Suomen perussanaston etymologiset kerrostumat

KAISA HÄKKINEN

Omaperäisen aineksen osuutta suomen kielen sanastossa on monissa yhteyksissä luonnehdittu huomattavan suureksi (esim. Vesikansa 1978: 15; Joki 1989: 12), ja omaperäisyyden leimaa vahvistaa se tosiseikka, että useilla kansainvälisillä sanoilla on suomessa kotoisista aineksista muodostettu käypä vastine (esimerkkejä ks. Hakulinen 1979: 474–475). Sanojen lainaamiseen on suhtauduttu torjuvasti etenkin 1800-luvulla ns. varhaisnykysuomen kaudella, jolloin luotiin pohjaa suomen nykyiselle yleis- ja kirjakielelle. Uutta sanastoa muodostettiin korostetun tietoisesti omaperäisistä aineksista sekä johtaen että yhdistäen, ja myös kielessä ennestään käytössä olleita vierasperäisiä aineksia yritettiin raivata pois mm. kehittämällä omaperäistä tieteen, taiteen ja yhteiskuntaelämän terminologiaa (ks. esim. Setälä 1921, erit. 85–91). Tästä huolimatta vieraiden kielten vaikutus on sanastossamme havaittavissa sekä selvinä lainoina että omaperäisestä poikkeavien sananmuodostusmallien noudatteluna. Sanoja on lainattu kaikkina aikoina, ja erityisesti indoeurooppalaisten kielten voidaan osoittaa vaikuttaneen suomeen ja sitä edeltäviin kielimuotoihin koko sen ajan, jonka osalta kehitystä voidaan tieteellisesti tutkia.

Suomen kielen omaperäisen ja lainatun sanaston kvantitiivisia suhteita on toistaiseksi pyritty valaisemaan vain muutaman otostutkimuksen avulla. Knut Cannelin on vuoden 1931 Virittäjässä esittänyt suomalais-ruotsalaisen taskusanakirjansa aineistoon perustuvan laskelman, jonka mukaan 26 000 hakusanan joukossa on yhteensä 3341 kantasanaa ja näistä lainoja 675. Tämän perusteella Cannelin ilmoittaa lainojen osuudeksi n. 2 prosenttia, mutta luku on selvästi virheellinen. Kuten Cannelin itsekin on myöhemmin huomauttanut, olisi prosentin pitänyt olla 20,2 (Hakulinen 1979: 479). Lauri Hakulinen selostaa Cannelinin laskelmia Suomen kielen rakenne ja kehitys (SKRK) -teoksessaan ja täydentää niitä omalla tutkimuksellaan, jonka mukaan n. 2000 sanan laajuisessa tekstissä (puoleksi poliittissävyistä proosaa, puoleksi Otto Mannisen runoja) omaperäisen sanaston osuus on n. 85 prosenttia. Lainojen osuus tekstissä siis on pienempi kuin sanastossa, ja Haku-

linen toteaa tämän olevan odotuksenmukaista: vanhimmat omaperäiset sanat nimeävät yleensä kielen keskeisimpiä käsitteitä, jotka toistuvat tekstissä tämän tästä. Sanakirjoissa sekä keskeiset että marginaalisemmat elementit ovat siinä mielessä tasa-arvoisessa asemassa, että kukin niistä mainitaan vain kerran.

Käsiteltyään eri lainasanakerrostumia Hakulinen ilmoittaa niiden yhteydessä mainittujen esimerkkisanojen määräksi noin 1350 tai 1400 (1979: 479; kaksi lukua samalla sivulla), mutta toteaa samalla lainojen määrän todellisuudessa olevan huomattavasti suurempi, sillä esimerkkiluetteloihin ei ole otettu erikoisalojen kielenkäyttöön tai pelkästään murteisiin rajoittuvaa sanastoa, ei liioin äänneasultaan kieleen sulautumatonta nuorta vierassanastoa. Cannelinin arviota omaperäisten ja lainattujen sanojen suhteesta hän pitää uskottavana ja olettaa tämän perusteella, että omaperäisten kantasanojen määrä suomen yleiskielessä olisi n. 5400.

Cannelinin tutkimus on vuodelta 1931, kuten yllä on jo todettu. Hakulisen kuuluisa SKRK on ilmestynyt ensimmäisen kerran kahtena niteenä vuosina 1941 ja 1946, ja vaikka teoksen myöhempiä laitoksia monissa suhteissa onkin uudistettu ja ajantasaistettu, ovat sanaston rakennetta koskevat laskelmat säilyneet muuttamattomina. Jo tämän perusteella on selvää, etteivät ne voi luotettavasti heijastaa sanahistorian tutkimuksen tämänhetkistä tilaa. Tarkistamiseen olisi aihetta, sillä etymologian alalla on viime aikoina tapahtunut paljon: erityisesti vanhojen germaanisten lainojen tutkimus on 1970-luvun alusta lähtien vilkastunut, ja myös uusia balttilaisia ja slaavilaisia lainaetymologioita on esitetty koko joukko. Indoeurooppalaisten ja suomensukuisten kielten vanhimpia kontakteja on arvioitu uudelleen. Tutkimus on paljastanut myös ennestään tunnistamattoman lainakerrostuman, jota edustavat levikiltään suppeat mutta äänneasultaan huomattavan arkaistiset lainasanat. Nämä todistavat suomalaisten esi-isien joutuneen indoeurooppalaisten naapurien vaikutuspiiriin Itämeren alueella huomattavasti aikaisemmin kuin kielitieteessä koko 1900-luvun alkupuolta hallinneen ns. uudisasutusteorian puitteissa on pidetty mahdollisena. Lainakerrostuman identifioinnin ja ns. jatkuvuusteorian kielitieteellisen perustelun suhteen läpimurtona on pidettävä Jorma Koivulehdon artikkeleita »Seit wann leben die Urfinnen im Ostseeraum» (1983a) ja »Suomalaisten maahanmuutto indoeurooppalaisten lainojen valossa» (1983b). Sekä uudisasutusteoria että jatkuvuusteoria ovat alun perin syntyneet arkeologian piirissä, ja siellä ensin mainittu on saanut vanhentuneena väistyä jo selvästi aikaisemmin kuin kielitieteessä (ks. esim. Huurre 1979, erit. 137—148, Meinander 1984).

