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

Firma vai *yritys*? Vierassanojen ja omaperäisten sanojen objektiivinen ja subjektiivinen frekvenssi selkokielessä

• Idastiina Valtasalmi, Informaatioteknologian ja viestinnän tiedekunta, Tampereen yliopisto

Kirjoittajan yhteystiedot: Idastiina Valtasalmi, idastiina.valtasalmi@tuni.fi

Artikkelissa tarkastellaan selkoteksteistä poimittujen vierassanojen (N = 50) ja niiden omaperäisten vastinesanojen (N = 50) objektiivista ja subjektiivista frekvenssiä. Objektiivista frekvenssiä tutkittiin vertailemalla sanojen esiintymistaajuutta selko- ja yleiskielisiä mediatekstejä sisältävissä korpuksissa. Subjektiivista frekvenssiä tutkittiin kyselyllä, jolla selvitettiin, kuinka usein kieliasiantuntijat (N = 25) kohtasivat sanoja ja mitä sanoja he käyttivät yleensä itse. Tutkimusaineiston omaperäiset sanat olivat korpusaineistoissa harvinaisempia kuin vierassanat, mutta omaperäisiä sanoja esiintyi selkokielessä enemmän kuin yleiskielessä. Yksittäisten sanojen esiintymistaajuudessa oli huomattavia eroja, jotka selittyivät esimerkiksi korpusaineistojen ja kielimuotojen eroilla. Kohtaamistaajuudeltaan yleisimpiä olivat objektiiviselta frekvenssiltään yleiset sanat, jotka esiintyivät monissa eri konteksteissa. Kohtaamistaajuus ja objektiivinen frekvenssi olivat samansuuntaisia etenkin yleisemmissä sanoissa, kun taas objektiiviselta frekvenssiltään harvinaiset sanat olivat kohtaamistaajuudeltaan monenlaisia. Tutkimusaineiston lyhyimmät sanat olivat objektiiviselta frekvenssiltään ja kohtaamistaajuudeltaan yleisimpiä. Kieliasiantuntijat suosivat omassa kielenkäytössään tutkimusaineiston sanaparien lyhyempiä sanoja ja vierassanoja. Esiintymistaajuus ei riitä ainoaksi sanojen yleisyyden arviointikriteeriksi selkokielessä. Yleisyyttä voitaisiinkin arvioida myös kohtaamistaajuudella ja käytöllä.

Avainsanat: korpuslingvistiikka, objektiivinen frekvenssi, selkokieli, subjektiivinen frekvenssi, verkkolomakekysely

1 JOHDANTO

Selkokieli on yksinkertaistettu kielimuoto, josta on hyötyä henkilöille, joille yleiskieli on liian vaikeaa (Leskelä, 2019, s. 93). Selkokielen kohderyhmät määritellään Suomessa selkokielen tarpeen pysyvyyden ja taustalla olevan syyn perusteella (Leskelä & Lindholm, 2012). Selkokielen tarve voi johtua pysyvistä ja synnynnäisistä neurobiologisista syistä (esim. kehitysvammaiset henkilöt), kielitaidon heikentymisestä elinaikana (esim. muistisairaat henkilöt) tai karttuvasta kielitaidosta (esim. kielenoppijat). Tilannekohtaisesti selkokielestä voi hyötyä kuka tahansa.

Selkokielen kirjoitusohjeet ovat perustuneet enimmäkseen käytännön kokemuksiin eivätkä tutkittuun tietoon (Uotila, 2020). Selkokielessä suositaan yleisiä ja tuttuja sanoja, jotka ovat mieluummin lyhyitä kuin pitkiä (Selkokeskus, 2022, kriteerit 28 ja 30). Vierassanoja vältetään, koska niitä pidetään vaikeina selkokielen käyttäjille (kansainvälisistä selko-ohjeista ks. Inclusion Europe, 2009, s. 10; suomesta ks. Virtanen, 2012, s. 83). Vierassanat korvataan mahdollisuuksien mukaan yleisillä, omaperäisillä vastineilla (Selkokeskus, 2022, kriteeri 38). Jos vaikeiksi oletettuja sanoja käytetään, ne selitetään (kansainvälisistä selko-ohjeista ks. IFLA, 2010, s. 11; suomesta ks. Leskelä, 2019, s. 134–136; Selkokeskus, 2022, kriteeri 29). Suomessa selkokielen kehittäminen

alkoi 1980-luvulla vammaisjärjestöjen piirissä, mutta tieteellinen tutkimus alkoi yleistyä 2010-luvulla (Uotila, 2020). Selkosuomen periaatteille etsittiin kuitenkin tieteellisiä perusteita jo kehittämistyön alkuvaiheessa.

Selkosuomen periaatteiden taustalla on viestinnäntutkija Osmo A. Wiion (1974) teoria kielellisestä ymmärrettävyydestä. Teorian mukaan harvinaiset, pitkät ja vierasperäiset sanat lisäävät tekstin vaikeustasoa (Wiio, 1974, s. 181–183). Kielen kansanomaistamisessa vierasperäiset sivistyssanat, kuten *energia* ja *planeetta*, korvataan suomenkielisillä vastineilla tai selitetään, jotta suuri yleisö voi ymmärtää ne (Wiio, 1974, s. 228). Kieliympäristö on kuitenkin muuttunut vuosikymmenten aikana, eikä vieraista kielistä peräisin olevia sanoja enää pidetä vain harvojen osaamina (Räsänen 2002). Myös selkokielen käyttäjien kieliympäristö on muuttunut, sillä he eivät elä yleiskielisestä kieliympäristöstä irrallaan. Siksi on tarpeellista tutkia vierassanojen käyttöä selkokielessä ja selvittää niiden yleisyyttä uudempien frekvenssiteorioiden pohjalta. Ruotsalainen kielentutkija Åsa Wengelin (2015) on peräänkuuluttanut näyttöön perustuvaa kielenhuoltoa ja havainnut, että selkoruotsin kirjoitusohjeet perustuvat osittain vanhentuneeseen tutkimustietoon. Selkosuomenkin kirjoitusohjeiden taustalla voi olla vanhentunutta tietoa.

Kielentutkimuksessa sanojen yleisyyttä kuvataan frekvensseillä (ks. esim. Milton, 2009, s. 22). Objektiivista frekvenssiä tutkitaan laskemalla sanojen esiintymistaajuuksia aineistoissa, kuten tekstikorpuksissa (McEnery & Hardie, 2011). Frekvenssiin vaikuttavat aineiston koko ja sisältö. Sanoja saatetaan käyttää vain tietyissä konteksteissa, joita voi olla aineistossa vain vähän tai ei lainkaan (Baayen, 2001, s. 164; Thompson & Desrochers, 2009, s. 453–454). Frekvenssi vaihtelee tekstilajeittain, sillä sanat voivat olla tyypillisempiä joissakin tekstilajeissa kuin toisissa (Desrochers & Thompson, 2009, s. 546–574). Sanojen frekvensseissä on eroja selko- ja yleiskielen välillä (Valtasalmi, 2021b; Vanhatalo & Lindholm, 2020). Frekvenssiin vaikuttaa myös ajankohta, sillä eri aikoina käytetään erilaisia sanoja. Frekvenssillä on vaikutusta sanojen oppimiseen, prosessointiin ja käyttöön. Tiedetään, että objektiiviselta frekvenssiltään yleisiä sanoja opitaan aiemmin, prosessoidaan paremmin ja käytetään useammissa eri konteksteissa kuin harvinaisia sanoja (Schmitt, 2010, s. 63–64).

Sanojen yleisyyttä tutkitaan usein korpuksista, vaikka sanoja voidaan kohdata kieliympäristössä monella eri tavalla. Jos sanat esiintyvät korpuksessa yhtä usein, ne ovat objektiiviselta frekvenssiltään samanlaisia, vaikka olisivat taajuudeltaan muussa kielenkäytössä erilaisia. Siksi on tutkittu subjektiivista frekvenssiä eli kohtaamistaajuutta (engl. *frequency of encounters*, Balota, Pilotti & Cortese, 2001). Subjektiivinen frekvenssi kertoo, kuinka usein kielenkäyttäjä kohtaa sanoja omassa ympäristössään. Kohtaaminen voi toteutua lukemisena, kuulemisena, kirjoittamisena tai puhumisena (Balota ym., 2001, s. 641). Subjektiivinen frekvenssi ei edellytä sanan merkityksen tuntemista, vaan se kuvaa ainoastaan sanan ilmiasun kohtaamistaajuutta. Kohtaamistaajuutta arvioidaan kyselyillä, joissa arviointiasteikon ääripäissä ovat sanan kohtaaminen usein tai ei koskaan (Ballot, Mathey & Robert, 2022; Balota ym., 2001; Chen & Dong, 2019). Objektiivinen ja subjektiivinen frekvenssi korreloivat keskenään, mutta eivät täydellisesti (Chen & Dong, 2019; Desrochers & Thompson, 2009).

Objektiivista ja subjektiivista frekvenssiä ei eroteta selkokielen teoriassa, vaan kumpaakin kutsutaan *yleisyydeksi*. Objektiivista frekvenssiä tarkoitetaan, kun selko-ohjeissa (Leskelä, 2019, s. 131; Virtanen, 2012, s. 82–83) neuvotaan hakemaan tietoa sanojen yleisyydestä Suomen kielen taajuussanastosta (Saukkonen, 1979) ja Parole-tekstikorpuksen sanamuotojen taajuuslistalta (Kotimaisten kielten keskus, ei pvm.). Subjektiivista frekvenssiä tarkoitetaan, kun selkomukauttajia neuvotaan opettelemaan tunnistamaan yleisin sanasto ja käyttämään sitä (Leskelä, 2019, s. 131). Oletuksena on, että selkomukauttajille kehittyy tuntuma, jonka perusteella he tunnistavat yleisiä sanoja ja arvioivat, sopivatko ne käytettäväksi selkokielessä (Virtanen, 2012, s. 85–86).

Selkokielen teoriassa yleisyyteen liitetään tuttuus. Ajatellaan, että vieraat sanat tulevat lukijoille tutummiksi, kun ne toistuvat selkoteksteissä (Virtanen, 2012, s. 84). Siksi toisto on suositeltavampaa selkoteksteissä kuin yleiskielisissä teksteissä. Tuttuuden ajatellaan myös riippuvan siitä, millaisia sanoja kielenkäyttäjät kuulevat omassa ympäristössään (Leskelä, 2019, s. 131). Tuttuutta ei määritellä selkokielen teoriassa tarkasti, mutta

se tarkoittanee jonkinlaista sanatietoa, joka voi kasvaa kohtaamistaajuuden tihenemisen myötä. Kielentutkimuksessa tuttuudella tarkoitetaan käsitteellistä tuttuutta, eli kielenkäyttäjä tuntee käsitteen, johon sanahahmo ja sen merkitys viittaavat (ks. esim. Gernsbacher, 1984). Tuttuutta on selvitetty kyselyillä, joissa arviointiasteikon ääripäissä ovat tutuiksi ja vieraiksi koetut sanat (ks. esim. Gernsbacher, 1984; Pajunen, Itkonen & Vainio, 2015; Valtasalmi, 2021a). Tuttuus ja objektiivinen frekvenssi korreloivat keskenään (Tanaka-Ishii & Terada, 2011). Subjektiivinen frekvenssi ei edellytä tuttuutta, sillä kielenkäyttäjät voivat kohdata tuntemattomia sanoja.

