

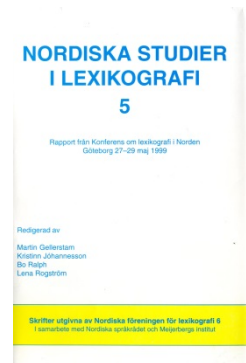
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Lagring av ordbok över äldre finskt skriftspråk i strukturerad form

Forfatter: Riitta Palkki

Kilde: Nordiska Studier i Lexikografi 5, 2001, s. 271-280
Rapport från Konferens om lexikografi i Norden, Göteborg 27.-29. maj 1999

URL: <http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive>



© Nordisk forening for leksikografi

Betingelser for brug af denne artikel

Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Søgbarhed

Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for 'optical character recognition' og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Riitta Palkki

Lagring av ordbok över äldre finskt skriftspråk i strukturerad form

The Dictionary of Old Finnish (VKS, *Vanhan kirjasuomen sanakirja*) is written using a structural program, FrameMaker+SGML (SGML = Standard Generalized Markup Language). The program is used in Unix environment. A Frame document consists of parts called elements, which are hierarchically organized in fields according to the Document Type Definition (DTD). In one field there can be several parallel or subordinate fields. The basic structure of the articles in VKS is as follows: index word with its variants, word class code, definition, data from old dictionaries, examples with sourcecodes, list of compound words. When writing an article for the VKS, several windows are needed. The window "Elements" contains all the elements of the dictionary. The "Structure View" shows a graphic outline of the document's structure. The Document window is used when writing the text. Other windows, "Terminal" and "Text Editor", which do not belong to the Frame program, are also available. The written text can also be viewed on the screen in SGML format. When finished, the dictionary can be used in research for example as a text corpus.

Bakgrund

Perioden för *äldre finskt skriftspråk* började med *Mikael Agricolas Abc-bok* ungefär år 1540 och räckte till slutet av svenska tiden, dvs. ungefär till år 1810. Materialet för *Ordbok över äldre finskt skriftspråk* (*Vanhan kirjasuomen sanakirja*, i fortsättningen kallad VKS) utgörs av tryckalster från den tiden och omfattande handskrifter, till exempel finska översättningar av lagar, som senare har blivit tryckta.

Ordbokens första del (bokstäverna *a* till *i*), som utkom år 1985, skrevs på gammaldags skrivmaskiner. Den andra delen (bokstäverna *j* till *k*) skrevs med ordbehandlingsprogrammet *WordPerfect* och kom ut år 1994. När arbetet på den tredje delen inleddes övergick vi samtidigt till ett program för strukturerad redigering, *FrameMaker+SGML* (i fortsättningen kallad *Frame*).

Varför Frame?

Frame fungerar i *Unix*-miljö, så redaktörerna har terminaler anslutna till en *Unix*-centralenhet. Bildskärmen fungerar som *arbetsbord*. Programmen öppnas som s.k. *arbetsfönster* på arbetsbordet. Utöver Frame erbjuder *Unix* flera användbara hjälpprogram för ordboksarbete. När det gäller VKS är det främst *Terminal* och *Text Editor* som är viktiga.

En strukturerad redigering av VKS diskuterades för första gången redan för ett tiotal år sedan. Då var ordboken exempelprojekt i ett forskningsprogram som gällde strukturering av texter och behandlingen av strukturerade texter. En av ledarna för projektet var *Erja Nikunen*, dåvarande chef för adb-avdelningen vid Forskningscentralen för de inhemska språken. Hon utarbetade en s.k. *kontextoberoende grammatik* för VKS, som i hög grad påminde om strukturbeskrivningen för Frame, och ordboken blev av denna orsak pilotprojekt för Frame vid forskningscentralen.

Det fanns flera praktiska skäl till att det nya systemet togs i bruk. Forskningscentralens *textkorpus* med äldre finskt skriftspråk omfattar över 2,5 miljoner ord, och en betydande del av nyckeltexterna i ordbokens stora källmaterial föreligger i elektronisk form. Det var därför viktigt att göra det så lätt som möjligt att utnyttja korpusmaterialet. En annan orsak var att man ville att redaktörerna skulle slippa tänka på sådana tekniska detaljer som maskinen sköter felfritt. Det var vidare önskvärt att ordboken skulle få en sådan form att den å sin sida skulle gå att utnyttja för så mångsidiga sökningar som möjligt. Av allt detta följde att artiklarnas struktur måste göras mer konsekvent än tidigare. När man skrev på maskin eller wp behövde så stor uppmärksamhet inte fästas vid att strukturen var absolut logisk.

