
#1

Onko suomen kielen sanasto poikkeuksellisen omaperäistä?

Kaisa Häkkinen

Abstract

Etymological stratification of the Finnish vocabulary

In his famous textbook "Suomen kielen rakenne ja kehitys", Lauri Hakulinen states that the vocabulary of the Finnish language is quite exceptional in its etymological composition. Only a fifth or even less can be proven loans from other languages.

As the vocabulary of a natural language is an open system, it can only be analyzed with a sample survey. If we take the word-stems included in the central vocabulary of Finnish, it turns out that about half are loans.

Similar results have been obtained from related languages. Thus, there is no reason to consider the Finnish vocabulary exceptional in any way.

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

1. Johdanto

Kautta aikojen tunnetuin yleisesitys suomen kielen rakenteesta ja kehityshistoriasta on Lauri Hakulisen *Suomen kielen rakenne ja kehitys*, jonka ensimmäinen painos ilmestyi kahdessa osassa vuosina 1941 ja 1946. Sittemmin teoksesta on julkaistu uusia painoksia, joihin on tehty lukuisia täydennyksiä ja korjauksia. Neljäs (1979) ja viides (2000) painos ovat sisällöltään samat.

Alusta alkaen on muuttamattomana säilynyt kohta, jossa Hakulinen kertoo suomen kielen perussanaston olevan valtaosaltaan omaperäistä. Sanakirjantekijä Knut Cannelinin esittämiin laskelmiin viitaten hän arvioi lainasanojen osuudeksi 26 000 hakusanaa sisältävässä sanakirjassa noin 20 prosenttia eli viidesosan.

Vuosikymmenien mittaan tätä arviota ei ole kyseenalaistettu, eikä asia näytä muutenkaan herättäneen suurta kiinnostusta tutkijapiireissä. Lainasanatutkimus on kyllä edistynyt etenkin germaanisten ja indoiranilaisten lainojen osalta, mutta yksittäistapausten tutkiminen ei ole johtanut uuteen kokonaiskuvaan. Ei myöskään ole itsestään selvää, miten näin laajaa kysymystä olisi mahdollista ja perusteltua tutkia. Kun Hakulinen esitti arvionsa, suomen kielestä ei ollut olemassa yhtään ainoaa etymologista sanakirjaa. Hakulinen ratkaisi ongelman viittaamalla Cannelinin havaintoihin ja tutkimalla itse manuaalisesti 2000 sanan laajuista tekstiotosta, josta puolet oli poliittissävyistä proosaa ja puolet Otto Mannisen runoja. Tässä aineistossa omaperäisten sanojen osuus oli vieläkin suurempi, lähes 85 prosenttia.

On monta syytä epäillä, etteivät Hakulisen esittämät tiedot päde nykysuomen yleiskieleen. Sanaston rakenteesta ja dynamiikasta on käytettävissä uutta tietoa, etymologian tutkimuksessa on saavutettu uusia tuloksia ja useita etymologisia sanakirjajulkaisuja on ilmestynyt 1950-luvulta alkaen. Tietotekniikan ansiosta pystytään käsittelemään suuria aineistoja, yhdistämään tekstisanojen esiintymät automaattisesti perusmuotoihin ja laskemaan sanojen frekvenssit miljoonia sanaesiintymiä sisältävistä korpuksista. Tältä pohjalta on mahdollista määritellä keskeinen sanasto ja tutkia sen etymologisia kerrostumia aiempaa luotettavammin.

"Sanaston rakenteesta ja dynamiikasta on käytettävissä uutta tietoa."

2. Sanaston ikäkerrostumat

Suomenkielisiä tekstejä on olemassa 1500-luvulta alkaen. Keskiajan puolelta tunnetaan vain nimistöä ja tekstifragmentteja. Keskiaikaisten tai tätä vanhempien sanojen etymologiaa on mahdollista tutkia vain vertailevan kielitieteen menetelmin. Nyrkkisääntö on se, että sana on sitä vanhempi, mitä etäisemmistä sukukielistä

sille löytyy äänteellisesti ja semanttisesti uskottavia, samaa alkuperää olevia vastineita. Vanhinta sanastoa ei voi luotettavasti tutkia perehtymättä uralilaiseen ja kantasuomalaiseen äännehistoriaan.

Nykysuomen sanaston voi karkeasti jakaa kolmeen ikäkerrostumaan. Vanhin eli uralilainen tai suomalais-ugrilainen osa on se, jolla on vastineita etäsukukielissä. Näistä useimpia puhutaan nykyisen Venäjän alueella Uralvuoriston molemmin puolin. Arkeologian aikakausiin suhteutettuna tämän osan ikä on kivikaudelta pronssikaudelle. Iältään keskimmäinen on se kantasuomalaiseksi nimitetty sanaston kerrostuma, joka on yhteinen Itämeren alueella puhuttujen lähisukukielten kanssa. Tämä kerrostuma ajoittuu pronssikaudelta viikinkiajalle. Nuorinta osaa edustaa se sanasto, joka esiintyy pelkästään suomessa. Siihen kuuluu rautakauden lopun, keskiajan ja uuden ajan sanasto. Kaikissa kerrostumissa on sekä omaperäistä että lainattua sanastoa.