Cannelinin ja Hakulisen edellä mainittuja tutkimuksia koskevat selostukset ovat niin ylimalkaisia, että niiden perusteella on mahdotonta päätellä,

miten tekijät ovat suhtautuneet väistämättä eteen tulleisiin ongelmatapauksiin. Sekä Cannelin että Hakulinen puhuvat koko ajan sanoista ja laskevat sanoja. Sanat ovat morfologiselta rakenteeltaan kuitenkin erilaisia: ne voivat olla joko jakamattomia perussanoja, johdoksia tai yhdyssanoja. Yhdyssanaan sisältyy aina vähintään kaksi vapaata morfeemia, ja nämä voivat olla keskenään eri alkuperää ja hyvinkin eri-ikäisiä (joissakin tapauksissa sama morfeemi voi esiintyä kahteenkin kertaan, esim. *pojanpoika*). Hakulinen ilmoittaa laskeneensa sanoja, mutta ei kerro, millä tavoin johdokset ja yhdyssanat on käsitelty. Johdoksista on saatettu ottaa huomioon vain niiden kanta eli vapaata morfeemia edustava osa. Toisaalta olisi myös mahdollista luokitella lainavartalosta omaperäisen suffiksin avulla muodostettu johdos omaperäisen sananmuodostusprosessin tulokseksi ja laskea se tämän nojalla omaperäisten sanojen kategoriaan, ei enää lainaksi. Yhdyssanoista voi määritellä jokaisen rakenneosan alkuperän erikseen tai sitten laskea koko sanan vasta suomessa yhdistämällä syntyneeksi.

Cannelin ilmoittaa laskeneensa kantasanoja, joilla hän tarkoittanee yksimorfeemisia, jakamattomia perussanoja. Sananmuodostusmorfologian nykyisessä terminologiassa tosin kantasanaksi on tapana nimittää sitä sanaa, josta tietty johdos välittömästi voidaan selittää muodostetuksi, mutta tästä ei Cannelinin artikkelissa voi olla kysymys. Tällainen välitön kantasanahan voi olla morfologiselta rakenteeltaan hyvinkin mutkikas: esim. verbi *ajanmukaistaa* on muodostettu kantasanasta *ajanmukainen*. Cannelin ei kuitenkaan ole erikseen puuttunut siihen ongelmaan, että suomessa samoin kuin muissakin kielissä on runsaasti sanavartaloita, johtomorfologian kannalta ajatellen kantoja tai kantavartaloita, jotka eivät reaalistu jakamattomina perussanoina lainkaan, ainoastaan johdosten ja yhdyssanojen rakenneosina. Esimerkiksi adjektiivit *kevyt* ja *keveä* ovat johdoksia samasta vartalosta kuin esimerkiksi verbi *keventää* ja substantiivi *keveys*, mutta yhteistä kantasanaa, kielessä yksimorfeemisena reaalistuvaa lekseemiä näillä ei ole.

Cannelin ilmoittaa laskeneensa kaikki *-eA*-loppuiset adjektiivit kantasanoiksi puuttumatta siihen ongelmaan, ettei tämän tuntomerkin avulla muodostettu ryhmä ole morfologisesti tai etymologisesti homogeeninen. Osa *-eA*-loppuisista sanoista on lainoja, jotka suomen näkökulmasta ovat ainakin etymologisesti jakamattomia perussanoja. Tällaisia ovat esim. *huokea*, *lakea*, *nopea*, jotka ovat lainaa suffiksia muistuttavaa loppuosaansa myöten. Melkoinen osa *-eA*-loppuisista adjektiiveista voidaan kuitenkin selittää omaperäisiksi johdoksiksi sen nojalla, että niiden vartalolle löytyy vastineita sukukielistä ja että samasta vartalosta on olemassa myös muita johdoksia, esim. *keveä* edellä tai *pimeä*, *pireä*, *sakea* ym. Vastaavalaisia ongelmia löytyy myös muista kuin *eA*-loppuisista sanoista. Esimerkiksi *-e'*-loppuisista sanoista osa

on jakamattomia, lainaperäisiä lekseemejä (*aarre*, *herne*, *käärme*), osa taas johdoksia (*jänsi* > *jänne*, *kuuma* > *kuume*, *lauta* > *laude*). Pelkästään nykykielisten vastineiden tarkastelu ei riitä sen seikan selvittämiseksi, onko kysymys alun perin johdoksesta vai johdoksen näköisestä jakamattomasta perussanasta, sillä johtimelta näyttävä loppuosa on voitu myöhemmin irrottaa ja korvata muilla johtimilla (vrt. esim. *huokea* — *huojeta*, *huojentaa*, *huojistaa* ym.).

