Unkarin sanoja uudessa järjestyksessä

Reverse-Alphabetized Dictionary of the Hungarian Language — A magyar nyelv szóvégmutató szótára. Compiled by FERENC PAPP. Akadémiai kiadó. Budapest 1969.

H. Grassmannin 1873 julkaisemaan Rigvedan sanakirjaan (Wörterbuch zur Rig-Veda, Leipzig) sisältyy perinteisen sanakirjaosan lisäksi luettelo, johon sanat on järjestetty sananlopusta lähtevän aakkostuksen mukaisesti. Tämä lienee ensimmäinen tieteellisiin tarkoitusperiin tähtäävä käänteissanakirja; samalle periaatteelle rakentuvia ns. riimisanakirjoja on mahdollisesti julkaistu aikaisemminkin. Vastaavanlaisen sanasto-osan sisällytti C. Bartolomae Altiranisches Wörterbuchiin (Strasbourg 1904), ja samana vuonna ilmestyi ensimmäinen kielen koko sanaston käsittävä käänteissanakirja, O. Gradenwitzin Laterculi vocum latinarum: Voces latinas et a fronte et a tergo ordinandas (Leipzig 1904). Tämän jälkeen ovat ilmestyneet ainakin seuraavat käänteissanakirjat: P. Kretschmer, Rückläufiges Wörterbuch der Griechischen Sprache (Göttingen 1944), C. D. Buck—W. Petersen, Rewerse Index of Greek Nouns and Adjectives (London 1944), L. Sadnik—R. Aitzetmüller Handwörterbuch zu den altkirchenslavischen Texten (La Haye — Heidelberg 1955), Romanian akatemian toimittama, ensimmäinen nykykielestä tehty käänteissanakirja Dictionar Invers (Bucarest 1957), H. H. Bielfeldt, Rückläufiges Wörterbuch der russischen Sprache der Gegenwart (Berlin 1958), R. Greeve — B. Kroesche (Max Vasmerin johdolla), Rückläufiges Russisches Wörterbuch (Berlin—Wiesbaden 1958—1959), M. L. Alinei, Dizionario Inverso Italiano (La Haye 1962), A. F. Brown, Normal and Reverse English Word List (Philadelphia 1963), A. Juilland, Dictionnaire inverse de la langue française (Haag 1965), E. Mater, Rückläufiges Wörterbuch der deutschen Gegenwartssprache (Leipzig 1965). Luetteloon voidaan nyt lisätä ensimmäinen suomalais-ugrilaisia kieliä koskeva käänteissanakirja, ja uranuurtajan kunnian saa Ferenc Pappin Reverse-Alphabetized Dictionary of the Hungarian Language (Budapest 1969).[1]

Käänteissanakirjan idea on siis lähtöisin indoeurooppalaisten ja klassillisten kielten piiristä, mutta epäilemättä tällaisen sanakirjan tarjoamat tutkimusmahdollisuudet on oivallettu monilla tahoilla, erityisesti agglutinoivia derivaatiokieliä tutkivien piirissä. Toteutuksen vaivalloisuus vain on ollut toimiin ryhtymisen esteenä. ATK-menetelmien kehitys ja niiden sovellukset kielenainesten käsittelyyn ovat ratkaisevasti muuttaneet tilannetta, ja tietääkseni kaikki viimeaikaiset käänteissanakirjat on laadittu tietojenkäsittelykoneita käyttäen. Näin on myös Pappin sanakirjan samoin kuin

[1] Leksikaalisia aineksia esittelevänä Pappin teos on todella ensimmäinen tämänlaatuinen työ kielikuntamme piirissä. On kuitenkin huomattava, että v. 1968 ilmestyi Hampurin yliopiston Societas Uralo-Altaic -sarjassa V. H. Veenkerin Verzeichnis der ungarischen Suffixe und Suffixkombinationen a tergo geordnet. Veenkerin teos on nimensä mukaisesti pelkkien suffiksien ja suffiksikombinaatioiden a tergo -luettelo. Se ei siis sisällä muita kielenaineksia eikä laskelmia siitä, missä määrin esim. suffiksikombinaatioilla on realisoitumia kielessä. — Meillä Vilho Setälä on tutkinut ja luetteloinut monipuolisesti Uuden testamentin sanastoa, vrt. Vir. 1967 s. 368—.

parhaillaan tekeillä olevan suomen kielen käänteissanakirjankin laita.