Tässä tutkimuksessa tarkastellaan selkoteksteistä poimittujen vierassanojen ja niiden omaperäisten vastinesanojen objektiivista ja subjektiivista frekvenssiä. Korpuksista kerätyillä objektiivisilla frekvensseillä kuvataan sanojen esiintymistaajuutta selko- ja yleiskielessä. Kielen ja viestinnän asiantuntijoilta (vastedes *kieliasiantuntijoilta*) kerätyillä kohtaamistaajuusarvioilla kuvataan sanojen subjektiivista frekvenssiä. Oletuksena on, että jos kieliasiantuntijat eivät lue, kuule, puhu tai kirjoita sanaa, sitä ei todennäköisesti kannata käyttää selkokielessä. Tarkemmin tutkitaan sanojen käyttöä, joka voi toteutua puhumisena tai kirjoittamisena. Sanat voivat olla kohtaamistaajuudeltaan samanlaisia, vaikka käytössä olisi eroa. Kielenkäyttäjä voi esimerkiksi kohdata viikoittain sanoja *bussi* ja *linja-auto* mutta käyttää itse sanaa *bussi*.

Tutkimuksen tavoitteena on selvittää, mitä *yleisyys* tarkoittaa selkokielen sanastokriteerinä. Selkokielen sanojen yleisyyttä kuvaillaan tässä tutkimuksessa ensimmäistä kertaa objektiivisella ja subjektiivisella frekvenssillä. Tutkimuksessa etsitään vastauksia seuraaviin kysymyksiin: Ovatko selkoteksteistä poimitut omaperäiset sanat yleisempiä kuin vierassanat, ja miten yleisyyttä voidaan arvioida, kun tehdään sanavalintoja selkokielessä? Ovatko objektiivinen ja subjektiivinen frekvenssi samansuuntaisia, ja miten yleisyys liittyy sanojen pituuteen ja käyttöön? Vastauksia kysymyksiin etsitään teksti- ja käyttäjälähtöisesti: korpustutkimus edustaa tekstilähtöistä näkökulmaa ja kieliasiantuntijoiden kysely edustaa käyttäjälähtöistä näkökulmaa (vrt. Hansen-Schirra & Maaß, 2020, s. 9).[1] Käyttäjälähtöisessä näkökulmassa on mukana myös selkokielen tuottajien näkökulmaa, sillä osa kyselyn vastaajista oli selkokielen asiantuntijoita (vrt. Hansen-Schirra & Maaß, 2020, s. 10).

2 AINEISTO JA MENETELMÄT

Tutkimusaineisto koostui selko- ja yleiskielen korpuksista (ks. lukua 2.1), niistä kerätyistä sanoista ja frekvensseistä (ks. lukua 2.2) sekä kieliasiantuntijoiden kyselyvastauksista (ks. lukua 2.3). Korpuksista tutkittiin sanojen objektiivista frekvenssiä selko- ja yleiskielessä. Kyselyllä tutkittiin subjektiivista frekvenssiä, eli kuinka usein kieliasiantuntijat kohtaavat sanoja. Lisäksi selvitettiin, mitä sanoja he käyttävät yleensä itse. Korpusaineistot analysoitiin vertailemalla sanojen frekvenssejä selko- ja yleiskielessä. Kyselyaineisto analysoitiin tarkastelemalla tyypillisimpiä ja yleisimpiä vastauksia (analyysimenetelmistä ks. lukua 3).

2.1 Korpusaineistot

Tutkittavana oli 50 vierassanaa ja 50 omaperäistä vastinesanaa[2] (esim. *demokratia* – *kansanvalta*) sekä niiden objektiiviset frekvenssit, jotka kerättiin kolmesta selkokielen korpuksesta. Selkosanomat/Selkouutiset -korpus (Helsingin yliopisto, 2017[3]) sisältää vuosina 2010–2013 julkaistuja sanomalehtitekstejä, ja Yle suomenkielisen uutisarkiston selkouutiset 2011–2018 -korpus (Yleisradio, ei pvm.b) sisältää vuosina 2011–2018 julkaistuja mediatekstejä. Kummankin korpuksen tekstit on suunnattu kaikille selkokielen kohderyhmille. Leija-korpus (Helsingin yliopisto, 2017) sisältää vuosina 2009–2016 julkaistuja aikakauslehtitekstejä, jotka on suunnattu erityisesti kehitysvammaisille selkokielen käyttäjille. Korpusten tekstit ovat vaikeustasoltaan perustason selkokieltä, joka on tarkoitettu henkilöille, jotka todennäköisesti

1 Hansen-Schirra ja Maaß (2020, s.10) rajaavat käyttäjälähtöisen näkökulman selkokielen kohderyhmiin, mutta tässä tutkimuksessa näkökulma käsitetään laajemmin. Käyttäjälähtöisiä näkökulmia selkokieleen voidaan saada myös kielenkäyttäjiltä, joilla ei ole pysyvää selkokielen tarvetta (ks. esim. Valtasalmi 2021a).

2 Omaperäisiksi laskettiin myös muista kielistä omaksutut sanat, jotka ovat täysin mukautuneet suomen kieleen (esim. *pelle*).

3 1990- ja 2000-luvun suomalaisia aikakaus- ja sanomalehtiä -korpus (Helsingin yliopisto, 2017) sisältää Selkosanomat/Selkouutiset- ja Leija-korpuksen.

lukevat itse (vaikeustasoista ks. Leskelä, 2019, s. 160–172). Selkokielen korpuksissa on yhteensä 1 428 158 sanetta.

Sanojen frekvenssit kerättiin myös Ylen suomenkielinen uutisarkisto 2019–2021 -korpuksesta (Yleisradio, ei pvm.a), joka sisältää enimmäkseen yleiskieltä. Korpusaineisto rajattiin vuoteen 2021, jotta saatiin selville, mitä sanoja on käytetty viime aikoina. Korpus sisältää pienen määrän Ylen Selkouutiset -aineistoa vuodelta 2021, mikä kuvastaa ympäristön kielisyötettä. Suomessa kieliympäristö on pääasiassa yleiskielinen, mutta laajalevikkisiä mediatekstejä julkaistaan myös selkokielellä. Kielenkäyttäjien kohtaamat sanat voivat siis esiintyä kohtaamishetkellä selko- tai yleiskielessä. Vuoden 2021 Ylen suomenkielinen uutisarkisto -korpuksessa on yhteensä 25 374 938 sanetta. Näistä 25 242 467 sanetta on peräisin yleiskielisistä mediateksteistä ja 132 471 sanetta vuoden 2021 Yle Selkouutisisista. Yle Selkouutiset 2021 muodostaa noin 0,5 % vertailuaineistosta, eli selkokielen osuus on hyvin pieni.

Koska korpusaineistot olivat kielimuodoiltaan, sisällöiltään ja julkaisuajoiltaan erilaisia, niistä ei saatu täysin vertailukelpoista tietoa sanojen frekvenssieroista. Niistä kuitenkin saatiin tarpeellisia näkökulmia sanojen arviointiin, kun tehdään sanavalintoja selkokielessä. Koska yleisyys on keskeinen selkokielen kriteeri, on hyödyllistä tarkastella sanojen frekvenssejä eri kielimuodoissa ja tekstilajeissa. Myös ajan myötä tapahtuneista sanaston muutoksista on hyvä saada tietoa, jotta selkokieleen sopivien sanojen yleisyyttä voidaan tarkastella muutenkin kuin melko vanhoista taajuussanastoista (vrt. lukuun 1).

2.2 Tutkittavat sanat ja objektiiviset frekvenssit

Tutkittavat sanat kerättiin hakemalla Selkosanomat/Selkouutiset, Leija ja Yle suomenkielisen uutisarkiston selkouutiset 2011–2018-korpuksista vierassanoja, joille annetaan vastineeksi samaa tarkoittava omaperäinen sana *eli*-selityksellä, kuten virkkeessä *Lauri kertoo pitävänsä muodista, jossa on* ***luksusta eli ylellisyyttä*** (Leija 4/2015) (eli-selityksistä ks. Kulkki-Nieminen, 2010, s. 135–138; Leskelä, 2019, s. 135; Virtanen, 2012, s. 85). Aineistoon ei otettu sanapareja, joissa omaperäinen vastine on vierassanan yläkäsite (esim. *gradu eli opinnäytetyö*), alakäsite (esim. *diplomaatti eli suurlähettiläs*) tai rinnakkaiskäsite (esim. *rasismi eli muukalaisviha*). Pois karsittiin selittävät tilapäismuodosteet (esim. *aerobic eli tanssijumppa*), sanaliitot (esim. *asteroidi eli avaruuden kivi*) ja sanaparit, joissa sanoilla on yhteinen yhdyssanan loppuosa (esim. *fantasia- eli mielikuvitusmaailma*). Aineistoon ei otettu esimerkinomaisia ilmauksia (esim. *digilaitteet eli älypuhelimet ja tietokoneet*) eikä parafraaseja (esim. *dokumentti eli mahdollisimman todenmukainen asiaohjelma*). Näillä kriteereillä valittiin 44 sanaparia. Lisäksi valittiin kuusi sanaparia, joissa vierassanaa ei selitetä omaperäisellä vastineella lainkaan. Oletuksena on, että sanoja *bussi – linjaauto, bändi – yhtye, frendi – kaveri, idea – ajatus, media – tiedotusväline* ja *turisti – matkailija* pidetään niin tuttuina, ettei niitä tarvitse selittää.

Sanojen synonyymit ja niiden merkitykset tarkastettiin Nykysuomen sanakirjasta (Nykysuomen Sanakirja I-VI, 2002) ja Kielitoimiston sanakirjasta (Kielitoimiston sanakirja 2021). Kahdessa sanaparissa merkitys oli hieman epätarkka. *Ekosysteemin* vastineeksi hyväksyttiin *eliöyhteisö*, vaikka ekosysteemi koostuu eliöistä ja elottomasta luonnosta. *Infektion* vastineeksi hyväksyttiin *tulehdus*, vaikka *tulehduksen* tarkempi vastine olisi *inflammaatio*. *Infektio* voi kuitenkin tarkoittaa tulehdusta, jonka aiheuttaa taudinaiheuttajan tunkeutuminen elimistöön (Suomen virtuaaliyliopisto, 2006). Synonymia oli joissakin sanapareissa tarkempaa kuin toisissa, sillä sanojen merkitykset saattoivat olla laajuudeltaan erilaisia (synonymiasta ks. esim. Larjavaara 2007, 138–141). Esimerkiksi sanaparissa *remontti – korjaus* omaperäinen sana on ekstensioltaan laajempi. Aineistossa oli myös monimerkityksisiä sanoja, kuten *virus, tabletti* ja *resepti*. Kaikki tutkimusaineiston sanat olivat substantiiveja, ja vierassanat olivat suurimmaksi osaksi erikoislainoja. Erikoislainat ovat nuoria lainasanoja, jotka ovat osittain mukautuneet suomen kieleen mutta poikkeavat äänteiltään tai äänneyhdistelmiltään omaperäisistä sanoista (Häkkinen, 1990, s. 260–265).

Selkokielen korpuksista kerätyn aineiston kahdessakymmenessä sanaparissa omaperäinen sana oli yleisempi kuin vierassana, 26 sanaparissa vierassana oli yleisempi kuin omaperäinen sana, ja neljässä sanaparissa sanat olivat yhtä yleisiä (ks. taulukkoa 1).