3. Mistä sanat tulevat

Vanhimpaan omaperäiseen kerrostumaan kuuluvat ensisijaisesti sellaiset keskeiset perussanat, jotka ovat tarpeen aina ja kaikkialla. Näitä ovat esimerkiksi perusverbit *elää, kuolla, mennä, olla* ja *tulla*. Tärkeää perussanastoa, jossa omaperäisen sanaston osuus on huomattavan suuri, ovat myös persoona- ja demonstratiivipronominit, pieniin lukuihin viittaavat numeraalit, keskeiset sukulaisuustermit, ruumiinosien nimitykset ja jokapäiväiseen elinympäristöön kuuluvien asioiden ja ilmiöiden nimitykset.

Vanhimpiin sanastokerrostumiin kuuluvista lainoista on pystytty luotettavasti tunnistamaan vain indoeurooppalaisista naapurikielistä saadut lainat. Sekä uralilaisen että indoeurooppalaisen kielikunnan muinaisten ydinalueiden oletetaan sijainneen Mustanmeren pohjoispuolisilla seuduilla siten, että lainakontaktit ovat olleet mahdollisia. Lainojen suunta on ollut indoeurooppalaisesta uralilaiseen. Vanhimmat lainat voivat edustaa indoeurooppalaisen kantakielen äännetasoa, mutta enemmistö on tullut indoiranilaisista kielistä, joita alan kirjallisuudessa on perinteisesti kutsuttu arjalaisiksi. Iranilaiskielistä osseettia puhutaan edelleen Kaukasuksella sekä Venäjän että Georgian alueella, ja monissa suomen etäsukukielissä on myös nuorempia iranilaisia lainoja, joita itämerensuomessa ei ole.

Vanhimmista indoeurooppalaisista lainoista on aineistoa melko vähän, ja useissa tapauksissa sitä voidaan myös äännehistoriallisesti tulkita eri tavoin. Esimerkiksi *mehiläinen, mesi* ja *porsas* ovat lainoja, joita on kirjallisuudessa esitetty sekä vanhoina indoeurooppalaisina että arjalaisina lainoina, ja sanoja *porsas* ja *salko* on arveltu myös indoeurooppalaisen kantakielen baltoslaavilaisesta haarasta lainatuiksi (tarkemmin esim. Holopainen, 2019, 2022). Äänteellisten tuntomerkkien perusteella melko varmasti indoiranilaisesta haarasta juontuviksi lainoiksi on katsottu esimerkiksi *arvo, jumala, marras, sarvi, sata* ja *vasara*.

Omaperäiseen kantasuomalaiseen sanastoon kuuluvat varmimmin sellaiset sanat, joilla on laaja levikki itämerensuomalaisissa lähisukukielissä mutta joille ei tunneta vastineita etäsukukielistä eikä myöskään uskottavaa lainaetymologiaa.

Näitä ovat esimerkiksi *jänis, latva, leski, neuvo, saari* ja *vilu*. Tämäntyyppisistä sanoista huomattavan monet on viime vuosikymmeninä pystytty osoittamaan lainoiksi. Osa sanoista on selvästi ääntä jäljittelevistä eli onomatopoieettisista vartaloista muodostettuja, esim. *paukkua, siristä, vinkua*. Tällaisetkin sanat voivat olla lainoja, mutta toisaalta samanlainen äänteellinen motivaatio voi synnyttää samantapaisia sanoja erikseen eri kielissä (tarkemmin esim. Kim, 2019).

Itämeren alueella kosketukset indoeurooppalaisia kieliä puhuneisiin naapureihin ovat olleet tiiviit, ja kontaktien välityksellä on opittu paljon uutta. Karjanhoidon ja maanviljelyn aloittaminen sekä taito käsitellä metalleja ovat uudistaneet koko elämäntapaa ja sen ulkoisia puitteita. Balttilaisista kielistä saatuja lainoja ovat esimerkiksi *halla, herne, morsian, oinas, paimen, silta, sisar, tytär* ja *vuohi*. Laaja kokonaisesitys balttilaisista lainoista on Santeri Junttilan väitöskirja (2015).

Vanhoiksi germaanisiksi lainoiksi on tapana sanoa niitä sanoja, jotka on saatu kantagermaanista tai sitä seuranneesta kantaskandinaavista. Tähän kerrostumaan kuuluvia lainoja ovat esimerkiksi vesiliikenteestä kertovat *airo, laiva, purje*, metalleihin ja metalliesineisiin liittyvät *miekka, rauta, rengas*, valtasuhteisiin liittyvät *kuningas, ruhtinas, valta* sekä maatalouskulttuuriin kuuluvat *lammas, nauta, pelto*. Vanhoja germaanisia lainoja on esitelty kattavasti vuosituhannen taitteessa valmistuneessa etymologisessa teemasanakirjassa (Kylstra ym., 1992–2012).