Sanaston etymologista rakennetta selvitettäessä ei voi lähteä siitä oletamuksesta, että jokaisen sanueen ytimenä ja edustajana olisi morfologisesti jakamaton perussana. Analyysi on suoritettava useassa vaiheessa siten, että ensin on tutkittava valitun korpuksen jokaisen sanan morfologinen rakenne ja etsittävä aineistoon sisältyvät kantavartalot, esiintyivätpä ne itsenäisenä perussanana tai kompleksisen lekseemin rakenneosana. Varsinainen etymologinen tutkimus voidaan sitten kohdistaa näin löydettyihin kantavartaloihin. Morfologisten rakennetyyppien suhteen voi tutkia erikseen, samoin sen, muodostetaanko kompleksisia lekseemejä yhtä lailla omaperäisistä kuin lainatuistakin vartaloista.

Kielen koko sanaston tutkiminen yhdellä kertaa on käytännössä mahdotonta, joten on tyydyttävä otostutkimukseen. Oman tarkasteluni kohteeksi olen valinnut aineiston, jonka Michael Branch, Pauli Saukkonen ja Antero Niemikorpi ovat koonneet sanakirjaansa A Student's Glossary of Finnish. Tekijöiden empiiristen havaintojen mukaan tämä sanasto on funktionaalisti keskeistä siten, että sen turvin on mahdollista lukea suomenkielistä kirjallisuutta ja suhteellisen nopeasti alkaa käyttää suomen kieltä suullisesti jokapäiväisissä puhetilanteissa (Branch ym. 1980: 14). Omaa tutkimustani varten olen muokannut aineistoa ainoastaan siten, että erisnimet ja interjektiot on jätetty pois ja mukaan on otettu ne lekseemit, jotka Saukkosen ym. 1979 toimittaman frekvenssisanakirjan mukaan kuuluvat suomen tuhannen yleisimmän sanan joukkoon mutta jotka syystä tai toisesta puuttuvat Glossarysta. Tällä tavoin muodostetun korpuksen lekseemien yhteismääräksi tuli 1888. Aiemmin olen vastavalla tavalla tutkinut suppeampaa korpusta, joka sisältää edellä mainitun frekvenssisanakirjan 1000 yleisintä sanaa (Häkkinen 1985, erit. 150).

Niin sanotut nuoret lainasanat jätetään usein leksikaalisten tutkimusten ulkopuolelle, koska niiden lainaperäisyyden osoittamiseksi ei tarvita erityistä tutkimusta ja koska niiden äänneasussa on yleensä piirteitä, jotka poikkeavat omaperäistä leksikkoa koskevista fonologisista ja fonotaktisista säännöistä. Käytännössä tällaisten nuorten lainojen rajaaminen on kuitenkin hankalaa. Fonologiset ja fonotaktiset säännöt voivat muuttua aikojen kuluessa, ja usein muutoksiin antavat sysäyksen alkuperältään vieraat mutta

ajan mittaan kieleen yhä paremmin integroituvat ainekset. Lainan todellinen ikäkään ei aina käy yksiin sen mielikuvan kanssa, joka sanan vierasperäisyyden asteesta äänneasun perusteella syntyy. Esimerkiksi omaan tutkimusaineistooni sisältyvät sanat *filosofia*, *historia*, *keisari*, *kristillinen* ja *teksti* tunnistetaan yleensä vaikeuksitta lainoiksi, vaikka ne ovat kuuluneet kirjakieleen Agricolasta alkaen. Sanat *artikkeli*, *eversti*, *kenraali*, *presidentti* ja *professori* saatetaan äänneasunsa perusteella lukea nuorten lainojen joukkoon, vaikka ne tunnetaan jo 1600-luvun kirjakielestä. Toisaalta on helppo osoittaa lukuisia sanoja, jotka äänneasultaan sulautuvat paljon paremmin omaperäiseen sanastoon mutta ovat silti varsin nuoria tulokkaita. Esimerkiksi *huumori*, *kilo*, *kuoro*, *litra*, *metri*, *miettiä*[1], *rooli*, *romaani*, *tyyli* ja *tyyppi* ovat tulleet suomen kirjakieleen vasta 1800-luvulla. Omaan tutkimukseeni olen ottanut mukaan kaikki mainitussa sanakirjassa esiintyvät, funktionaalisti keskeisiksi määritellyt sanat niiden mahdollisesta lainaperäisyydesta ja lainauksen iästä riippumatta. Näin ollen on käynyt mahdolliseksi myös selvittää, mikä osuus lainautumisella on ollut keskeisen sanaston kartuttamisessa muutaman viimeksi kuluneen vuosisadan aikana.