TAULUKKO 1

Sanaparit ja sanojen objektiivinen frekvenssi tutkimusaineistossa (Selkokieli 2009–2018)[4]

Sanaparin tyyppi	**Sanaparit ja sanojen frekvenssit miljoonaa sanetta kohti**
1. Omaperäinen sana on yleisempi kuin vierassana	eduskunta (478,2) – parlamentti (173); laitos (158,2) – instituutio (0,7); aihe (261,9) – teema (133,7); kaveri (124,6) – frendi (24,5); näytelmä (76,3) – draama (4,9); maakunta (39,9) – provinssi (3,5); mielenilmaus (32,9) – protesti (7); pelle (23,1) – klovni (8,4); yhtye (168,7) – bändi (154,7); ajatus (106,4) – idea (93,8); tulehdus (8,4) – infektio (1,4); pikajuoksija (8,4) – sprintteri (2,8); viestintä (9,1) – kommunikaatio (3,5); liiketoiminta (9,8) – bisnes (5,6); mielikuvitus (6,3) – fantasia (2,8); elinalue (4,2) – reviiri (1,4); rahoittaja (5,6) – sponsori (2,8); tilauslento (3,5) – charterlento (1,4); tuotemerkki (2,8) – brändi (1,4); tunnus (41,3) – logo (40,6)
2. Vierassana on yleisempi kuin omaperäinen sana	firma (448,1) – yritys (205,2); media (176,5) – tiedotusväline (16,8); bussi (180) – linja-auto (27,3); turisti (200,3) – matkailija (75,6); diabetes (94,5) – sokeritauti (9,8); meteorologi (70) – ilmatieteilijä (0,7); finaali (107,1) – loppukilpailu (46,9); rasismi (53,2) – rotusyrjintä (0,7); budjetti (58,1) – talousarvio (9,1); fani (55,3) – ihailija (8,4); tsunami (46,2) – hyökyaalto (4,9); demokratia (38,5) – kansanvalta (6,3); tabletti (32,2) – taulutietokone (5,6); remontti (43,4) – korjaus (18,9); projekti (102,9) – hanke (80,5); abiturientti (20,3) – ylioppilaskokelas (0,7); konkurssi (23,1) – vararikko (6,3); resepti (14) – ruokaohje (1,4); virus (16,1) – haittaohjelma (4,9); geeni (9,8) – perintötekijä (2,1); luksus (4,2) – ylellisyys (1,4); osteoporoosi (3,5) – luukato (0,7); mikrobi (4,2) – pieneliö (2,1); delta (2,8) – suistomaa (0,7); skeittaus (4,9) – rullalautailu (3,5); zeppeliini (1,4) – ilmalaiva (0,7)
3. Vierassana ja omaperäinen sana ovat yhtä yleisiä	gallup (6,3) – mielipidetiedustelu (6,3); printteri (2,1) – tulostin (2,1); ekosysteemi (1,4) – eliöyhteisö (1,4); trendikkyys (0,7) – muodikkuus (0,7)

4 Taulukoissa 1 ja 2 sanaparit on järjestetty yleisemmän ja harvinaisemman sanan frekvenssien erotuksen perusteella. Kohtien 1 ja 2 luettelojen alussa vierassanojen ja omaperäisten vastinesanojen frekvenssien erot ovat suurimmillaan ja tasoittuvat luettelojen loppua kohti.

Pääasiassa yleiskieltä sisältäneen vertailuaineiston 21 sanaparissa omaperäinen sana oli yleisempi kuin vierassana ja 29 sanaparissa vierassana oli yleisempi kuin omaperäinen sana (ks. taulukkoa 2).

TAULUKKO 2
Sanaparit ja sanojen objektiivinen frekvenssi vertailuaineistossa (Yleis- ja selkokieli 2021)

Sanaparin tyyppi	Sanaparit ja sanojen frekvenssit miljoonaa sanetta kohti
1. Omaperäinen sana on yleisempi kuin vierassana	yritys (635) – firma (22,7); aihe (359,3) – teema (69,1); maakunta (285,3) – provinssi (9,4); laitos (195,7) – instituutio (8,7); hanke (238,9) – projekti (58,4); ajatus (220,9) – idea (73,3); eduskunta (197,6) –parlamentti (58,4); kaveri (96,9) – frendi (0,6); matkailija (73,1) – turisti (36,8); viestintä (38,7) – kommunikaatio (4,1); liiketoiminta (42,3) –bisnes (17); näytelmä (35,6) – draama (18,9); yhtye (38,1) – bändi (30,3); pelle (4,7) – klovni (0,6); rahoittaja (12,1) – sponsori (8); tilauslento (3,1) – charterlento (0,1); pikajuoksija (6,1) – sprintteri (3,4); mielikuvitus (6,6) – fantasia (4,5); ilmalaiva (0,6) – zeppeliini (0,1); tulostin (0,8) – printteri (0,4); muodikkuus (0,2) – trendikkyys (0,1)
2. Vierassana on yleisempi kuin omaperäinen sana	media (270,2) – tiedotusväline (14,2); virus (151,1) – haitta-ohjelma (4,8); finaali (145,2) – loppukilpailu (8,5); fani (51,5) – ihailija (0,8); konkurssi (44,8) – vararikko (0,7); demokratia (44,8) – kansanvalta (1,5); bussi (71,3) – linja-auto (30); rasismi (29,5) – rotusyrjintä (0,3); budjetti (63,8) – talousarvio (36,8); meteorologi (18,2) – ilmatieteilijä (0,3); remontti (69,7) – korjaus (51,9); abiturientti (18,6) – ylioppilaskokelas (1,6); brändi (16,7) – tuotemerkki (2,6); resepti (13,7) – ruokaohje (0,2); geeni (8,6) – perintötekijä (0,2); diabetes (8,5) – sokeritauti (0,2); tabletti (8,3) – taulutietokone (0); infektio (12,4) – tulehdus (5,5); ekosysteemi (7) – eliöyhteisö (0,2); mikrobi (6,9) – pieneliö (0,5); delta (5,8) – suistomaa (0); skeittaus (6,2) – rullalautailu (1); protesti (34,4) – mielenilmaus (29,8); reviiri (5,9) – elinalue (2); luksus (4,3) – ylellisyys (0,4); gallup (4,3) – mielipidetiedustelu (1,1); logo (43,7) – tunnus (42,1); tsunami (2) – hyökyaalto (0,9); osteoporoosi (0,3) – luukato (0)

Sanoja *frendi* ja *tiedotusväline* käsiteltiin tutkimuksessa muusta aineistosta poikkeavalla tavalla: *Frendin* perusmuodoksi on annotoitu selkokielen korpuksissa monikkomuoto *frendit*, mutta tässä tutkimuksessa perusmuotona käytettiin yksikkömuotoa. Korpuksissa sana kuitenkin esiintyy vain monikossa ja liittyy *Toisenlaiset frendit* -televisiosarjaan (Salmi, 2010–2014), jonka päähenkilöinä on kehitysvammaisia ihmisiä. *Tiedotusvälineen* perusmuotona oli korpuksissa yksikkömuoto, mutta sana esitettiin kieliasiantuntijoiden kyselyssä monikkomuodossa *tiedotusvälineet. Media* voidaan tulkita kollektiivikäsitteeksi, jonka merkityksenä on *viestimet, tiedotusvälineet* (Mantila, 1996). Selkokielen korpuksissa *tiedotusväline* esiintyy yhtä poikkeusta lukuun ottamatta monikkomuodossa.

2.3 Verkkolomakekysely kieliasiantuntijoille

Tutkimuksessa tehtiin verkkolomakekysely suomea äidinkielenään puhuville kieliasiantuntijoille. Kysely toteutettiin huhti- ja toukokuussa 2022 Microsoft Forms -verkkolomakkeella. Verkkolomakekyselyssä vastaajat arvioivat tutkittavien sanojen kohtaamistaajuutta. Kysymyksissä sanat esitettiin aakkosjärjestyksessä ja vastaajia pyydettiin arvioimaan, kuinka usein he kohtaavat eli lukevat, kuulevat, puhuvat tai kirjoittavat niitä (vrt. Balota ym, 2001). Arviointiasteikon vaihtoehdot olivat *päivittäin, viikoittain, kuukausittain, vuosittain* ja *harvemmin kuin vuosittain tai ei koskaan.* Kyselyssä oli myös sanojen käyttöä eli puhumista ja kirjoittamista koskevia kysymyksiä, joissa samaa tarkoittavat sanat esitettiin sanapareina aakkosjärjestyksessä kunkin sanaparin vierassanan perusteella.

Vastaajia pyydettiin kertomaan, kumpaa sanaa he käyttävät yleensä itse. Vastausvaihtoehtona oli myös *En käytä kumpaakaan sanaa*.

Koska verkkolomakekyselyssä oli pelkkiä suljettuja kysymyksiä, siihen oli periaatteessa mahdollista vastata pohtimatta kysymyksiä lainkaan. Siksi kyselyyn lisättiin keksityt kontrollisanat *ostepriaali* ja *pyrstäs*. Oletuksena oli, että vastaajat eivät olleet kohdanneet tai käyttäneet niitä. Odotuksenmukaisina pidettiin vastauksia, joissa vastaajat arvioivat kohdanneensa kontrollisanoja *harvemmin kuin vuosittain tai ei koskaan* ja kertoivat, etteivät käytä kumpaakaan sanaa. Kyselyn alussa osallistujille kerrottiin, että mukana saattoi olla keksittyjä sanoja.

Verkkolomakekyselyyn haettiin vastaajia jakamalla kyselylinkkiä Selkokeskuksen ja tutkijan omien verkostojen kautta. Taustatietoina kysyttiin äidinkieltä, ikää, koulutusta ja asiantuntijuuden tyyppiä (*kieliasiantuntija, viestinnän asiantuntija, muu kuin kielen tai viestinnän asiantuntija*). Lisäksi kysyttiin, onko vastaaja selkokielen asiantuntija. Vastaajien katsottiin kuuluvan kyselyn kohderyhmään, jos he kertoivat olevansa äidinkieleltään suomea puhuvia kielen tai viestinnän asiantuntijoita tai selkokielen asiantuntijoita. Selkokielen asiantuntijoita työskentelee muillakin kuin kielen ja viestinnän aloilla.

Verkkolomakekyselyyn vastasi 26 henkilöä. Yhden vastaukset poistettiin aineistosta, koska hän kertoi äidinkielensä olevan muu kuin suomi. Tutkimusaineistoon jäi vastaukset 25 suomea äidinkielenään puhuvalta henkilöltä (ks. taulukkoa 3).