Kantasuomalaisen ajan lopulla on saatu myös slaavilaisia tai muinaisvenäläisiä lainoja (ajoituksista esim. Kallio, 2006), joista äännekriteerien perusteella vanhimpiin kuuluvat esimerkiksi *akkuna, lusikka, palttina, papu, risti, saapas, sirppi, taltta, tuska* ja *värttinä*.

Suomen kielen itsenäinen kehitys on päässyt vauhtiin toisella kristillisellä vuosituhannella. Keskiaika toi mukanaan suuria muutoksia sekä henkiseen että materiaaliseen elämään. Kristinusko levisi ja vakiintui, ja sen myötä Suomen alue liittyi monin sitein osaksi läntistä eurooppalaista kulttuuriverkostoa, jonka yhteinen sivistyskieli oli latina. Itämeren kauppa kehittyi, ja Suomen ensimmäiset kaupungit perustettiin rannikoille. Hallinnollisesti Suomesta tuli Ruotsin Itämaa, ja maahan asettui pysyvästi ruotsinkielisiä asukkaita. Suomessa astuivat voimaan emämaan lait, joita oli erikseen maaseutua ja kaupunkeja varten. Järjestäytymisestä kertovat esimerkiksi ruotsalaisperäiset lainasanat *laki, lääni, tuomari*, jotka esiintyvät täysin vakiintuneina jo 1500-luvun kielessä. Toisaalta jo vanhimmassa lakikielessä käytetään sellaisia omakielisistä aineksista muodostettuja termejä kuin *käräjät, pitäjä, rikos*, jotka viittaavat omaehtoiseen järjestäytymiseen.

Vain suomen murteissa tunnettuja suppealevikkisiä murresanoja on runsaasti, mutta suomen yleiskieleen kotiutuneita, täysin vailla etymologista selitystä olevia sanoja on vain vähän, sillä useimmiten etymologisesti epäselvillä sanoilla on vastineita ainakin joissakin lähisukukielissä. Lisäksi jotkin tähän ryhmään aiemmin lasketuista sanoista ovat viimeaikaisessa tutkimuksessa osoittautuneet lainoiksi. Tuntematonta alkuperää olevia suomalaisia sanoja, jotka ovat esiintyneet kirjakielessä Agricolasta alkaen, ovat esimerkiksi *askare, aurinko, hankkia, harha* sekä *rymä*, josta nykyisessä yleiskielessä tosin käytetään takavokaalista

varianttia *ruma* (tarkemmin Agricolan sanakirja, 2025).

Keskiajalla tärkeimmäksi lainanantajakieleksi kohosi ruotsi. Lisäksi merkittävä rooli oli keskialasaksalla, joka oli Itämeren alueen kaupan yhteinen kieli. Alasaksaa käytettiin myös keskiaikaisten kaupunkien raadeissa, koska suuri osa niiden jäsenistä oli kauppaporvareita. Alasaksa vaikutti voimakkaasti ruotsiin, ja suoran vaikutuksen lisäksi alasaksalaisia lainoja virtasi suomeen myös ruotsin kautta. Kielimuodot ovat niin samankaltaisia, että usein on vaikea ratkaista, kummalta taholta tietty sana on lainautunut suomeen. Vanhastaan ruotsalaiset ja alasaksalaiset lainat on niputettu yhteen ja niitä on nimitetty nuoremmiksi germaanisiksi lainoiksi. Erottelua on pyritty tietoisesti kehittämään vasta aivan viime vuosikymmeninä (Bentlin, 2008). Ruotsalaisia lainoja ovat esimerkiksi *muori* ja *vaari, katti* ja *rotta*, *katu* ja *kaupunki*, *penkki* ja *sänky*. Jokseenkin luotettavasti alasaksalaisiksi voidaan tunnistaa esimerkiksi *ammatti, hovi, häät, nöyrä, rouva, touvi* ja *öykkäri*.

Venäläisiä lainoja on monisatavuotisen naapuruuden ansiosta tullut etenkin itämurteisiin, mutta osa on esiintynyt kirjakielessäkin Agricolasta alkaen. Varsinkin 1700- ja 1800-luvulla venäjän vaikutus yleiskieleen lisääntyi selvästi. Venäläisiä lainoja ovat esim. *kapakka, kasarmi, kolpakko, leima, meteli, miettiä, määrä, pätsi, ravita, rosvo, rotu, rusakko, saapas, sääli, tappara, tyrmä, ukaasi, vaino, velho* ja *viesti*. Autonomian ajan lainoissa on monia, joita nykysuomalainen voi ajatella Stadin slangille ominaisina sanoina, esimerkiksi *lusia, mesta, murju, narikka, pirssi, pomo, putka, rokuli* ja *safka* (Paunonen, 2016).