Lagring av ordboken på Frame

Ett Frame-dokument består av delar som kallas *element*. Dessa ordnas i enlighet med den logiska *strukturbeskrivningen* för en ordboksartikel som *fält* som ligger i följd, jämsides eller innanför

varandra, så att de tillsammans bildar textens stomme. Med tanke på SGML (förkortning för *Standard Generalized Markup Language*) talas det om en DTD (*Document Type Definition*), en *strukturgrammatik*, som klargör vilka element som är tillåtna och i vilken ordning dessa kan användas (se t.ex. Lehtinen / Karvonen / Rahikainen 1995:55). Strukturbeskrivningen för VKS utarbetades ursprungligen av Erja Nikunen. Redaktörerna har i över två års tid prövat och finlipat strukturen tillsammans med forskningscentralens adb-expertter, och nu börjar det fungera någorlunda felfritt. Strukturbeskrivningen har lagrats i *master*-filen, som tas fram som underlag för varje nytt dokument.

1 2 3 4

hyypiö, -ä s. huuhkaja (S *hyypiä* bubo, buff, der
 Franck | VR *hyypiö* bubo, wff | F *hyypiä* = VR |
 J, JllsK *hyypiä* aboens. ulula, uf; lisäksi J *hyyp-*
piö | G 223a *hyypiä*, *hyypiö*, *hyypiä* aboens. uuf,
 ulula; stenuggla; ~ *huhkaja*) | A III 359 (B¹⁻⁴ Ps 5
 102:7) Mine olen ninquin iocu Hypie njisse häuite-
 tyis Caupungeissa (A I 300 hwhkaija) | ib. 707 (B¹
 Sef 2:14) Cacarit ia Hypiet pite heiden Tornisans
 asuman (B²⁻⁴ tarhapöllöit) | B¹⁻⁴ 3 Ms 11:17 ei
 teidän pidä syömän [lintuja], cuin owat -- Kiwi-
 rautia, Luicoi, Hyypiä | B¹ 5 Ms 14:16 Hyypiö,
 Riutahyypiö, Yöcköi (B² Linduhaucka, Hyypiö,
 Yöcköi, B¹⁻⁴ Huhkaja, Hyypiö, Yöcköi) | Pet-
 raeus 1649 A3a *hyypiä*, *hyypiän*, bubo | Pacch
 1667 N3b [Monien pitää olla] niijn quin Hyypiän
 nijsä häwitetyisä Caupungeis | RWL 1759 106
 (RWL 1808 112) haucan, *hyypiän*, cockolinnun, ja
 muita raatelewia linduja, mahta joca mies syytöin-
 nä ambua (Forseen 1738 137 tarhapöllön) | Ran
 1803 35 Se pyhä maa ei ole -- kärmetten koto eikä
hyypiän pesä. — *Yhd.* maki-, riutahyypiö.

6

BILD 1. Strukturen för ordartikeln i VKS.

Strukturen för ordartiklarna i VKS är i huvudsak följande (som exempel ordet *hyypiö* (berguv, *Bubo bubo*) (bild 1): (1) *uppslagsordet* med varianter; (2) *ordklassangivelse*; (3) *betydelsebeskrivning*; (4) uppgifter ur *gamla ordböcker*; (5) *exempel* med källhänvisningar; och (6) förteckning över *sammansatta ord*. Det finns många variationsmöjligheter, och en ordartikel behöver inte nödvändigtvis innehålla alla dessa delar. Ordklassangivelse och

betydelsebeskrivning kan saknas. Om betydelsebeskrivningen är indelad i moment, kommer uppgifterna ur gamla ordböcker genast efter ordklassförkortningen. Betydelsemomenten markeras med arabiska siffror och undermomenten med små bokstäver. Romerska siffror och, i undantagsfall, stora bokstäver används för större övermoment. Olika *användningsområden*, till exempel *erik.*, som betyder specialbetydelse, specialanvändning, *kuv.* (bildligt), *lak.* (i lagspråk), kan bilda egna underavdelningar, liksom stående uttryck och andra *sublemman* (*alihakusana*). Anvisningarna för artikelförfattarna har på sin tid gjorts vaga och flytande, och det har därför inte varit alldeles lätt att skapa en fungerande strukturbeskrivning utgående från dem.