TAULUKKO 3
Kieliasiantuntijoiden (N = 25) demografiset tiedot

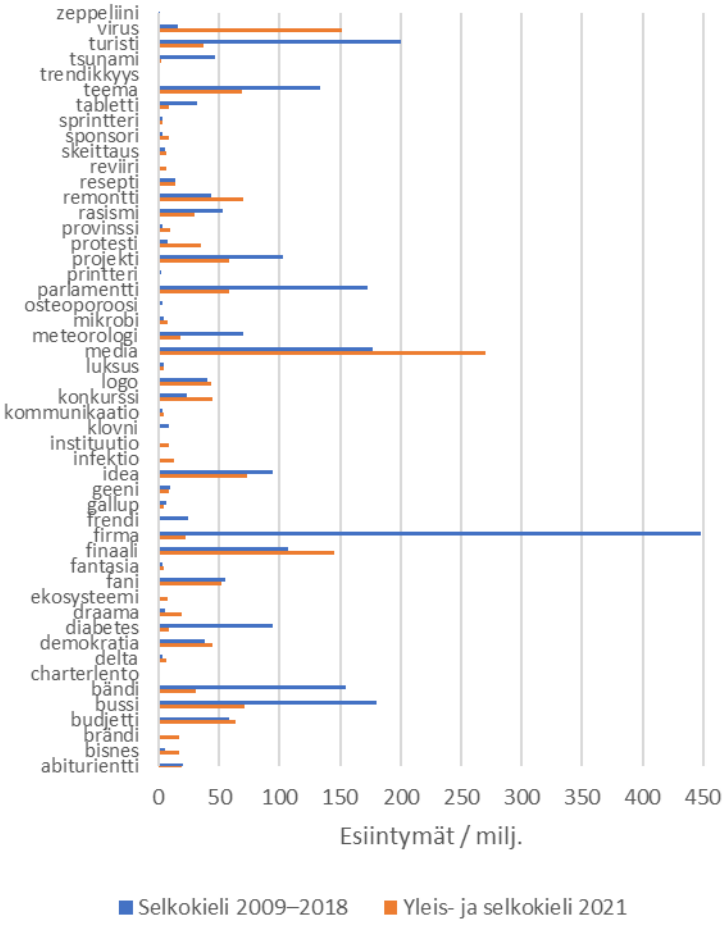
Demografinen kategoria	**Vastaajien määrä**
Sukupuoli	
Nainen	23
Mies	2
Ikäryhmä	
20–29-vuotiaat	4
30–39-vuotiaat	3
40–49-vuotiaat	10
50–59-vuotiaat	2
60–69-vuotiaat	5
yli 70-vuotiaat	1
Koulutus	
Ylemmän asteen yliopistotutkinto	21
Alemman asteen yliopistotutkinto	1
Ylempi ammattikorkeakoulututkinto	1
Ammattikorkeakoulututkinto	2
Asiantuntijuus	
Kieliasiantuntija (joista selkokielen asiantuntijoita)	16 (10)
Viestinnän asiantuntija (joista selkokielen asiantuntijoita)	6 (4)
Muu asiantuntija (joista selkokielen asiantuntijoita)	3 (3)

3 VIERASSANOJEN JA OMAPERÄISTEN SANOJEN OBJEKTIIVINEN FREKVENSSI, KOHTAAMISTAAJUUS JA KÄYTTÖ

Sanojen objektiivista frekvenssiä tutkittiin vertailemalla esiintymistaajuutta selkokielen korpuksissa ja vertailuaineistossa (ks. lukua 3.1). Subjektiivista frekvenssiä kuvattiin kieliasiantuntijoiden kohtaamistaajuusarvioiden mediaanilla (ks. lukua 3.2). Kohtaamistaajuusarvioiden mediaaneja tarkasteltiin suhteessa objektiivisiin frekvensseihin ja selvitettiin, ovatko ne samansuuntaisia (ks. lukua 3.3). Tutkittavaksi otettiin etenkin kohtaamistaajuudeltaan vastinettaan yleisemmät sanat ja objektiiviselta frekvenssiltään harvinaiset sanat, joita kohdattiin päivittäin, viikoittain, kuukausittain, vuosittain ja harvemmin kuin vuosittain tai ei koskaan. Sanojen käyttöä tutkittiin selvittämällä, kumpaa sanaparin sanaa kieliasiantuntijat käyttivät yleensä itse (ks. lukua 3.4). Lisäksi selvitettiin, miten sanojen pituus liittyy niiden käyttöön (ks. lukua 3.4). Lopuksi sanojen merkkimäärät ja frekvenssit taulukoitiin ja selvitettiin, miten pituus liittyy objektiiviseen ja subjektiiviseen frekvenssiin (ks. lukua 3.5).

3.1 Vierassanojen ja omaperäisten sanojen objektiivinen frekvenssi

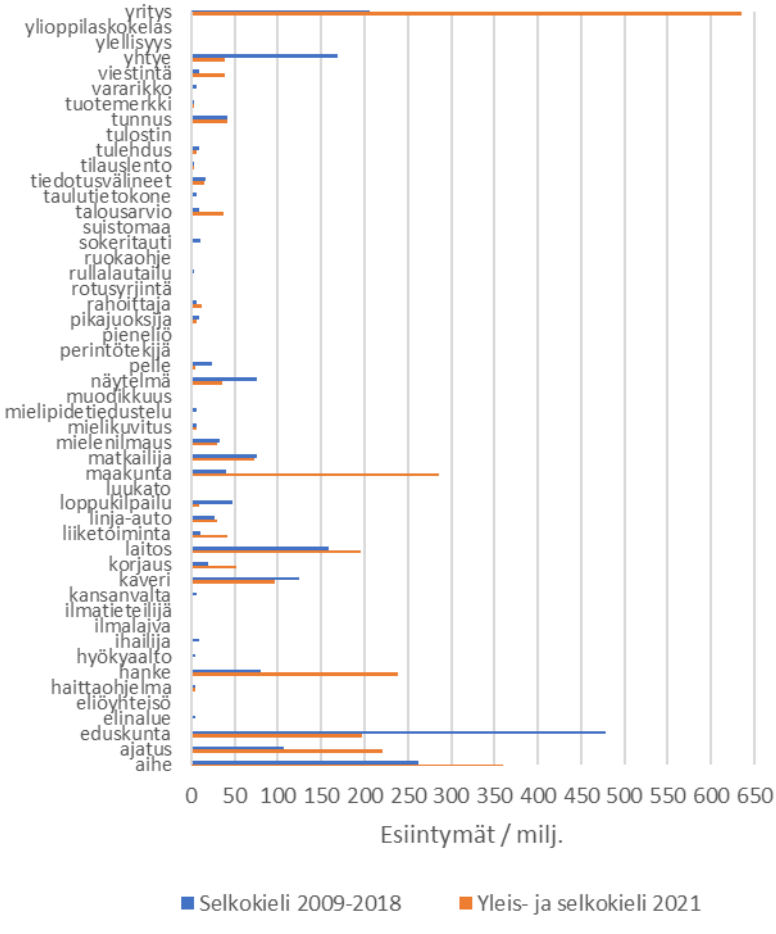
Sanojen objektiivinen frekvenssi oli selkokielen korpuksissa 0,7–478,2 esiintymää miljoonaa sanetta kohti ja mediaani oli 8,4, eli tutkimus painottui harvinaisempiin sanoihin. Pääasiassa yleiskieltä sisältäneessä vertailuaineistossa frekvenssi vaihteli välillä 0–635 ja mediaani oli 8,5, eli sanojen yleisyydessä ei ollut mediaanien perusteella juurikaan eroa. Yksittäisten sanojen frekvensseissä sen sijaan oli tutkimus- ja vertailuaineistossa merkittäviäkin eroja (ks. kuviota 1).



KUVIO 1
Vierassanojen objektiivinen frekvenssi tutkimusaineistossa (Selkokieli 2009–2018) ja vertailuaineistossa (Yleis- ja selkokieli 2021)

Selkokieliaineistossa vierassanojen frekvenssi vaihteli välillä 0,7–448,1 ja mediaani oli 11,9, kun vaihteluväli oli vertailuaineistossa 0,1–270,2 ja mediaani oli 13,5. Kun aineistoja vertailtiin keskenään, selkokieliaineistossa selvästi yleisempiä vierassanoja olivat *bussi, bändi, diabetes, firma, frendi, idea, meteorologi, parlamentti, projekti, rasismi, tabletti, teema, tsunami* ja *turisti* (ks. kuviota 1). Vertailuaineistossa selvästi yleisempiä vierassanoja olivat *bisnes, brändi, draama, finaali, infektio, konkurssi, media, protesti, remontti* ja *virus* (ks. kuviota 1).

Omaperäiset sanat olivat selkokieliaineistossa hieman harvinaisempia kuin vierassanat, sillä omaperäisten sanojen frekvenssi vaihteli välillä 0,7–478,2 ja mediaani oli 7,5. Vertailuaineistossa omaperäisten sanojen frekvenssi vaihteli välillä 0–635 ja mediaani oli 4,75, eli omaperäiset sanat olivat tässäkin aineistossa vierassanoja harvinaisempia (ks. kuviota 2).



KUVIO 2

Omaperäisten sanojen objektiivinen frekvenssi tutkimusaineistossa (Selkokieli 2009–2018) ja vertailuaineistossa (Yleis- ja selkokieli 2021)

Kun aineistoja vertailtiin keskenään, selkokieliaineistossa selvästi yleisempiä omaperäisiä sanoja olivat *eduskunta, kaveri, loppukilpailu, näytelmä, pelle* ja *yhtye* (ks. kuviota 2). Vertailuaineistossa selvästi yleisempiä olivat *aihe, ajatus, hanke, korjaus, laitos, liiketoiminta, maakunta, talousarvio, viestintä* ja *yritys* (ks. kuviota 2).

Edellä esitetyissä yleisemmissä sanoissa oli mukana abstrakteja sanoja, jotka esiintyivät monenlaisissa konteksteissa. Tällaisia olivat selkokieliaineistossa yleisemmät *teema* (selkokielessä 133,7; vertailuaineistossa 69,1; vastedes frekvenssit esitetään järjestyksessä selkokieli; vertailuaineisto) ja *idea* (93,8; 73,3) sekä vertailuaineistossa yleisemmät *aihe* (261,9; 359,3) ja *ajatus* (106,4; 220,9). Monet vertailuaineiston sanoista liittyivät aihepiiriltään liiketoimintaan, talouteen ja viestintään. Vertailuaineistossa yleisempiä olivat *bisnes* (5,6; 17), *liiketoiminta* (9,8; 42,3), *talousarvio* (9,1; 36,8), *brändi* (1,4; 16,7), *media* (176,5; 270,2) ja *viestintä* (9,1; 38,7). Mielenkiintoista on, että selkokielessä *yritystä* (205,2; 635) kutsuttiin yleensä *firmaksi* (448,1; 22,7).[5] Selkokielessä myös suosittiin sanaa *projekti* (102,9; 58,4), kun vertailuaineistossa suosittiin sanaa *hanke* (80,5; 238,9).

Selkokieliaineistossa korostuivat kulttuuriin, urheiluun ja säähän liittyvät sanat. Niitä olivat *bändi* (154,7; 30,3), *yhtye* (168,7; 38,1) ja *näytelmä* (76,3; 35,6). Vertailuaineistossa *draama*-sanalla (4,9; 18,9) kuvattiin näytelmien lisäksi tunteita kuohuttavia tapahtumia. Selkokieliaineistossa *frendi* (24,5; 0,6) viittasi televisiosarjaan *Toisenlaiset frendit* (ks. lukua 2.2.). *Pelle* (23,1; 4,7) viittasi muun muassa sirkuksesta tuttuihin hahmoihin ja esiintyi erisnimissä. Urheiluun ja kulttuuriin liittyvää *finaali*-sanaa suosittiin vertailuaineistossa, mutta sana oli yleinen selkokielessäkin (107,1; 145,2). *Loppukilpailu* oli kuitenkin yleisempi selkokielessä kuin vertailuaineistossa (46,9; 8,5). Säähän liittyvistä sanoista *meteorologi* (70; 18,2) esiintyi selkokielessä etenkin määritteenä (esim. *meteorologi Kerttu Kotakorpi*).