Tutkimuksen ensimmäisessä vaiheessa olen analysoinut lekseemit morfologisesti ja järjestänyt ne etymologisesti. Hakusanoiksi on valittu kyseiseen sanueeseen kuuluvan jakamattoman perussanan vartalo tai tällaisen puuttuessa rekonstruoitu kantavartalo. Hakusanan muodolla ei tässä yhteydessä ole merkitystä, se on ainoastaan materiaalin jäsentämistä palveleva apuväline. Kunkin hakusanan kohdalle on erikseen koottu sitä aineistossa edustavat lekseemit, sekä jakamattomat, johdetut että yhdistetyt. Useimmat yhdyssanat tulevat mainituiksi kahden hakusanan kohdalla, mutta tämä on lekseemien määriä koskevissa laskelmissa otettu huomioon. Jokaisesta lekseemistä on kirjattu sanaluokka ja morfologinen rakennetyyppi (perussana, johdos, yhdyssana), kirjakielinen ikä ja levikki. Vartalon ikä ja levikki on kirjattu erikseen hakusanan kohdalle. Vartalon alkuperä on selvitetty niin pitkälti kuin se etymologisen kirjallisuuden nojalla on ollut mahdollista. Etymologisista selityksistä on erikseen kirjattu ne, jotka esiintyvät yleisesti tunnetuissa etymologisissa hakuteoksissa (FUV, SKES, MSzFE, MTESz, UEW), ja ne uudet etymologiat, jotka ovat löydettävissä vain erillisistä artikkeleista ja esitelmistä (näistä suurin osa on Jorma Koivulehdon). Sanojen kirjakielisen iän selvittämisessä on käytetty sekä Kotimaisten kielten tutkimuskeskuksen kokoelmia että omia alkuperäislähteisiin pohjautuvia poimintoja.

[1] Sana esiintyy Kristfrid Gananderin sanakirjan käsikirjoituksessa 1787 asussa *miehtiä*.

Yhdyssanojen määrä tutkitussa aineistossa on suhteellisen vähäinen, vain 133 lekseemiä. Tämä on odotuksenmukaista, sillä vaikka yhdyssanoja absoluuttisesti onkin yleiskielen sanoista suurin osa, keskeisimmässä sanastossa lyhyet lekseemit ovat selvästi suosituimpia. (Sanojen pituuden ja frekvenssin välistä suhdetta on äskettäin selvitellyt perinpohjaisesti Antero Niemikorpi (1991)). Yhdyssanojen osuus riippuu sekä otoksen laadusta että koosta. Mitä suurempaa otosta tutkitaan, sitä marginaalisempia lekseemejä aineistoon tulee mukaan ja sitä suuremmaksi yhdyssanojen osuus kasvaa (ks. esim. Vesikansa 1978: 26). Esimerkiksi Cannelinin edellä mainitussa taskusanakirjassa (26 000 lekseemiä) yhdyssanoja on n. 44 %, Nykysuomen sanakirjan aineistossa (n. 210 000 lekseemiä) n. 65 %. Eroa näyttää olevan myös siinä, tutkitaanko murteita vai yleiskieltä. Yleiskielessä yhdistäminen on suositumpi sananmuodostuskeino kuin murteissa (Tuomi 1989: 28–29).

Jakamattomien perussanojen ja johdosten rajaaminen on vaikeampaa kuin ennalta voisi kuvitella. Typologisesti suomea pidetään yleensä agglutinoivana kielenä, jossa vartalon ja affiksien rajat näkyvät selvästi, ja useimmissa tapauksissa asia onkin näin. Ongelmia koituu kuitenkin esimerkiksi lainoista, etenkin lainaperäisistä verbeistä, joissa vieraan kielen vartalo on sopeutettu suomen morfologiseen systeemiin lisäämällä siihen sinänsä merkityksetön mutta johtimelta näyttävä aines (*mainitse-* < germ. **mainjan-*). Myös omaperäisiltä näyttävien sanojen joukosta löytyy vartaloita, jotka historiallista fonotaksia koskevien tietojen valossa eivät voi olla jakamattomia perusvartaloita mutta joita ei varsinaisesti ole todistettu johdoksiksikaan. Tällaisia ovat esim. pelkästään itämerensuomeen rajoittuvat verbit *eksyä*, *syöksyä*, *tuoksua*. Mikäli sekvenssi *-ksU-* käsitettäisiin johtimeksi, verbit voitaisiin selittää vanhastaan tunnettujen verbinvartaloiden johdoksiksi (kieltoverbi *e-*, *syö-*, *tuo-*; viimeksi mainitun osalta on huomattava, että *tuoksua*-verbin aistihavaintoon liittyvä merkitys on myöhäinen, alkuperäisempi on SKES:n mukaan 'pöllytä, tupruta; heilimöidä' ym.), mutta näin ei etymologisessa kirjallisuudessa toistaiseksi ole tehty.

Omassa tutkimuksessani olen laskenut johdoksiksi kaikki ne lekseemit, joiden johdosperäisyydestä on positiivisia todisteita. Todisteeksi on kelpuutettu omaperäisten vartaloiden osalta saman vartalon rinnakkaisjohdokset tai pelkän vartalon etymologiset vastineet sukukielissä. Omaperäisten sanojen segmentoinnissa on noudatettu SKES:n kantaa. Lainasanoissa perussanaksi on katsottu originaalia lähinnä vastaava muoto. Johdoksiksi ei ole tulkittu niitä lekseemejä, joissa johtimen tapainen aines palvelee pelkästään vartalon morfologista sopeuttamista eikä vartalo reaalistu erillisenä, yksinkertaisempana lekseeminä.

