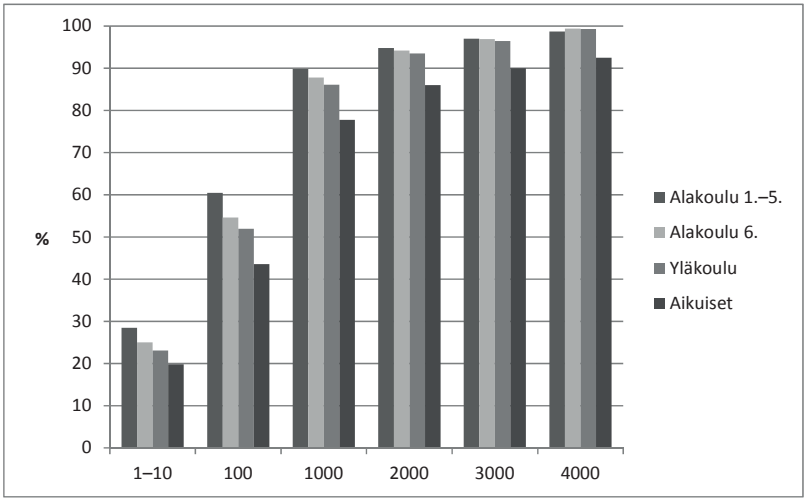
yleisyystasoja sekä taivutusperheitä ja suhteutamme eri frekvenssitekijöitä myös esimerkiksi sanojen rakenteeseen.

Yleisyystiedot laskettiin HS2000-korpuksen avulla. Haut tehtiin sekä WordMill- että ContextMill-ohjelmilla. Ensimmäinen antaa tiedoksi sana- ja taivutusmuotofrekvenssit ja erittelee sanan kaikki ne esiintymämuodot, mitkä lukuihin on sisällytetty, ja jälkimmäinen antaa sanojen käyttökontekstista halutulla tavalla rajatun tiedon. Esiintymämuotoihin sisältyvät muun muassa *minen*-johdokset ja eri funktioissa käytetyt partisiippimuodot. Saatuja frekvenssitietoja verrattiin Tieteellisen laskennan (CSC) antamiin yleisyyksiin. Frekvenssitiedot on siten koostettu yhteensä lähes 70 miljoonan sanaesiintymän aineistosta. Koska tiedot ovat peräisin sanomalehtikielestä, niiden perusteella saadaan luotettavaa tietoa suomen yleisestä sanastosta mutta ei aina tavoiteta arkikielen sanastoa: vaikka frekvenssilistojen kärki on tekstilajista riippumatta hyvin samankaltainen, ei asiatekstistä koostettuun frekvenssilistaan luonnollisesti juurikaan nouse puhutun kielen rekistereille tyypillistä sanastoa. Sellaisenaan tiedot eivät myöskään anna luotettavaa tietoa koko kielen harvinaisesta sanastosta. Tässä tutkimuksessa kuitenkin riittää, että saadaan tehtyä ero yleisen ja harvinaisen välillä.

Käsittelemme tuloksia luokittelemalla esiintymäyleisyyttä pääasiassa ryhmiin hyvin yleiset ja yleiset (yleisyystaso 1, yleisyystasot 2–4), tavalliset (yleisyystasot 5–9) ja harvinaiset (yleisyystaso >10). Lisäksi on puuttuvien joukko, johon kuuluvat ne, jotka eivät satu esiintymään HS2000-korpuksessa, ja ne, jotka eivät ole suomen vakiintuneita sanoja vaan esimerkiksi satunnaismuodosteita. Ikäryhmiä käsittelemme hieman eri tarkkuudella. Tarkin jako erottaa alakoulun luokat 1.–5. ja 6., yläkoulun ja aikuiset. Monet erot näkyvät jo jaottelulla ala- ja yläkouluun, mutta kuudes luokka on analyysissämme ikään kuin siirtymätyyppiä. Koululaisten kirjoitelmissa kertovan tekstin luonne ja kerrontastrategia ovat melko samanlaisia etenkin aikuisteksteihin verrattuna, joten monet peruserot ovat osoitettavissa sanastotasolla jo vertaamalla koululaisia (melko hyviin) aikuiskirjoittajiin.

Hieman yleistäen voidaan sanoa, että yleisimmät kielen lekseemit opitaan ensimmäiseksi, ja toisaalta, että yleisimmät lekseemit ovat monikäyttöisyytensä vuoksi kommunikoinnissa ne hyödyllisimmät (ks. sanas-

https://doi.org/10.21435/skst.1472



Kaavio 2. Suomen kielen yleisimpien sanojen kattavuus kaikista sanaesiintymistä eri-ikäisten Unelmakirjoitelmissa: 10–4 000 yleisintä sanaa (= lekseemiä).

ton yleisyyden ja toiston vaikutuksesta oppimiseen esimerkiksi Honko 2013, 29–31). Lisäksi tarvitaan toki myös harvinaisempaa sanastoa, jonka osaamisen tarpeet ja siksi myös hyödyt määrittyvät tiiviimmin kielenkäytön yksilöllisten ja tilanteisten tarpeiden mukaan. Lekseemejä, joiden esiintymiä miljoonassa sanassa on tuhansia tai edes tuhat, on yleensä hyvin vähän; suomessa muun muassa *olla* ja *ja* ovat yleisyyskärjessä.

Yleisimpien lekseemien merkitys on tärkeä myös tekstin kattavuuslukujen kannalta. Kattavuus tarkoittaa sitä prosenttiosuutta, minkä tietty määrä lekseemien esiintymiä kattaa tekstin kokonaissanamassasta (ks. Laufer & Ravenhorst-Kalovski 2010). Miltonin (2009, 46) mukaan jo runsas kymmenkunta lekseemiä voi englannissa kattaa neljänneksen korpuksen sanaesiintymistä, sata noin puolet ja tuhat kolme neljännestä. Englanti on analyyttinen kieli, joten kattavuuteen vaikuttavat myös kieliopilliset sanat. Unelmasanastossa kymmenen yleisintä lekseemiä kattaa 20–30 prosenttia tekstimassasta, sata noin puolet ja tuhat yleisintä jo 80–90 prosenttia (ks. kaavio 2). Nuorimpien ja vanhimpien kirjoittajien välinen ero on yleisyysryhmittäin kymmenisen prosenttiyksikköä, ja kattavuus laskee lineaarisesti iän myötä. Yleisimpien sanojen osuus aikuisaineiston koko tekstimassasta muistuttaa englannin suhteita,

https://doi.org/10.21435/skst.1472

mutta tähän vaikuttaa myös tekstin luonne ja teeman yhteneväisyys. Moninaisemmassa aineistossa peittävyys olisi suomessa englantia alempi, koska suomessa ei ole englannin tapaisia prepositioita eikä esimerkiksi artikkeleita.

Kaikissa ikäryhmissä kymmenen yleisimmän lekseemin joukkoon kuuluvat *ja*, *olla*, *se*, *minä* ja *päivä*. Ne kuvastavat toisaalta tekstityyppiä ja toisaalta kirjoitelman aihetta. Sen sijaan muun muassa relatiivipronomini *joka* nousee yleisimpien listalle vasta kuudennella luokalla, jolloin puolestaan sanat *sitten* ja *kun* eivät enää esiinny listalla. Yleiskielessä yleinen *ei* taas ilmestyy listalle vasta kuudennen jälkeen; unelmista ei niinkään kirjoiteta kielteisessä mielessä. Kuudennella *mennä*-verbin rinnalle ilmestyy *lähteä*-verbi, mutta aikuisten aineistossa yleisimpien sanojen listalla ei enää ole verbien *olla* ja *ei* lisäksi muita verbejä. Yleisimpien lekseemien esiintymäjärjestys toisin sanoen tuntuu aika hyvin kuvaavan syntaksin taitojen kehittymistä: esimerkiksi relatiivilauseet lisääntyvät kehityksellisesti, kuten kirjan loppuluvusta käy ilmi. Nuorimmilla yleisimmät kymmenen lekseemiä kattavat lähes kolmanneksen tekstimassasta, mikä osoittaa, että näiden lekseemien toistuvuus on erittäin korkea. Aikuisilla yleisimmät sanat ovat joko kieliopillisessa funktiossa käytettäviä konjunktioita, verbejä tai teemaan liittyviä (*unelma*, *päivä*).[15]

Kaikissa ikäryhmissä sadan yleisimmän lekseemin joukosta noin kolmannes on verbejä. Tekstikattavuuden kannalta siirtymä yleisimmistä iän myötä harvinaisempiin käy jokseenkin lineaarisesti. Yleisyyslähteinä toimivien sanomalehtikorpusten sana-aineksista puuttuvat lekseemit ovat nuorimmilla sadunomaisia satunnaismuodosteita tai arkikielisiä (*suklaatulivuori*, *älämölänpoistoaine*, *mörköpeli*, *porkkanaylläri*, *rentoilu*), aikuisilla valtaosaan nousevat spesifiä merkitystä ilmaisevat satunnaismuodosteet tai tilanteiset yhdyssanamuodosteet (*aamurapsutus*, *helikopterinraato*). Lisäksi kirjoitelmissa on joitain arkikielen sanoja (*mussutella*, *pörräily*) tai uudissanoja, joita 2000-luvun alun sanomalehdissä ei vielä esiintynyt (*peukutus*).

Unelma-aineiston sanasto on luonnollisesti paitsi teemaan ja tekstityyppiin sidoksissa myös pieni: kussakin koululaisryhmässä eri leksee-

15 Kirjoitelmien otsikointia (Unelmieni päivä) ei ole laskettu sanamassaan mukaan.

https://doi.org/10.21435/skst.1472

mien määrä jää alle 5 000:n (kokonaissanamäärä 34 209–60 860), vain aikuisaineistossa lekseemejä on enemmän. Siten sanasto ei kuvasta suomen yleiskielen yleisyyssuhteita riittävästi muiden kuin kaikkein yleisimpien sanojen osalta. Varsinainen yleisyystasoluokitus suomen kielestä vielä puuttuu. Lekseemien jakauma yleisyystasoihin on ryhmittäin melko tasainen (ks. taulukko 5). Erot ovat suurimmat koululaisten ja aikuisten välillä. Koululaisilla kirjoitelmissa käytetyistä lekseemeistä on hyvin yleisiä (yleisyystaso 1) tai yleisiä (yleisyystasot 2–4) lähes 30 prosenttia ja tavallisia (yleisyystasot 5–9) noin kuudennes. Koululaisryhmissä yleisimpien määrä hieman kasvaa eikä laske iän mukaan, kuten odottaisi, mutta on mahdollista, että ero kuvastaa syntaksin taitojen kehitystä. Aikuisilla hyvin yleisiä tai yleisiä lekseemejä on selvästi vähemmän kuin koululaisilla eli noin viidennes. Harvinaisia lekseemejä on kaikilla määrällisesti paljon, vaikka esiintymiä harvinaisella lekseemillä on tyypillisesti vain yksi. Tässä koululaisten ja aikuisten ero on suurin eli yli kymmenen prosenttiyksikköä. Massakorpuksista eri syistä puuttuvia lekseemejä on kymmenisen prosenttia. Aineistossa on luonnollisesti vähän suljettujen sanaluokkien jäseniä ja niiden joukossa on paljon hyvin yleisiä lekseemejä, avoimien luokkien tilanne on taas päinvastainen. Harvinaisista lekseemeistä suurin osa kuuluu substantiiveihin, verbeissä on enemmän toistuvasti käytettyjä, yleisiä lekseemejä. Käsittelemme sanaluokkien suhdetta yleisyyteen semanttisen analyysin yhteydessä tarkemmin.

Taulukko 5. Unelmakirjoitelmasanaston jako yleisyystasoittain eri ikävaiheissa: eri yleisyysluokkien lekseemien osuus kaikista eri lekseemeistä.

Lekseemin yleisyys / Ikäluokka %	Alakoulu 1.–5.	Alakoulu 6.	Yläkoulu	Aikuiset
Hyvin yleinen	12,2	13,8	14,2	7,2
Yleinen	14,2	15,0	15,8	10,6
Tavallinen	18,9	18,4	20,6	17,2
Harvinainen	44,3	44,6	43,0	56,9
Puuttuu korpuksesta	10,5	8,2	6,5	8,1

https://doi.org/10.21435/skst.1472

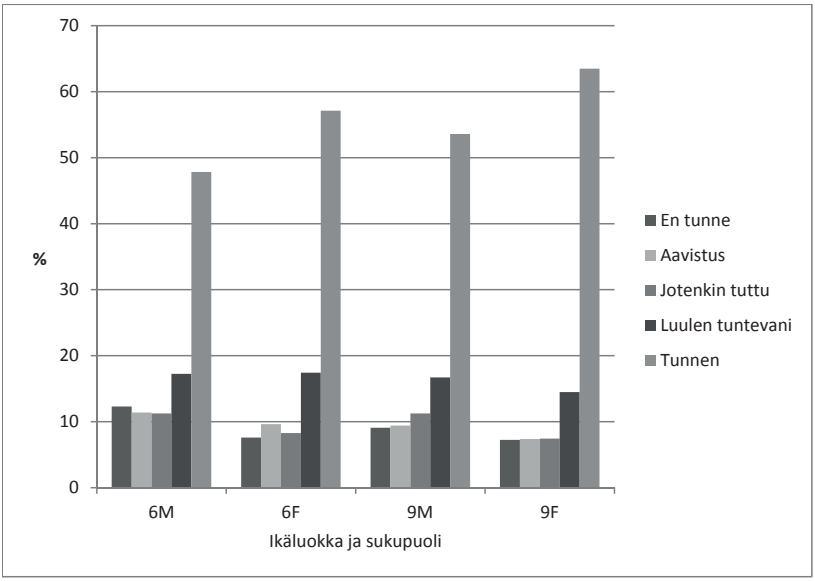
Unelmasanaston sanaesiintymistä kaksi kolmannesta kuuluu suomen tuhannen yleisimmän lekseemin joukkoon eli ne ovat hyvin yleisiä (ks. taulukko 6). Luku on peruskoululaisilla lähes 20 prosenttiyksikköä matalampi kuin kattavuuslaskelmissa (vrt. kaavio 2) ja tämä ero liittyy lekseemien toisteisuuteen. Esiintymistä noin 15 prosenttia kuuluu yleiseen sanastoon (yleisyystasot 2–4), kymmenisen prosenttia tavalliseen (yleisyystasot 5–9) tai harvinaiseen sanastoon. Peruskoululaisten välillä yleisyyssuhteissa ei ole suuria eroja, mutta trendi on lievästi laskeva. Nuoret aikuiset eroavat selvemmin peruskoululaisista kuin nämä toisistaan.