Keskiajalta lähtien on varsinkin kirkolliseen kieleen saatu latinalaisia ja latinan kautta kreikkalaisia ja heprealaisia lainoja. Näitä kieliä kutsuttiin pyhiksi kieliksi sen vuoksi, että heprea ja kreikka edustivat Raamatun alkukieliä ja latina oli Raamatun yleisimmin käytetyn käännöksen *Versio Vulgatan* kieli ja samalla myös roomalaiskatolisen kirkon ja sen tarjoaman koulutusjärjestelmän kieli. Osa tähän kerrostumaan kuuluvista lainoista on tullut suoraan latinasta, osa ruotsin kautta. Esimerkiksi seemiläiskielistä juontuvat *aamen, mammona* ja *saatana*, kreikkalaisperäiset *historia, koulu, planeetta* ja *testamentti* sekä latinasta lainatut *palatsi, palmu, salvia* ja *temppeli* ovat esiintyneet jo Agricolan teksteissä. Reformaatioajalta lähtien pyhistä kielistä saatujen lainojen välittäjäkielenä on toiminut ruotsin ohella myös yläsaksa, joka oli käännösesikuvina arvostettujen uskonnollisten teosten kieli.

Lainaperäinen uskonnollinen perussanasto on kerran kieleen vakiinnuttuaan säilynyt suurin piirtein ennallaan, mutta etenkin tieteen, taiteen ja kulttuurin aloilla antiikin sivistyskielten välillinen vaikutus on jatkunut ja voimistunut yleiseurooppalaisen sanaston muodossa. Suomesta tuli opetuksen, kulttuurin, taiteen, tieteen ja yhteiskuntaelämän kieli vasta 1800-luvulla, ja monet sanat, jotka muualla Euroopassa olivat olleet aktiivisessa käytössä jo vuosisatojen ajan, omaksuttiin vasta tässä vaiheessa osaksi suomen yleiskielen sanastoa. Esimerkiksi vanhaan aikaan viittaava *antiikki* ja hengenviljelyä merkitsevä *kulttuuri* ovat kotiutuneet kieleen nykymerkityksessään vasta 1800-luvun jälkipuoliskolla. Latina on uudistetussa muodossa säilynyt etenkin tieteen kielenä, josta on saatu lainoja kansankieliin, ja 1700-luvun puolimaissa

syntynyt lajien tieteellinen nimistö on toiminut lainalähteenä myös joillekin suomeen omaksutuille sanoille, esim. *kantarelli* (< *Cantharellus cibarius*), *manuli* (< *Felis manul*), *suula* (< *Sula bassana*).

Varsinkin 1700- ja 1800-luvulla huomattava osa yleiseurooppalaisista sanoista on tullut aikakauden muodikkaasta sivistyskielestä ranskasta joko suoraan tai ruotsin välityksellä. Tähän joukkoon kuuluvat esimerkiksi *ateljee, flanelli, insinööri, kalossi, kaneli, kaniini, kapteeni, kasarmi, kersantti, klisee, komitea, kulissi, luutnantti, marssia, miljardi, miljoona, mineraali, mitali, muoti, novelli, paketti, paneeli, parketti, pataljoona, pommi, pusero, ramppi, raportti, riimi, rooli, tunneli, vaneri* ja *viulu*.

Suomen sukukielistä lainattuja sanoja tunnetaan melko vähän, ja varsinkin vanhempien lainojen ongelmana on usein se, että niitä on vaikea erottaa vanhasta yhteisestä perintösanastosta. Nuorimmissa sanastokerrostumissa ero näkyy selvemmin, ja lainautumisesta voi parhaassa tapauksessa olla kirjallista taustatietoa. Saamelaisia lainoja ovat esimerkiksi *jänkä, mursu, naali, piekana, raanu, uivelo* ja *väylä*. Jo Agricolan kielessä esiintynyt *norsu* on *mursu*-sanan variantti, joka Himangan murteessa on nykyaikoihin asti säilynyt merkityksessä 'mursu'. Viro on välittänyt suomeen alkuaan muista kielistä juontuvia sanoja kuten *isota* 'tuntea nälkää', *leiviskä, raamattu* ja *turku*, jotka ovat esiintyneet jo Agricolan kielessä. Nuorempia virolaisia lainoja ovat esim. *lavaste* ja *lennokki*. Huomattavasti vanhempi uskonnolliselle kielelle ominainen virolaislaina on verbi *nuhdella*.

Nykyajan tärkein lainanantajakieli englanti on alkanut vaikuttaa suomen sanastoon melko myöhään. Vanhin englannin välityksellä mutta ilmeisesti lopulta ruotsin kautta saatu laina on vuonna 1637 ilmestyneessä tulkkisanakirjassa mainittu *tupakki*. 1700-luvun lainoja ovat *halli, kutteri* ja *punssi*. 1800-luvulla on saatu monet purjealustyyppien nimet, kuten jaala, *klipperi, kuunari, priki*, sekä merenkulkuun liittyvien tuotteiden ja tarvikkeiden nimet, esimerkiksi *mahonki, moppi, muki* ja *toti*. Suomenkielisissä sanomalehdissä ja käännöskirjallisuudessa kerrottiin ulkomaiden oloista, matkoista, maisemista ja ylellisestä elämästä käyttäen asiaan kuuluvaa lainasanastoa, esimerkiksi *booli, kaktus, klubi, kriketti, lady, lordi, party* 'kutsut; puolue', *pihvi, preeria, pyjama, rommi, tennis* ja *toteemi*.

