Inledning

Nusvensk frekvensordbok 1 är resultatet av en kvalitativ och kvantitativ undersökning av nusvenskans vokabulärsystem, representerat av ett skriftspråksmaterial hämtat ur svensk morgonpress. En meningsfull kvantifiering förutsätter en systematisk strukturanalys där de operativa enheterna noggrant definieras. Undersökningar av autentiska material ger i sin tur bättre underlag för analyser av språksystemet. Kvalitativa och kvantitativa studier bör alltså gå hand i hand.

Tidningsmaterialet får i väsentliga avseenden anses vara representativt för det nusvenska standardspråket i dess skrivna form. Tidningsprosan är av speciellt intresse eftersom den sprids genom ett viktigt massmedium. Därtill kommer att den via uppläsningsspråk också i viss utsträckning förmedlas av etermedierna.

Ett autentiskt textmaterial är emellertid bundet till ämne, genre, skrivsituation. Det kan aldrig exakt avspegla språksystemet, i detta fall språkets lexikon. Ett genomsnittsspråk, där alla enheter har en given sannolikhet att realiseras, finns inte i den språkliga verkligheten. Det är därför viktigt att till läsarens ledning noga beskriva materialets sammansättning.

Korpusen omfattar en miljon löpande ord. Ett material av denna omfattning är å ena sidan tillräckligt för att ge tillförlitliga data beträffande en väsentlig del av ordskatten (liksom beträffande flera andra språkliga kategorier). Å andra sidan är det för litet för att ge en säker uppfattning — eller över huvud taget upplysningar — om en rad ovanliga ord. Läsaren finner kanske att flera ord eller ordformer i hans egen vokabulär är orepresenterade. För att i någon större utsträckning öka den belagda ordmängden skulle krävas en mångdubbling av materialet. Härvid skulle analysen av allt att döma inte kunna göras lika ingående som i föreliggande undersökning. Tidningskorpusens storlek kan ses som en kompromiss.

Ordboken innehåller tjugoen olika bearbetningar av det textmaterial som den grundar sig på. Arton av dem gäller ord av olika slag betraktade från olika synpunkter. Av dessa arton har tolv formen av lexikon eller ordlistor och sex formen av tabeller. De tre återstående bearbetningarna gäller grafem (dvs. bokstäver, siffror, skiljetecken o. d.) och propriella syntagmer (dvs. egennamn som består av mer än ett ord).

De olika bearbetningarna är ordnade efter ett mönster som ansluter sig till den beskrivningsmodell som har utformats för undersökningen. Modellen har presenterats i uppsatserna En undersökning av nusvenskans vokabulärsystem (i Språklig databehandling. En introduktion utgiven av S. Allén och J. Thavenius, 1970, s. 31–59) och Vocabulary data processing (The Nordic languages and modern linguistics. Proceedings of the In-

1*** – 703405 NFO 1

ternational conference of Nordic and general linguistics. Edited by H. Benediktsson. Reykjavik 1970, s. 235–261). En mera utförlig presentation ingår i en otryckt monografi med arbetstiteln Studier över nusvenskans vokabulärsystem, som planeras ingå i samma serie som denna ordbok.

Som exempel på bearbetningar som tar upp mer speciella aspekter på materialet kan nämnas Martin Gellerstams doktorsavhandling Jämförande vokabulärstudier i svensk morgonpress 1965, 1970 (dupl.) — se även Språkvård 1969: 3 — och Sture Bergs licentiatavhandling Homografi i nusvenskan, 1970 (dupl.).

Frekvensundersökningar har med olika metoder genomförts sedan rätt lång tid. Av äldre arbeten kan det räcka att nämna The teacher's word book of 30 000 words av Edward L. Thorndike och Irving Lorge (1944). Metodiken har sedan dess vidareutvecklats i olika avseenden. Automatisk databehandling har blivit ett viktigt hjälpmedel och har inte minst möjliggjort nya typer av undersökningar. Två utländska arbeten från senare tid får representera denna utveckling: Frequency dictionary of Spanish words av Alphonse Juilland och E. Chang-Rodriguez (1964) och Computational analysis of present-day American English av Henry Kučera och W. Nelson Francis (1967).

Av de arbeten som gäller svenskan skall först nämnas Olof Werling Melins två ordräkningar publicerade i Stenografiens historia 2 (1929), s. 565–577. De gäller 100 000 löpande ord vardera (riksdagsprotokoll respektive affärsbrev) och ger bland annat frekvenser för de 100 vanligaste orden.