Selkokieliaineistossa yleisempiä olivat arjenläheiset sanat *bussi* (180; 71,3) ja *kaveri*[6] (124,6; 96,9) sekä syrjintää kuvaava *rasismi* (53,2; 29,5). Mielenkiintoista on, että selkokieliaineistossa *matkailijaa* (75,6; 73,1) kutsuttiin yleensä *turistiksi* (200,3; 36,8). Vertailuaineistossa yleisempiä olivat *remontti* (43,4; 69,7) ja *korjaus* (18,9; 51,9). Jälkimmäistä käytettiin vertailuaineistossa myös tietona teksteihin tehdyistä muutoksista.

Jotkut sanoista liittyivät ajankohtaisiin tapahtumiin ja tilanteisiin. Selkokieliaineistossa *tsunami* (46,2; 2) esiintyi tyypillisesti lauseissa, joissa kerrottiin Japanissa vuonna 2011 ja Indonesiassa vuonna 2018 tapahtuneista luonnonkatastrofeista. Vertailuaineistossa korostuivat vuonna 2019 alkaneeseen koronapandemiaan ja sen yhteiskunnallisiin vaikutuksiin liittyvät sanat *infektio* (1,4; 12,4), *virus* (16,1; 151,1), *konkurssi* (23,1; 44,8), *protesti* (7; 34,4), *laitos* (158,2; 195,7) ja *maakunta* (39,9; 285,3) (ks. myös lukuja 3.2 ja 3.3).

Selkokieliaineistossa muutamiin sanoihin liittyi valistavia sävyjä. Etenkin sanojen *eduskunta* (478,2; 197,6) ja *parlamentti* (173; 58,4) käyttöä lisäsivät selitykset, kuten *Eduskuntatalossa tekee työtä **Suomen parlamentti eli eduskunta*** (Yle Selkouutiset 2017-11-18). *Tabletilla* (32,2; 8,3) tarkoitettiin selkokieliaineistossa yleensä teknistä laitetta, jonka ulkonäköä ja käyttötarkoitusta kuvailtiin esimerkiksi virkkeellä *Lukulaitteet ja tabletit ovat mukana kannettavia litteitä tietokoneita* (Leija 1/2015).[7] Valistavia *sävyjä liittyi selkokielessä etenkin diabetes*-sanaan (94,5; 8,5). Lukijaa saatettiin myös puhutella suoraan, kuten *Jos **sinulla** on diabetes, **voit** hakea diabeteskurssille* (Leija 2/2013). Suora puhuttelu on selkokielen kriteerien mukaista, sillä se suuntaa tekstin lukijalle ja kertoo, miten hänen on mahdollista toimia (Selkokeskus, 2022, kriteerit 12 ja 13).

3.2 Vierassanojen ja omaperäisten sanojen kohtaamistaajuus

Verkkolomakekyselyssä kieliasiantuntijat arvioivat sanojen kohtaamistaajuutta asteikolla *päivittäin*, *viikoittain*, *kuukausittain*, *vuosittain* ja *harvemmin kuin vuosittain tai ei koskaan*. Sanat jakautuivat viiteen ryhmään kohtaamistaajuusarvioiden mediaanin perusteella. Suurinta osaa

5 *Firma*-sanaa suosittiin Yle Suomenkielisen uutisarkiston selkouutiset 2011–2018 -korpuksessa, kun taas *yritys*-sanaa suosittiin Leija ja Selkosanomat/Selkouutiset -korpuksissa.
6 *Kaveri* esiintyi usein Leija-lehden kirjeenvaihtoilmoituksissa.
7 *Tabletti*-sanalla viitattiin joskus lääkkeisiin.

sanoista kohdattiin kuukausittain (37 sanaa) ja viikoittain (36 sanaa). Loppuja sanoista kohdattiin vuosittain (15 sanaa), päivittäin (8 sanaa) ja harvemmin kuin vuosittain tai ei koskaan (4 sanaa). Kaikki vastaajat arvioivat kohtaavansa myös kontrollisanoja harvemmin kuin vuosittain tai ei koskaan (ks. taulukkoa 4).

TAULUKKO 4

Kieliasiantuntijoiden (N = 25) kohtaamistaajuusarviot 50 vierassanalle ja 50 omaperäiselle sanalle. Vastinettaan yleisemmät sanat on merkitty tähdellä (*).

Kohtaamistaajuus-arvioiden mediaani	**Sanaparit, vierassanat ja omaperäiset sanat**	**Sanojen määrä**
Päivittäin	idea–ajatus	2
	media*, virus*	2
	aihe*, kaveri*, viestintä*, yritys*	4
Viikoittain	bisnes–liiketoiminta, brändi–tuotemerkki, bussi–linja-auto, infektio–tulehdus, logo–tunnus, parlamentti–eduskunta, projekti–hanke, remontti–korjaus, turisti–matkailija	18
	budjetti*, bändi*, demokratia*, draama*, fani*, firma, kommunikaatio, meteorologi*, rasismi*, resepti*, tabletti*, teema	12
	laitos*, maakunta*, mielikuvitus*, rahoittaja*, tiedotusvälineet, tulostin*	6
Kuukausittain	diabetes–sokeritauti, finaali–loppukilpailu, gallup–mielipidetiedustelu, geeni–perintötekijä, konkurssi–vararikko, luksus–ylellisyys, mikrobi–pieneliö, protesti–mielenilmaus, reviiri–elinalue, trendikkyys–muodikkuus	20
	ekosysteemi*, fantasia, frendi, instituutio, osteoporoosi*, printteri, skeittaus*, sponsori	8
	haittaohjelma, ihailija, näytelmä, pelle*, pikajuoksija*, rotusyrjintä, ruokaohje, talousarvio, yhtye	9
Vuosittain	abiturientti–ylioppilaskokelas, tsunami–hyökyaalto	4
	delta*, klovni, provinssi, sprintteri	4
	eliöyhteisö, ilmalaiva*, ilmatieteilijä, kansanvalta, luukato, rullalautailu, tilauslento*	7
Harvemmin kuin vuosittain tai ei koskaan	ostepriaali–pyrstäs (kontrollisanat)	-
	charterlento, zeppeliini	2
	suistomaa, taulutietokone	2

Tutkimusaineiston 22 sanaparissa vierassanan ja omaperäisen vastineen kohtaamistaajuudessa ei ollut eroa, sillä kumpaakin kohdattiin yhtä usein. Tällaisia olivat esimerkiksi *idea – ajatus* (päivittäin)*, bisnes – liiketoiminta* (viikoittain)*, diabetes – sokeritauti* (kuukausittain) ja *abiturientti – ylioppilaskokelas* (vuosittain) (ks. taulukkoa 4).

Viidessätoista sanaparissa vierassanaa kohdattiin useammin kuin omaperäistä sanaa. Näistä etenkin vierassanat *virus* (päivittäin), *demokratia* (viikoittain), *meteorologi* (viikoittain) ja *tabletti* (viikoittain) olivat kohtaamistaajuudeltaan selvästi yleisempiä kuin omaperäiset sanat *haittaohjelma* (kuukausittain), *kansanvalta* (vuosittain), *ilmatieteilijä* (vuosittain) ja *taulutietokone* (harvemmin kuin vuosittain tai ei koskaan). Kolmessatoista sanaparissa omaperäistä sanaa kohdattiin useammin kuin vierassanaa. Näistä etenkin *kaveri* (päivittäin) ja *maakunta* (viikoittain) olivat kohtaamistaajuudeltaan selvästi yleisempiä kuin *frendi* (kuukausittain) ja *provinssi* (vuosittain).

Sanojen *virus* ja *maakunta* kohtaamistaajuuteen todennäköisesti vaikutti ajankohtainen tilanne, sillä *virus* (päivittäin; selkokielessä 16,1; vertailuaineistossa 151,1; vastedes frekvenssit esitetään järjestyksessä subjektiivinen frekvenssi; objektiivinen frekvenssi selkokielessä; objektiivinen frekvenssi vertailuaineistossa) on vuoden 2021 yleis- ja selkokielisessä korpuksessa selvästi yleisempi kuin vuosien 2009–2018 selkokielisissä korpuksissa. Vuoden 2021 aineiston esiintymiskontekstit osoittavat, että *virus*-sanan objektiivisten frekvenssien nousu liittyy vuonna 2019 alkaneeseen koronapandemiaan, eikä tietokoneen *haittaohjelmaan* (kuukausittain; 4,9; 4,8). Koronauutisointi vaikuttaa lisänneen myös *maakunta*-sanan käyttöä (viikoittain; 39,9; 285,3), kun on kerrottu tartuntatilanteesta eri maakunnissa.

Maakunnan ja *provinssin* (vuosittain; 3,5; 9,4) kohtaamistaajuuden eroon liittyy todennäköisesti kulttuurisidonnaisuutta, sillä Suomessa maakuntia ei kutsuta provinsseiksi.[8] Eräänlaista kulttuurisidonnaisuutta voidaan nähdä myös sanaparissa *kaveri* (päivittäin; 124,6; 96,9) ja *frendi* (kuukausittain; 24,5; 0,6), sillä *frendi* on slangisana (ks. esim. Forsberg, 2021). *Kaveri* on yleiskielisempi ja objektiiviselta frekvenssiltään *frendiä* yleisempi, mikä selittää sen korkeaa kohtaamistaajuutta.

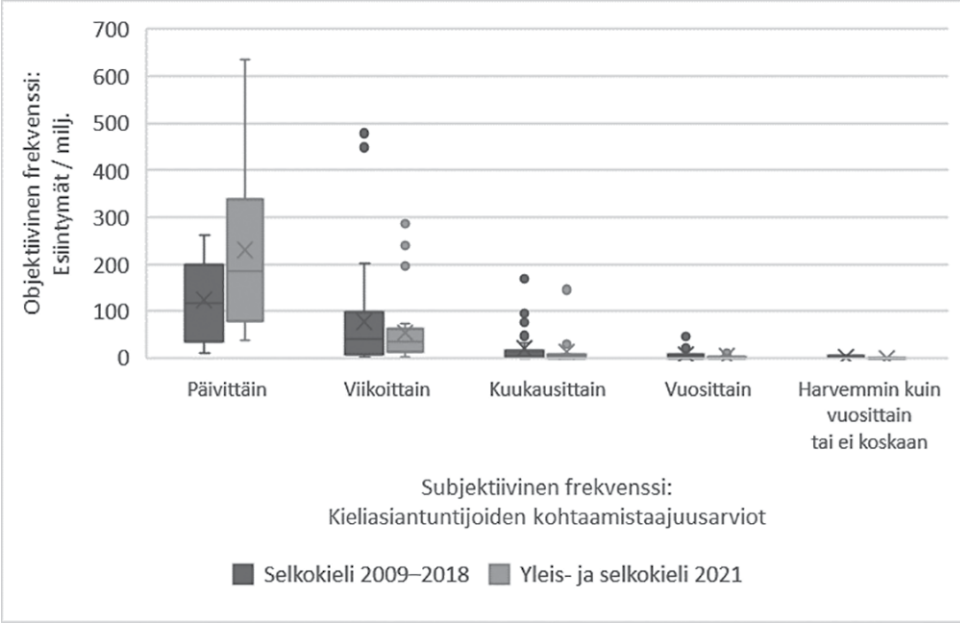
Joissakin tapauksissa kohtaamistaajuuden eron syynä on vierassanan yleistyminen tarkoitteen nimeksi. *Meteorologi* (viikoittain; 70; 18,2) on subjektiiviselta ja objektiiviselta frekvenssiltään yleisempi kuin *ilmatieteilijä* (vuosittain; 0,7; 0,3). Sama pätee myös sanapareihin *demokratia* (viikoittain; 38,5; 44,8) ja *kansanvalta* (vuosittain; 6,3; 1,5) sekä *tabletti* (viikoittain; 32,2; 8,3) ja *taulutietokone* (harvemmin kuin vuosittain tai ei koskaan; 5,6; 0).