Ongelmatapauksia jää vielä tämänkin jälkeen, sillä osa etymologioista on

epävarmoja tai kiistanalaisia. Näin ollen johdosten ja perussanojen määrää ei voi ilmoittaa yhdellä luvulla. Jakamattomien perussanojen osuudeksi näyttää muodostuvan 29—32 prosenttia ja johdosten määräksi 61—64 prosenttia. Morfologisten tyyppien suhde poikkeaa selvästi Cannelinin (12 : 44 : 44) ja Nykysuomen sanakirjan aineiston (8,6 : 26,6 : 64,8; Niemikorpi 1991: 154) vastaavista suhteista, mutta tätä on pidetävä odotuksenmukaisena, kun muistetaan tutkitun aineiston keskeisyys ja sen tästä johtuva poikkeuksellinen luonne.

Aineistoon sisältyvät lekseemit jakautuvat kirjakielisen ikänsä suhteen eri vuosisatojen osalle seuraavasti:

(1)	1500-luku	58 %	1096	lekseemiä
	(Agricolalla	56 %	1056	lekseemiä)
	1600-luku	9 %	167	lekseemiä
	1700-luku	13 %	245	lekseemiä
	1800-luku	18 %	343	lekseemiä
	1900-luku	2 %	37	lekseemiä
yht.		100 %	1888	lekseemiä

Kuten edellä olevasta taulukosta käy ilmi, on enemmän kuin puolet nykysuomen keskeisimmistä sanoista ollut käytössä jo suomen vanhimmassa kirjakielessä. Agricolaa voidaan myös tämän tilaston valossa pitää suomen kirjakielen isänä; muiden 1500-luvun lähteiden osuus on aivan vähäinen. 1800-luvun sanastonuudistus ei prosentuaalisesti arvioiden yllä lähellekään Agricolan saavutusta. Huomiota kannattaa kiinnittää myös 1900-luvun uusien keskeisten lekseemien vähäiseen määrään. Tätä edeltävien kolmen vuosisadan aikana vallinut kasvusuuntaus on kääntynyt laskuksi, mikä on selvä merkki keskeisimmän sanaston vakiintumisesta 1800-luvun lopulle tultaessa. Toisaalta tietysti uusien, joskin suhteellisen harvojen lekseemien ilmaantuminen vasta 1900-luvun kuluessa todistaa, ettei kielen keskeisinkään sanasto pysy täysin muuttumattomana.

Tutkitussa suppehakossa aineistossa enimmät kantavartalot reaalistuvat vain yhden lekseemin välityksellä. Tästä huolimatta produktiivisimmat vartalot erottuvat selvästi. Suosituimmuuslistan alkupää näyttää seuraavalta:

(2)	vartalo	levikki/lähde	yhdistämättömien lekseemien määrä
	esi	suom.-ugr.	21
	yksi	suom.-ugr.	17
	se	ural.	16
	koke-	< ieur. (JK)	14

(2)

vartalo	levikki/lähde	yhdistämättömien lekseemien määrä
näke-	suom-ugr.	13
o(le)-	ural.	13
tosi	< esigerm. (JK)	13
erä	< ieur. (JK)	12
jo- (pron.)	suom.-volg., < ieur.?	11
keski	suom.-perm.	11
mi- (pron.)	ural.	11
perä	suom.-ugr., ?ural.	11
tunte-	ural.	11
tuo (pron.)	ural.	11
ilma	suom.-ugr.	10
ku- (pron.)	ural., < ieur.?	10
käy-	< germ. (JK)	10
muu	suom.-ugr.	10
pää	ural.	10
ase-	ural.	9
osa	< arj.	9
pise-	< suom.-volg.	9
täve-	< germ. (JK)	9
ylä	suom.-ugr., ?ural.	9
ala	ural.	8
oppe-	suom.-ugr.	8
taka	ural.	8
tie	suom. perm.	8

(JK = Jorma Koivulehdon etymologia)

Huno Rätsep on tutkinut vastaavalla tavalla, joskin huomattavasti laajemmasta aineistosta, viron kantavartaloiden produktiivisuutta, ja edellä oleva laskelma noudattelee samoja linjoja kuin Rätsepin tulokset (ks. esim. 1986). Produktiivisimpien kantojen joukossa ei ole yhtään nuorta lainaa, ja erityisen suosittuja johdosten kantoina näyttävät olevan vanhimpiin omaperäisiin sanastokerrostumiin kuuluvat elementit. Tämä on sinänsä odotuksenmukaista: vanhimmat kielelliset käsitteet nimeävät ihmiselämän keskeisimpiä ilmiöitä, jotka ovat aktuaalisia aina ja kaikkialla.