P. G. Widegren utgav 1935 en ordbok med titeln Frekvenser i nusvenskans debattspråk 1. Hans material omfattar 530 000 löpande ord från riksdagsprotokoll o. d. Lexikonet ger förutom ordfrekvenser även uppgifter om frekvenser för vissa fraser.

Carita Hassler-Göransson har sammanställt resultatet av flera olika undersökningar omfattande sammanlagt 500 000 löpande ord i lexikonet Ordfrekvenser i nusvenskt skriftspråk (1966). Materialet, som är spritt över femtio år, är sammansatt av litterär prosa, tidningstext, skolbarnsuppsatser osv.

Med utgångspunkt från i första hand Carita Hassler-Göranssons då föreliggande resultat sammanställde Martin S. Allwood och Inga Wilhelmsen 1947 en Basic Swedish word list. I denna ges inga frekvensuppgifter för orden, men de graderas efter vanlighet.

Vissa uppgifter om frekvenser i såväl tal som skrift lämnas slutligen av Gunnar Fant i ett (duplicerat) Kompendium i talöverföring (1967). Undersökningen gäller 22 000 löpande ord ur 50 tidningsartiklar och 63 000 löpande ord ur telefonsamtal. Kompendiet innehåller bland annat en lista över de 50 vanligaste orden i talspråksmaterialet.

Material

Frekvensordboken bygger på en slumpmässigt vald artikelsamling ur fem morgontidningar med politisk och regional spridning: Svenska Dagbladet, Stockholms-Tidningen, Dagens Nyheter, Göteborgs Handels- och Sjöfarts-

Tabell I. *Materialets procentuella fördelning på tidningar och genrer.*

	GHT	SvD	ST	DN	SDS	*Totalt*
A	5,8	13,6	6,2	3,9	13,1	42,6
K	9,4	12,9	4,5	6,8	12,4	46,0
U	2,4	3,1	1,4	1,7	2,8	11,4
Totalt	17,6	29,6	12,1	12,5	28,3	100,0

Tidning och Sydsvenska Dagbladet Snällposten. Tidningssätterierna levererade år 1965 under vardera fem fjortondagarsperioder all text på hålremsa med undantag av nyhetsbyråtelegram, annonser, insändare, anonyma bidrag, sportartiklar och kåserier. Vid genomgången av materialet bortsorterades artiklar med metaspråkliga inslag, artiklar av utländska författare och artiklar med längre citat. Samtidigt klassificerades de återstående artiklarna i tre genrer: utrikeskorrespondenternas rapporter (U), kultursidesartiklar (inklusive recensioner på annan plats än kultursida; K) och allmänna reportage (A). Korpusen omfattar i definitivt skick 1 000 669 löpande ord fördelade på 1 387 artiklar av 569 författare (en källförteckning över artikelsamlingen föreligger i stencilerad form). Den procentuella fördelningen på tidningar och genrer framgår av tabell I.

Genreindelningen var avsedd att ge ett grovt mått på fördelningen av artiklarnas huvudtyper. Den återspeglar för övrigt en redaktionell bedömning av materialet. Senare genomfördes en indelning på innehållsliga grunder. Denna omfattar sex ämnessfärer med följande karakteristika:

(1) Naturvetenskap med tillämpningar (Na), omfattande artiklar med naturvetenskaplig, medicinsk och teknisk tyngdpunkt.
(2) Kulturvetenskap med tillämpningar (Ku), omfattande artiklar hänförliga till de humanistiska, samhällsvetenskapliga, juridiska och teologiska fälten.
(3) Politik och samhällsfrågor (Po), omfattande artiklar om aktuell in- och utrikespolitik, kommunalpolitik, politiska organisationer och samhällsfrågor.
(4) Näringsliv och samfärdsel (Nä), omfattande artiklar rörande industri, hantverk, jordbruk, handel, ekonomi och samfärdsel.

Tabell II. *Materialets procentuella fördelning på tidningar och ämnessfärer.*

	GHT	SvD	ST	DN	SDS	*Totalt*
Na	1,2	3,3	,5	,8	2,6	8,4
Ku	4,6	8,8	3,1	4,6	7,9	29,0
Po	2,6	5,2	3,5	2,3	4,8	18,5
Nä	2,4	1,9	1,0	1,0	2,8	9,1
Mä	2,8	4,3	2,3	2,1	5,2	16,7
Ko	4,0	6,0	1,6	1,7	5,0	18,3
Totalt	17,6	29,6	12,1	12,5	28,3	100,0

(5) Människa och miljö (Mä), omfattande artiklar om hem, familj, mat, kläder, enskilda personer (typen »human interest»), trivselfrågor och fritidsverksamhet.
(6) Konstnärlig verksamhet (Ko), omfattande artiklar rörande fiktionslitteratur, teater, film, musik, balett, bildande konst och konstnärligt foto.

