VILHO SETÄLÄ

Tilastollisia tietoja Uuden testamentin suomennoksen sanastosta

Sanastotilastollinen työni, johon olen saanut avustusta Suomalaiselta Tiedeakatemialta ja Opetusministeriöltä ja jonka tietokonekäsittelyn on tehnyt mahdolliseksi Yliopiston Suomen kielen laitos yhteistoiminnassa Sovelletun matematiikan laskentakeskuksen kanssa, tähtää ensitavoitteena eräiden tietojen selvittämiseen UT:n suomennoksen sanastosta. Tässä esitän ne tulokset, jotka ilmenevät välittömästi tietokonekäsittelyn antina.

Sanastollista

Tilastollisen käsittelyn vuoksi on välttämätöntä pitää erillään *sane* (pienimmät erilliset kielelliset ilmaukset) ja *sana* (käsite, sanakirjallinen merkitys, jota edustavat samaan paradigmaan kuuluvat saneet). Käyttäessäni ilmausta *sane* tarkoitan aina muodoltaan ja merkitykseltään toisistaan eroavia saneita; kun on kysymys saman saneen eri esiintymien lukumäärästä, käytän sanaa *useus*.

Sanapesyeen (johdossarjan) yhteistä perusosaa nimitän lyhyyden vuoksi *juureksi*; tätä ilmaisua ei ainakaan tässä yhteydessä tarvita muuhun merkitystehtävään. *Johdoksiksi* luetaan tässä tutkimuksessa sellaiset sanat, jotka suomen kirjakieltä hallitseva tajuaa johdetuiksi, ja juuriksi ne, jotka hän tajuaa johtamattomiksi, riippumatta siitä mikä tilanne kielen historiallisen kehityksen valossa on oikea.

Työn lähtökohdaksi on valittu Kaarlo Hartevan julkaisema UT:n aakkosellinen hakemisto. Tämä siitä käytännöllisestä syystä, että hakemisto sisältää viittauksen UT:n jokaisen saneen kaikkiin esiintymiin, alkaen tavallisimmasta saneesta *ja* (useus 9354) aina niihin lähes 1000 sanaan asti, jotka esiintyvät kukin vain kerran yhden ainoan saneen edustamina. Täten on kortistoa varten ollut tarpeen vain laskea kunkin saneen useus. Tämä hakemisto on toiselta puolen erinomainen lähtökohta, koska se on tehty v. 1913 hyväksytyn ns. Ahon komitean laatimasta suomennoksesta, joka on suomennoksista lähinnä nykykieltä.

On todettava, että hakemisto on miltei virheetön. Kun siinä ei kuitenkaan ole mitenkään erotettu toisistaan suomen kielen lukuisia homonyymejä, on ollut välttämätöntä tarkistaa tuhansia esiintymiä itse UT:sta. Tällöin on hakemistosta löytynyt vain joitakin merkityksettömiä kompastuksia, epäjohdonmukaisuutta ison alkukirjaimen käytössä jms. tulokseen vaikuttamatonta. Ainoat silmään osuneet harhaviittaukset ovat 1) *täytyi*-saneen toinen esiintymispaikka; sivulta 400 on nähtävästi pudonnut yksi rivi, jolla tämä on mainittu, ja edellisestä rivistä näkyy vain viittaus

Pietarin toiseen kirjeeseen, josta se sitten löytyikin; 2) *sitten*-saneen rinnakkaismuotona esiintyy *sitte*, mutta UT:sta ilmenee, että tämä onkin vain *saisitte*-saneen (toiselle riville siirtynyt) jatke ja siis jätettävä huomiotta.

Reikäkortit

Laskentaa varten on tehty kaksi reikäkorttisarjaa.

Ensimmäiseen sarjaan lävistettiin kutakin *sanaa* varten kortti, josta käy ilmi sanan *juuri* sekä siihen sisältyvät *johtimet* kirjaimin. Numeroin merkittiin eräiden indeksien ohella mm. sanan pituus tavuina, sanaluokan tunnus, sen muotojen (saneiden) luku ja näiden yhteisuseus.

Koska näistä korteista ei ollut mahdollista saada selville saneiden pituutta koskevia tietoja, kun näet samaan paradigmaan yleensä kuuluu eripituisia saneita, valmistettiin toinen kortisto, johon kustakin *saneesta* merkittiin tiedot mm. saneen pituudesta tavuina, sen sanaluokasta ja useudesta.

Korttien lävistämisen ohjeita laadittaessa on inhimillisen erehdyksen vaara aina lähellä. Tarkkuuden mittana voimme pitää sitä, että kaikkien saneiden yhteisuseus edellisestä kortistosta (jossa erehtymisen vaara, varsinkin homonyymien kohdalla, oli suurempi) saatiin 0,7 % pienempänä kuin jälkimmäisestä.