Taulukko 6. Unelmasanaesiintymien jako yleisyystasoittain.

Sanaesiintymien yleisyys / Ikäluokka %	Koululaiset	Aikuiset
Hyvin yleiset	67,7	60,9
Yleiset	14,3	15,9
Tavalliset	8,2	10,0
Harvinaiset	8,7	12,2
Puuttuu korpuksesta	1,1	1,0

Peruskoululaisilla sanasto on melko tavanomaista ja tuttua, ja ero aikuisiin on lekseemi- ja esiintymätasolla hyvin selvä. Eroa selittää ainakin osaksi se, että koululaisilla on vielä suuria puutteita sanastossaan. Peruskoululaisille tehdyssä tuttuustestissä, jossa aineistona oli 40 melko harvinaista liikeverbiä, sanoista tunnettiin hyvin noin puolet (ks. kaavio 3). Loppuja tunnettiin heikosti tai ei ollenkaan. Esimerkiksi verbit *kuhista*, *parveilla* ja *rämpiä* olivat monille tuntemattomia. Aineisto on testattu myös nuorilla aikuisilla ja heillä tuntemus on ollut yli 90 prosenttia. Luetun ymmärtämisen kannalta katsotaan, että noin 95 prosenttia sanaesiintymistä pitäisi olla tuttuja, ennen kuin luettu oikeasti ymmärretään (Milton 2009, 52). Prosentti ei ehkä suoraan sovellu suomeen, jossa ymmärtämisongelmia voivat tuoda myös harvinaisemmat taivutus- ja etenkin johdinainekset sekä näiden käyttöön liittyvä morfofonologinen sanavartaloiden vaihtelu, mutta lienee silti viitteellinen.

https://doi.org/10.21435/skst.1472

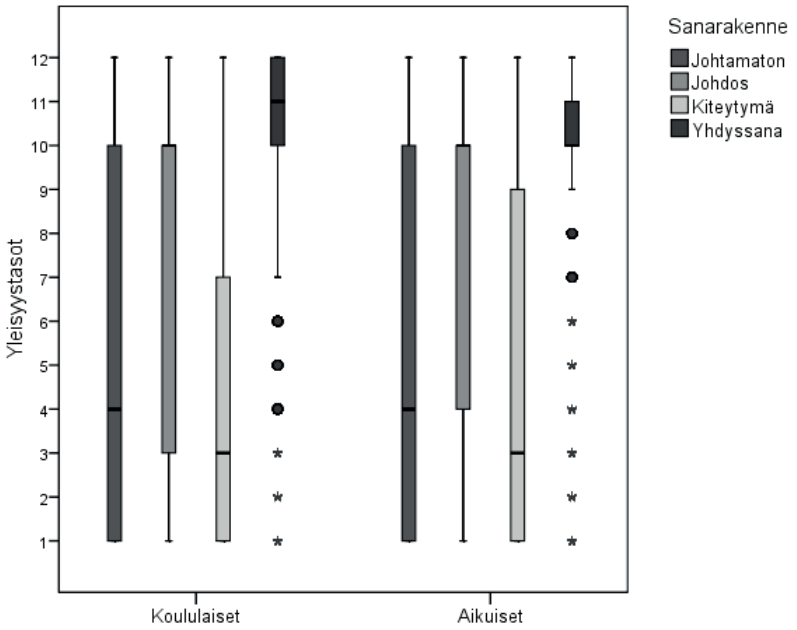


Kaavio 3. Melko harvinaisten liikeverbien tuttuustesti kuudes- ja yhdeksäsluokkalaisilla (N = 157).

Yleisyystasoluokitukseen heijastuu suomen tapaisessa kielessä myös lekseemien sanarakenne (ks. kaavio 4).[16] Johtamattomien sanojen yleisyystaso on keskimäärin matalampi kuin johdettujen ja johdettujen matalampi kuin yhdyssanojen. Kiteytymät eli lähinnä substantiivien tai adjektiivien paradigmasta irronneet sanat ovat tyypillisesti adverbeja, ja ne sijoittuvat alimpiin yleisyystasoluokkiin. Yhdyssanoissa mediaani on korkein. Yleisyystaso nousee kehityksellisesti jokaisessa sanarakennetyypissä eli sanasto harvinaistuu. Esimerkiksi johdetuista lekseemeistä keskeisin osa sijoittuu koululaisilla yleisyystasoille 3–10, aikuisilla tasoille 4–10. Koululaisilla on enemmän vertailukorpuksesta puuttuvia yhdyssanoja, jotka aineistossa ovat satunnaismuodosteita.

16 Niin sanotussa boxplot-kaaviossa pylvään keskiviiva ilmaisee keskimmäisen arvon eli mediaanin, leveä osuus sen ympärille sijoittuvat havainnot (yhteensä 50 % kaikista havainnoista) ja viivat ala- tai yläkvartiilin havainnot. Pallot ja tähdet viittaavat poikkeaviin havaintoihin. Kaaviossa 4 yleisyystaso 11 ja 12 viittaavat korpuksesta puuttuviin sanoihin ja satunnaismuodosteihin, epäsanat on laskelmista poistettu.

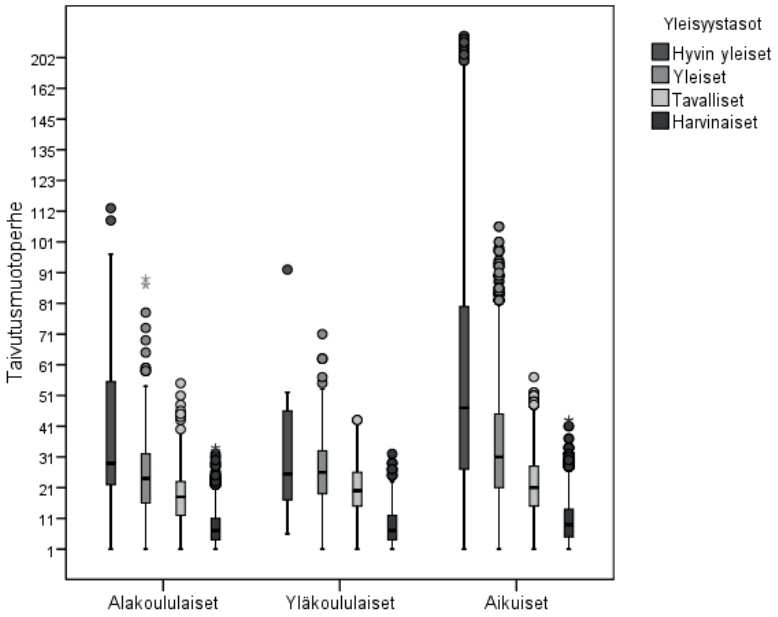
https://doi.org/10.21435/skst.1472



Kaavio 4. Sanan yleisyystason ja sanarakenteen riippuvuus ikäluokittain.

Taivutusperheiden ja yleisyystasojen ristiintaulukointi osoittaa, että yleisyys ja taivutusmuotojen määrä korreloivat yleisimmissä lekseemeissä hyvin vahvasti (ks. kaavio 5): kaikki lekseemit, joille massakorpuksista löytyy yli sata eri taivutusmuotoa tai niiden kombinaatiota kuuluvat suomen yleisimpien lekseemien joukkoon. Tuhannen yleisimmän joukkoon kuuluvilla taivutusmuotoja voi olla kolmekin sataa, mutta mediaani on huomattavasti matalampi, noin 50. Suurin osa paljon taivutusmuotoja saaneista lekseemeistä on verbejä, ja yksittäisellä verbillä on sanomalehtitekstissä aina enemmän taivutusmuotoja kuin yksittäisellä substantiivilla toisteisuuserojen takia. Erosta voi päätellä, että verbisyntaksin opetus voisi hyödyttää, jos kirjoittamistaitoja halutaan edistää. Monella koululaisella ongelma on lauseenmuodostuksessa, ja juuri verbit ovat lauseenmuodostuksen ydin. Eri verbimuodoilla tuotetaan paitsi leksikaaliset myös temporaaliset ja modaaliset merkitykset sekä muutetaan lauseen argumenttien asemaa.

https://doi.org/10.21435/skst.1472



Kaavio 5. Massakorpuksen taivutusperheet ja Unelmasanaston lekseemit yleisyystasojakauman mukaan ikäluokittain.

Lehtikielessä lekseemien taivutusperheiden koko laskee yleisyystasoittain siten, että harvinaisempia taivutetaan vähemmän kuin yleisempiä. Nuorempien ja vanhempien unelmasanasto näyttää olevan hiukan erilaista taivutusperheiden suhteen, koska aikuisilla käytössä olevia sanoja taivutetaan lehtikielessä enemmän kuin niitä sanoja, joita koululaiset käyttävät (ks. kaavio 5). Ero näkyy monista yleisyysluokista ja syntyy pitkälti aikuisten melko abstraktimerkityksisistä ja pääosin johdetuista verbeistä. Yleisimmistä verbeistä voi esimerkkeinä mainita verbit *käsitellä*, *korostaa*, *arvostaa*, *sallia* ja *poistaa* ja harvinaisista muun muassa verbit *sovitella*, *turhautua*, *kontrolloida*, *aloitella*, *puhutella*; peruskoululaiset eivät näitä verbejä käytä lainkaan. Ero viittaa siihen, että aikuisten sanasto vastaa paremmin lehtikielen sanastoa kuin koululaisten, ja siihen, että aikuisilla sekä leksikaalinen että syntaktinen variaatio on suurempi kuin koululaisilla. Peruskoululaisten ryhmissä taivutusperheiden koko kasvaa iän mukaan tasaisesti ja osoittaa muutoksia sanastossa. Yläkoululaissanastossa

https://doi.org/10.21435/skst.1472

on muun muassa verbit *arvioida*, *uudistaa*, *hävittää*, *lakkauttaa*, *miellyttää*, *teloittaa* ja *sulattaa*, jotka kokonaan puuttuvat alakoululaisilta.

Massakorpuksesta laskettujen taivutusperheiden koko ilmaisee myös sen, miten moninaisesti kielenkäyttäjät altistuvat tietylle lekseemille. Tällainen luku sopii käyttöpohjaisen kielenoppimisteorian (ks. esimerkiksi Tomasello 2003; myös MacWhinney 2001, 464) mukaisiin näkemyksiin, vaikka ei kerrokaan koko totuutta siitä, mitä yksittäisiä sanoja eri yksilöt kohtaavat tai merkityksellistävät ja todennäköisimmin oppivat. Tulokset osoittavat selvästi, että Unelmasanastossa on eniten esiintymiä sanoista, joiden taivutus on massakorpuksissa moninaista. Vähemmän taivutettujen lekseemien esiintymät kasvavat kuitenkin lineaarisesti. Leksikon ja sen käyttökontekstien jatkuva laajeneminen ja harvinaisempiin lekseemeihin ja muotoihin kehittyminen on vastoin radikaaleimpia – ja yksinkertaisimpia – frekvenssitekijöiden roolia korostavia näkemyksiä, joiden mukaan ilmaukset automatisoituvat yli perinteisten morfeemi- ja sanarajojen ja muuttuvat leksikaalisiksi kollokaatioiksi (vrt. esimerkiksi Bybee & Hopper 2001). Tällainen näkemys ei ylipäätään sovellu suomen tapaiseen taivutuskieleen.

Unelmasanaston diversiteetti

Sanaston diversiteettiä koskevaan tarkasteluun poimittiin otos eri-ikäisten aineistoa. Alkuvaiheessa otettiin mukaan kaikki kolmannen, kuudennen ja yhdeksännen vuosiluokan koululaisten sekä ammattikorkeakoululaisten kirjoitelmat eli yhteensä 729 tekstiä, joista 67 poistettiin seuraavassa vaiheessa liian lyhyinä (tekstin sanaesiintymiä oli alle 42). Lopullisen aineiston kokonaissanamäärä oli yhteensä 105 942 sanetta ja 662 kirjoitelmaa (ks. taulukko 7).