Pappin teos sisältää yksityiskohtaisen unkarin- ja englanninkielisen johdannon (s. 7—45), sanaluettelo-osan koodeineen (s. 49—537) sekä joukon käsiteltyä ainesta koskevaa statistiikkaa, mm. unkarin sananloppuisten foneemien esiintymistaajuutta esittävän taulukon (s. 541—544), kahden, kolmen ja neljän sananloppuisen (kirjain)sekvenssin esiintymiä ja hierarkiaa (s. 545—588), kirjoitettujen sanojen pituutta (s. 589), sanojen merkitysmäärän jakaumaa (s. 590—591) koskevat laskelmat ja määrätavalla laaditun tyyliarvon mukaisen sanaluettelon (s. 592—594) eli varsin paljon perusluonteista tietoa.

Sanakirjan pohjana on A magyar nyelv értelmező szótára (ÉrtSz), jonka hakusanat on kaikki otettu mukaan. Käsiteltyjen sanojen kokonaismäärä on 58 323. Varsinainen sanasto-osa esittelee hakusanat oikeinpäin mutta sananlopusta alkaen aakkostetussa järjestyksessä siten, että sananloput ovat samalla pystysuoralla sarakkeella, minkä ansiosta teos muuten poikkeaa edukseen kaikista edeltäjistään. Lisäksi kaksipalstaiseen sivuun sisältyy seitsensarakkeinen (A—G) koodiosa, joka koostuu yksinumeroisesta kolminumeroiseen vaihtelevasta koodista. A-pylvääseen on kooditettu sanan rakenne: sana koostuu yhdestä, kahdesta jne. kannasta, kannasta ja prefiksistä jne. Sarakkeessa B on esitetty sanaluokka (Szófaj ~ Part of Speech) 15 erilaisella koodilla. Tavanomaisen sanaluokkajaon lisäksi on käytetty muutamaa uuttakin kategoriaa. Koodi 0 tarkoittaa, ettei aineksella ole lainkaan sanaluokkaa, koodi 11, että kyseessä on lausesana (sentence word ~ egyeb mondatszó) ja koodi 999, että kyseinen kielenaines kuuluu johonkin mainitsemattomaan ryhmään. Sarakkeessa C on nominien taivutusvartalot (10 erilaista koodia). Sarakkeissa D, E ja F on substantiivien ja substantiivisten pronominien taivutuspäätteiden kooditus, ja samoihin sarakkeisiin on sijoitettu myös adjektiivien ja niiden kaltaisten pronominien taivutuksen, vokaaliharmonian, adverbiaalimuotojen ja komparaation koodit. Unkarin monimuotoinen verbintaivutus on kooditettu pylväisiin B, C, D_1, D_2, E ja F.

Kooditettu informaatio perustuu osaksi suoraan ÉrtSz:n tarjoamiin tietoihin (B-sarake), osaksi koostajan ja hänen työryhmänsä analysoimiin tosioihin (A- ja G-sarakkeet) ja osaksi ÉrtSz:n kannanottoihin, joita tekijä ja työryhmä ovat paikoin analysoineet toisin (C-, D-, E- ja F-sarakkeet). Lisäksi työn pohjana olleeseen reikäkorttimateriaaliin on kooditettu muutakin tietoa aineksesta, mutta sen käsittely ja tulokset on säästetty myöhemmin ilmestyviä erillistutkimuksia varten. Sanakirjaosan palsta näyttää esimerkiksi seuraavalta:

	A	B	C	D	E	F	G
		A					
A^1	1	2	1	03	03	02	
A^2	1	2	1	03	03	02	
A^3	1	522		00	00	00	
A^4	1	10		00	00	00	
A^5	1	10		00	00	00	
BABA	1	2	2	03	03	02	
PRÓBABABA	2	2	2	03	03	02	
FABABA	2	2	2	03	03	02	
FÖLDBABA	2	2	2	03	03	02	
KUGLIBABA	2	2	2	03	03	02	
JÁTÉKBABA	2	2	2	03	03	02	2
ALVÓBABA	2	2	2	03	03	02	2
KISBABA	2	2	2	03	03	02	
RONGYBABA	2	2	2	03	03	02	
BÁBA	1	2	2	03	03	02	
FELIBE-HARMADÁBA	2	6		00	00	00	9
HIÁBA	1	6		00	00	00	9

Mitä tietoa hankalilta tuntuviin numerosarjoihin sisältyy? Sanasta *hegyközség* 'viininviljely-yhdyskunta' voi koodia koskevaa selvitystä seuraten todeta, että siihen sisältyy kaksi sanakantaa (A 2), se on substantiivi (B 2), vartalo on vaihtelematon (C 1), yksikön akkusatiivi

	A	B	C	D	E	F	G
hegyközség	2	2	1	04	04	04	1
sötétedik	1	101	31	20	00	00	1
örök	1	326	1	05	05	45	
hal[1]	1	1	1	10	00	00	

muodostetaan suffiksilla *-et* (D 04), monikon nominatiivi suffiksilla *-ek* (E 04), yksikön 3. p. possessiivisuffiksina sanassa käytetään *-e*:tä (F 04) ja että sana on johdos (G 1). Vastaavasti verbistä *sötétedik* 'pimetä' ovat tiedot seuraavat: yhdistämätön (A 1), yksipersoonainen verbi (B 101), konjugaatio ja vokaaliharmonia epätäydellinen (D_2 E F 0 00 00), sanan lopussa on johdin (G 1). *örök* 'ikuinen': yhdistämätön (A 1), adjektiivi, substantiivi ja adverbi (B 326), vaihtelematon vartalo (G 1), taivutus substantiivina tyyppiä *-öt*, *-ök*, *-e* ~ *-je* (D, E, F 05 05 45) eikä sanaan liity johdinta (G 0). *hal*[1] 'kuolla', joka vastaa ÉrtSz:n hakusanaa *hal*[1], on yhdistämätön (A 1), täydellisesti taipuva verbi (B 1), monikon toisen persoonan pääte, samoin preteritin yksikön ensimmäisen ja kolmannen persoonan pääte liittyy ilman sidevokaalia vartaloon ja konditionaalin *n* ja imperatiivin *j* liittyvät samoin suoraan vartaloon (C 1), sana on takavokaalinen (D_1 1), se ei vaadi mitään määritteitä (D_2, E, F 0 00 00) eikä sisällä johdinta (G 0).

Johdannon päätteeksi tekijä esittelee eräitä sanakirjan käyttömahdollisuuksia ja tähdentää erityisesti sitä, että tavanomaisesta sanakirjasta poiketen myös kooditus muodostaa käyttökelpoisen hakuaparaatin. Esim. neliosaiset yhdyssanat löytyvät sarakkeessa A olevan 4:n kohdalta. Sanat, joilla on useamman sanaluokan piiriin kuuluvaa käyttöä, ovat saaneet sarakkeeseen B kaksi- tai useampinumeroisen koodin. Siten esimerkiksi adjektiivina, adverbina ja substantiivina esiintyvällä sanalla on kooditus B 362 jne. Tämä on kaikki tietoa, jonka hankkiminen ei ole vaikeaa perinnäisistä kielenkuvauksista, normaalimuotoisista sanakirjoista ja kieliopeista. Kiintoisampia näkymiä sen sijaan avaa esim. kysymys, onko mitään säännöllisyyttä siinä, että *b*-loppuisten substantiivien yksikön 3. persoonan omistusliitteinen muoto on joko *j*:llinen, tai *j*:tön (*láb* : *lába* ~ *comb* : *combja*)? Mahdollinen säännönmukaisuus on löydettävissä siten, että *b*-loppuisten substantiivien ryhmästä (substantiivin koodi B 2) ryhmitetään yhteen tapaukset, joissa F-sarakkeessa on koodi 1 (*-a*) tai 4 (*-e*) ja toisaalta tapaukset, joissa F-sarakkeessa on koodi 2 (*-ja*) tai 5 (*-je*). Kyseisessä tapauksessa ryhmittely käy nopeasti, koska *b*-loppuisia sanoja on kaikkiaan vain neljän palstan verran, yhteensä noin 250 sanaa ja näistäkin vain vajaat kaksi kolmannesta substantiiveja. Tämänlaatuiset taivutuksen ja etenkin derivaation lainalaisuuksien ja anomalioiden selvittelymahdollisuudet ovat epäilemättä käänteissanakirjojen merkittävin anti kielentutkimukselle.