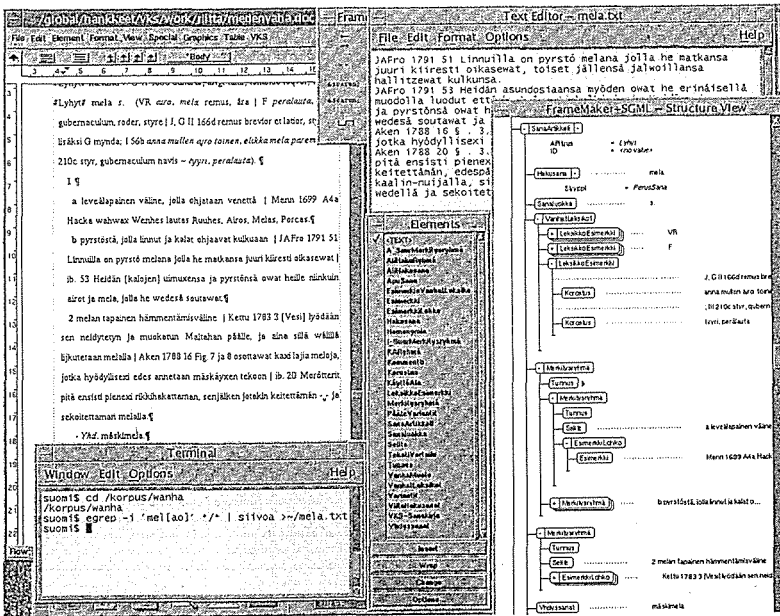


BILD 2. Arbetsfönster i VKS.

I arbetet på VKS behövs flera olika *arbetsfönster* (bild 2). Fönstret *Elements* innehåller alla de ungefär trettio element som förekommer i ordboken. I fönstret *Structure View* finns ett *struk-*

turdiagram, som åskådliggör den inbördes hierarkin mellan artikelns delar. Redaktören väljer ett element och placerar det på rätt plats i strukturdiagrammet. De lodräta strecken visar på vilken nivå i diagrammet man rör sig. Elementförteckningen förändras beroende på var i strukturdiagrammet eller artikeln man befinner sig. Överst i förteckningen finns alltid de element som är de rätta på det aktuella stället. Programmet förhindrar inte att man väljer fel element, men ett felaktigt val syns i strukturdiagrammet som en streckad röd linje. En röd fyrkant i diagrammet visar att något element saknas.

Stora fönstret till vänster är *dokument*-fönstret, det egentliga skrivfönstret. I översta balken finns namnet på dokumentet och uppgifter om i vilken katalog det finns. Strukturtaggar eller elementhierarki syns i allmänhet inte i detta fönster. I ordboken används specialtecken som inte finns på tangentbordet, och dem tar man fram med musen i det lilla fönstret *Frame*. *Terminal* och *Text Editor* är hjälpfönster som inte hör till *frame*. I *Terminal* ges kommandon, bl.a. när det gäller korpussökning. Resultaten av sökningen sammanförs i allmänhet till en fil, och det är lätt att bearbeta dem i texteditorn.

Några exempelartiklar

Som exempel tar jag artikeln *mela* (bild 2). Först väljer vi elementet *sana-artikkeli* (ordartikel). Vi får då fram ett fönster där artikelns längd ska bestämmas. Det vanliga värdet är *#lyhyt#* (kort). Om artikeln kommer att innehålla flera betydelsemoment och bli så lång att den måste indelas i stycken väljer vi här alternativet *#pitkä#* (lång). Valet leder till att elementet *hakusana* (uppslagsord) framträder på strukturdiagrammet. Uppslagsordet *mela* skrivs nu i dokumentfönstret. Sedan väljer vi ordklass (*sanaluokka*), här alltså *s.* (substantiv). I VKS intar exempelfraserna en nyckelställning. Beläggen för exempelfraser på ordet *mela* visar att ordet inte är entydigt, och artikeln måste därför indelas i betydelsemoment. Då väljer vi alltså i detta skede *vanhat leksikot* (gamla lexikon). Uppgifterna ur varje ordbok skrivs

som ett eget lexikonexempel (*leksikkoesimerkki*). Det finskspråkiga materialet markeras med taggen *korostus* (markering) och blir kursiverat.