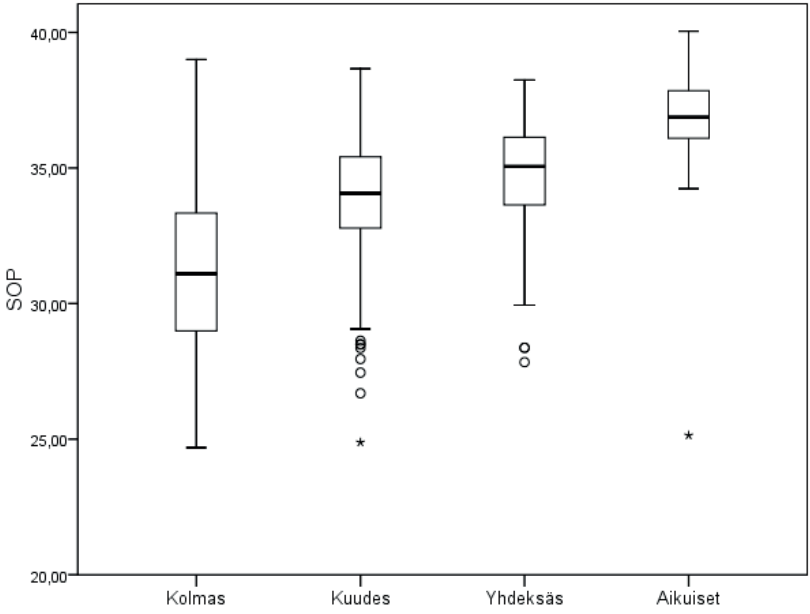
Unelmakirjoitelmien sanaston diversiteettiä on verrattu kirjoittajien ikään (kaaviot 6–7). Kaaviot on piirretty koko siitä aineistosta, josta käytetyn diversiteettitunnusluvun (*sum of probabilities*, SOP) laskeminen on mahdollista (kirjoitelman sanamäärä on vähintään 42).[17] Käytetty

17 Koska kaaviot on pyritty pitämään luettavassa muodossa, yksi yhdeksäsluokkalaisen kirjoitelman poikkeavan korkea arvo (lähes 50) jää näkymättömiin.

https://doi.org/10.21435/skst.1472

Taulukko 7. Unelma-aineiston määrä diversiteettilaskelmissa.

Ikäluokka	Kirjoitelmat	Kirjoittajan sukupuoli	Poistot	Sanamäärä
Kolmas	154	69 F, 85 M	30	10 422
Kuudes	313	160 F, 153 M	27	41 543
Yhdeksäs	161	88 F, 72 M	10	24 983
AMK	101	43 F, 58 M	Ei	28 994



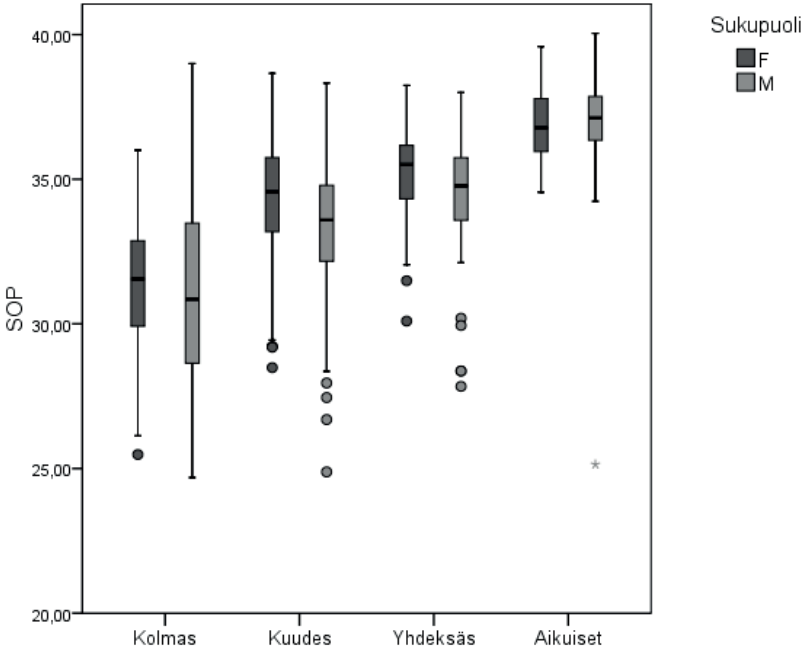
Kaavio 6. Kirjoittajan iän yhteys kirjoitelmien sanaston monimuotoisuuteen (SOP) ikäluokittain.

mittaustapa on arvioitu tekstipituuden suhteen melko neutraaliksi ja siten turvalliseksi hyvinkin eripituisten tekstien arvioinnissa (McCarthy & Jarvis 2007, 460; ks. myös Malvern, Richards, Chipere & Durán 2009), mutta kaikkein lyhimmissä teksteissä yleisimmätkään sanat eivät ehdi toistua yhtä todennäköisesti kuin hieman pidemmissä. Tarkastellussa aineistossa sanaston keskimääräinen diversiteetti kasvaa lähes lineaarisesti ikäluokasta seuraavaan (ks. kaavio 6). Samalla yksilöiden välinen

https://doi.org/10.21435/skst.1472

hajonta pienenee eli vanhimpien kirjoittajien ryhmissä yksittäisten kirjoitelmien sanaston diversiteetti on lähempänä ikäluokan keskimääräistä tasoa. Kehitys on selkeä ikäryhmien ylintä neljännestä lukuun ottamatta: vaikka sanastollisesti köyhiä kirjoitelmia esiintyy nuorimmilla kirjoittajilla selvästi enemmän kuin vanhemmilla kirjoittajilla, jo kaikkein nuorimpien aineistossa on myös sanastoltaan vaihtelevia tekstejä. Tätä selittää osittain leksikaalisten sidoskeinojen vähäinen käyttö ja nuorimpien koululaisten kirjoitelmien lyhyys, mutta osittain myös nuorimmilla on sanastollisesti moninaisia ja muutenkin tasokkaita kirjoitelmia.

Sukupuolen ja sanaston diversiteetin yhteys ei ole yksiselitteinen, sillä vaikka peruskoululaisten kirjoitelmissa sanasto näyttää olevan vaihtelevampaa tyttöjen kuin poikien teksteissä (ks. myös Honko 2013), nuorten aikuisten kirjoitelmissa tilanne on päinvastainen (kaavio 7). Poikien ryhmissä yksilölliset erot sanaston diversiteetissa ovat hieman suurem-



Kaavio 7. Kirjoittajan sukupuolen yhteys kirjoitelmien sanaston monimuotoisuuteen (SOP) ikäluokittain.

https://doi.org/10.21435/skst.1472

mat kuin tyttöjen ryhmissä, mutta sekä pojilla että tytöillä yleistaso on muutamaa poikkeusta lukuun ottamatta jo peruskoulun yhdeksännellä luokalla vähintään kohtalaista tasoa (SOP >30, englannin kielestä McCarthy & Jarvis 2007, 464).

Samaan aikaan kun kirjoitelmien sanaston diversiteetti kirjoittajan iän mukaan lisääntyy, myös sanojen keskimääräinen pituus kasvaa. Molempia piirteitä voidaan pitää kehityksellisinä siinä mielessä, että ryhmätasolla ne kuvaavat kehittyvää kieli- ja kirjoitustaitoa, joka tarkastellussa aineistossa on kiinteässä suhteessa kirjoittajan ikään. Myös tietyn ikäluokan sisällä tekstipituuden ja sanaston diversiteetin välillä on tarkastellussa aineistossa löyhä yhteys: pisimmät kirjoitelmat ovat sanastoltaan moninaisempia hieman todennäköisemmin kuin lyhyet. Sanaston diversiteetti ei yksilötasolla kuitenkaan ole voimakkaasti riippuvainen tekstipituudesta, mikä on odotuksenmukaista myös aiemman tutkimuksen valossa (Honko 2013). Esimerkiksi yhdeksäsluokkalaisten kirjoitelmissa korrelatiivinen yhteys tekstipituuden ja sanastollisen diversiteetin välillä on heikko, eikä sitä voi pitää kirjoittamisen ennustettavuuden kannalta merkityksellisenä. Siitä huolimatta heikkokin (tilastollisesti merkitsevä) positiivinen yhteys vahvistaa näkemystä leksikaalisesta diversiteetistä kehityksellisenä piirteenä.

On huomionarvoista, että vanhempien kirjoittajien tekstit ovat paitsi sanastoltaan vaihtelevampia myös pisimpiä, vaikka tekstipituuden kasvaessa korkean sanastollisen diversiteetin ylläpitäminen ja kasvattaminen käy yhä haastavammaksi. Sanaston vaihtelevuuteen vaikuttaa luonnollisesti myös aiheiden määrä. Heinosen (2018) mukaan kuudes- ja yhdeksäsluokkalaisten sekä ammattikorkeakoululaisten Unelmakirjoitelmissa on käsittelyssä keskimäärin 3–4 aihetta, mutta kaikkein nuorimmilla aiheita on enemmän kuin aikuisilla. Yleisimmät aiheet ovat harrasteet, ravinto, ystävät ja hyödykkeet. Nuorimmat kirjoittavat mielellään perheestä ja kuuluisuudesta, vanhimmat urasta, arvoista ja parisuhteesta. Kun aikuiset kirjoittavat pidempiä kirjoitelmia kuin kouluikäiset, he samalla käsittelevät harvempia aiheita mutta kutakin aihetta monipuolisemmin. Toisin sanoen nuorimmilla sanaston monipuolisuus syntyy useammin pinnallisista luettelonomaisista kirjoitelmista, aikuisilla monipuolisesta saman aiheen käsittelystä.

https://doi.org/10.21435/skst.1472

Aineistosta on lisäksi laskettu leksikaalinen tiheys (sisältösanojen osuus sanamassassa) ja substantiivi–verbi-indeksi eli substantiivien osuus substantiivien ja verbien summasta (NTVR eli *noun-to-verb ratio*). Näiden lukujen on ajateltu kertovan tekstien informaatiotiheydestä ja tiiviydestä, ja niiden suhde sanastolliseen diversiteettiin on siksi kiinnostava. Leksikaalinen tiheys on Unelmakirjoitelmissa 75–80 prosenttia kokonaissanamäärästä ja nousee hieman iän myötä, kuten Pajunen ja Vainio tässä kirjassa osoittavat. Leksikaalisen tiheyden ja sanaston diversiteetin välinen yhteys koko tekstiaineistossa on heikko mutta positiivinen. Sanastoltaan vaihtelevissa kirjoitelmissa vaihtelu saavutetaan todennäköisimmin käyttämällä runsaasti avointen eli sisältösanaluokkien sanoja, ja kaikkein toisteisimmissa kirjoitelmissa puolestaan kieliopillisten sanojen osuus on hieman keskimääräistä suurempi. Mitä enemmän leksikaalinen tiheys jää alle kirjoitelmista laskettujen mediaaniarvojen, sitä varmemmin myös sanaston diversiteetti jää alle ikämediaanin. Tämä tulos sopii jo aiemmin esiin tulleeseen tendenssiin, jonka mukaan ikäryhmittäisissä mediaaneissa pyörivät arvot viittaavat ikäluokan osaamistasolla onnistuneeseen tekstiin.

Substantiivi–verbi-indeksi on kirjoitelmissa ja oletettavasti kertovassa tekstissä ylipäätään keskimäärin 1,1:1 ja nousee iän mukaan muutaman prosenttiyksikön, tavanomainen vaihteluväli on Pajusen ja Vainion mukaan 50–57 prosenttia eli keskimäärin hieman enemmän substantiiveja kuin verbejä (ks. myös Biber, Conrad & Reppen 1998). Jos arvo on matala, teksti koostuu peruslauseista ja on lähellä puhutun kielen rakennetta, mutta jos arvo on korkea, teksti on rakenteeltaan kompleksista (paljon alisteisia verbirakenteita) ja/tai abstraktia (paljon substantiiveja). Siten oletus on, että substantiivi–verbi-mediaanissa olevissa kirjoitelmissa sanasto olisi melko vaihtelevaa. Toisaalta oletus myös on, että mitä korkeampi substantiivi–verbi-indeksi on, sitä vaihtelevampaa sanasto on. Nämä oletukset näyttäisivät pitävän paikkansa yhdeksäsluokkalaisilla ja etenkin aikuisilla. Kuudesluokkalaisilla hajonta kummassakin arvossa on suurta ja viittaa toisaalta puhutun kielen kaltaiseen yksinkertaiseen rakenteeseen, toisaalta lauseenmuodostamisongelmiin.

https://doi.org/10.21435/skst.1472

Unelmasanaston semanttinen analyysi

Unelmasanaston semanttinen analyysi alkaa sanaluokkajaolla, joka on perustaltaan ontologis-semanttinen: substantiiveilla nimetään olioita, adjektiiveilla niitä kuvaillaan ja verbeillä predikoidaan substantiiveilla nimettävistä olioista paitsi tapahtumia myös tiloja. Tämä kuvaus edustaa tietysti sanaluokkia prototyyppeinä (Croft 1991), ja se korreloi kielittäin muun muassa merkinnän, kuten taivutuksen ja johtamisen, luonteeseen ja määrään (suomesta ks. Pajunen 1994; 1998; 2010). Substantiivit, verbit ja adjektiivit edustavat niin sanottuja pääsanaluokkia, ja ne voidaan määritellä paitsi ontologis-semanttisesti myös fonologisin, morfologisin, syntaktisin ja pragmaattisin kriteerein. Keskitymme tässä yhteydessä analysoimaan semanttisesti erityisesti substantiiveja ja vain sivuamme muita sanaluokkia, koska juuri substantiivien semanttinen luokitus paljastaa parhaiten kirjoitelma-aiheen käsittelyn iänmukaiset muutokset. Substantiivien luokituksesta ja niiden roolista Unelmakirjoitelmissa ei myöskään ole aiempaa tietoa (Koulukorpuksen adjektiiveista, ks. Kuisma 2011; adverbeista, ks. Koivunen 2012; verbeistä ks. Pajunen 2012).[18] Kieliopillisia sanaluokkia ei tässä yhteydessä analysoida määriä pidemmälle (Koulukorpuksen konjunktioista ks. Nahkola 2012; modaalisuuden ilmaisuista ks. Keskinen 2012; pronomineista Palonen 2013).