Fördelningen på tidningar och ämnessfärer återges i tabell II.

Samtidigt som artikelsamlingen uppenbarligen spänner över ett vitt ämnesområde är korpusen språkligt väl sammanhållen. Språket torde kunna karakteriseras som svenskt standardskriftspråk.

Kvalitativa aspekter

De kvalitativa aspekter som läggs på materialet gäller beskrivningsnivå, ordtyp, enhet och sortering. De skall kortfattat behandlas i detta avsnitt.

Beskrivningsnivå

Det är viktigt att eftersträva sådana beskrivningsnivåer som medger att mesta möjliga grammatiska information kombineras med största möjliga konsekvens. Detta betyder att de operativa enheterna bör definieras på basis av formella språkkriterier. En konsekvent genomförd frekvensundersökning baserad på semantiskt definierade enheter ligger på forskningens nuvarande ståndpunkt inte inom räckhåll.

Det bör i detta sammanhang påpekas att vissa moment i undersökningen, såsom satsanalysen och beläggens inplacering i flexionsserier, förutsätter att man förstår texten. Den väsentliga skillnaden mellan en semantiskt och en formellt baserad analys framträder exempelvis i fall som de följande.

I den här genomförda undersökningen bildar *djup* och *hög* vardera två enheter, ett substantiv och ett adjektiv. Orden *mal* 'ett slags insekt; ett slags fisk; grusfält' och *magasin* 'förrådshus; handelsbod; patronrum (på vapen); tidskrift' bildar vardera en enhet på grundval av flexionen (pluralis enbart *malar* respektive *magasin*). De oböjliga orden *icke* och *inte* bildar liksom *ned* och *ner* olika enheter. I en semantiskt baserad analys råder osäkerhet beträffande antalet enheter, vilket man lätt kan övertyga sig om genom att slå upp dessa och andra ord i gängse ordböcker. Det skall framhävas att vad som här har sagts innebär, att undersökningen i vårt fall gäller formellt karakteriserade enheter — det innebär under inga omständigheter att en motsvarande undersökning av semantiskt karakteriserade enheter, så långt den kunde föras, vore ointressant.

De beskrivningsnivåer som urskiljs i undersökningen illustreras i figur I.

Operativ enhet på den grundläggande ordnivån är *grafordet* (det grafiska ordet), en enhet som bestäms såsom ett segment begränsat av två på varandra följande spatier (mellanrum). Den överväldigande mängden graford svarar mot vad man normalt menar med ord: STACK, FEM OSV. I vissa speciella fall, t. ex. KULTUR-, MUMMA-REVYN, ingår grafordet i en kombination som i fullbordat skick bildar en enhet på *frasnivå: kultur-*

Figur I. *Beskrivningsnivåer.*

och samhällsliv, Kar de Mumma-revyn. Beträffande enstaka mer svårtydda graford av den nämnda typen lämnar ordboken upplysningar i noter. Enheterna på frasnivå utgörs normalt av återkommande förbindelser som *över huvud taget, ligga bakom* osv. En lista över propriella fraser (*Artur Lundkvist, Dagens Nyheter* osv.) ges i bearbetning 3.3.