Arkisempaa englannista juontuvaa sanastoa saatiin Amerikkaan lähteneiden siirtolaisten välityksellä. Tähän joukkoon kuuluvat esimerkiksi *hoopo, kaara, kämppä, lokari, maili, mainari, pokeri* ja *taala*.

Lainojen virta on jatkunut yhä voimistuen 1900-luvulta eteenpäin. Erotuksena entiseen on se, ettei sanoja välttämättä sopeuteta suomen kielen äännerakenteeseen (esim. *trawl* > *trooli*, *wire* > *vaijeri*), vaan niitä käytetään sitaattilainojen tapaan alkuperäisessä asussaan (*know how, workshop*).

4. Miten sanoja muodostetaan

Rakenteensa perusteella suomen sanat jaetaan yleensä kolmeen pääryhmään. Nämä ovat jakamattomat perussanat (*kala*), johdokset (*kalastaa*) ja yhdyssanat (*kalamies*). Käytännössä rajat eivät aina ole selvät. Tavallinen kielenkäyttäjä ei

välttämättä huomaa, että esimerkiksi *toinen* on *tuo*-pronominin johdos ja *tällainen* on johdos pronominimääritteen (*tämä*) ja pääsanan (*laji*) muodostamasta rakenteesta. Kompleksisen sanan rakenne voi aikojen kuluessa hämärtyä, eivätkä sen osat enää erotu selvästi, vaikka sana olisi alkuaan täysin algoritmisen sananmuodostusprosessin tulos.

Käytännössä sanoja tuotetaan myös muilla tavoin, esimerkiksi varioimalla olemassa olevaa sanaa tai yhdistämällä aineksia muiden periaatteiden mukaan. Esimerkiksi slangissa tai hypokoristisessa nimistönmuodostuksessa on tavallista, että alkuperäisestä sanasta hyödynnetään vain alkuosa ja loppuun lisätään yleismerkityksinen johdin (*huoltoasema -> huoltis*, *huoltsikka*; *Reino -> Reiska, Repe*). Sanoja tai sananosia voidaan risteyttää keskenään (*maanärhi -> marhi*), ja sanoja muodostetaan myös typistämällä (*alennusmyynti -> ale*) tai lyhenteiden pohjalta (*luonnonmukainen -> luomu*).

Rakennetyyppien (perussana, johdos, yhdyssana) määräsuhteet suomen nykyisessä yleiskielessä ovat *Nykysuomen sanakirjan* aineistosta laskettuina prosentteina noin 8 : 27 : 65. Yhdyssanoja on siis ylivoimainen enemmistö, ja johdoksiakin on huomattavasti enemmän kuin jakamattomia perussanoja. Viimeksi mainitut edustavat kuitenkin sanaston ydinosaa siinä mielessä, että niistä voi muodostaa rajattoman määrän uusia johdoksia ja yhdyssanoja. Etenkin ikivanhaan perussanastoon kuuluvat sanat ovat tässä suhteessa osoittautuneet hyvin produktiivisiksi. Etymologisen tutkimuksen kohteina on perinteisesti pidetty jakamattomia perussanoja ja niissä edustuvia sanavartaloita (esim. *käsi*, *käte-*).

5. Etymologiset kerrostumat keskeisessä sanastossa

Keskeinen sanasto voidaan määritellä eri tavoin. Yksi mahdollinen tapa on laskea sanojen esiintymistiheyttä suurista tekstikorpuksista ja pitää keskeisinä sanoina niitä, jotka sijoittuvat frekvenssitilaston alkupäähän aineiston aihepiiristä riippumatta. Suomen yleiskielen monipuolisimmat lingvistiset frekvenssianalyysit perustuvat ns. Oulun korpukseen, joka on koostettu Pauli Saukkosen johdolla Oulun yliopistossa ja jonka pohjalta on julkaistu frekvenssisanakirja (Saukkonen ym., 1979).

Toinen mahdollisuus on luottaa kokemukseen ja käytännön opetustyössä tehtyihin havaintoihin. Suomea vieraana kielenä opiskeleville laadituissa oppikirjoissa esiintyvä sanasto pyritään yleensä valitsemaan niin, että sen turvin pystyy ainakin auttavasti seuraamaan mediaa ja selviämään jokapäiväisistä kielenkäyttötilanteista.