Substantiivien analyysi nojaa paitsi semanttiseen kirjallisuuteen (Miller 1998; Frawley 1992; Ravid 2006; Gärdenfors 2014) myös subjektiivisia semanttisia muuttujia arvioivien testien tuloksiin (vrt. Stavroula-Thaleia, Vigliocco, Del Campo, Vinson & Andrews 2011; Söderholm, Häyry, Laine & Karrasch 2013; Yap, Lim & Pexman 2015, Pollock 2018). Adjektiivien luokitus noudattaa Dixonin (1977, 2004) tunnettua luokitusta (ks. myös Pajunen 1994), verbien luokitus perustuu suomesta tehtyihin analyyseihin (Pajunen 1999; 2001; 2006), ja adverbit on luokiteltu lähinnä perinteisen kieliopin mukaisesti. Subjektiivisilla testeillä

18 Kouluikäisten sanastonhallintaa hankkeessa ovat erilaisin testein tutkineet muun muassa Härkönen (2012, verbi-idiomit), Jääskeläinen (2008, verbisanasto), Rantala (2012, olotilanilmaukset), Routama (2008, spatiaalinen sanasto) ja Turunen (2012, abstraktisanat).

https://doi.org/10.21435/skst.1472

saadaan esiin kielenkäyttäjän näkemykset esimerkiksi sanan konkreettisuudesta, tuttuudesta tai vaikkapa sen herättämän tunteen voimakkuudesta tai merkityksen kielteisyydestä. Lisäksi luokittelussa välttämättä sivutaan Popperin esittämää todellisuuden kolmijakoa fysikaalisten ja mentaalisten olioiden ja tapahtumien maailmaan (maailmat 1 ja 2) sekä abstraktien entiteettien maailmaan (maailma 3), koska olioista voidaan puhua monilla eri todellisuustasoilla. Esimerkiksi sana *koulu* voi viitata tiettyyn koulurakennukseen, tietyn asteen koulutusta antavaan kouluun sekä instituutioon. Tämän kanssa risteää käsitteen ja sen toteutuman tai esiintymän suhde, joka voi todellisuudessa niin ikään olla monikerroksinen (ks.Wetzel 2009). Esimerkiksi aineesta voidaan puhua abstraktimmalla (*hyvä ruoka*) tai sen toteutuman konkreettisemmalla tasolla (*jäätelötuutti*). Unelmasanaston vertailussa olisi hyväksi, jos aineisto olisi kauttaaltaan voitu luokitella tyyppi–esiintymä-tasoisesti, mutta käytännössä tällainen tarkkuus on isossa aineistossa mahdoton. Substantiiviaineisto on luokiteltu myös suhteessa hierarkiatasoihin, joita tässä yhteydessä käsitellään vain oletuksella, että perustason jäsenet ovat johtamattomia sanoja (*kissa, koira*), geneerisen tason jäsenet voivat olla sekä johtamattomia että johdettuja (*eläin, elolliset*), ja spesifin tason jäsenten sanarakenne voi olla mikä tahansa mutta tyypillisimmin johdos tai yhdyssana (vrt. Taylor 1995, 46–51). Lisäksi aineistoon on koodattu erikseen lekseemit, joiden tarkoite on jossain määrin opaakki (*puuha, kamppeet, vehje, kamat*); tähän luokkaan kuuluvat myös relationaaliset sanat (*siivu, viipale, lohko*).

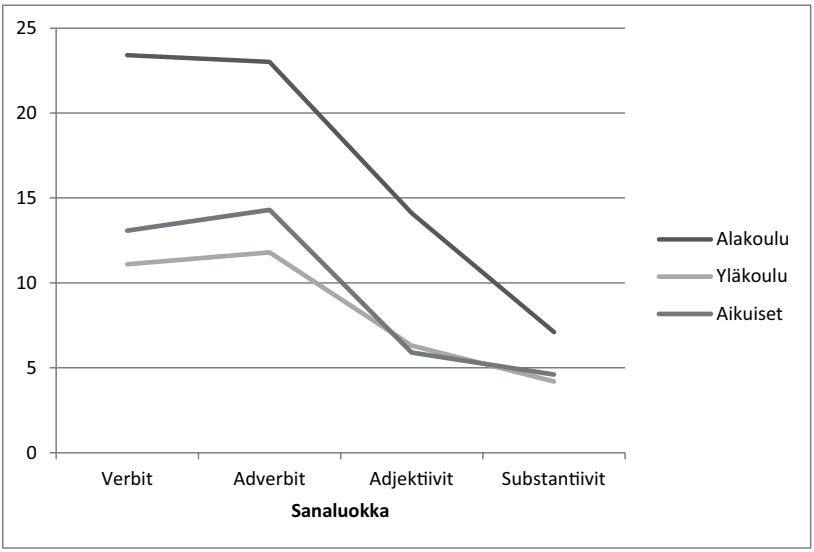
Kielenhallinnan ja kirjoittamisen kehitys näkyy jo sanaesiintymien sisäisestä jakaumasta sanaluokkiin (ks. taulukko 8). Adjektiivi- ja substantiiviesiintymien määrä nousee lineaarisesti, verbien, konjunktioiden ja pronominien määrä laskee yhtä lailla lineaarisesti. Erot selittyvät joko kerronnan monipuolistumisen kautta tai sitten niitä selittää tendenssi yksinkertaisesta kompleksiin. Substantiivien osuus kasvaa määrällisesti eniten.

Lekseemimäärät ovat substantiiveissa suurimmat ja muissa leksikaalisissa luokissa – verbeissä, adjektiiveissa ja adverbeissa – pienempiä mutta muissa luokissa yksittäisillä lekseemeillä on enemmän esiintymiä. Korkein toisteisuus on verbi- ja adverbilekseemeillä (ks. kaavio 8). Toistei-

https://doi.org/10.21435/skst.1472

Taulukko 8. Unelmakirjoitelmien sanaesiintymät sanaluokittain ja ikäluokittain.

Sanaesiintymät sanaluokittain / Ikäluokka %	Alakoulu	Yläkoulu	Aikuiset
Adjektiivit	5,8	7,0	9,4
Adverbit	13,4	13,7	12,2
Konjunktiot	11,0	10,2	7,0
Numeraalit	1,3	1,1	0,8
Pronominit	10,4	9,3	7,4
Substantiivit	27,5	28,7	36,8
Verbit	30,6	29,8	26,2



Kaavio 8. Saman lekseemin toisteisuus (kpl keskimäärin) ikäluokittain ja sanaluokittain.

suus laskee kehityksellisesti, ja ero on suurin alakouluikäisten ja muiden välillä.[19] Substantiivit voivat olla sanarakenteeltaan johtamattomia, johdettuja tai yhdyssanoja, muut ovat pääasiassa joko johtamattomia tai johdettuja.

19 Murrosikä tai kehitystaantuma saattaa näkyä U-muotoisena kehityksenä lineaarisen sijaan (vrt. esimerkiksi McNeil 2007), ulkoisen motivaation puute pelkästään poikkeuksena. Mainittakoon, että ulkoisen motivaation puute heijastuu eniten tuotettujen sanojen määriin, ei niinkään osaamistasoon (vrt. Kyösti 2018; Pajunen & Vainio, tämä teos).

https://doi.org/10.21435/skst.1472

Myös yhden esiintymän lekseemien eli HL-lekseemien (*hapaks legomenon*) määrä muuttuu kehityksellisesti siten, että niiden määrä on aikuisilla aina suurempi kuin koululaisilla. Kehityksellinen muutos on suurin adjektiivilekseemeissä eli 13 prosenttiyksikköä. HL-lekseemien määrä on adjektiivi- ja substantiiviesiintymistä 6–13 prosenttia ja verbi- ja adverbiesiintymistä vain 1,3–3 prosenttia; ero liittyy sanaluokkien lekseemimääräeroihin.

SUBSTANTIIVIEN SEMANTTINEN LUOKITUS

Substantiiveja voi semanttisesti jaotella monella tavalla ja joka luokituksessa on rajatapauksia. Lähtökohta voi olla käsitteellinen WordNet-ajattelun tapaan (ks. Miller 1998, 29; Vossen 2002, 5). WordNet-luokitus koostuu niin sanotuista synonyymijoukoista (*synset*), joissa sanat ovat oletuksen mukaan keskenään vaihdettavissa. Luokituksessa on 25 niin sanottua aloittajakategoriaa, ja entiteetit on jaoteltu kategorioittain toisaalta ominaisuuksia periyttäväksi hierarkiaksi, toisaalta yksittäisiksi. Hierarkiassa alaluokkia ovat organismit (eläimet, inhimilliset, kasvit), objektit (artefaktit, luonnonilmiöt, aine), abstraktiot (ominaisuus, määrä, suhde, aika) sekä psykologiset ominaisuudet (kognitio, emootio ja motivaatio). Alaluokittamattomia entiteettejä ovat aktiviteetti ja tapahtuma, paikka ja tila, ryhmä, omistus ja muoto. Millerin luokitusta on sovellettu lukuisissa WordNetin kielisovelluksissa (ks. esimerkiksi Fellbaum 1998; Vossen 2002). Substantiivit on Unelmaluokituksessa jaoteltu Millerin luokitusta soveltaen neljään pääryhmään a)–d) mutta pääasiassa käsittelemme substantiiveja näiden ryhmien alaryhminä, joissa on jonkin verran otettu tyyppi–toteutuma-jakoa huomioon:

a) Organismit: inhimilliset (yleiset ja roolinimet), elolliset (eläimet, kasvit; elollisista predikoitavat tilat ja tapahtumat; ruumiinosanimet)
b) Objektit ja muut: elottomat (konkreettiset ja laskettavat esineet; konkreettiset spatiaaliset rakennelmat) sekä ainesanat (aine; ruokanimi ja ruoka-annos), luonnontapahtumat, kollektiivit ja instituutiot; spatiaaliset alueet ja luonnolliset muodostelmat; tapahtumanimet

https://doi.org/10.21435/skst.1472

c) Abstraktiot: abstraktioasteeltaan vaihtelevat asianimet, temporaaliset nimet
d) Luokittelun ulkopuoliset

Substantiivien luokittelu heijastaa paitsi Millerin luokitusta myös elollisuus- ja konkreettisuusasteikkoa (ks. myös Frawley 1992; Ravid 2006; Gärdenfors 2014, 115–134). Osaksi luokittelu on riippuvainen aineistosta. Elollisuusasteikolla oliot jakautuvat inhimillisiin, elollisiin, elottomiin konkreettisiin sekä abstrakteihin. Elollisuus voidaan määritellä joko biologian (kädelliset ja muut eläimet, kasvi, bakteeri ja niin edelleen) tai kielen kriteerein (ihminen, korkeaelollinen eli inhimillistettävä, mielikuvan luova eläin, muut eläimet ja niin edelleen) (ks. esimerkiksi Frawley 1992, 89). Konkreettisuusasteikolla konkreettisuus vähenee suunnilleen suhteessa elollisuusasteikkoon mutta alaluokissa vaihtelevammin. Konkreettisimpia ovat niin elolliset, havaittavat, kosketeltavat kuin laskettavat ja manipuloitavat materiaaliset objektit. Konkreettisuus laskee, jos elottomalla oliolla on suuri ulottuvuus, koko tai vastaava (esimerkiksi vuori vs. käteen mahtuva kivi), mikä estää sen käsiteltävyyttä, kosketeltavuutta ja kerralla havaitsemista (ks. Wetzel 2009; Pajunen, Itkonen & Vainio 2016; Rosen 2017). Oliot – ja niiden nimet – eivät kuitenkaan sijoitu suoraan kummallekaan asteikolle, joten aineisto on ensin luokiteltu pienempiin semanttisiin luokkiin, joiden elollisuus- ja konkreettisuusaste on arvioitu myöhemmin erikseen. Niin sanottujen subjektiivisten semanttisten ominaisuuksien – kuten konkreettisuus, kuviteltavuus, polaarisuus ja niin edelleen – arviointi perustuu koehenkilötesteihin, ja ne arvioidaan asteikollisina (ks. esimerkiksi Pajunen & Itkonen 2019).