Tietokonekäsittelyn yhteydessä on suoritettu myös joitakin statistisia laskelmia, joiden tulokset on esitetty liitteissä. Eniten tilaa on uhrattu sanaloppuisille ykkös-, kakkos-, kolmos- ja nelosleikkauksien (final di-, tri- ja tetragrams) tuloksille, ts. tapauksille, joissa on laskettu sananloppuisten 1—4-jäsenisten sekvenssien esiintymistaajuudet. Tulostuksesta on ilmoitettu absoluuttinen ja prosenttinen määrä sekä kumulatiivinen prosentti. Ykkösleikkaus on tehty foneemeista, muut kirjainsekvensseistä. Sananloppuisista foneemeista yleisin on ʃ (esiintymiä kaikkiaan 9639; 16,53 %) ja harvinaisin *z* (2; 0,00 %.)[2] Erikseen on laskettu sananloppuisten vokaalien taajuudet ja tulokset esitetään monipuolisesti taulukoituna. Todettakoon vain, että yleisin sananloppuinen vokaali on *a* (4201; 36,6 %) ja harvinaisin *ö* (2; 0,02 %). Lyhyiden takavokaalien osuus sananlopussa on 37.90 % ja pitkien 28.25 % eli yhteensä

[2] Kirjoittaja on käyttänyt foneemien merkinnässä kansainvälistä fonologista merkintää (IPA-transcription). ʃ vastaa grafeemia *s* ja *z* *zz*:ää.

takavokaaleja 66.15 %. Etuvokaalien kokonaisosuudeksi jää siis 33.85 %, ja näistä 25 % on pitkiä, 8.85 % lyhyitä. Lyhyiden vokaalien osuus sananlopussa on 46.75 % ja vastaavasti pitkien 53.25 % jne. Näitä kuten muitakin teoksen statistisia tuloksia käytettäessä on syytä muistaa, että ne pohjautuvat sanakirjallisesta perusmuodosta tehtyihin laskelmiin. Puheen tai tekstin virrasta tehdyt laskelmat antaisivat olennaisesti toisenlaisia arvoja; niihin vaikuttavat sanojen ja suffiksien vaihtelevat esiintymätaajuudet. Toisaalta sanakirjallisesta perusmuodosta tehdyissä laskelmissa eräillä taivutuspäätteillä, suomessa esim. I infinitiivin päätteellä, on korosteisen voimakas vaikutus. Eräitä vähäisiä poikkeuksia (*heläjää, humajaa*) lukuun ottamatta verbit lisäisivät koko määrällään jäännöslopukkeen esiintymiä. Enemmän informaatiota tarjoaisivat laskelmat, joissa tietyllä tavalla ryhmitellen laskettaisiin verbeistä $-a^x \sim ä^x$ -sekvenssiä edeltävien äänteiden taajuus. Se on kuitenkin puolestaan sidoksissa verbien vartalotyyppien frekvensseihin. Ts. laskentaperusteiden yksityiskohtaisempi ennakkoanalyysi olisi epäilemättä lisännyt tulosten käyttöarvoa, mutta eräänlaisena konstituenttien statistiikkana sanakirjallisesta perusmuodosta kaavamaisesti tehdyt laskelmatkin ovat hyödyllisiä.