Kun itse ryhdyin selvittämään suomen yleiskielen keskeisimmän sanaston etymologista taustaa, otin laskelmien pohjaksi sekä frekvenssisanakirjan että suomea vieraana kielenä opiskeleville tarkoitetun oppikirjan sanaston (Häkkinen, 1992). Tämä aineisto sisältää frekvenssisanakirjan tuhat yleisintä sanaa, ja tähän aineistoon on lisätty ulkomaalaisopetuksen tarpeisiin laaditusta sanakirjasta *A Student's Glossary of Finnish* (Branch ym., 1980) kaikki ne sanat, jotka eivät ole tulleet otokseen vielä tuhannen yleisimmän sanan joukossa. Sanaston dynamiikkaa koskevissa tutkimuksissa (esim.

Niemikorpi, 1991) on todettu, että suomen yleiskielessä 350 yleisintä sanaa kattaa keskimäärin 50 prosenttia otoksen eri tekstien sanastosta. Puhuttua kansankieltä edustavissa murteissa samaan kattavuuteen riittää jo 25 yleisintä sanaa, jotka ovat *se, olla, ja, niin, sitten, kun, ne, ei, minä, että, siellä, semmoinen, mutta, tulla, siinä, mennä, sanoa, nyt, no, tehdä, kyllä, saada, vaan, me, jotta* (Jussila ym., 2002). Näistäkin osa edustaa keskenään samaa sanavartaloa, esim. *se*-pronominin lisäksi joukkoon kuuluvat siitä muodostetut *semmoinen, siellä, siinä, sitten*.

Perusaineiston rajauksen jälkeen sanat on analysoitu rakenteellisesti ja niistä on poimittu esiin sanavartalot, joita tässä tapauksessa löytyi 844. Vartaloiden alkuperä on selvitetty ajantasaisen etymologisen kirjallisuuden avulla. Tulokset näkyvät taulukossa 1.

6. Omaperäisten ja lainattujen sanojen suhde

Suomen yleiskielen keskeistä sanastoa koskeva etymologinen laskelma osoittaa, että aineistoon sisältyvistä sanavartaloista hieman vajaa puolet on omaperäisiä. Tämä on olennaisesti vähemmän kuin Lauri Hakulinen edellä mainitussa käsikirjassaan esitti. Sen sijaan tulos on lähempänä unkarin kielestä tehtyä vastaavaa laskelmaa

Omaperäiset sanat	410	48,6 %
Vanhat indoeurooppalaiset ja arjalaiset	38 + 20?	4,5 % – 6,9 %
Vanhat germaaniset	141 + 38?	16,7 % –21 %
Balttilaiset	46 + 7?	5,5 % –6,3 %
Slaavilaiset	19 + 1?	2,3 % – 2,4 %
Nuoret germaaniset	63 + 13?	7,5 % – 9 %
Yleiseurooppalaiset	61	7,2 %
Muut, kiistanalaiset	45	5,3 %

Taulukko 1. Keskeisten sanavartaloiden (yht. 844) etymologinen tausta. Useimmista lainakerrostumista on ilmoitettu erikseen varmat ja epävarmat tapaukset (merkitty kysymysmerkillä), jolloin prosenttiluvut ilmoittavat vähimmäis- ja enimmäisosuuden koko aineistosta.

(Tolnai, 1928), johon Hakulinen kirjassaan viittasi. Sen mukaan unkarin kielen kantasanoista 65 prosenttia on omaperäisiä. Tarkempiin vertailuihin ei ole syytä mennä, koska lainakerrostumat ovat unkarissa historiallisista ja maantieteellisistä syistä suurimmaksi osaksi toiset kuin suomessa.

Omaperäisten ja lainattujen sanavartaloiden määräsuhteita on laskettu myös virosta, joka on suomen lähisukukieli ja maantieteellisesti läheinen naapurikieli (Rätsep, 2002). Laskentaperusteet ovat sikäli erilaiset kuin suomea koskevassa laskelmassa, että kiistanalaisia tapauksia ei ole erotettu omaksi ryhmäkseen, vaan jokaisesta ryhmästä on erikseen kirjattu varmat ja epävarmat tapaukset. Tämän perusteella viron sanastossa on omaperäisiä sanavartaloita 46,67–60,93 prosenttia, lainavartaloita 40,48–48,71 prosenttia ja oppitekoisia sanavartaloita 0,94–1,01 prosenttia. Suomessa sanastoa on etenkin 1800-luvulla tietoisesti kartutettu muodostamalla oppitekoista sanastoa ennestään käytössä olleiden sanavartaloiden pohjalta, mutta virossa on pyritty myös keksimään kokonaan uusia sanavartaloita, esimerkiksi *küülik* 'kaniini', *relv* 'ase'.