Sanavartaloiden määrä käsitellyssä aineistossa on merkitsevästi suurempi kuin jakamattomien perussanojen määrä, ts. on runsaasti vartaloita, jotka esiintyvät ainoastaan kompleksisten lekseemien osina. Suurin osa vartaloista on ollut käytössä jo 1500-luvun kirjakielessä, joskaan ei aina juuri sen lekseemin muodossa, joka sisältyy tutkimuksen varsinaisena kohteena olevaan keskeisimpään sanastoon. Vartaloiden kirjakielisen iän selvittämiseksi olen Kotimaisten kielten tutkimuskeskuksen ensiesiintymäarkistosta käynyt läpi paitsi kaikki tutkimusaineistoon sisältyvät lekseemit myös näihin sisältyvien vartaloiden muut, mahdollisesti vanhemmat esiintymät. Vartaloita on yh-

teensä 844, ja niiden ensiesiintymät jakautuvat eri vuosisatojen osalle seuraavasti:

					lainoja
(3)	1500-luku	81,2 %	685	vartaloa	
	1600-luku	7,1 %	60	vartaloa	60 %
	1700-luku	6,6 %	56	vartaloa	45 %
	1800-luku	4,5 %	38	vartaloa	63 %
	1900-luku	0,6 %	5	vartaloa	100 %

Laskelmasta näkyy, että "uusien" keskeiseen sanastoon kuuluvien vartaloiden määrä on jatkuvasti vähentynyt, ja yli puolet "uusista" vartaloista on lainattuja. Vartaloiden inventaari näyttää pysyvän vakaampana kuin reaalisten lekseemien joukko. Erityisesti kannattaa panna merkille, että 1800-luvun sanastonuudistuksen merkitys näyttää keskeisen vartaloinventaarin osalta vähäiseltä. Tämä voidaan tulkita siten, että enin osa uusista sanoista muodostettiin vanhojen vartaloiden pohjalta.

Toinen kysymys on, ovatko laskelmassa uudelta näyttävät vartalot todella uusia suomen kielessä. Omaperäisistä vartaloista eivät absoluuttisesti uusia voi ilmeisestikään olla ne, joilla on vastineita sukukielissä. Periaatteessa olisi tietysti mahdollista, että suomen sana ei olisikaan etymologisesti yhteinen sana vaan myöhempää lainaa sukukielistä, mutta nyt tutkitun aineiston osalta tätä mahdollisuutta ei nähdäkseni tarvitse ottaa lukuun. Ylipäänsä tilanne näyttää yksinkertaiselta siinä mielessä, että kaikki sukukielten vastineiden perusteella uralilaisiksi tai suomalais-ugrilaisiksi katsottavat sanavartalot ovat olleet mukana jo 1500-luvun kirjakielessä. Ainoat poikkeukset ovat 1600-luvulla ilmaantuneet *koivu* ja *suksi*, joiden puuttuminen 1500-luvun voittopuolisesti uskonnollisesta kirjallisuudesta on semanttis-funktionaalisista syistä aivan ymmärrettävää.

Jos jätetään pois kaikki sukukielten nojalla omaperäisiksi katsottavat ja toisaalta selvästi lainaperäisiksi selittyvät vartalot, jää jäljelle vielä kourallinen omaperäisiltä näyttäviä vartaloita, joiden esiintyminen rajoittuu pelkästään suomeen. Osa näistä voidaan selittää äänteellisesti motivoiduiksi, deskriptiivisiksi tai onomatopoeettisiksi vartaloiksi (*sävV-* > *sävy*, *vauva*; ensiesiintymä 1700-luvulla). Muutamat ovat kuitenkin toistaiseksi jääneet vaille minkäänlaista etymologista selitystä. Nämä ovat 1600-luvulla kieleen ilmaantunut *vane-* (> *vankka*), 1700-luvun *heti*, *ihta* (> *ihan*) ja *pelkkä* sekä 1800-luvun tulokas *aito*. Näistä huolimatta voidaan koko aineiston perusteella selvästi todeta, että enin osa kirjakieleen »uusina» ilmaantuneista sanavartaloista on lainaperäisiä.

Koko aineiston sanavartaloista on 410 eli 49 % mitä ilmeisimmin omaperäisiä. Tähän lukuun sisältyvät myös ns. protoeurooppalaiset elementit, ts. sellaiset sanavartalot, joiden vastineet rajoittuvat itämerensuomeen ja joiden on arveltu lainautuneen Baltian alueen tuntemattomiksi jääneistä alkuperäiskielistä (ks. tarkemmin esim. Ariste 1981). Vartaloista 389 (46 %) on varmoja lainoja, joskaan kaikissa tapauksissa ei ole mahdollista sitovasti osoittaa, mistä (indoeurooppalaisesta) kielimuodosta sanavartalo on lainattu. Kiistanalaisia tapauksia jää 45 (5 %).

Lainattujen vartaloiden määrä on selvästi suurempi kuin Cannelinin laskelmissa ja Hakulisen arvioissa (ks. edeltä). Lainavartalot jakautuvat eri kerrostumiin seuraavasti (varmat ja epävarmat etymologiat on ilmoitettu erikseen):

			etymologioista	
(4)	indoeurooppalaiset ja		vanhoja	uusia
	arjalaiset	38 + 20?	16 + 10?	22 + 10?
	vanhat germaaniset	141 + 38?	70 + 27?	71 + 11?
	nuoremmat germaaniset	63 + 13?		
	balttilaiset	46 + 7?	28 + 5?	18 + 3?
	slaavilaiset/venäläiset	19 + 1?		
	yleiseurooppalaiset	61		

On huomattava, että varmojen lainojen yhteen laskettu määrä tässä laskelmassa ei ole sama kuin varmasti lainaperäisiksi tiedettyjen vartaloiden määrä yleensä, sillä monista sinänsä varmoista lainoista on mahdotonta sanoa, mihin kerrostumaan ne kuuluvat (esim. *varsi* voi olla balttilainen tai germaaninen laina).