Elollisuusasteikon elolliset lukeutuvat organismeihin, ja ne on melko helppo yksilöidä. Organismeihin sijoittuvat tässä yhteydessä myös elollisiin osa–kokonaisuussuhteessa olevat ruumiinosanimet ja edelleen elollisesta predikoitavat fysiologiset ja psykologiset tilat ja tapahtumat (*uni*, *stressi*; *alitajunta*, *innostus*), jotka elollisuusasteikolla edustavat abstraktiota. Elottomia objekteja ovat kaikki (melko) konkreettiset oliot, joita ei ole luettu elollisiin. Kaavioissa elottomia eritellään tarpeen mukaan eri alaluokkiin. Rajatuimmin elottomilla viitataan esineisiin (*kynä*, *pöytä*, *ovi*), rakennelmiin (*laituri*, *kivitalo*) sekä ontologialtaan useammantasoisiin

https://doi.org/10.21435/skst.1472

instituutioihin (*koulu*, *kirkko*, *häät*), jotka kirjoitelmissa edustavat sekä olioita (esimerkiksi koulurakennus) että instituutioita. Nämä elottomat eivät ole merkittäviä alueellisilta ulottuvuuksiltaan toisin kuin varsinaiset spatiaaliset alueet ja niihin viittaavat sanat (*suo*, *viljelmä*). Unelmasanaston ainesanoista suuri osa liittyy ruokaan, ja erityisesti nimetään ruokien nimiä ja ruoka-annoksia, koska syöminen kuuluu oleellisesti päiväskriptiin. Ruoka edustaa käsitteen tasoa (*type*) ja ruoka-annos sen toteutumaa (*token*). Näitä ruoan toteutumia on erityisen paljon koululaiskirjoittajilla, mutta eroa ei ole erikseen taulukoitu.

Konkreettisuusasteikon jaossa on oleellista erottaa havaittavat (*pöytä*, *juosta*, *vieri*) ja materiaaliset (*pöytä*) havainnontakaisista (*ajatella*, *ajatus*, *päätös*) ja immateriaalisista (*ajatella*, *juosta*, *vieressä*). Havaittavia ja materiaalisia voi todennäköisesti koskettaa ja liikuttaa, ne sijaitsevat jossakin ja niillä on muoto, hahmo tai muu sellainen. Periaatteessa kaikkein konkreettisimpia ovat ne, jotka voidaan havaita kaikilla aisteilla (näkö, kuulo, haju, maku, tunto ja tuntoaistin alatyypit kosketus, asentoaisti ja niin edelleen) eli esimerkiksi laskettavat ja kosketeltavat esineet. Havaittavia ja immateriaalisia ovat konkreettiset teot ja esimerkiksi läheisyyden ja etäisyyden relaatiot. Brysbaertin, Warrinerin ja Kupermanin (2014) 40 000 sanan konkreettisuusarvioinneissa koehenkilöt nojautuivat useimmiten visuaaliseen ja haptiseen aistiin. Sellaisten substantiivien konkreettisuus onkin yleensä korkea, joiden kuviteltavuuskin on korkea, joista on melko helppo saada mielikuva.

Suomalaiskoehenkilöt (yhteensä 28) arvioivat eri semanttisiin luokkiin kuuluvia sanoja sekä kuviteltavuuden että konkreettisuuden suhteen. Arviontiasteikko oli 1–5. Tulosten mukaan objekti- eli esinesanojen sekä luonto- ja ainesanojen kuviteltavuus on hyvin korkea ja niiden konkreettisuus on joko hyvin korkea (esineet) tai melko korkea (luontosanat). Relationaaliset sanat jakaantuivat sekä kuviteltavuuden että konkreettisuuden suhteen pääasiassa niiden moniontologisuuden mukaan: esimerkiksi sanoilla *raja*, *kolkka* tai *pieli* voi olla sekä spesifi merkitys (*mökki lähellä Venäjän rajaa*) että yleinen, abstraktimpi merkitys (*rajana vain mielikuvitus*). Abstraktisubstantiivien kuviteltavuus on matala, mutta konkreettisuusasteikoille ne sijoittuvat koehenkilönäkemysten mukaan joko melko abstrakteihin (*haju*, *tauko*, *huhu*), abstrakteihin (*huoli*,

https://doi.org/10.21435/skst.1472

harmi) tai hyvin abstrakteihin (*usko*). Tämä ero luonnollisesti kuvastaa taas kerran eroa tyypin ja sen toteutuman välillä (vertaa *hieno tuoksu* ja *laittaa tuoksu*).[20]

On syytä huomata, että vaikka konkreettisuus voidaan perustaa ontologisiin kriteereihin, abstraktius liittyy sekä kieleen että mieleen. Lisäksi abstraktit merkitykset voivat olla kulttuurisia ja osin yksityisiäkin. Oppimisen kannalta ero on huomattava ja vaikuttaa oppimisen lisäksi käyttöön. Konkreettisuusasteikko on kehityksellisestä kannalta mielenkiintoinen ja tiedetään, että erityisesti substantiiviluokissa kielen ja kirjoittamisen kehitys näkyy tekstin abstraktistumisena (Ravid 2006). Alakouluikäisillä onkin vielä paljon puutteita abstraktisanaston hallinnassa (ks. esimerkiksi Turunen 2012). Puutteet heijastuvat hyvin todennäköisesti myös abstraktisanaston käyttöön, koska konkreettisten sanojen motivoituva syntaksi hallitaan paremmin kuin motivoitumattoman abstraktisanaston. Ero näkyy niin kokeellisesti (ks. esimerkiksi Barsalou 2008) kuin erityisryhmilläkin (Kukkonen & Pajunen 1985).

Koko Unelma-aineistossa on analysoitavia substantiivilekseemejä 8 863 ja niiden esiintymiä yhteensä 63 651.[21] Substantiivien semanttisten luokkien jakaumat ovat eri ikäryhmissä aika samanlaisia, mutta selvä siirtymä konkreettisista abstrakteihin on näkyvissä. Vastaava kehityksellinen trendi on vielä selvempi substantiiviesiintymissä (ks. taulukko 9). Keskeinen on 5–10 prosenttiyksikön siirtymä inhimillisistä ja elollisista sekä konkreettisista elottomista abstraktimpiin luokkiin. Toisin sanoen aivan odotuksenmukaisesti iän myötä kirjoitetaan samasta aiheesta selvästi abstraktimmin. Ainesanojen osuus pienenee; ruoasta puhutaan hieman vähemmän. Inhimillisten roolinimistä (*opettaja, kirjailija*) ja ryhmistä (*porukka, ryhmä, pelijoukkue*) kirjoitetaan samoin tai hieman enemmän vanhemmissa ikäryhmissä, samoin aluesanoista ja spatiaalisista ulottuvuuksista (*metsä*; *kulkureitti, matka*), mutta elollisista – kuten lemmikeistä – kirjoitetaan vähemmän. Ajan ilmaukset (*ikä, kuukausi, tulevaisuus, vapaapäivä*) lisääntyvät. Suurimmat prosentuaaliset muutokset

20 Testin jälkeen koehenkilöt kommentoivat *raja*-tyypin sanojen luokittelua sanomalla, että oli vaikea tietää, pitikö luokitella yleistasolla vai ei.

21 Substantiiviluokituksen on tehnyt kaksi henkilöä toisistaan riippumatta ja tulokset on myöhemmin yhtenäistetty.

https://doi.org/10.21435/skst.1472

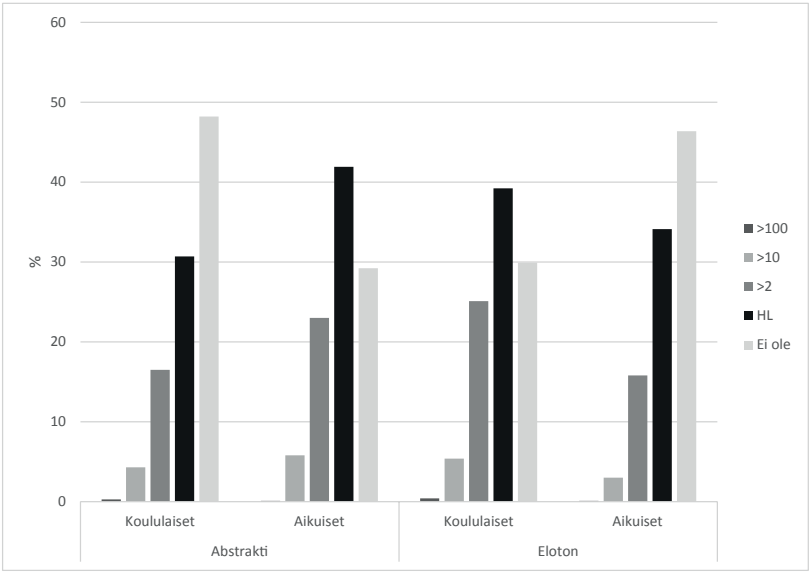
Taulukko 9. Substantiiviesiintymien jakauma semanttisiin luokkiin ikäluokittain.

Substantiiviesiintymien semanttinen luokitus / Ikäluokka %	Alakoulu n = 22 560	Yläkoulu n = 9 779	Aikuiset n = 26 477
Inhimillinen	10,4	7,5	6,6
Inhimillisten rooli	2,3	2,5	2,3
Elolliset	6,5	4,6	2,5
Inhimillisen tai elollisen tila tai osa	3,2	4,4	8,3
Ryhmä	1,2	1,5	1,5
Aine	8,1	9,0	7,0
Eloton	30,8	28,0	22,5
Alue	6,5	7,8	6,7
Ulottuvuus	2,2	2,7	2,5
Relaatio	0,8	1,2	1,1
Aika	14,5	15,5	16,0
Asia	13,5	15,3	22,9

koskevat elottomia konkreettisia objektisanoja (31 % > 22 %) ja elottomia abstraktisanoja (13 % > 23 %). Elottomien alaluokissa muutos näkyy parhaiten esinenimien vähenemisenä ja teon nimitysten lisääntymisenä. Esinenimet (*kumi*, *kynä*) ovat konkreettisuusasteikolla korkeammalla kuin teonnimet (*vihellys*, *potku*), joten elottomiin viittaavan substantiiviluokan sisälläkin näkyy iänmukainen abstraktistumismuutos, vaikka luokan koko sinänsä pienenee iän mukaan. Jos luokituksessa olisi otettu huomioon myös se, viittaako kirjoittaja käsitteeseen tai instituutioon niiden toteutumien sijaan, kehityksellinen ero näkyisi selvemmin.

Suurimmat kehitykselliset muutokset esiintymätasolla näkyvät siis objektisanoissa ja abstraktisanoissa. Jos eroa tarkastellaan lekseemitasolla ja esiintymäluokkatasolla (kaavio 9), nähdään, että koululaisilla ja aikuisilla näiden sanojen luokat koostuvat osin samoista lekseemeistä, mutta erot kasvavat siten, että koululaisilta puolet koko aineiston yhteenlasketuista abstraktisanoista puuttuu, aikuisilta vain noin 30 prosenttia. Lisäksi aikuisilla on abstrakteja HL-lekseemejä enemmän kuin koululaisilla. Objektisanoissa tilanne on päinvastainen. Erot näkyvät vielä sellaisissa yleisyysluokkien lekseemeissä, joissa esiintymiä on muutama (>2 kaaviossa 9). Hyvin yleisiä luokan jäseniä on molemmilla vähän ja eroja ei niissä sanottavasti ole.

https://doi.org/10.21435/skst.1472



Kaavio 9. Elottomiin ja abstrakteihin viittaava sanasto koululaisilla ja aikuisilla jaoteltuna yhteiseen sanastoon (sanan esiintymiä luokittain yli sata, yli kymmenen, yli kaksi, vain yksi) ja erilliseen sanastoon (sana ei esiinny molemmilla ryhmillä eli ei ole yhteinen).

Substantiivit viittaavat olioihin, jotka voi luokitella konkreettisuusasteikolle hieman eri kohtiin riippumatta siitä, mihin semanttisiin luokkiin ne kuuluvat. Esimerkiksi konkreettisen, ihmisen tekemän, laskettavan ja manipuloitavan kokoisen esineen (*kynä, kuppi*) konkreettisuus on korkeampi kuin spatiaalisen alueen (*piha, metsä, saari*), kuten jo todettiin ja kuten koehenkilöiden testaus myös osoitti. Vastaavasti fysiologinen, elollisesta predikoitava tapahtuma (*sydänkohtaus, haukotus*) on konkreettisempi kuin vastaava mentaalinen tila tai tapahtuma (*ajatuksenjuoksu, uteliaisuus*). Lisäksi vaikka sanan tarkoite olisi konkreettinen, sana ei välttämättä kielennä tarkoitetta kovin konkreettisena. Esimerkiksi sanat, jotka viittaavat yksilöitäviin olioihin (*mies*) ovat asteikolla konkreettisempia kuin yksilöistä koostuviin kollektiiveihin viittaavat sanat (*miesryhmä, miesjoukko*), vaikka inhimilliset sinänsä kuuluvat asteikolla ykkössijalle. Vastaavasti sanan merkityksen granulariteetti eli yksityiskohtaisuus voi

https://doi.org/10.21435/skst.1472

Taulukko 10. Substantiiviesiintymien jakauma konkreettisuusasteikolle ikäryhmittäin.