Yleisesti voidaan sanoa, että kohtaamistaajuudeltaan yleisimpiä olivat objektiiviselta frekvenssiltään yleiset ja melko lyhyet sanat, jotka esiintyivät useissa eri konteksteissa (*ajatus, idea, aihe* ja *kaveri*; frekvenssien vaihteluväli on selkokielessä 93,8–261,9 ja vertailuaineistossa 73,3–359,3). Mukana oli myös sanoja, jotka mahdollisesti liittyivät kieliasiantuntijoiden työhön (*viestintä, yritys* ja *media*) ja koronapandemiaan (*virus*). Kohtaamistaajuudeltaan harvinaisimmissa sanoissa oli spesifeissä konteksteissa käytettyjä melko pitkiä vierassanoja, joiden kirjoitus- ja äänneasu eivät vastaa toisiaan suomessa (*charterlento*, *zeppeliini*). Kohtaamistaajuudeltaan harvinaisimmissa sanoissa oli myös omaperäinen yhdyssana, jota käytetään vain rajatuissa konteksteissa (*suistomaa*), ja kolmiosainen omaperäinen yhdyssana, jonka tarkoitteen nimeksi on yleistynyt vierassana (***taulutietokone*** > *tabletti*). Kohtaamistaajuudeltaan harvinaisimmat sanat olivat vuosien 2009–2018 selkokieliaineistossa objektiiviselta frekvenssiltään harvinaisia (0,7–5,6). Vuoden 2021 yleis- ja selkokieliaineistossa ne olivat erittäin harvinaisia, tai esiintymiä ei ollut lainkaan (0–0,1).

3.3 Vierassanojen ja omaperäisten sanojen kohtaamistaajuuden suhde objektiiviseen frekvenssiin

Kohtaamistaajuusarvioiden mediaaneja tarkasteltiin suhteessa sanojen objektiivisiin frekvenssiehin. Tarkastelu osoitti, että kieliasiantuntijat arvioivat objektiivisilta frekvensseiltään yleisimmät sanat kohtaamistaajuudeltaan yleisiksi (ks. kuviota 3).

8 Vastaavanlaista kulttuurisidonnaisuutta on sanoissa *eduskunta* (viikoittain; 478,2; 197,6) ja *parlamentti* (viikoittain; 173; 58,4), sillä Suomen kansanedustuslaitosta kutsutaan *eduskunnaksi*.



KUVIO 3

Kieliasiantuntijoiden (N = 25) kohtaamistaajuusarvioiden suhde sanojen (N = 100) objektiiviseen frekvenssiin eli korpusesiintymiin miljoonaa sanetta kohti.

Objektiivisten frekvenssien mediaanit laskivat kohtaamistaajuuden harvenemisen myötä seuraavasti: *päivittäin* (selkokielessä 115,5; vertailuaineistossa 186), *viikoittain* (selkokielessä 40,25; vertailuaineistossa 33,55), *kuukausittain* (selkokielessä 4,9; vertailuaineistossa 4,7), *vuosittain* (selkokielessä 3,5; vertailuaineistossa 1,5) ja *harvemmin kuin vuosittain tai ei koskaan* (selkokielessä 1,4; vertailuaineistossa 0,05). Tulos osoittaa, että objektiiviselta frekvenssiltään yleisimpiä sanoja kohdattiin keskimäärin useammin.

Objektiivisten frekvenssien minimi- ja maksimiarvot laskivat kohtaamistaajuusarvioiden harvenemisen myötä, mutta lasku oli selvempää maksimiarvoissa. Objektiivisten frekvenssien vaihteluvälit olivat suurimmat päivittäin kohdatuissa sanoissa (selkokielessä 9,1–261,9; vertailuaineistossa 38,7–635) ja viikoittain kohdatuissa sanoissa (selkokielessä 1,4–478,2[9]; vertailuaineistossa 0,8–285,3). Vaihteluvälit olivat selvästi pienemmät kuukausittain kohdatuissa sanoissa (selkokielessä 0,7–168,7; vertailuaineistossa 0,1–145,2) ja vuosittain kohdatuissa sanoissa (selkokielessä 0,7–46,2; vertailuaineistossa 0–18,6). Harvemmin kuin vuosittain tai ei koskaan kohdatuissa sanoissa vaihteluvälit olivat lähes olemattomat (selkokielessä 0,7–5,6; vertailuaineistossa 0–0,1). Maksimiarvot ja mediaanit osoittivat, että objektiiviselta frekvenssiltään yleiset sanat olivat tyypillisesti kohtaamistaajuudeltaan yleisiä. Minimiarvot osoittivat, että objektiiviselta frekvenssiltään harvinaiset sanat saattoivat olla kohtaamistaajuudeltaan lähes millaisia tahansa (ks. taulukkoa 5).

9 Maksimiarvo on tässä korkeampi kuin päivittäin kohdatuissa sanoissa, koska mukana ovat poikkeavat arvot *eduskunta* (viikoittain; 478,2) ja *firma* (viikoittain; 448,1).

TAULUKKO 5
Objektiivisten frekvenssien minimiarvot kohtaamistaajuusarvioittain ja sanoittain

Kohtaamistaajuus	Objektiiviselta frekvenssiltään harvinaiset sanat, Selkokieli 2009–2018	Objektiiviselta frekvenssiltään harvinaiset sanat, Yleis- ja selkokieli 2021
Päivittäin	viestintä (9,1)	viestintä (38,7)
Viikoittain	brändi (1,4) infektio (1,4)	tulostin (0,8)
Kuukausittain	instituutio (0,7) muodikkuus (0,7) rotusyrjintä (0,7) trendikkyys (0,7)	trendikkyys (0,1)
Vuosittain	ilmalaiva (0,7) ilmatieteilijä (0,7) luukato (0,7) ylioppilaskokelas (0,7)	luukato (0)
Harvemmin kuin vuosittain tai ei koskaan	suistomaa (0,7)	suistomaa (0) taulutietokone (0)

Minimiarvojen ainoa päivittäin kohdattu sana oli objektiiviselta frekvenssiltään muita minimiarvojen sanoja selvästi yleisempi *viestintä* (9,1; 38,7), joka todennäköisesti liittyi kieliasiantuntijoiden työhön. Työhön liittyviä olivat mahdollisesti myös *brändi* (viikoittain; 1,4; 16,7) ja *tulostin* (viikoittain; 2,1; 0,8). On siis mahdollista, että osallistujien tausta vaikutti kyselyvastauksiin.

Objektiiviselta frekvenssiltään harvinaisimmista sanoista koronapandemiaan liittyi *infektio* (viikoittain; 1,4; 12,4). Myös *instituutio* (kuukausittain; 0,7; 8,7) on osittain yhdistettävissä koronaan, sillä sitä on käytetty vuoden 2021 yleis- ja selkokieliaineistossa politiikkaan ja koronaan liittyvissä konteksteissa. Selvemmin koronaan kuitenkin liittyy *instituution* vastinesana *laitos* (viikoittain; 158,2; 195,7), joka esiintyy vertailukorpuksessa usein *Terveyden ja hyvinvoinnin laitos* -nimen osana.

Minimiarvojen perusteella vuosittain ja harvemmin kuin vuosittain kohdattiin sanoja, jotka ovat objektiiviselta frekvenssiltään harvinaisia ja joita käytetään vain tietyissä konteksteissa. Näitä olivat *ilmalaiva* (vuosittain; 0,7; 0,6) ja *suistomaa* (harvemmin kuin vuosittain tai ei koskaan; 0,7; 0). Sana voi liittyä myös vuosittain toistuvaan tapahtumaan, kuten ylioppilaskirjoituksiin. *Ylioppilaskokelaan* (vuosittain; 0,7; 1,6) vastinetta *abiturientti* kohdattiin myös vuosittain, mutta se oli objektiiviselta frekvenssiltään yleisempi (20,3; 18,6).

Minimiarvoissa oli myös muutama muu omaperäinen sana, joiden tarkoitteen nimeksi on yleistynyt vierassana. Näitä olivat *rotusyrjintä* (kuukausittain; 0,7; 0,3), *luukato* (vuosittain; 0,7; 0) ja *ilmatieteilijä* (ks. lukua 3.2). *Rotusyrjinnän* vastinetta *rasismi* kohdattiin viikoittain, ja se oli myös objektiiviselta frekvenssiltään vastinettaan yleisempi (53,2; 29,5). *Luukadon* vastinetta *osteoporoosi* puolestaan kohdattiin kuukausittain, vaikka se oli objektiiviselta frekvenssiltään harvinainen (3,5; 0,3).

Minimiarvojen ainoa sanapari, jonka kohtaamistaajuudessa ja objektiivisissa frekvensseissä ei ollut selvää eroa olivat *muodikkuus* (kuukausittain; 0,7; 0,7) ja *trendikkyys* (kuukausittain; 0,7; 0,1). Kumpaakin kohdattiin kuukausittain, vaikka ne olivat objektiiviselta frekvenssiltään hyvin harvinaisia.

3.4 Vierassanojen ja omaperäisten sanojen käyttö

Kieliasiantuntijat arvioivat, kumpaa sanaparin sanaa he käyttivät yleensä itse. Vastausvaihtoehtona oli myös *en käytä kumpaakaan sanaa*. Vierassanaa käytettiin yleisemmin 33 sanaparissa ja omaperäistä sanaa 16 sanaparissa (ks. taulukkoa 6).