Omaperäiset vartalot on käytännöllisintä kerrostaa niiden levikin nojalla, ja kuva muodostuu seuraavanlaiseksi:

(5)	uralilaiset	61 + 23?
	suomalais-ugrilaiset	59 + 47?
	suomalais-permiläiset	12 + 26?
	suomalais-volgalaiset	23 + 26?
	varhaiskantasuomalaiset	21 + 26?
	itämerensuomalaiset	134 + 52?
	suomalaiset	9 + 3?

Varmojen tapausten yhteen laskettu määrä ei tässäkään laskelmassa ole sama kuin omaperäisten vartaloiden summa, sillä monien vartaloiden levikin täsmällinen määrittely on joidenkin sukukielten vastineiden epävarmuuden takia mahdotonta, vaikka vartalon omaperäisyyttä sinänsä ei olisikaan syytä asettaa kyseenalaiseksi.

Omaperäisten vartaloiden määrä ei näiden laskelmien valossa näytä ainutlaatuisen suurelta (vrt. Hakulinen 1979: 479 ja tässä mainitut tutkimukset). Huomiota herättää itämerensuomeen rajoittuvien etymologioiden runsaus. Osa tähän kerrostumaan kuuluvista sanoista on varmasti äänteellisesti motivoituja uudismuodosteita, ja näistä jotkut voivat olla myös rinnakkaiskehityksen tulosta. Kaikkia ei kuitenkaan voi edes pyrkiä selittämään tällä tavoin, joten ilmeiseltä näyttää, että joukkoon sisältyy vielä tunnistamattomia lainoja. Uudet sanavartalot eivät ilmesty kieleen itsestään, vaan niillä on jokin järkevä etymologinen selitys. Useimmiten tämä selitys on vanhan, kompleksisen sanan hämärtyminen perussanan kaltaiseksi, lainautuminen tai äänteellinen motivaatio (ks. tarkemmin Häkkinen 1990, 86 alk.).

Lainasanatutkimuksen viimeaikainen kehitys antaa aihetta olettaa, että juuri lainautumisen osalta selitysmahdollisuuksia ei ole ammennettu vielä läheskään tyhjiin. Nyt tutkitun keskeisimmän sanaston lainaetymologioista peräti 111 eli n. 30 % on uusia siinä mielessä, etteivät ne ole ehtineet mukaan tätä tutkimusta laadittaessa käytettävissä olleisiin etymologisiin sanakirjoihin. Suurin osa näistä uusista selityksistä koskee vanhoja germaanisia lainoja, joiden määrä on kaksinkertaistunut (70 → 141). Lisäksi on tietysti esitetty vielä runsaasti sekä germaanisia että muita lainaetymologioita, jotka eivät ole tulleet esille nyt tutkitussa aineistossa. Uusia omaperäisiä etymologioita ei sen sijaan juurikaan ole pystytty esittämään lisää. Etymologisen tutkimuksen painopiste näyttää viime aikoina kallistuneen vahvasti lainasanatutkimuksen puolelle.

LÄHTEET

Ariste, Paul 1981: Keelekontaktid. Eesti keele kontakte teiste keeltega. Tallinn.

Branch, Michael — Saukkonen, Pauli — Niemikorpi, Antero 1980: A Student's Glossary of Finnish. Porvoo.

Cannelin, Knut 1931: Sanansyntyopin asema äidinkielen opetuksessa. Virittäjä 35 s. 186—195.

Collinder, Björn 1977: Fenno-Ugric Vocabulary. An Etymological Dictionary of the Uralic Languages. 2. painos. Hamburg.

FUV = Collinder 1977.

Gallén, Jarl (toim.) 1984: Suomen väestön esihistorialliset juuret. Bidrag till kännedom av Finlands natur och folk. H. 131. Helsinki.

Hakulinen, Lauri 1979: Suomen kielen rakenne ja kehitys. Neljäs, korjattu ja lisätty painos. Helsinki.

Huurre, Matti 1979: 9000 vuotta Suomen esihistoriaa. Keuruu.

Häkkinen, Kaisa 1985: Suomen kielen sanaston historiallista taustaa. Fennistica 7. Turku.