Konkreettisuusasteikko / Ikäluokka %	Alakoulu	Yläkoulu	Aikuiset
1, konkreettinen	37,8	30,3	25,0
2	29,8	32,1	26,8
3	21,5	24,1	24,9
4	10,1	11,5	18,5
5, abstrakti	0,9	2,0	4,8

vaikuttaa konkreettisuuteen. Esimerkiksi hyponyymi–hyperonyymisuhteisessa hierarkiassa hyperonyymi (*laji, eläin*) luokittuu vähemmän konkreettiseksi kuin perustason (*koira*) tai spesifitason (*susikoira*) sana. Meronyymi–holonyymisuhteiset sanat voivat olla vähemmän konkreettisia kuin hyponyymi–hyperonyymisuhteiset, mutta erityisesti relationaaliset, elottoman osan ja kokonaisuuden suhdetta ilmaisevat sanat ovat tilanteisia tarkoitteitaan vähemmän konkreettisia. Esimerkkinä voi mainita suomen lukuisat erilaisia reunaosia kielentävät sanat (*sivu, laita, syrjä, pääty* jne.).

Sekä substantiivilekseemien että -esiintymien (ks. taulukko 10) yhteydessä konkreettisuus vähenee kehityksellisesti lähes lineaarisesti. Erot alakoululaisten ja aikuisten välillä ovat asteikon ääripäissä yli 10 prosenttiyksikön luokkaa eli myös tilastollisesti merkitseviä. Konkreettisuusluokituksessa inhimilliset ja elolliset sekä etenkin konkreettiset esinesanat edustavat konkreettisimpia, kollektiivisanat, ainesanat ja aluesanat pääosin luokkaa 2. Luokkiin 3 ja 4 sijoittuu vielä kuviteltavissa oleviin tarkoitteisiin viittaavia jo jonkin verran abstrakteja sanoja ja erilaisia verbaalisiin tarkoitteisiin viittaavia sanoja. Kumpaankin ryhmään kuuluvia sanoja voidaan ajatella joko yksilöivästi tai yleistasolla (esimerkiksi *allekirjoitus* todistuksessa tai todisteena), tai niihin voi liittyä voimakkaita negatiivisia tunnereaktioita (*sää, kirous*) ja siten konnotaatioita käytönkin tasolla. Puhtaan abstrakteja ovat sanat, jotka eivät herätä mielikuvia tai (yleensä) kielteisiä tunnereaktioita (esimerkiksi *hyvyys, aitous*) (ks. Söderholm ja muut 2013).

Substantiivit eroavat muiden pääsanaluokkien sanoista sikäli, että lekseemien määrä on suuri ja yksittäisen lekseemin yleisyys pieni. Suurin osa Unelmasanaston substantiiveista kuuluukin yleisyystasolle 10 (mediaani), ja aika monissa semanttisissa luokissa neljännes substantiivi-

https://doi.org/10.21435/skst.1472

lekseemeistä on harvinaisia tai puuttuu lehtikorpuksesta. Kirjoitelman näkökulma aiheeseen ja kulloinenkin teema vaikuttavat substantiivisanastoon enemmän kuin muiden sanaluokkien valintoihin. Ero etenkin verbeihin on huomattava. Yleisimmät substantiivilekseemit löytyvät luokista inhimillisen rooli, kollektiivi, aika ja abstraktisana. Substantiivisanastossa näkyy yleisyysmuutoksia semanttisen luokan ja iän mukaan, muun muassa inhimillisiin viittaavat lekseemit sekä aika ja abstrakti -lekseemit harvinaistuvat iän mukaan. Koululaisilla roolinimet ovat harvinaisempia, ja toisaalta sekä elollisiin että elottomiin viittaavat sanat ovat yleisempiä kuin aikuisilla. Kehitykselliset erot eivät tosin ole kovin suuret. Aika usein ero näkyy siten, että kaikkein yleisin tietyn luokan sanasto puuttuu aikuisilta.

Koululaisilla johtamattomat sanat hallitsevat kaikissa semanttisissa substantiiviluokissa. Tämä on odotuksenmukaista, koska vielä yhdeksäsluokkalaisillakin on puutteita derivaation hallinnassa (ks. Vainio ja muut 2019). Aikuisilla johtamattomia on paljon luokissa inhimillinen, eloton ja asiasana sekä myös elollisen tilaa tai osaa nimeävissä. Kummallakin ikäluokalla yhdyssanoja on eniten elottomiin ja asioihin viittaavissa sanoissa. Nämä ovatkin aineiston suurimmat semanttiset luokat. Myös saman lekseemin toisteisuus siirtyy abstraktimpiin sanoihin iän myötä. Leksikaalisessa hierarkiassa johtamattomat sanat ilmaisevat tyypillisesti perustasoa (*kissa*, *koira*) ja johdetut ja yhdyssanat spesifiä tasoa (*siamilainen*, *kisuliini*). Siten spesifin tason sanoilla kielennettäessä granulariteettiaste on korkeampi kuin perustason sanoilla. Koululaisten ja aikuisten suurimmissa substantiiviluokissa eli objekti- ja asiasanoissa granulariteettiero tulee näkyviin erityisen hyvin, jos verrataan sanarakennemuutosta kunkin semanttisen luokan sisällä: aikuisilla johtamattomat sanaesiintymät laskevat koululaisiin verrattuna elottomiin viittaavissa ja asiasanoissa yli 10 prosenttiyksikköä ja inhimillisiin viittaavissa yli 30 prosenttiyksikköä johdettujen hyväksi. Aikuisilla toisin sanottuna substantiivisanasto harvinaistuu sekä abstraktistuu ja ilmaus samalla täsmentyy.

https://doi.org/10.21435/skst.1472

MUIDEN LEKSIKAALISTEN SANALUOKKIEN SEMANTTINEN LUOKITUS

Verbit on luokiteltu toisaalta aikapaikkaisia asiantiloja kielentäviin ja toisaalta niihin, joilla koodataan asiaintiloihin suhtautumista tai vastaavaa. Adjektiivit puolestaan on luokiteltu niiden kielentämän ominaisuuden mukaan. Adverbit jakautuvat ajan, tavan, paikan ja määrän, intensiteetin sekä kommenttien ilmaisuihin. Vertailemme tuloksia ensin semanttisten luokkien perusteella ja siirrymme sitten yleisyystasoittaiseen analyysiin.

Asiantiloja kielentävissä verbeissä on teko- ja tapahtumaverbejä, tilaverbejä ja liikeverbejä, suhtautumista tai vastaavaa ilmaisevissa mentaaliverbejä eli aistihavainto-, emootio-, kognitio- ja kommunikaatioverbejä. Lisäksi on erotettu muutama pienryhmä. Luokkien erottelukriteerit ovat paitsi semanttisia myös syntaktisia (ks. Pajunen 2001). Merkitykseltään melko konkreettisilla verbeillä (havaittavat immateriaaliset) ilmaistaan tekoja, tapahtumia ja tiloja mutta iän myötä samoja verbejä aletaan käyttää abstraktimmin. Esimerkiksi yleiset liikeverbit, kuten *mennä* ja *tulla*, voivat liikkeen sijaan ilmaista aineiston vanhemmilla koehenkilöillä muutosta. Mentaaliverbeillä ilmaistaan asiaintiloihin kohdistuvia propositionaalisia asenteita ja referoidaan. Niiden käyttö lisääntyy kehityksellisesti (ks. myös Pajunen 2012).

Verbilekseemeissä tapahtumaverbien määrä nousee kehityksellisesti ja liikeverbien laskee. Mentaaliverbeissä alakoululaisilla on enemmän kommunikaatioverbejä ja aikuisilla kognitioverbejä. Esiintymien osalta kehityksellinen muutos on selvempi: mentaaliverbien esiintymämäärät – erityisesti kognitioverbien – lisääntyvät, vastaavasti liike- ja tilaverbien esiintymämäärät vähenevät (ks. taulukko 11). Teko- ja tapahtumaverbien esiintymät sen sijaan nousevat. Muutokset ovat toisin sanottuna ennakkoodotuksen mukaisia, ja niitä selittää muun muassa se, että skriptimäinen, siirtymiä ja rutiineja korostava päiväskeemakirjoittaminen vähenee.

Taulukko 11. Verbiesiintymien jakauma semanttisiin luokkiin ikäryhmittäin.

Verbien semanttiset luokat / Ikäluokka %	Alakoulu n = 30 664	Yläkoulu n = 10 184	Aikuiset n = 25 048
Tapahtuma	26,4	28,8	29,4
Liike	25,7	22,7	17,6
Tila	23,9	21,8	19,9
Mentaali	19,0	20,4	26,0
Kieliopillinen	5,0	6,4	7,1

https://doi.org/10.21435/skst.1472

Lekseemitasolla johdettuja verbejä on enemmän kuin johtamattomia muissa luokissa paitsi kieliopillisissa. Verbiyhdyssanat ovat suomen kielessä harvinaisia ja niitä esiintyy harvoissa verbiluokissa (ks. myös Pajunen 2006). Esiintymätasolla johtamattomat vallitsevat kaikissa verbiluokissa. Johdettujen verbien määrä kasvaa ikäryhmittäin kaikissa muissa verbiluokissa paitsi tilaverbeissä ja kieliopillisissa luokissa. Eniten ne lisääntyvät tapahtuma- ja tekoverbeissä, joissa johtaminen muuttaa argumenttirakennetta eli on syntaktista (*kulua* > *kuluttaa*; *aueta* > *aukaista*) tai spesifioi merkitystä (*aloittaa* > *aloitella*). Verbimerkityksetkin siis spesifioituvat iän myötä. Kognitioverbeissä esiintymät kasvavat sekä johtamattomissa että johdetuissa. Verbijohdosten määrän pysyminen kuitenkin suhteellisen vähäisenä lienee tekstityyppi- ja aiheominaisuus.

Unelmasanaston adjektiivit on luokiteltu Dixonin (1977, 2004) perusluokituksen mukaisesti seuraavasti: arvoa tai arvottamista ilmaisevat (*hyvä, huono, huikea, ihana, merkityksellinen*), dimensiota ilmaisevat (ulottuvuus ja muoto: *pitkä, pyöreä*), fyysistä ominaisuutta ilmaisevat (*kova, pehmeä*), inhimillistä taipumusta ilmaisevat (*avulias, ilkeä, rehti*), ikää tai aikaa ilmaisevat (*nuori, viisivuotias*; *keväinen, eilisiltainen*), materiaalista ominaisuutta ilmaisevat (*arkullinen, kurainen, pilvetön*) sekä väriä ja nopeutta ilmaisevat. Lisäksi on eroteltu suomessa adjektiiviluokkaan kuuluvat järjestys (*viimeinen, seuraava*), kansalaisuus (*ulkomaalainen, ruotsalainen*; *ruotsinkielinen*), pronominaalit ja muut (satunnaismuodosteet).

Adjektiivilekseemit jakautuvat näihin semanttisiin luokkiin odotuksenmukaisesti siten, että fyysistä ja materiaalista ominaisuutta ilmaisevia on paljon, inhimillistä taipumusta ilmaisevia suhteellisen paljon. Sen sijaan arvoa ja arvottamista ilmaisevia on paljon odotusarvoon nähden, mutta ero johtuu osaksi siitä, että aineisto on koodattu ennemmin käyttötapaa kuin yksittäistä lekseemiä painottaen. Arvoa ilmaisevia lekseemejä on suomen kielessä melko vähän, mutta arvottamista voi ilmaista subjektiivisin painotuksin. Osaksi arvolekseemien yleisyys johtuu myös kirjoitelmien aiheesta, sillä monet kirjoittajat ovat pyrkineet evaluoimaan unelmapäivän käsitettä. Evaluointipyrkimys nostaa myös värilekseemien määrää. Evaluointi nousee vielä merkittävämmäksi adjektiiviesiintymien osalta (ks. taulukko 12): arvoa ja arvottamista ilmaisevia

https://doi.org/10.21435/skst.1472

adjektiiveja käytetään eniten sekä koululaisilla että aikuisilla (yli kolmannes esiintymistä). Dimensiota ja fyysistä ominaisuutta ilmaistaan lähes yhtä usein, muissa luokissa esiintymämäärät jäävät melko vähiin.

Koululaisten ja aikuisten välillä on havaittavissa tässäkin yhteydessä kehityksellinen ero siten, että koululaiset arvottavat adjektiiveilla useammin kuin aikuiset ja ilmaisevat myös ulottuvuuksia useammin. Nuorimmilla arvottaminen voi tosin olla kiva/hyvä-toteamusta (vrt. Kuisma 2011) sen sijaan, että teksti rakennettaisiin arvoja perustellen. Aikuiset ilmaisevat koululaisia useammin fyysisiä ja materiaalisia ominaisuuksia sekä inhimillisiä taipumuksia. Aikuisilla on lähes jokaisessa semanttisessa luokassa enemmän lekseemejä kuin koululaisilla, toisin sanottuna diversiteetti on korkeampi. Koululaisilla on enemmän värilekseemejä ja -esiintymiä.

Taulukko 12. Adjektiiviesiintymien jakauma semanttisiin luokkiin ikäluokittain.