TAULUKKO 6

Kieliasiantuntijoiden (N = 25) vierassanojen ja omaperäisten sanojen käyttö

% vastaajista (N = 25) käyttää sanaa itse	**Sanaparit: Vierassanaa käytetään yleisemmin (% vastaajista)**	**Sanaparit: Omaperäistä sanaa käytetään yleisemmin (% vastaajista)**
100 %	demokratia (100), kansanvalta (0); diabetes (100), sokeritauti (0); rasismi (100), rotusyrjintä (0); tabletti (100), taulutietokone (0)	eduskunta (100), parlamentti (0); kaveri (100), frendi (0); maakunta (100), provinssi (0)
90–99 %	konkurssi (96), vararikko (0); osteoporoosi (96), luukato (0); budjetti (96), talousarvio (4); gallup (96), mielipidetiedustelu (4); meteorologi (96), ilmatieteilijä (4); remontti (96), korjaus (4); reviiri (96), elinalue (4); abiturientti (92), ylioppilaskokelas (0); bändi (92), yhtye (8); fani (92), ihailija (4); geeni (92), perintötekijä (4); resepti (92), ruokaohje (8)	mielikuvitus (92), fantasia (4)
80–89 %	mikrobi (88), pieneliö (0); brändi (84), tuotemerkki (16); bussi (84), linja-auto (16); logo (84), tunnus (16); media (84), tiedotusvälineet (16); turisti (84), matkailija (16); virus (84), haittaohjelma (12); skeittaus (80), rullalautailu (16); tsunami (80), hyökyaalto (20)	näytelmä (84), draama (12); tulehdus (84), infektio (16); yritys (84), firma (16); aihe (80), teema (20); tulostin (80), printteri (20)
70–79 %	ekosysteemi (76), eliöyhteisö (4); finaali (76), loppukilpailu (20); luksus (76), ylellisyys (12); protesti (76), mielenilmaus (16); projekti (72), hanke (28)	pelle (76), klovni (16); pikajuoksija (76), sprintteri (8); viestintä (76), kommunikaatio (20)
60–69 %	bisnes (64), liiketoiminta (32); idea (60), ajatus (40); sponsori (60), rahoittaja (40)	-
50–59 %	-	tilauslento (52), charterlento (16)
40–49 %	-	ilmalaiva (44), zeppeliini (20); muodikkuus (44), trendikkyys (40)
30–39 %	-	-
20–29 %	-	suistomaa (20), delta (16)

Muista sanapareista poiketen *instituutio* ja *laitos* -sanoja käytettiin yhtä yleisesti. 48 % kieliasiantuntijoista kertoi käyttävänsä jompaakumpaa sanaa yleensä itse.

Omaperäisyys ei ollut keskeinen piirre kieliasiantuntijoiden yleisemmin käyttämissä sanoissa, vaan tärkeämpi oli pituus. He siis suosivat sanavalinnoissaan merkkimäärältään lyhyempiä sanoja. Sanaparien yleisemmin käytettyjen sanojen (49 kpl ilman sanoja *instituutio* ja *laitos*) mediaanipituus oli 8 merkkiä, ja sanat olivat pituudeltaan 4–12 merkkiä. Sanaparien harvemmin tai ei lainkaan käytettyjen sanojen (49 kpl) mediaanipituus oli 10 merkkiä, ja sanat olivat pituudeltaan 5–19 merkkiä. Suomenkaltaisessa agglutinoivassa kielessä sanojen taivuttaminen, yhdistäminen ja johtaminen lisäävät sanan pituutta. Kieliasiantuntijat suosivatkin yhdistämättömiä vierassanoja kaksi- tai useampiosaisten omaperäisten yhdyssanojen sijaan (esim. ***bussi*** – *linja-auto,* ***gallup*** – *mielipidetiedustelu*) ja lyhyempiä vierassanoja, kun omaperäisenä vastineena oli yhdistämätön johdos (esim. *fani* – ***ihailija****, luksus* – ***ylellisyys***).

Sanojen käyttö ja kohtaamistaajuus olivat odotuksenmukaisesti samansuuntaisia, sillä kohtaaminen saattoi toteutua käyttönä. Kieliasiantuntijat käyttivät useammin kohtaamiaan sanoja yleensä itse, ja jos sanaparin sanoja kohdattiin yhtä usein, käyttöön valittiin yleensä lyhyempi sana. Käyttö ja kohtaamistaajuus menivät ristiin vain muutamissa sanapareissa. Kieliasiantuntijat käyttivät yleensä sanoja *näytelmä* ja *sponsori,* vaikka kohtasivat niitä harvemmin (kuukausittain) kuin vastinesanoja *draama* ja *rahoittaja* (viikoittain). He myös käyttivät yleisemmin sanaa *suistomaa*, vaikka sitä kohdattiin harvemmin (harvemmin kuin vuosittain tai ei koskaan) kuin vastinesanaa *delta* (vuosittain).

3.5 Objektiivisen frekvenssin ja kohtaamistaajuuden suhde sanojen pituuteen

Tutkimusaineiston sanojen objektiivista frekvenssiä ja kohtaamistaajuutta tarkasteltiin suhteessa sanojen merkkimääräiseen pituuteen. Aiemmasta tutkimuksesta tiedetään, että objektiiviselta frekvenssiltään yleisemmät sanat ovat tyypillisesti lyhyempiä ja harvinaisemmat sanat ovat pidempiä (Schmitt 2010, s. 64). Näin oli myös tässä tutkimuksessa (ks. taulukkoa 7).

TAULUKKO 7
Sanojen pituus ja objektiivinen frekvenssi

Sanapituus merkkeinä	Sanojen määrä	Frekvenssien mediaani: Selkokieli 2009–2018	Frekvenssien mediaani: Yleis- ja selkokieli 2021	Frekvenssien vaihteluväli: Selkokieli 2009–2018	Frekvenssien vaihteluväli: Yleis- ja selkokieli 2021
4	4	74,55	62,4	40,6–261,9	43,7–359,3
5	11	133,7	38,1	2,8–448,1	4,7–270,2
6	12	16,45	17,95	1,4–205,2	0,6–635
7	9	18,9	13,7	0,7–200,3	0–145,2
8	16	8,4	8,4	1,4–102,9	0,5–285,3
9	10	4,2	3,45	0,7–478,2	0–197,6
10	9	4,9	3,4	0,7–75,6	0,1–73,1
11	11	3,5	3,1	0,7–173	0,1–58,4
12	7	6,3	6,1	0,7–32,9	0,1–29,8
13	6	4,2	4,45	2,1–46,9	0,2–42,3
14	2	3,15	0,15	0,7–5,6	0–0,3
16	1	16,8	14,2	16,8	14,2
17	1	0,7	1,6	0,7	1,6
19	1	6,3	1,1	6,3	1,1

Tutkimusaineiston lyhyimmät eli neljän ja viiden merkin pituiset sanat olivat objektiivisen frekvenssin mediaanilla ilmaistuna yleisimpiä selkokielen korpuksissa ja vertailuaineistossa. Viisi merkkiä pitkissä sanoissa selkokieliaineiston mediaania nostivat sanat *firma* (448,1 esiintymää / milj.) sekä *bussi, bändi, media, teema* ja *yhtye,* joiden frekvenssit vaihtelivat välillä 133,7–180 esiintymää miljoonassa. Muuten frekvenssien mediaanit olivat tutkimus- ja vertailuaineistossa samantyyppisiä. Pitkistä sanoista yleisin oli kummassakin kielimuodossa *tiedotusvälineet,* jonka pituus oli monikkomuodon vuoksi 16 merkkiä.

Odotuksenmukaisesti myös kohtaamistaajuudeltaan yleisimmät sanat olivat merkkimäärältään lyhyimpiä. Päivittäin kohdattujen sanojen mediaanipituus oli 5,5 merkkiä, viikoittain kohdattujen 8 merkkiä, kuukausittain kohdattujen 9 merkkiä, vuosittain kohdattujen 10 merkkiä ja harvemmin kuin vuosittain tai ei koskaan kohdattujen 11 merkkiä. Yleisyys ja lyhyys olivat siis leksikaalisia piirteitä, jotka esiintyivät tutkittavissa sanoissa tyypillisesti yhdessä.

4 YHTEENVETO JA POHDINTA

Tässä artikkelissa on tarkasteltu selkoteksteistä poimittujen vierassanojen ja omaperäisten sanojen objektiivista ja subjektiivista frekvenssiä. Sanojen yleisyyttä on tarkasteltu eri näkökulmista ja selvitetty, ovatko omaperäiset sanat yleisempiä kuin vierassanat ja miten yleisyyttä voidaan arvioida, kun tehdään sanavalintoja selkokielessä. Lisäksi on tutkittu, ovatko objektiivinen ja subjektiivinen frekvenssi samansuuntaisia ja miten yleisyys liittyy sanojen pituuteen ja käyttöön.

Selkokielen korpuksista ja pääasiassa yleiskieltä sisältäneestä vertailukorpuksesta kerättyjen objektiivisten frekvenssien vertailusta selvisi, että omaperäiset sanat olivat kummassakin aineistossa harvinaisempia kuin vierassanat. Omaperäisiä sanoja käytettiin kuitenkin selkokielessä enemmän kuin vertailuaineistossa. Myös yksittäisten sanojen frekvensseissä oli eroja aineistojen välillä. Osa eroista selittyi korpusten sisällöillä: selkokieliaineistossa korostuivat kulttuuriin, urheiluun ja säähän liittyvät sanat, ja vertailuaineistossa korostuvat liiketoimintaan, talouteen ja viestintään liittyvät sanat.[10] Frekvenssien eroihin vaikuttivat ajankohtaiset tapahtumat ja tilanteet. Vaikutusta oli myös kielimuotojen eroilla: selkokielessä selitetään vieraiksi arvioituja sanoja synonyymeillä.

Kieliasiantuntijat kohtasivat suurinta osaa sanoista mediaanilla ilmaistuna kuukausittain ja viikoittain. On mahdollista, että mediaani vääristi tulosta kohti arviointiasteikon keskikohtaa tai vastaajat valitsivat keskimmäisen vaihtoehdon, jos heillä ei ollut selkeää mielikuvaa kohtaamistaajuudesta. Mediaani oli kuitenkin informatiivinen, koska vastaajat eivät yleensä olleet yksimielisiä. Arviot olivat kuitenkin samansuuntaisia. Kohtaamistaajuudeltaan yleisimpiä olivat melko lyhyet ja objektiiviselta frekvenssiltään yleiset sanat, jotka esiintyivät monissa eri konteksteissa. Kohtaamistaajuudeltaan harvinaisimpia olivat pitkähköt objektiiviselta frekvenssiltään harvinaiset sanat, joita käytettiin spesifeissä konteksteissa. Kohtaamistaajuudeltaan harvinaisimmissa sanoissa oli myös objektiiviselta frekvenssiltään harvinaisia vierassanoja, joiden äänne- ja kirjoitusasu eivät vastaa toisiaan suomen kielessä. Sanojen kohtaamistaajuuteen vaikuttivat todennäköisesti vastaajien tausta, kulttuuri- ja aikasidonnaiset tekijät sekä sanaston muuttuminen.

Sanojen kohtaamistaajuus ja objektiivinen frekvenssi olivat samansuuntaisia etenkin objektiiviselta frekvenssiltään yleisissä sanoissa. Objektiiviselta frekvenssiltään harvinaiset sanat puolestaan olivat kohtaamistaajuudeltaan monenlaisia. Objektiiviselta frekvenssiltään yleiset sanat olivat siis todennäköisemmin myös subjektiiviselta frekvenssiltään yleisiä, minkä vuoksi niitä voidaan käyttää varmemmin selkokielessä. Objektiiviselta frekvenssiltään harvinaisten sanojen käyttö selkokielessä vaatii kuitenkin tarkempaa pohdintaa. Pohdinnan apuna voitaisiin käyttää taajuussanastojen lisäksi korpuksia, joissa esitetään sanojen frekvenssejä eri kielimuodoissa ja tekstilajeissa.[11] Lisäksi voitaisiin pohtia subjektiivista frekvenssiä, eli kuinka usein sanoja kohdataan.

Lyhyys ja yleisyys olivat leksikaalisia piirteitä, jotka liittyivät tutkituissa sanoissa selvästi

10 Selkotekstien aihevalintoja on aiemmin säännelty selkokielen kriteereillä (Selkokeskus, 2018, kriteeri 6). Aihevalintoja kannattaisikin tutkia myös sääntelyn näkökulmasta.