HÄKKINEN, KAISA 1990: Mistä sanat tulevat. Suomalaista etymologiaa. Tietolipas 117. Helsinki.
JOKI, AULIS J. 1989: Sanastomme perusainekset. — Vesikansa (toim.) 1989.
KOIVULEHTO, JORMA 1983a: Seit wann leben die Urfinnen im Ostseeraum? Zur relativen und absoluten Chronologie der alten idg. Lehnwortschichten im Ostseefinnischen. — Symposium Saeculare Societatis Fenno-Ugricae. Suomalais-ugrilaisen Seuran toimituksia 185. Helsinki.
KOIVULEHTO, JORMA 1983b: Suomalaisten maahanmuutto indoeurooppalaisten lainojen valossa. — Suomalais-ugrilaisen Seuran aikakauskirja 78. Helsinki.
A magyar nyelv történeti-etimológiai szótára 1—3. Budapest 1967—1976.
A magyar szókészlet finnugor elemei 1—3. Budapest 1967—1978.
MEINANDER, C. F. 1984: Kivikautemme väestöhistoria. — Gallén (toim.) 1984.
MSzFE = A magyar szókészlet finnugor elemei 1—3.
MTESz = A magyar nyelv förténeti-etimológiai szótára 1—3.
NIEMIKORPI, ANTERO 1991: Suomen kielen sanaston dynamiikkaa. Acta Wasaensia No 26; Kielitiede 2. Vaasa.
RÄTSEP, HUNO 1986: Eesti kirjakeele sõnatüvede tuletuskoormus. Keel ja kirjandus 11/1986.
SAUKKONEN, PAULI — HAIPUS, MARJATTA — NIEMIKORPI, ANTERO — SULKALA, HELENA 1979: Suomen kielen taajuussanasto. Porvoo.
SETÄLÄ, E. N. 1921: Oikeakielisyydestä suomen kielen käytäntöön katsoen. — E. N. Setälä: Kielentutkimus ja oikeakielisyys. Helsinki.
SKES = Suomen kielen etymologinen sanakirja.
Suomen kielen etymologinen sanakirja I—VII. Y. H. Toivonen (I—II), Erkki Itkonen (II—VI), Aulis J. Joki (II—VI), Reino Peltola (V—VI). Hakemiston (VII) koostaneet Satu Tanner ja Marita Cronstedt. Helsinki 1955—1981.
TUOMI, TUOMO 1989: Yleiskielemme murrepohjainen sanasto. Vesikansa (toim.) 1989.
UEW = Uralisches etymologisches Wörterbuch.
Uralisches etymologisches Wörterbuch. Lieferung 1—7. Toim. Károly Rédei. Wiesbaden 1986 —.
VESIKANSA, JOUKO 1978: Miljoona sanaa. Porvoo — Helsinki.
VESIKANSA, JOUKO (toim.) 1989: Nykysuomen sanavarat. Porvoo — Helsinki.

The etymological strata of the Finnish lexicon

KAISA HÄKKINEN

In an article in Virittäjä in 1931 Knut Cannelin estimated, from the material included in his pocket dictionary, that the basic Finnish vocabulary is about 20 % loanwords. Cannelin's estimate has been cited unchanged in later studies, although his methods of calculation do not entirely meet the demands of etymological research. Moreover, Cannelin's results are now more than 60 years old, so they can by no means reflect the present state of etymological knowledge.

For my own research I have taken the functionally central vocabulary contained in *A Student's Glossary of Finnish*, edited by M. Branch, P. Saukkonen and A. Niemikorpi. The only modifications have been the omission of proper names and interjections, and the inclusion of lexemes which according to the Finnish Frequency Dictionary (Saukkonen et al.) come within the most frequent 1000 words but which are absent from the Glossary (the vast majority of the material is found in both these sources). The total extent of the corpus is 1888 lexemes.

The vocabulary was first analysed morphologically and grouped etymologically in such a way that each word-

stem is accompanied by all the lexemes derived from it. The age of the lexemes in the standard language was determined via the archives of the Research Centre for Domestic Languages. Table 1 in the article gives the distribution of the first occurrences of the lexemes by century. This shows that over half of the most frequent words in modern Finnish were already present in Mikael Agricola's texts.

The material contains 844 stem-forms altogether, most of which are realised in a single lexeme only. The greatest number of complex words are formed from the old original stems *esi* 'pre' and *yksi* 'one'. Table 2 shows the beginning of the productivity ranking of the word-stems; in addition to old original words there are also some loanwords belonging to the older strata.

An examination of the age of word-stems in the standard language indicates that the inventory of the most central stems renews itself more slowly than the set of actual lexemes. Table 3 shows that over four-fifths of the most frequent stems in modern Finnish were already in use in the standard language of the 16th century. The same calculation also shows that over half of the new stems appearing in the standard language have been loans.

Of the total number of word-stems in the material, 410 seem to be clearly original (49 %), although a precise age cannot always be determined because of the partial uncertainty of the equivalents in related languages. Indisputable loans amount to 389 (46 %), debatable ones 45 (5 %). The proportion of loans is clearly higher than Cannelin's estimate.

Because the loans have come from Indo-European languages that were themselves inter-related, it is not always possible to pinpoint the donor language unambiguously. Table 4 illustrates the distribution of loan-stems according to language of origin; certain and uncertain loans are given separately.

Original stems are grouped according to area of occurrence, from most extensive to most restricted; in the absence of other criteria this can also be interpreted as the order of age from oldest to youngest. The distribution is given in Table 5. The largest individual stratum is the Baltic-Finnic one, which suggests that this stratum contains many hitherto unrecognized loans. Phonologically unmotivated basic words do not spring up from nowhere, and a large number of new loan etymologies have recently been proposed in precisely the area of vocabulary restricted to Baltic-Finnic. This is also to be expected since contacts with both Baltic and Germanic languages are known to have been particularly lively during the shared Baltic-Finnic period.

My research deals separately with etymologies contained in well-known reference works and recent individual studies. Although the research only covers the most central vocabulary of modern Finnish, a total of 844 word-stems, the number of new etymologies is as high as 110. In practice this means that a third of the loan etymologies are new. The greatest increase is in the number of old Germanic loans (from 70 to 141). On the other hand, not one new etymology extending to distantly related languages could be suggested for any of the Baltic-Finnic words in the corpus. The focus of research into word history is clearly shifting towards loanword research.