Adjektiivien luokka / Ikäluokka %	Koululaiset n = 7 680	Aikuiset n = 8 197
Arvo	39,7	36,1
Dimensio	17,3	13,5
Fyysinen	13,3	16,7
Inhimillinen	5,6	9,0
Järjestys	4,7	3,9
Materiaalinen	2,7	4,7
Temporaalinen	6,6	7,3
Väri	4,5	3,0
Muu	5,7	5,8

Adverbit on luokiteltu aikaa (*jo*, *päivittäin*, *kauan*), paikkaa (*kotiin*, *loitolla*, *taakse*, *ääreen*), tapaa (*alasti*, *hellästi*, *hissukseen*, *jalan*), määrää (*kerrallaan*, *lähes*, *osaksi*, *vailla*) ja suhtautumista ilmaiseviin (*muka*, *kenties*, *varmasti*, *valitettavasti*) adverbeihin sekä intensiteetin ilmaisuihin (*sangen*, *perin*, *kunnon*) ja pronominaaleihin (*jotenkin*, *miltei*, *silloin*), jotka voivat ilmaista kaikkia adverbien tavallisia funktioita ja lisäksi toimia lauseita yhdistävinä elementteinä. Koululaiset käyttävät aikuisia enemmän kaikkien muiden adverbiluokkien lekseemejä paitsi kommentti- ja tapa-adverbeja. Erot ovat melko suuria lähinnä paikan adverbeissa koululaisten hyväksi ja tavan adverbeissa aikuisten hyväksi (noin 5 prosenttiyksikköä ja yli

https://doi.org/10.21435/skst.1472

10 prosenttiyksikköä). Esiintymien osalta aikuiset käyttävät enemmän ajan, määrän ja tavan ilmaisuja ja myös kommenttiadverbeja (taulukko 13). Aineisto siten osoittaa, että aikuiset paitsi kehystävät kerrontaansa enemmän kuin koululaiset, he myös evaluoivat sitä melko monipuolisesti. Koululaisilla on enemmän proadverbeja (yli 10 prosenttiyksikköä), joista huomattava osa syntyy lauseketjutyyppisestä *sitten*-sanan käytöstä, myös paikan adverbeja on koululaisilla enemmän. Intensiteettiadverbeja on molemmilla ryhmillä yhtä paljon.

Taulukko 13. Adverbiesiintymien jakauma semanttisiin luokkiin ikäluokittain.

Adverbien luokka / Ikäluokka %	Koululaiset n = 17 844	Aikuiset n = 11 372
Intensiteetti	7,4	7,5
Kommentti	3,0	6,9
Määrä	9,7	10,3
Paikka	20,4	18,5
Proadverbi	25,0	14,9
Temporaalinen	24,2	27,1
Tapa	10,3	14,9

Verbien, adjektiivien ja adverbien semanttinen analyysi osoittaa selvää siirtymää arvioivampaan suuntaan. Asenteiden ilmaisun ja kommenttien osuus lisääntyy ja toisaalta subjektiivisen arvottamisen osuus vähenee. Kehityksellinen ero näkyy sekä yleisyystasojen että ilmaisujen granulariteetin nousussa. Tapahtumien, arvojen ja ajan runsas ja lisääntyvä kielentäminen puolestaan liittyy kertovaan tekstiin ja kertomusrakenteen kehittymiseen.

Lopuksi

Unelmasanaston analyysi osoittaa sanojen valintaan liittyvän useanlaisia tekijöitä. Osa on tekstin ominaisuuksia, osa liittyy kirjoittajan tyylivalintoihin ja osa on selvästi kehityksellistä. Aineiston hallitseva tekstityyppi on kertomus ja kasvava osa niistä on omaelämäkertatyyppisiä. Kertomuksille ominaisessa sanastossa on paljon tapahtumaverbejä sekä kehystäviä ajan ja paikan adverbeja. Kerronta on usein minämuotoista tai re-

https://doi.org/10.21435/skst.1472

feroivaa; tapahtumia koetaan usein yhdessä perheen tai ystävien kanssa ja kyseiset tapahtumat pyritään evaluoimaan positiivisiksi. Kirjoitelmien aihe Unelmieni päivä tuo paljon päiväskriptisanastoa, ja valitut teemat nostavat tiettyjä tapahtumasarjoja esiin: esimerkiksi hääpäivänä toteutuva unelma koostuu häille tyypillisestä tapahtumasarjasta pukeutumisesta vihkimiseen, onnitteluihin, hääkakun leikkuuseen, tanssimiseen, kukkapuketin heittoon ja häämatkaan. Näihin eri tapahtumiin liittyvä sanasto ei vielä yksittäisiä sananvalintoja tarkastelemalla tuo esiin sinänsä juuri kertomukselle tyypillisiä piirteitä eikä myöskään paljasta kehityksellisiä trendejä. Kehityksellisiä piirteitä tulee näkyviin erityisesti frekvenssitekijöiden, sanaston diversiteetin ja sen granulariteettiasteen kautta sekä elollisuus- ja konkreettisuusasteikon muutoksista.

Myös kertomusrakenteen hallinta on kehityksellistä. Suomalaislapset oppivat hallitsemaan kertomusskeemaa keskimäärin ala- ja yläkoulun vaihteessa, nuoruudessa kertomusskeema monipuolistuu kuvailuun ja arviointiin. Nuorilla aikuisilla kirjoitelma saattaa olla tyylipuhdas kertomus unohtumattomasta kokemuksesta tai nautinnollisesta päivästä, mutta se voi myös olla tunnelmia kuvaileva ja maalaileva tai unelmia arvioiva – tai vähän kaikkea. Ehkä vähäinen suunnittelu pitää kertomukset edelleen vaikka ajallisesti aiempaa tiiviimpinä kuitenkin melko kronologisesti etenevinä. Vasta keski-ikäisillä kirjoittajilla tämä kronologia näyttää rikkoutuvan, ja kertomus voidaan laatia hyvin monella tapaa, kuten Petälä (2018) osoittaa. Kertovaan tekstiin liittynee myös ainakin osaksi sanaston suhteellinen yksinkertaisuus, suhteellisen suuri johtamattomien ja melko yleisten sanojen määrä.

Tyylilliset valinnat tai strategiat heijastuvat myös sanastoon. Kirjoitelmista voi erottaa kaksi vallitsevaa tyylistrategiaa, joista toista voisi nimittää kuvailemaan pyrkiväksi ja toista pelkistyneeksi. Edellinen tuottaa lapsilla värikkäitä adjektiiveja ja erikoisia juonen käänteitä, vanhemmilla kirjoittajilla melko yksityiskohtaista kuvailua, jälkimmäinen enemmän toteavaa kerrontaa, jossa päättely jää lukijan kontolle. On myös rönsyilevää ja tiivistä kerrontaa. Tyylilliset valinnat peittyvät usein tilastojen ulottumattomiin mutta on syytä muistaa, että osa koehenkilö- ja ikäluokkakohtaisesta variaatiosta johtuu tyylillisistä valinnoista. Yhtä lailla on kuitenkin syytä panna merkille, että vaikka tyylilliset valinnat vaikuttavat

https://doi.org/10.21435/skst.1472

enemmän aikuis- kuin lapsikerrontaan, tutkimus silti paljastaa selviä iänmukaisia sanastollisia kehitystrendejä.

Koululaiset ovat kirjoitelma-aineiston sana-analyysin perusteella keskimäärin vielä peruskoulun päättyessäkin aika kaukana nuorten aikuisten tasosta. Tämä näkyy erityisen hyvin niin sanotuista peittoluvuista: 2 000–4 000 lekseemiä kattaa koululaisten sanamassasta lähes kaiken mutta aikuisilta vain 90 prosenttia. Koululaiskirjoitelmissa on hyvin yleisiä ja tavallisia lekseemejä merkitsevästi enemmän kuin aikuiskirjoitelmissa; aikuisilla taas harvinaisia lekseemejä on vastaavasti huomattavasti enemmän. Ero tarkoittaa, että aikuiset ilmaisevat täsmällisempiä merkityksiä kuin koululaiset. Sanaesiintymien tasolla massa koostuu molemmilla hyvin yleisistä sanoista, mutta yleisyysero näkyy silti. Aikuisilla myös johdetut sanat ja kiteytymät ovat yleisyydeltään harvinaisempia kuin koululaisilla mutta koululaisilla on enemmän harvinaisia yhdyssanoja. Ero kuitenkin syntyy koululaisten suosimista satunnaismuodosteista, kyse ei ole merkityksen ilmaisun täsmällisyydestä. Taivutusmuotoanalyysi perustui lekseemeille ominaiseen taivutusmuotoprofiiliin sanomalehtikielessä. Kun tätä profiilia verrattiin unelmasanaston yleisyystasoihin, kävi ilmi, että aikuissanasto liittyy vaihtelevampaan syntaksiin kuin koululaissanasto. Tämän voinee tulkita niin, että aikuiset kirjoittavat yleiskielen tapaan ja koululaiset poikkeavat siitä enemmän. Vastaava kehitys tulee esiin sanaston diversiteettianalyysistä. Nuorilla aikuisilla leksikaalinen erilaisuus – eri lekseemien määrä – on selvästi korkeampi kuin yhdeksäsluokkalaisilla, jotka taas peittoavat kuudesluokkalaiset. Suurin ero on kolmasluokkalaisten ja kuudesluokkalaisten välillä.

Diversiteettiä mitattiin myös sukupuolten välillä – ainoana tässä tutkimuksessa. Tulos oli, että ainoastaan kolmasluokkalaisten tyttöjen ja poikien diversiteetti eroaa: mediaani on tytöillä korkeampi mutta pojilla keskeisten havaintojen hajonta on suurempi. Eron vähäisyys on odotuksen mukaista, koska muun muassa älykkyystutkimukset eivät osoita semanttisen tiedon sukupuolittuneisuutta. Se on linjassa myös sanastotestiemme tulosten kanssa. Jos sitten sanastotaidot katsotaan luku- ja kirjoitustaitojen edellytykseksi, tulos on vähintään mielenkiintoinen. Peruskoululaisten aineisto edustaa koko ikäluokkaa, nuorten aikuisten

https://doi.org/10.21435/skst.1472

vain ikäluokan koulutetumpaa osaa. Koululaisten tuloksissa on paljon poikkeavia havaintoja sekä tytöillä että pojilla mutta aikuisryhmästä nämä puuttuvat. Koululaisten aineistosta nouseva suuri poikkeavien havaintojen määrä viittaa erityistuen tarpeeseen. Kaikkein poikkeavin osa koululaisista eli ne, jotka eivät ole saaneet tekstiä sanottavasti aikaiseksi, nostaa tuettavien joukkoa.

Semanttisesti sanastosta näkyy hyvin aiheen käsittelyn yleinen abstraktistuminen, joka tulee esiin erityisesti substantiivisanastosta mutta myös verbisanastosta. Inhimillisistä, elollisista ja konkreettisista elottomista kirjoitetaan ikäryhmittäin aina vähemmän mutta asioista aina enemmän. Tämä tarkoittaa, että vaikka kirjoitelman aihe on kaikilla sama, aihe käsitteellistetään korkeammalle abstraktiotasolle ja siitä kirjoitetaan tiivistetymmin. Kun suunnitteluaikaa ei ollut millään kirjoittajaryhmällä, ero on kehityksellinen. Substantiivien semanttisista luokista näkyy selvä siirtyminen konkreettisista objektisanoista abstrakteihin asiasanoihin, mutta muutos tapahtuu myös luokittain siten, että muun muassa objektisanoista kirjoitetaan abstraktiotasoltaan korkeammalla tyyppien tasolla niiden yksittäisten esiintymäesimerkkien sijaan. Sanaston harvinaistuminen ja nimeämisen diversiteetin nousu taas osoittavat granulariteettiasteen nousun kehitykselliseksi. Yhteenvetona voi siis todeta, että sanamerkitysten abstraktistuminen, harvinaistuminen ja täsmentyminen ovat kehityksellisiä.

Lopuksi voi vielä pohtia, miksi abstraktistuminen on kehityksellistä. Eihän abstrakti teksti aina vaikuta edes tavoiteltavalta. Kehityksellisessä kontekstissa eroa voi luonnehtia seuraavasti: Konkreettisten sanojen oppiminen ja muistaminen on helpompaa kuin abstraktisanojen, konkreettismerkityksistä sanoista keksii paremmin assosiaatioita ja tarkoitteet, joihin ne viittaavat, ovat kuviteltavissa ja niistä voi muodostaa mielikuvia. Tämä kaikki tukee näiden sanojen käyttöä ja kontekstin hallintaa. Abstraktit sanat viittaavat kulttuurisiin ja yksilöllisiin konstruktioihin ja niiden syntaksi opitaan kontekstista – tai analogiasuhteesta konkreettisten sanojen käyttöön –, joten niiden syntaksi ei ole motivoitua (ks. esimerkiksi Barsalou 2008). Abstraktin sanaston oppiminen ja käyttöön ottaminen vaatii siis aikaa. Taitava tekstin rakentaminen edellyttää paitsi yleisyystason nostoa myös kykyä täsmentää ilmaisua.

https://doi.org/10.21435/skst.1472

Lähteet

Baayen, H. 2014. Experimental and psycholinguistc approaches. Julkaisussa: R. Lieber & P. Štekauer (toim.) *The Oxford handbook of derivational morphology*. Oxford: Oxford University Press, 95–117.

Barsalou, L. W. 2008. Grounded cognition. *Annual Review of Psychology* 59, 617–645.

Bauer, L. & Nation, P. 1993. Word families. *International Journal of Lexicography* 6: 4, 253–279.

Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Brysbaert, M., Warriner, A. B., Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46, 904–911.

Bybee, J. & Hopper, P. 2001. Introduction to frequency and the emergence of linguistic structure. Julkaisussa: J. Bybee & P. Hopper (toim.) *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, 1–24.