Olen laskenut etymologisten kerrostumien suhteita myös Mikael Agricolan teosten suomenkielisissä teksteissä esiintyvästä sanastosta (Häkkinen, 2023). Tämä aineisto sisältää kaikkiaan runsaat 9000 lekseemiä. Omaperäisten sanavartaloiden osuus on 39,6 prosenttia ja lainojen yhteismäärä 57,3 prosenttia. Tämä on olennaisesti vähemmän kuin Lauri Hakulinen (1979) edellä mainitussa käsikirjassaan esitti. Loput ovat epäselviä tapauksia. Etymologisena lähdeaineistona on pääosin käytetty vuonna 2020 päivitettyä *Nykysuomen etymologista sanakirjaa* (Häkkinen, 2020). Osa Agricolan sanastosta on ennestään etymologioimatonta, joten tältä osin taustat on selvitetty omien etymologisten tutkimusten ja yksittäisiä lekseemejä koskevan lähdekirjallisuuden avulla.

Agricolan sanastossa lainojen osuus on hieman suurempi kuin nykysuomen sanastossa, mikä on ymmärrettävää, koska teosten aihepiiri on kansainvälinen eikä lainoja ole pyritty tietoisesti välttämään niin kuin on tehty varsinkin 1800-luvulla nykysuomen sanastoa kehitettäessä.

7. Lopuksi

Edellä esitetyt laskelmat osoittavat, ettei suomen kielen sanaston etymologinen koostumus ole mitenkään poikkeuksellinen verrattuna tarjolla olevaan sukukielten aineistoon. Suomea on kieliä koskevissa keskusteluissa pidetty outona ja vaikeana kielenä ilmeisesti sen takia, että suomi ei kuulu indoeurooppalaiseen kielikuntaan niin kuin valtaosa muista Euroopan kielistä. Näin ollen siinä ei voi olla sitä yhteistä perintösanastoa, joka yhdistää Euroopan romaanisia, germaanisia, balttilaisia ja slaavilaisia kieliä keskenään.

Tieteen luonteeseen kuuluu jatkuva edistyminen ja aiempien tutkimustulosten päivittäminen. Elävällä tieteenalalla on selviö, etteivät sata vuotta sitten saavutetut tulokset voi olla ajan tasalla. Minkään luonnollisen kielen sanastoa ei voi kuvata tai tutkia kokonaisuutena sen laajuuden ja aineiston jatkuvan vaihtuvuuden vuoksi, ja kun siitä tehdään erilaisia otostutkimuksia, saadaan vaihtelevia tuloksia otoksen laadun ja laajuuden mukaan. Laajan sana-aineiston etymologinen kartoittami-

nen on kuitenkin niin työläs prosessi, että on olennaisesti helpompaa julkaista uudestaan vanhoja tutkimustuloksia kuin tehdä kokonaan uusi tutkimus. Esimerkiksi Hakulisen kirjan eri laitoksissa yksittäisten esimerkkisanojen etymologista statusta on päivitetty, mutta kokonaislaskelma on säilytetty ennallaan.

Etymologista kokonaiskuvaa kannattaa ryhtyä selvittämään sellaisessa tutkimuksen vaiheessa, kun käytettävissä on juuri valmistunut, laaja ja ajantasainen perusaineisto. Tällaista ei ole tarjolla esim. suuria sanastoja ja sanatietokantoja tuottavassa Kotimaisten kielten keskuksessa. Verkossa julkaistu *Suomen etymologinen sanakirja* on käytännössä vuonna 2000 valmistuneen Suomen sanojen alkuperä teoksen sähköinen versio. Päivitystä edustavat useiden sana-artikkelien perään lisätyt kommentit, jotka eivät sisällä itsenäistä etymologista tutkimusta eivätkä lähdeviitteitä. Lukijan on yleensä mahdotonta tietää, mihin kommentti perustuu, eikä kommentti myöskään tarjoa aineistoa etymologian itsenäistä arviointia varten.

Uusi ikkuna kokonaiskuvan luomiselle avautuu vuonna 2025, kun *Nykysuomen etymologisesta sanakirjasta* ilmestyy uusi, päivitetty versio. Sanakirja toimii laskelmien pohjana olevan aineiston yksityiskohtaisena esittelynä. Vaikka nykyaikainen kieliteknologia tarjoaa monia mahdollisuuksia suurten aineistojen kokoamiseen ja työstämiseen, etymologista algoritmia ei ole olemassa. Sanavartaloiden identifiointi ja niiden taustan selvittäminen äännehistorian, semantiikan, levikkitiedon ja kontaktilingvistiikan kannalta vaatii paljon manuaalista työtä, aikaa ja vaivaa. Tältä pohjalta on ymmärrettävää, että etymologinen tutkimus on perinteisesti keskittynyt mieluummin johonkin sanaston osaan kuin kokonaiskuvan luomiseen.

Kirjoittaja

Kaisa Häkkinen

Kaisa Häkkinen on suomen kielen emeritaprofessori Turun yliopistossa. Hänen keskeisimpiä tutkimusalojaan ovat suomen kielen ja erityisesti sanaston historia, Mikael Agricolan kieli, suomalais-ugrilainen etymologia ja kielentutkimuksen oppihistoria. Hänet on nimitetty tieteen akateemikoksi 2020 ja promovoitu Greifswaldin yliopiston kunniatohtoriksi 2021. .