Liitteessä 2 Papp esittelee unkarin sanojen jakauman pituuden mukaan. Kuten on odotettavissakin jakauman kuvaaja on lähes Gaussin kellokäyrän mukainen. Runsaimmin unkarissa on kahdeksankirjaimisia sanoja (8662; 14.85 %; kumulat. % 50.51), sitten yhdeksänkirjaimisia (8267; 14.17 %; k. % 64.68) ja edelleen seitsenkirjaimisia (7702; 13.21 %; k. % 35.65).

Monessakin mielessä kiintoisalta vaikuttaa sanojen merkityksen määrää koskevan jakautuman selvittely. Siinä on pyritty laskemaan, kuinka monella unkarin sanalla on yksi, kaksi jne. merkitystä. Laskelmat perustuvat ÉrtSz:n kannanottoihin; merkitykseksi on hyväksytty siis ÉrtSz:n erillisiksi merkityksiksi esittämät. Yksimerkityksisten sanojen osuus koko sanavarastosta on puolet (50.37 %), yksi- ja kaksimerkityksisten yhteensä kaksi kolmannesta (76.87 %) ja yksi — kolmimerkityksisten sanojen yhteisosuus 93.78 %. 25 tai useampia merkityksiä omaavat sanat luetellaan erikseen. Niitä on kaikkiaan 46. Runsaimmin merkityksiä (yht. 101) on noteerattu partikkelille *is* (perusmerkitys 'myös'), ja yleensäkin partikkelien osuus on tässä ryhmässä huomattava. Kuriositeettina todettakoon, että esimerkiksi adjektiivit *szabad* (perusmerkitys 'vapaa') ja *szépen* (perusmerkitys 'kaunis') ovat joutuneet tähän ryhmään. Vastaavanlaisia laskelmia on H. H. Josselsson tehnyt venäjästä (Automatization of Lexicography. Abstract of paper presented at the 1965 International Conference on Computational Linguistics, New York). Kun Josselsson ei toistaiseksi ole yksityiskohtaisemmin julkaissut tuloksiaan, on Papp, voidakseen verrata unkarin arvoja jonkin muun kielikunnan kieleen, tehnyt käsin [!] laskelmia venäjästä. Pohjana on ollut Ušakovin sanakirja (Толковый словарь языка. Москва 1934—1940). Näiden laskelmien mukaan yksimerkityksisten sanojen osuus venäjässä on peräti 78.26 %, yksi- ja kaksimerkityksisten 93.90 %, ja kolmimerkityksisten kumulatiivinen % on 97.88. Vaikka jakauman kuvaajien yleishahmo on likipitäen sama (s. 501), ovat erot yksi- ja kaksimerkityksisten distribuution osalta unkarin ja venäjän välillä huomattavat, siten että venäjän sanastollinen entropia on merkittävästi suurempi kuin unkarin.

Kirjan päättää luettelo, johon on ryhmitelty normaalista poikkeavan tyyliarvon omaavat sanat tyylialan mukaan. Luettelossa on slangin, lastenkielen, deskriptiivisten, onomatopoieettisten ja uudissanojen ryhmät. Mukaan on otettu vain sellaiset sanat, joissa tyyliä koskeva luonnehdinta koskee koko sanaa, ei siis esim. sellaista yleiskielen sanaa, jota jossakin merkityksessä käytetään slangissa.

Pappin sanakirja tuskin kuluu kansan käsissä, mutta jokaisen unkaria tutkivan kirjastoon se epäilemättä kuuluu jo sisältämänsä ensiarvoisen tietoaineksen vuoksi. Teoksen tarjoamat moninaiset tutkimusmahdollisuudet ehkä toteuttavat koostajan toivomuksen, että sanakirjaa voitaisiin käyttää sellaisiinkin tarkoituksiin, joita hän ei ole osannut kirjaa laatiessaan edes kuvitella. Toiveen toteutuminen riippuu ensisijaisesti tietysti tekijän kuvittelukyvystä; sinkoileehan täällä miehen ja miesten aatokset.

Tuomo Tuomi