11 Osa Kielipankin (Kielipankki, FIN-CLARIN & CSC, 2015–2021) aineistoista on julkisesti saatavilla.

yhteen. Kieliasiantuntijat käyttivät kohtaamistaajuudeltaan yleisempiä sanaparien sanoja yleensä myös itse. Jos he kohtasivat sanaparin sanoja yhtä usein, he käyttivät mieluummin lyhyempää sanaa. He siis suosivat omassa kielenkäytössään lyhyempiä yhdistämättömiä vierassanoja pidempien omaperäisten yhdyssanojen sijaan. Tutkimusaineistossa sanojen pituus ja yleisyys liittyivät muutenkin yhteen: Kohtaamistaajuudeltaan yleisimmät sanat olivat merkkimäärältään lyhyimpiä, ja sanat pitenivät kohtaamistaajuuden harvenemisen myötä. Lyhyet sanat olivat tyypillisesti myös objektiiviselta frekvenssiltään yleisiä.

Tulokset osoittavat, että sanaston selkokielisyyttä kannattaa tarkastella monesta eri näkökulmasta. Selko-ohjeiden mukaan selkoteksteissä ei käytetä vierassanoja, jos niille on yleinen, kotoisempi vastine. Ajatus on toistunut selko-ohjeissa selkokielen kehittämisen alkuajoista lähtien, ja taustalla on viestinnäntutkimuksen teoria, joka kirjoitettiin noin puoli vuosisataa sitten. Kieliympäristö on kuitenkin muuttunut, ja viimeaikaiset sanaston muutokset näkyvät tämän tutkimuksen aineistossakin. Useimmat tutkimusaineiston vierassanoista olivat objektiivisen frekvenssin, kohtaamistaajuuden ja käytön perusteella varsin yleisiä. Jatkotutkimuksissa olisi hyvä tehdä kyselyjä selkokielen käyttäjille ja selvittää, mitä sanoja he kohtaavat ja käyttävät yleensä itse. Tämä tutkimus tarjoaa jatkotutkimuksiin vertailuaineistoa.

AINEISTOLÄHTEET

Helsingin yliopisto (2017). *1990- ja 2000-luvun suomalaisia aikakaus- ja sanomalehtiä -korpus, versio 2* [tekstikorpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2017091901

Yleisradio (ei pvm.a). *Ylen suomenkielinen uutisarkisto 2019–2021, sekoitettu, Korp,* [tekstikorpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2022031703

Yleisradio (ei pvm.b). *Ylen suomenkielisen uutisarkiston selkouutiset 2011–2018*, *sekoitettu, Korp* [tekstikorpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2019121204

LÄHTEET

Baayen, R. H. (2001). *Word frequency distributions*. Kluwer.

Ballot, C., Mathey, S. & Robert, C. (2022). Age-related evaluations of imageability and subjective frequency for 1286 neutral and emotional French words: ratings by young, middle-aged, and older adults. *Behavior Research Methods 54*, 196–215. https://doi.org/10.3758/s13428-021-01621-6

Balota, D. A., Pilotti, M. & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*(4), 639–647. https://doi.org/10.3758/BF03200465

Chen, X. & Dong, Y. (2019). Evaluating objective and subjective frequency measures in L2 lexical processing. *Lingua*, *230*, 102738-https://doi.org/10.1016/j.lingua.2019.102738

Desrochers, A. & Thompson, G. L. (2009). Subjective frequency and imageability ratings for 3,600 French nouns. *Behavior Research Methods*, *41*(2), 546–557. https://doi.org/10.3758/BRM.41.2.546

Forsberg, U.-M. (2021). *Stadin slangin etymologinen sanakirja*. Gaudeamus.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281. https://doi.org/10.1037/0096-3445.113.2.256

Hansen-Schirra, S. & Maaß, C. (2020). *Easy language research: Text and user perspectives*. Frank & Timme.

Häkkinen, K. (1990). *Mistä sanat tulevat: Suomalaista etymologiaa*. Suomalaisen Kirjallisuuden Seura.

IFLA (2010). *Guidelines for easy-to-read materials*. International Federation of Library Associations and Institutions IFLA Professional Reports 120. Haettu 18.5.2022 osoitteesta www.ifla.org/files/hq/publications/professio-nal-report/120.pdf

Inclusion Europe (2009). *Make your information accessible! European standards to make information easy to read and to understand*. Haettu 18.5.2022 osoitteesta https://www.inclusion-europe.eu/wp content/uploads/2017/06/EN_Information_for_all.pdf

Kielipankki, FIN-CLARIN & CSC (2015–2021). *Kielipankki*. Haettu 1.8.2022 osoitteesta https://www.kielipankki.fi/

Kielitoimiston sanakirja (2021). Helsinki. Kotimaisten kielten keskuksen verkkojulkaisuja 35. URN:NBN:fi:kotus-201433. Päivitettävä julkaisu. Päivitetty 11.11.2021.

Kulkki-Nieminen, A. (2010). *Selkoistettu uutinen. Lingvistinen analyysi selkotekstin erityispiirteistä.* Tampere University Press.

Kotimaisten kielten keskus (ei pvm.). *Kirjoitetun suomen kielen sanojen taajuuksia*. Haettu 26.7.2022 osoitteesta https://kaino.kotus.fi/sanat/taajuuslista/parole

Larjavaara, M. (2007). *Pragmasemantiikka*. Suomalaisen Kirjallisuuden Seura.

Leskelä, L. (2019). *Selkokieli: Saavutettavan kielen opas*. Kehitysvammaliitto ry.

Leskelä, L. & Lindholm, C. (2012). Lukijalle. Teoksessa L. Leskelä & C. Lindholm (toim.) *Haavoittuva keskustelu: Keskustelunanalyyttisia tutkimuksia kielellisesti epäsymmetrisestä vuorovaikutuksesta*. (s. 7–11). Kehitysvammaliitto ry.

Mantila, H. (1996). Mitä viestivät media, viestin ja tiedotusväline? *Kielikello*, *2*. Haettu 27.6.2022 osoitteesta https://www.kielikello.fi/-/mita-viestivat-media-viestin-ja-tiedotusvaline-

McEnery, T. & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Milton, J. (2009). *Measuring second language vocabulary acquisition* (Vol. 45). Multilingual Matters.

Nykysuomen sanakirja I-VI (2002). 15. painos, näköispainos. WSOY.

Pajunen, A., Itkonen, E. & Vainio, S. (2015). Sanamerkityksen hallinta nuorilla aikuisilla. *Virittäjä*, *119*(2), 160–188. Haettu 30.12.2022 osoitteesta https://journal.fi/virittaja/article/view/41381

Räsänen, M. (2002). Vierassanojen kirjoitusongelmat. *Kielikello, 3*. Haettu 6.9.2022 osoitteesta https://www.kielikello.fi/-/vierassanojen-kirjoitusongelmat

Salmi, L. (Ohjaus ja käsikirjoitus). (2010–2014). *Toisenlaiset frendit*. Metronome Film & Television Oy A Shine Group Company.

Saukkonen, P. (1979). *Suomen kielen taajuussanasto = A frequency dictionary of Finnish*. WSOY.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Palgrave Macmillan. https://doi.org/10.1057/9780230293977

Selkokeskus (2018). *Selkomittari*. Haettu 22.1.2023 osoitteesta https://selkokeskus.fi/wp-content/uploads/2018/10/SELKOMITTARI_2018_11.10.18.pdf.

Selkokeskus (2022). *Selkomittari 2.0*. Haettu 18.5.2022 osoitteesta https://selkokeskus.fi/wp-content/uploads/2022/04/Selkokielen-mittari-2.0.pdf.

Suomen virtuaaliyliopisto (2006). Tulehdus (Inflammaatio). *Solunetti*. Haettu 26.7.2022 osoitteesta https://www.solunetti.fi/fo/patologia/tulehdus

Tanaka-Ishii, K. & Terada, H. (2011). Word familiarity and frequency. *Studia Linguistica*, *65*(1), 96–116. https://doi.org/10.1111/j.1467-9582.2010.01176.x

Thompson, G. L. & Desrochers, A. (2009). Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency. *Behavior Research Methods, 41*(2), 452–471. https://doi.org/10.3758/BRM.41.2.452

Uotila, E. (2020). Selko Suomessa: Selkokielen kehitys ja sovelluksia. *Puhe ja kieli, 39*(4), 307–324. https://doi.org/10.23997/pk.74581.

Valtasalmi, I. (2021a). Lukutaitoisten kehitysvammaisten aikuisten sanamerkitysten hallinta. *Finnish Journal of Linguistics*, *34*, 301–332. Haettu 26.7.2022 osoitteesta https://journal.fi/finjol/article/view/103226

Valtasalmi, I. (2021b). Selkoa ihmisestä: *Ihminen*-sanan merkitykset ja käyttö selkokielisissä sanomalehtiteksteissä. *Sananjalka*, *63*(63). https://doi.org/10.30673/sja.107345

Vanhatalo, U. & Lindholm, C. (2020). Prevalence of NSM primes in easy-to-read and standard Finnish: Findings from newspaper text corpora. Teoksessa L. Sadow, B. Peeters, & K. Mullan (toim.), *Studies in ethnopragmatics, cultural semantics, and intercultural communication. Vol. 3. Minimal English (and beyond)*, (s. 213–234). Springer.

Virtanen, H. (2012). *Selkokielen käsikirja*. 2. uudistettu painos. Oppimateriaalikeskus Opike.

Wengelin, Å. (2015). Mot en evidensbaserad språkvård? En kritisk granskning av några svenska klarspråksråd i ljuset av forskning om läsbarhet och språkbearbetning. *Sakprosa, 7*(2). Haettu 25.10.2022 osoitteesta www.journals.uio.no/index.php/sakprosa/article/view/983.

Wiio, O. A. (1974). *Ymmärretäänkö sanomasi? Viestintä – tiedonvälitys.* 6. painos. Weilin+Göös.

OBJECTIVE AND SUBJECTIVE FREQUENCY OF LOANWORDS AND NATIVE WORDS PICKED FROM EASY FINNISH TEXTS

• Idastiina Valtasalmi, Faculty of Information Technology and Communication Sciences, Tampere University

The article examines the objective and subjective frequency of loanwords (N = 50) and corresponding native words (N = 50) picked from Easy Finnish texts. Objective frequency was examined by comparing frequencies of word occurrence in corpora of Easy Finnish and standard Finnish media texts. Subjective frequency was examined with a survey to estimate how often language experts (N = 25) encounter the words and which words they usually use themselves. Native words occurred less frequently than loanwords in both corpora, but native words occurred more frequently in Easy Finnish than in standard Finnish. Significant differences were observed in the frequency of occurrence of individual words, which were explained, for example, by differences between corpora and language forms. Highest frequency of encounter estimates were given to words with high objective frequency that occurred in many different contexts. Frequency of encounter estimates ran in parallel with objective frequencies especially in words with high objective frequency. Words with low objective frequency varied in their frequency of encounter. Shortest words of the data set had highest objective frequency and were encountered most often. Language experts preferred to use shorter words of the word pairs in the data set. Most of the preferred words were loanwords. Frequency of occurrence is not sufficient as the only criterion for assessing word frequency in Easy Finnish. Frequency of encounter and use could be used as additional criteria.

Keywords: corpus linguistics, Easy Finnish, objective frequency, online survey, subjective frequency