Sedan går vi över till ordets betydelser. Då vi väljer elementet *merkitysryhmä* (betydelsemoment) får vi automatiskt beteckningen *1*. När samma element väljs på nytt får vi *2* osv. Efter siffran följer betydelsebeskrivningen (*selite*); vi väljer element och skriver in beskrivningen i dokumentfönstret. Därefter följer exempelblocket (*esimerkkilohko*) med exemplen i kronologisk ordning, dvs. så att det äldsta belägget kommer först. För att hitta exempel gjorde jag en korpussökning med programmet *egrep*. Söksträngen syns i terminalfönstret. Den teckensträng som jag sökte var *'mel[ao]'*. Jag skrev *[ao]* därför att *a* i vissa former av ordet *mela* ersätts av *o*. Resultatet av sökningen fick utgöra en fil i min egen hemkatalog, och denna fil tog jag in i texteditorn för bearbetning. På bilden syns bara en del av resultatet av sökningen. Endast Mennander har använt ordet i den egentliga betydelsen, 'styransordning i båt'. Frosterus har använt det för att beskriva en fågelstjärt. Hos Aken avser det ett paddelliknande redskap att röra om med. Korpusexemplen har mekaniskt förts in i artikeln med hjälp av musen. Exemplet från Kettu har tagits ur det gamla ordsedelmaterialet. Sådana exempel skriver redaktören själv in i texten och granskar samtidigt ortografin.

Det sista elementet i artikeln är förteckningen över sammansatta ord (*yhdys-sanaluettelo*). Här finns det bara ett sammansatt ord, *mäskimela*.

Uppslagsformen av *mela* är entydig. Av många andra ord finns det desto fler varianter (bild 3). Till exempel i artikeln *melkein* ingår den nuspråkliga uppslagsformen *melkein* och dessutom formerna *melkeen*, *melke*, *melkeinkin*, *melkäinkin*. När det gäller formen *melkeen* är det endast slutet som avviker från huvuduppslagsordet. Då väljer man elementet *päätevariantti* (ändelsevariant) och skriver *-en*. Man får inte glömma lodstrecket som avskiljer ändelsen. Programmet för in kommatecknet. De övriga formvarianterna skrivs sedan som *variantelement* (*variantti*); programmet ger automatiskt parentes och tilde. I elementförteckningen ingår också elementet *vanha muoto* (gammal form).

Så kallade varianterna nämligen om nuspråkets form av ordet, uppslagsformen, inte alls har förekommit. Ett exempel på detta är artikeln *melkeittäisin*: i alla exempelfraserna står det *melkeettäisin*. Varianter och gamla former skrivs in i ordboken som hänvisningsartiklar (*viiteartikkeli*) på sin egen plats i alfabetet, till exempel *melke* och *melkee-*. När man väljer elementet *viitehaku-sanat* (hänvisningslemman) lägger programmet automatiskt in en hänvisningspil.

The image shows a screenshot of a FrameMaker SGML Structure View. On the left is a text window showing a document with Swedish text and SGML tags. On the right is the Structure View showing a hierarchical tree of elements. The text window contains the following text and tags:

```

#Lyyhyt melata v. J brevi remo utor, brukar styrären | JlisP 210 pian
soottu soucka salmi. melattu (melattu) matala meri prov. | G II 167 a - J;
myndär, håller styr på båten, guberno cymbam. - melata, meloita.¶
#Lyyhyt melke adv. -> melkein.¶
#Lyyhyt melkee -> melkein.¶
#Lyyhyt melkein, -en (- melke, melkein, melkänkin) ¶
#Lyyhyt melkeittäisin (melkeettäisin) adv. suurin piirtein, jotenkin,
jokseenkin | Liz 1756 56 Carl Fleischer toisi melkettäisin lukea Kirjaa
sisältää, ja lupais nyt virisää opetella ulcoa | Wirz 1768 735 Kusa
orjantappurat joukkoittain kasowawat, sijinä on melkettäisin hedelmätöin
maa.¶
#Lyyhyt melkeä, -iä a. (J melkea aliguantus, tämlig | JlisP 210, JlisK

```

The Structure View on the right shows the following elements and their attributes:

- Hukusana melke
 - Sivupää PerusSana
- Sanalukka adv.
 - Viitehaku-sanat melkein
- SanaArtikkelit melkee-
 - APLus ID Elyhyt
 - ID eno vales
- SanaArtikkelit
 - APLus ID Elyhyt
 - ID eno vales
- Hukusana melkein
 - Sivupää PerusSana
- Pääkirjain -en
- Varmitt melke, melkein, melkänkin
- SanaArtikkelit
 - APLus ID Elyhyt
 - ID eno vales
- Hukusana melkeittäisin
 - Sivupää PerusSana
 - Vainio Muoto melkeettäisin
 - Sanalukka adv.
 - Selin suurin piirtein, jotenkin, jokseenkin
 - Esimerkki Liz 1756 56 Carl Fleischer lätin

BILD 3. Varianter, gamla former och hänvisningsartiklar.

Pronomenet *me* (vi) är ett bra exempel på en lång artikel. På bilden (bild 4) ses strukturen i huvuddrag. Trots att exempelfraserna har utelämnats är det bara en del av artikeln som har fått plats. Indelningen börjar med övermomentet (*suurmerkitysryhmä*) *I*; sedan kommer betydelsemomenten med arabiska siffror och undermomenten med små bokstäver. Det finns också avdelningar för användningsområden och sublemman på olika nivå.

Till höger på samma bild finns texten i SGML-form, uppsjälkad i element eller fält, och taggarna är som synes många. I

VKS är taggarna hela ord och alltså mycket åskådliga. Starttaggen för ett fält pekar till höger och sluttaggen till vänster. Det kan finnas flera fält innanför varandra, och inuti ett fält kan det finnas flera jämbördiga fält. På bilden syns betydelsemomentet *l 2* ur artikeln *me*. Inom den finns användningsområdet (*käyttöalaryhmä*) *erik*. (specialanvändning), som har indelats i undermomenten *a*, *b* och *c*. Inom *c* finns en sublemmagrupp (*alihakuryhmä*), som innehåller ett sublemma (*alihakusana*), dvs. *meitin*, en förklaring (*selite*) och ett exempelblock (*esimerkkilohko*), som alla är på samma nivå i hierarkin. Under *b* finns en textjämförelse (*tekstivertailu*) inom ett exempel (*esimerkki*). Längst ner finns sex sluttaggar efter varandra, först taggen för det innersta fältet och sist taggen för det yttersta fältet. Det är lätt att ta fram strukturfönstret på arbetsbordet, och man har stor hjälp av det om man misstänker att man har gjort något fel. När man skriver en artikel behöver man naturligtvis inte hålla hela strukturhierarkin i minnet; det räcker med att man har överblick över det avsnitt som man håller på med.

Bearbetningen och finslipningen av artiklar

En artikel som är skriven i strukturerad form är lätt att bearbeta. När man gör ändringar har man nytta av strukturdiagrammet som visar på vilken nivå man arbetar. Om ett element flyttas följer alla element med som finns lägre i hierarkin. Det är därför också lätt att flytta hela artiklar. Ibland måste man flytta betydelsemomenten, och då sörjer programmet automatiskt för att numreringen blir riktig. Innan artiklarna trycks måste naturligtvis taggarna och andra överflödiga detaljer avlägsnas, och texten måste bearbetas för tryckning. Denna finslipning sköter programmet om. Som exempel på färdiga artiklar finns här artikeln *kummitus* (bild 5) ur ordbokens andra del som redan har kommit ut.