Tuloksia

Jälkimmäisestä, tarkemmasta laskelmasta saimme tuloksena, että UT:n kaikkien saneiden yhteisuseus on 132 748 ja että saneita (erilaisia) on 19 343. Edellisestä kortistosta saimme sanojen lukumääräksi n. 6 300 ja näiden juurien luvuksi n. 2 360. Sanomme 'noin', koska kummankin lukumäärä riippuu siellä täällä harkinnasta — esimerkiksi, onko saneet *alla*, *alta*, *alle* luettava saman sanan muodoiksi vai erillisiksi sanoiksi.

Saneiden lukumäärä ja pituus tavuina sanaluokittain

Taulukosta 1 selviää, kuinka monta sanetta kustakin sanaluokasta on pituudeltaan 1, 2, · · · 12 tavua sekä kunkin pituisten saneiden summa, niin myös kuhunkin sanaluokkaan luettavien saneiden summa, absoluuttisena ja prosentteina.

Sanaluokista huomattakoon, että erisnimet on erotettu omakielisistä substantiiveista; verbaalinomineihin on luettu kaikki ne johdokset, joiden sanaluokkaa on mahdoton todeta sanakirjallisesta merkinnöstä. Ryhmään 'muita' on luettu mm. adverbit. Näistä mainittakoon johdetut ja päätteelliset adverbit: johdin -sti, -ttain, -isin (113 sanetta/useus 433), nominipäätteiset (371/5463) ja kieltosanan verbinen paradigma (54/3014), yhteensä 538 sanetta, useus 8910.

Taulukosta 2 selviävät samat seikat kaikkien saneiden *useudesta*.

Näihin taulukkoihin sisältyvistä luvuista voidaan laskea saneiden keskipituus sekä sanakirjalliselta pohjalta (erilaisten saneiden lukumäärästä) että koko tekstin pohjalta (kokonaisuseudesta). Saamme täten taulukon 3.

Taulukko 1. Saneiden pituus sanaluokittain

Tavuja Silab	Subst	Propr	Adj	Pronom	Num	Verb	Vnom	Cet	Sum
1	32	9	0	13	0	51	0	29	134
2	1135	166	221	210	40	957	95	233	3057
3	2535	338	636	188	47	2128	847	316	7035
4	1904	313	504	63	51	1282	1038	135	5290
5	1289	123	342	8	41	291	490	38	2622
6	468	46	119	0	11	45	149	18	856
7	160	5	37	0	11	4	42	2	261
8	36	2	8	0	2	3	8	0	59
9	15	0	5	0	2	0	2	0	24
10	3	0	1	0	0	0	0	0	4
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	1	0	0	0	1
13	0	0	0	0	0	0	0	0	0
Σ saneita	7577	1002	1873	482	206	4761	2671	771	19343
Σ tavuja	28349	3549	7168	1289	819	15254	10603	2295	69326
Saneita %	39,0	5,2	9,7	2,5	1,1	25,0	13,5	4,0	100 %

Taulukko 2. Saneiden useus pituuden ja sanaluokan mukaan

Tavuja Silab	Subst	Propr	Adj	Pronom	Num	Verb	Vnom	Cet	Sum
1	479	20	0	6960	0	3929	0	16475	27863
2	7403	1095	1723	16924	662	11693	1281	11995	52776
3	10653	1918	1888	3132	281	8743	3016	2652	32283
4	4943	1086	1121	258	205	2567	2176	590	12946
5	2528	349	612	16	120	418	690	64	4797
6	898	127	151	0	21	58	193	79	1527
7	282	6	56	0	21	5	66	2	438
8	49	2	15	0	2	3	10	0	81
9	19	0	7	0	3	0	2	0	31
10	3	0	1	0	0	0	0	0	4
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	2	0	0	0	2
13	0	0	0	0	0	0	0	0	0
Σ useudet	27257	4603	5574	27290	1317	27416	7434	31857	132748
Σ tavut	87611	14873	18145	51316	3927	66309	25482	51589	319252
Useus %	20,5	3,5	4,2	20,5	1,0	20,7	5,6	24,1	100 %

Taulukko 3. Saneiden keskipituus

	a) sanekohtainen	b) tekstikohtainen
Substantiivit	3,7	3,2
Erisnimet	3,5	3,2
Adjektiivit	3,8	3,2
Pronominit	2,7	1,9
Lukusanat	3,9	3,0
Verbit	3,2	2,4
Verbinominit	4,0	3,5
Muut	3,0	1,6
Kaikki saneet	3,58	2,41

Lopuksi olen tehnyt pienen pistokokeen arvioidakseni saneiden keskipituuden yksikkönä kirjain. Tätä varten olen merkinnyt tavujen pituudet Apostolien tekojen 11. luvusta, jakeista 8—18 ja laskenut näistä tavun keskipituuden kolmesta sadan tavun ja yhdestä 50 tavun perättäisestä ryhmästä sekä koko tavustosta. Viimeksi mainitussa tapauksessa on otettu huomioon myös *äänteiden* luku. Kertolaskulla saamme siten arvion saneen keskipituudeksi kirjaimina ja äänteinä (pitkä vokaali = yksi äänne). Taulukko 4.