Kuvaaja: Mikko Suominen

Lähteet

Bentlin, M. (2008). Niederdeutsch-finnische Sprachkontakte. Der lexikalische Einfluß des Niederdeutschen auf die finnische Sprache während des Mittelalters und der frühen Neuzeit. *Mémoires de la Société Finno-Ougrienne* 256. Suomalais-Ugrilainen Seura.

Holopainen, S. (2019). *Indo-Iranian borrowings in Uralic. Critical overview of the sound substitutions and distribution criterion*. Väitöskirja, Helsingin yliopisto. http://hdl.handle.net/10138/307582

Branch, M. & Saukkonen, P. & Niemikorpi, A. (1980). *A Student's Glossary of Finnish*. Werner Söderström Osakeyhtiö.

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

Holopainen, S. (2022). Entlehnte Verben arischen Ursprungs. Teoksessa Holopainen, S. & Junttila, S. (toim.), *Die alten arischen und baltischen Lehnwörter der uralischen Sprachen* (s. 22–57). Münchener Studien zur Sprachwissenschaft, Beiheft 33. J. H. Röll.

Häkkinen, K. (1992). Das etymologische Gesamtbild des finnischen Wortschatzes. Teoksessa L. Honti, S.-L. Hahmo, T. Hofstra, J. Jastrębska & O. Nikkilä, *Finnisch-ugrische Sprachen zwischen dem germanischen und dem slavischen Sprachraum* (s. 25–35). Rodopi.

Häkkinen, K. (2020). *Nykysuomen etymologinen sanakirja*. Päivitetty laitos. Kielikone. www.sanakirja.fi/fin_etymology

Häkkinen, K. (2023). Mikael Agricolan sanakirja. Teoksessa K. Häkkinen & T. Toropainen (toim.), *Mikael Agricola, Turku ja Suomi* (s. 188–203). Mikael Agricola -seura.

Häkkinen, K. (2025). *Nykysuomen etymologinen sanakirja*. MOT Kielipalvelu.

Häkkinen, K. & Toropainen, T. (2025). *Agricolan sanakirja*. Mikael Agricola -seura & Turun yliopisto. https://edition.fi/mikaelagricolaseura

Junttila, S. (2015). *Tiedon kumuloituminen ja trendit lainasanatutkimuksessa. Kantasuomen balttilaislainojen tutkimushistoria*. Väitöskirja, Helsingin yliopisto. https://hdl.handle.net/10138/158777

Jussila, R. & Nikunen, E. & Rautoja, S. (2002). *Suomen murteiden taajuussanasto*. VAPK-kustannus, Kotimaisten kielten tutkimuskeskus.

Kallio, P. (2006). On the Earliest Slavic Loanwords in Finnic. Teoksessa J. Nuorluoto (toim.), *The Slavicization of the Russian North. Mechanisms and Chronology* (s. 154–166). Helsinki University Press.

Kim, J. (2019). *Hulisemisesta hulinaksi. Onomatopoieettisuuden haalistuminen suomen fonesteemisten substantiivien valossa*. Väitöskirja, Helsingin yliopisto. http://hdl.handle.net/10138/304891

Kylstra, A. & Hahmo, S.-L., Hofstra, T. & Nikkilä, O. (1992–2012). *Lexikon der älteren germanischen Lehnwörter in den ostseefinnischen Sprachen*. I–III. Rodopi.

Niemikorpi, A. (1991). *Suomen sanaston dynamiikkaa*. Väitöskirja, Acta Wasaensia No 26, Kielitiede 2. Universitas Wasaensis.

Nykysuomen sanakirja (1951–1961). Valtion toimeksiannosta teettänyt Suomalaisen Kirjallisuuden Seura. Werner Söderström Osakeyhtiö.

Paunonen, H. (2016). *Sloboa stadissa. Stadin slangin etymologiaa*. Docendo.

Rätsep, H. (2002), Eesti kirjakeele tüvevara päritolu. Teoksessa H. Rätsep. *Sõnaloo raamat* (s. 59–77). Ilmamaa.

Saukkonen, P. & Haipus, M. & Niemikorpi, A. & Sulkala, H. (1979). *Suomen kielen taajuussanasto*. Werner Söderström Osakeyhtiö.

Suomen etymologinen sanakirja. 2024. Kotimaisten kielten keskuksen verkkojulkaisuja 72. URN:NBN:fi:kotus-202259. Päivitetty 15.8.2024 https://kaino.kotus.fi/suomenetymologinensanakirja

Suomen sanojen alkuperä 1–3. (1992–2000). Päätoim. E. Itkonen & U.-M. Kulonen. Suomalaisen Kirjallisuuden Seura. Kotimaisten kielten tutkimuskeskus.

Tolnai, V. (1928) Halhatlan magyar nyelv. Teoksessa Kosztolányi, D. (toim.) *Vérző Magyarország – Magyar írók Magyarország területéért* (s. 52–62). [Pallas.] https://turul.info/napok/halhatlanmagyarnyelv