Carlisle, J. F. 2000. Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing* 12, 169–190.

Croft, W. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago: Chicago University Press.

CSC = Tieteellinen laskenta, yleisyyskorpus (44 miljoonaa sanaesiintymää). https://korp.csc.fi/suomen-sanomalehtikielen-taajuussanasto-B9996.txt. Viitattu12.12.2014.

Dixon, R. M. W. 1977. Where have all the adjectives gone? *Studies in Language* 1, 19–80.

Dixon, R. M. W. 2004. Adjective classes in typological perspective. Julkaisussa: R. M. W. Dixon & A. Y. Aihkenvald (toim.) *Adjective classes: A cross-linguistic typology*. Oxford: Oxford University Press, 1–49.

Fellbaum, C. (toim.) 1998. WordNet. An electronic lexical database. Cambridge: The MIT Press.

Frawley, W. 1992. *Linguistic semantics*. Hillsdale: Lawrence Erlbaum Ass.

Gärdenfors, P. 2014. *The geometry of meaning. Semantics based on conceptual spaces*. Cambridge: The MIT Press.

Hadley, P. A. & Rispoli, M. 2012. Lifespan perspective on individual differences in grammatical abilities. *Linguistic Approaches to Bilingualism* 2: 3, 269–272.

Hakulinen, L. 1979. *Suomen kielen rakenne ja kehitys*. Neljäs, korjattu ja lisätty painos. Helsinki: Otava.

Harjunen, E. & Rautopuro J. 2015. Kielenkäytön ajattelua ja ajattelun kielentämistä. Äidinkielen ja kirjallisuuden oppimistulokset perusopetuksen päättövaiheessa 2014: keskiössä kielentuntemus ja kirjoittaminen. Helsinki: Kansallinen koulutuksen arviointikeskus.

Hart, B. & Risley, T. R. 1995. *Meaningful differences in the everyday experience of young American children*. Baltimore: Brooks.

Hart, B. & Risley, T. R. 2003. The early catastrope: The 30 million word gap by age 3. *American Educator* 27: 1, 4–9.

Heinonen, S. 2018. *Narratiivisuuden sisällöllinen kehittyminen aiheanalyysin kautta tarkasteltuna: Kvantitatiivinen tutkimus aiheiden esiintymisen vaihtelusta iän ja sukupuolen mu-*

kaan "Unelmieni päivä" -kirjoitelmissa. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.

Honko, M. 2013. *Alakouluikäisen leksikaalinen tieto ja taito: Toisen sukupolven suomi ja S1-verrokit.* Tampere: Tampereen yliopisto.

Härkönen, P. 2012. *"Jopas taas melkoisen sopan keitin." Verbi-idiomit Aku Ankassa ja kuinka nuoret niitä ymmärtävät*. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.

Jarvis, S. 2013. Capturing the diversity in lexical diversity. *Language Learning* 63, 87–106.

Jääskeläinen, Heikki. 2008. *Puhekuplat täyteen. Peruskoululaisen luoman kertomuksen loogisuus ja mielikuvituksellisuus sekä kertomuksessa käytetyn verbisanaston monipuolisuus.* Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.

Kauppinen, M., Tarnanen, M. & Aalto, E. 2014. "Voisin pyytää oppilaita alleviivaamaan kaikki adjektiivit" – luokanopettajaopiskelija kielitietoisen aineenoppimisen ohjaajana. Julkaisussa: M. Mutta, P. Lintunen, I. Ivaska & P. Peltonen (toim.) *AFinLA-e*. Soveltavan kielitieteen tutkimuksia 7, 81–100.

Keskinen, J. 2012. *"Jos joskus pitäisi keksiä joku mielikuvitusmaailma, niin aivan varmasti jokaisella ihmisellä olisi ainakin omat unelmat: siitähän jokainen unelmoi." Modaalisuus alakouluikäisten lasten unelmakirjoitelmissa*. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.

Keuleers, E., Stevens, M., Mandera, P. & Brysbaert, M. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology* 8, 1665–1692.

Koivunen, J. 2012. *Adverbien käyttö alakoululaisten kirjoitelmissa*. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.

Kuisma, J. 2011. *Ominaisuuden ilmausten käytön kehittyminen alakouluikäisten kirjoitelmissa*. Pro gradu -tutkielma. Tampereen yliopisto: Suomen kieli.

Kukkonen, P. & Pajunen, A. 1986. Rektio ja agrammatismi. *Virittäjä* 90: 22–45.

Kusnetsoff, T. 2017. *Suomenkielisten nuorten morfologinen tietoisuus ja sen yhteys luetun ymmärtämiseen.* Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.

Kuusela, J. 2011. Kun kirjoittaminen ei suju. Julkaisussa: E. Harjunen, R. Juvonen, J. Kuusela, B. Silén, M. Sääkslahti & M. Örnmark (toim.) *Miten peruskoululaiset kirjoittavat? Näkökulmia ja kysymyksiä. Perusopetuksen 9. luokan äidinkielen ja kirjallisuuden oppimistulosten seuranta-arvioinnin aineistoa 2010.* Raportit ja selvitykset 2011: 2, 13–22. www.oph.fi/julkaisut. Viitattu 22.4.2020.

Kyösti, J. 2018. *Motivoinnin vaikutuksia kuudesluokkalaisten kirjoittamiin kertomuksiin.* Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.

Laasanen, M., Pajunen, A. & Häikiö, T. Arvioitavana. Kielen natiivihallinnan variaatio ja sosiolingvistinen meneillään olevan kielenmuutoksen tutkimus.

Laine, M. & Virtanen, P. 1999. *WordMill. Lexical Search Program. Center for Cognitive Neuroscience*. Turku: University of Turku.

Laine-Leinonen, J. 2013. *Koulukorpuksen leksikko: 1.–6.-luokkalaisten aktiivinen sanavarasto*. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.

Laufer, B. & Nation, P. 1999. A vocabulary size test of controlled productive ability. *Language Testing* 16: 1, 33–51.

Laufer, B. & Ravenhorst-Kalovski, G. C. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* 22: 1, 15–30.

MacWhinney, B. 2001. Emergentist approaches to language. Julkaisussa: J. Bybee & P. Hopper (toim.) *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, 449–470.

Malvern, D. D., Richards, B. J., Chipere, N. & Durán, P. (toim.) 2009 [2004]. *Lexical diversity and language development. Quantification and assessment.* New York: Palgrave Macmillan.

McCarthy, M. & Scott, J. 2007. Vocd: A theoretical and empirical evaluation. *Language Testing* 24: 4, 459–488.

McNeil, N. 2007. U-shaped development in math: 7-year-olds outperform 9-year-olds on equivalence problems. *Developmental Psychology* 43, 687–695.

Michaels, S. 2013. Déjà Vu all over again: What's wrong with Hart & Risley and a "linguistic deficit" framework in early childhood education? *Learning Landscapes 1,* 23–41.

Miller, G. A. 1998. Nouns in WordNet. Julkaisussa: C. Fellbaum (toim.) *WordNet. An electronic lexical database*. Cambridge: The MIT Press, 23–46.

Miller, G. A. & Fellbaum, C. 1991. Semantic networks in English. *Cognition* 41, 198–229.

Milton, J. 2009. *Measuring second language vocabulary acquisition*. Bristol: The Multilingual Matters.

Nahkola, T. 2012. *Yhdyslauseiden kehitys kouluikäisen kielessä: Tutkimus 1.–6.-luokkalaisten kirjoitelmista*. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.

Nation, I. S. P. & Coxhead, A. 2021. *Measuring native-speaker vocabulary size*. Amsterdam: John Benjamins.

Nippold, M. A. 2006 [3. painos]. *Later language development: School-age children, adolescents and young adults*. Austin: Pro-Ed.

Olinghouse, N. G. & Wilson, J. 2013. The relationship between vocabulary and writing quality in three genres. *Reading and Writing* 26, 45–65.

OPH 2014. Perusopetuksen opetussuunnitelman perusteet. Helsinki: Opetushallitus.

Qian, D. D. 2002. Investigating the relationship between vocabulary knowledge and academic reading comprehension: An assessment perspective. *Language Learning* 52: 3, 513–536.

Pajunen, A. 1994. Adjektiivikategorian universaaliudesta. *Virittäjä* 98: 4, 513–542.

Pajunen, A. 1998. Adjectives in spoken language discourse. *Word* 49: 3, 341–368.

Pajunen, A. 1999. *Suomen verbirektiosta*. Turku: Turun yliopisto.

Pajunen, A. 2001. *Argumenttirakenne. Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Pajunen, A. 2006. Verbisanasto uusiutuu. *Puhe ja kieli* 26: 4, 205–219.

Pajunen, A. 2010. Sanojen synteettisyysasteesta suomen kielessä. *Virittäjä* 114: 4, 481–501.

Pajunen, A. 2012. Kirjoittamistaitojen kehitys 8–12-vuotiailla: Alakoululaisten unelmakirjoitelmat. *Virittäjä* 116: 1, 3–34.

Pajunen, A. 2016. Sukupuolierot kielen hallinnassa (etenkin kirjoittamisessa). Esitelmä. Språkets funktion 6.–7.6.2016, Åbo Akademi.

Pajunen, A. 2017. Tapahtumien nimeämisen normitus ja variaatio. Esitelmä. MYK 15.12.2017. Turun yliopisto.

Pajunen, A. & Virtanen, P. 2002. Helsingin Sanomat 2000–2001 [= HS2000-korpus]. Morfosyntaktisesti analysoitu tietokanta, 24 miljoonaa sananmuotoa.
Pajunen, A. Puranen, M. & Yli-Paavola, A. 2010. Koulukorpus. Alakoululaisten kirjoitelmia aiheesta Unelmieni päivä. Excel-tietokanta. Tekijöiden hallussa.
Pajunen, A. Itkonen, E., Vainio, S. 2015. Sanamerkityksen hallinta nuorilla aikuisilla. *Virittäjä* 119: 2, 160–187.
Pajunen, A. Itkonen, E. & Vainio, S. 2016. Nuorten aikuisten kyky määritellä sanoja. *Virittäjä* 120: 4, 477–515.
Pajunen, A. & Itkonen, E. 2019. Intuition and beyond: A hierarchy of descriptive methods. Julkaisussa: A. Mäkilähde, V. Leppänen & E. Itkonen (toim.) *Norms and normativity in language and in linguistics*. Amsterdam: John Benjamins, 213–234.
Palonen, O. 2013. *Pronominit 8–12-vuotiaiden kirjoitelmissa*. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.
Pollock, L. 2018. Statistical and methodological problems with concreteness and other semantic variables. A list of memory experiment case study. *Behavioral Research Methods* 50: 3, 1198–2016.
Rantala, J. 2012. *Ihmisen olotilaa kuvaavien ilmausten sanastonhallintataidot yläkoulussa ja lukiossa*. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.
Ravid, D. 2006. Semantic development in textual contexts during the school years: Noun scale analyses. *Journal of Child Language* 33, 791–821.
Rosen, G. 2017. *Abstract objects*. Stanford: Center for the study of language and information. https://plato.stanford.edu/entries/abstract-objects. Viitattu 1.3.2018.
Routama, K. 2008. *Peruskoulun 2. ja 3.-luokkalaisten spatiaalisen sanaston hallinta. Konkreettiset ja metaforiset ilmaukset*. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.
Saarela, L. 1997. *Peruskoululaisten kirjoitelmien kehittyminen sanastotutkimuksen valossa*. Oulu: Oulun yliopisto.
Schmitt, N. 2010. *Researching vocabulary. A vocabulary research manual*. London: Palgrave Macmillan.
Stavroula-Thaleia, K., Vigliocco, G., Del Campo, E., Vinson, D. P. & Andrews, M. 2011. The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General* 140: 1, 14–34.
Söderholm, C., Häyry, E., Laine, M. & Karrasch, M. 2013. Valence and arousal ratings for 420 Finnish nouns by age and gender. *Plos One* 8: 8, e72859.
Taylor, J. R. 1995 [1989]. *Linguistic categorization. Prototypes in linguistic theory*. Oxford: Oxford University Press.
Tomasello, M. 2003. *Constructing a language. A usage-based theory of language acquisition*. Cambridge: Harvard University Press.
Treffers-Daller, J. & Milton, J. 2013. Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistic Review* 4: 1, 151–172.
Turunen, R. 2012. *Viides- ja seitsemäsluokkalaisten abstraktien substantiivien ja verbien hallinta*. Suomen kielen pro gradu -tutkielma. Tampereen yliopisto.
Vainio, S., Pajunen, A. & Häikiö, T. 2019. Acquisition of Finnish derivational morphology: School-age children and young adults. *First language* 39: 2, 139–157.

https://doi.org/10.21435/skst.1472

Wagner, R. K., Muse, A. E. & Tannenbaum, K. R. (toim.) 2007. *Vocabulary acquisition: Implications for reading comprehension*. New York: The Guilford Press.

Wetzel, L. 2009. *Types and tokens: On abstract objects*. Cambridge: The MIT Press.

Vossen, P. (toim.) 2002. EuroWordNet. General document. http://www.hum.uva.nl/~ewn. Viitattu 1.3.2018.

Yap, M. J., Lim, G. Y. & Pexman, P. M. 2015. Semantic richness effects in lexical decision: The role of feedback. *Memory & Cognition* 43: 1148–1167.