Taulukko 4. Tavun ja saneen keskipituus

	a) kirjaimina	b) äänteinä
1. sata tavua	2,50	
2. sata tavua	2,57	
3. sata tavua	2,59	
+ 50 tavua	2,68	
350 tavua	2,57	2,49
saneen (2,41 tavua) keskipituus	6,19	6,00

Ottaen huomioon tavun pituuden pienen vaihtelun voimme katsoa suomenkielisen saneen keskipituudeksi 6,2 kirjainta eli 6,0 äännettä.

Vokaaliharmonia

Suomen kielen vokaaliharmonian kannalta saamme kiintoisaa tietoa taulukosta 5. Taulukko perustuu useuslukuihin — muullahan ei ole sanottavaa merkitystä. Yllättävältä tuntuu, että äänteiden mukaan laadittu tilasto (kirjoitelmani Virittäjässä n:o 2, 1951) osoittaa etuvokaaleille lievää ylivoimaa, kun tässä ilmenee, että saneista (kaikkine esiintymineen) peräti 62 % on takavokaalisia. — Sekavokaalisiksi on luettu yhdyssaneet ja vieraat nimet, joissa esiintyy sekä etu- että takavokaalisia tavuja, neutraalisiksi taas ne, joissa esiintyy ainoastaan i ja e.

Taulukko 5. Saneiden etu- ja takavokaalisuus

	Subst	Propr	Adj	Pron	Num	Verb	Vnom	Cet	Sum	%
Etuvok	5526	55	1630	13768	592	4956	1818	8090	36435	27,5
Takavok	18756	4393	3670	7791	616	21274	5467	20312	82279	62,0
Sekavok	826	88	109	1	68	17	55	101	1265	0,96
Neutr	2149	67	165	5730	41	1169	94	3354	12769	9,6
Σ	27257	4603	5574	27290	1317	27416	7434	31857	132748	100,0

Vortara statistiko pri la Nova Testamento en finna traduko

DE VILHO SETÄLÄ

La objekto estas finna traduko de 1913, kiu prezentas relative modernan formon de finna literatura lingvo. La tuta teksto enhavas 132748 vortojn (aparte skribatajn grupojn de literoj) el 19 343 diversaj vortoj. La nombro de glosoj (paradigmoj), al kiuj la vortoj apartenas, estas ĉ. 6 000, kaj iliaj radikoj ĉ. 2360.

Tabelo 1 montras la nombron de malsamaj vortoj en ĉiu vortklaso, dividite laŭ longeco de la vortoj (en silaboj), en absolutaj nombroj kaj procentoj. *Tabelo 2* donas la respondajn faktojn pri la frekvencoj (oftecoj) de la vortoj. En *tabelo 3* estas kalkulita la meza longo de la vortoj en ĉiu vortklaso, a) kalkulite baze de diversaj vortoj, b) kalkulite laŭ la frekvencoj.

Por proksimuma taksado de la vortlongeco a) en literoj, b) en sonoj (longa vokalo skribata per 2 vokalsignoj) estas kalkulita en *tabelo 4* el mallonga tekstprovo la meza longeco de silabo (el tri grupoj po 100 silaboj kaj unu po 50 silaboj kaj el tuto de 355 silaboj). La rezultoj — 2,57 literoj aŭ 2,49 sonoj — multiplikite per la silaba longeco 2,41 donas la mezan vortlongecon de 6,2 literoj aŭ 6,0 sonoj.

Kaŭze de la vokalharmonio en finna lingvo, kiu ne permesas aperon de antaŭaj vokaloj (ä, ö, y) en sama vorto kiel la malantaŭaj (a, o, u), dum e kaj i estas neutralaj, estas kalkulitaj en *tabelo 5* la oftecoj de la vortoj antaŭvokalaj kaj malantaŭvokalaj en ĉiuj vortklasoj, absolute kaj procente. La miksvokalaj (»sekavok») estas aŭ kunmetitaj vortoj aŭ fremdlingvaj nomoj.