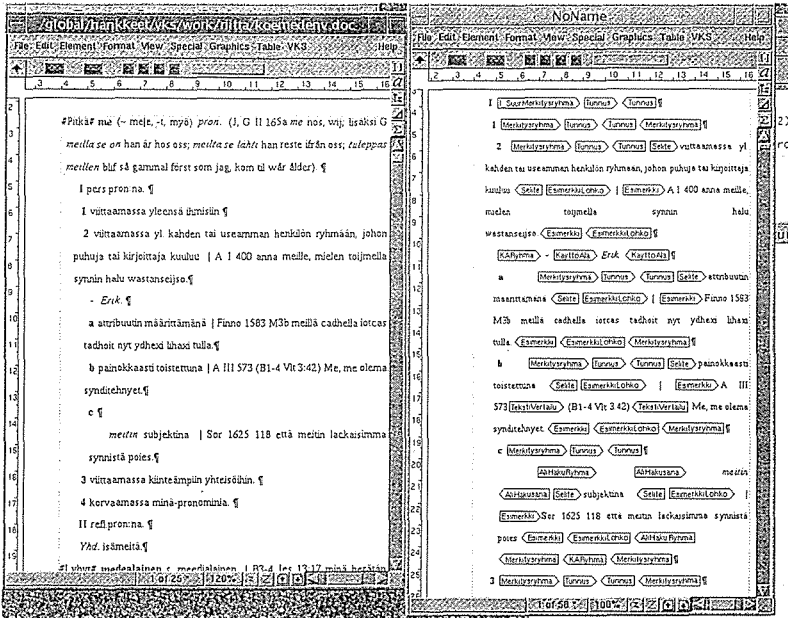


BILD 4. Pronomenet *me* som exempel på en lång artikel.

kummitus s. (JlisP 62 spöke, Sotk. | G 496b spö-
kande, spökerin ~ kummitteleminen). I outo ja
selittämätön olento tai ilmiö, aave; myös kuv. |
Fors 1704 83 [Maailman menot] silmisiä sinulla
ovat culla kijldäwäinen, Kijldäwäiset cummi-
tuxet | Weg 1747 867 cummituxet ja Piruin näöt
enimmitten aina näkywät Kirkoig aidoisa | Weg
1749 953 taita hän sairautten parannuxisa ja mui-
sa cummituxisa paljon ihmedeldäwiä asioita toi-
mitta | SAchr 1766 20 Caicki cuollon cummituxet
Jätti julkinen Jumala | JAFro 1791 28 senkaldai-
set [tulet metsässä] ei ole mitkän cummituxet,
mutta heillä owat heidän luonnolliset syynsä.
2 tarpeeton, kauhistuttava teko tai ilmiö, hullu-
tus, turhuus | GJCal 1734 16 [Maanpetturij] Teki
wisi willitystä, cuusi muuta cummitusta | Weg
1747 esip. 9 wierasten waatetten ylöllisyydesä ja
muodottomisa keweisä cummituxisa ja waihetuxi-
sa ja muusa lihan menosa | ib. 415 [Velhot] Isä
meidän Rucousta edes tacaisin lukewat? ja mon-
da muuta edes tacaista cummitusta sijnä edesäns
pitäwät. — Yhd. ihmekummitus.

BILD 5. Ordet *kummitus* ur ordbokens andra del.

Till slut

En ordbokstext som har lagrats i strukturerad form är en ypperlig källa för fortsatt forskning. Som undersökningsobjekt kan man välja vilket enskilt element eller vilken elementkombination som helst. Man kan till exempel plocka ut enbart uppslagsorden eller betydelsebeskrivningarna eller varför inte alla verbartiklar. Möjligheterna är många, och det är viktigt. Forskningsarbetet blir ju avsevärt lättare och ändamålsenligare när det finns material och program som kan bidra med detaljerad och mångsidig information.

Litteratur

A. ORDBÖCKER

VKS = *Vanhan kirjasuomen sanakirja*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 33. 1. Päätoim. Esko Koivusalo. Helsinki 1985. 2. Päätoim. Maija Länsimäki. Helsinki 1994. Valtion painatuskeskus.

B. ÖVRIG LITTERATUR

- Adobe FrameMaker+SGML*. Using FrameMaker+SGML. 1996.
 – Quick Reference. 1996.
 – Getting Started. 1996.
- Hietanen, Risto / Syrjänen, Seppo. 1995. *Unix-käyttäjän alkeisopas*. 3. painos. Helsinki: Yliopistopaino.
- Lehtinen, Marja / Karvonen, Pirjo / Rahikainen, Tarmo. 1995. *Tekstikorpuksset*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 81. Helsinki.
- Nikunen, Erja. 1988. *Vanhan kirjasuomen sanakirjan rakenne*. Moniste.
- Nikunen, Erja / Rahikainen, Tarmo / Lounela, Mikko. 1998. *Vanhan kirjasuomen sanakirjan rakennekuvaus*. Uusin versio elokuu.

Vanhan kirjasuomen sanakirjan artikkelinkirjoitusohjeet 1977.

Toim. Esko Koivusalo / Irene Hellemaa / Riitta Suhonen /
Anja Lehtimäki. Helsinki.

Vanhan kirjasuomen sanakirjan toimitusperiaatteet 1997. Toim.

Riitta Palkki. Moniste.

Översättning från finska av Monica Äikäs.