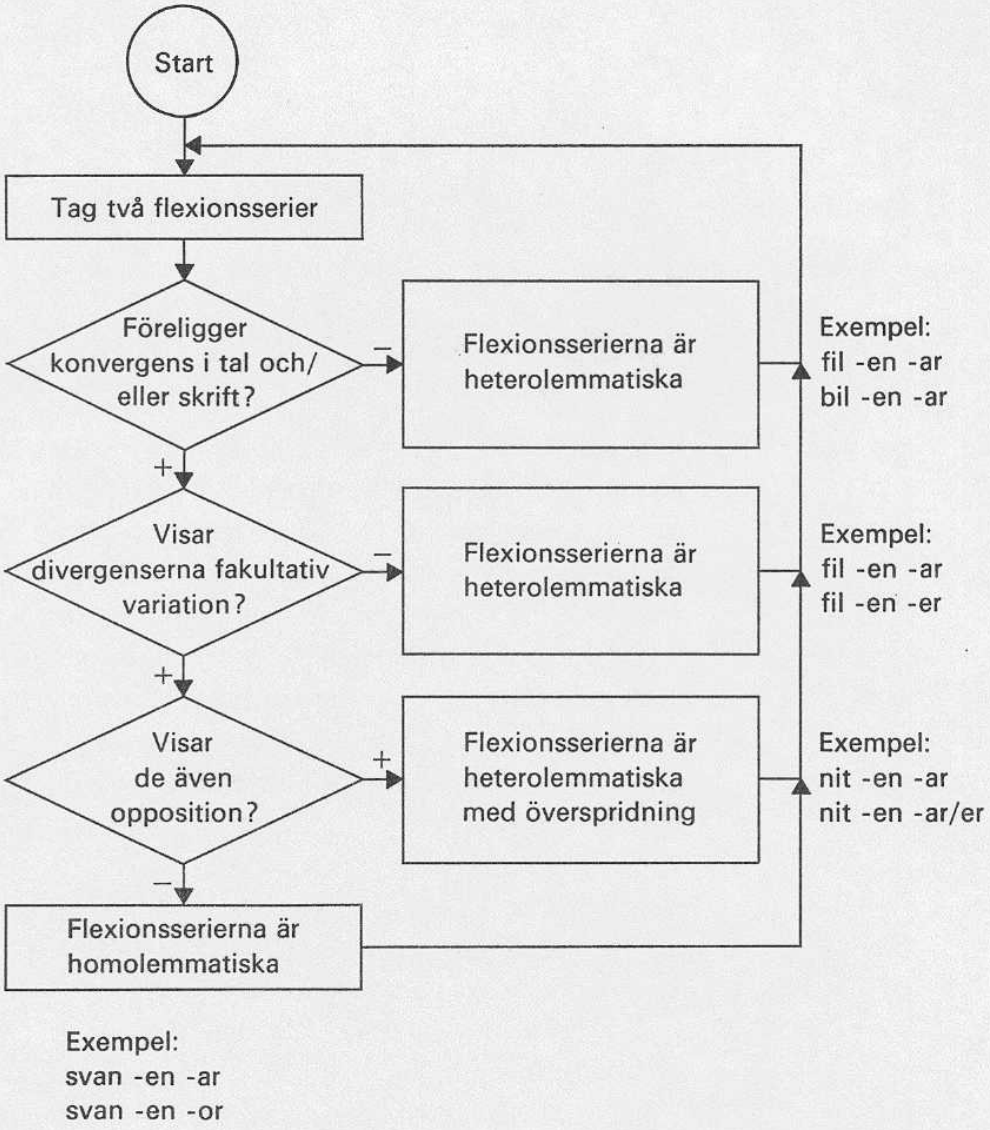
Graforden är uppbyggda av *grafem:* a, b, 1, 2, / osv. (se närmare nedan under Ordtyp och jämför bearbetningarna 3.1 och 3.2). Som har framgått neutraliseras distinktionen mellan versaler och gemena vid beskrivningen på grafordsnivån: enheterna återges med kapitäler. Detta motiveras dels av att de omfattande problem som knyter sig till versaliteten bör behandlas i den grammatiska analysen vid beskrivningen på nästa ordnivå, dels av att det från språkvetenskaplig synpunkt är intressantare att föra samman belägg som *Han* (i början på mening) och *han* i ordkroppen HAN än att särhålla dem.

Ett graford som STACK kan uppdelas i två homografkomponenter: substantivformen *stack* och verbformen *stack*. De uppträder som separata enheter på närmast högre nivå, *homografkomponentnivån,* som också omfattar de icke-homografa enheterna (heterograferna). En synnerligen omfattande homografseparering har krävts för att nå denna nivå. Teoretiskt sett kan separeringen indelas i tre steg.

Det första steget innebär att komponenter som tillhör olika ordklasser skiljs från varandra. Proceduren förutsätter att ett ords ordklasstillhörighet kan bestämmas. För detta ändamål har en flödesgrammatik utarbetats som med hjälp av ett antal ordnade kriterier fördelar enheterna på ordklasser. Grammatiken bygger i första hand på syntaktiska kriterier, i andra hand på morfologiska, grafematiska och fonematiska kriterier. Proceduren är en vidareutveckling och precisering av den traditionella ordklassindelningen.

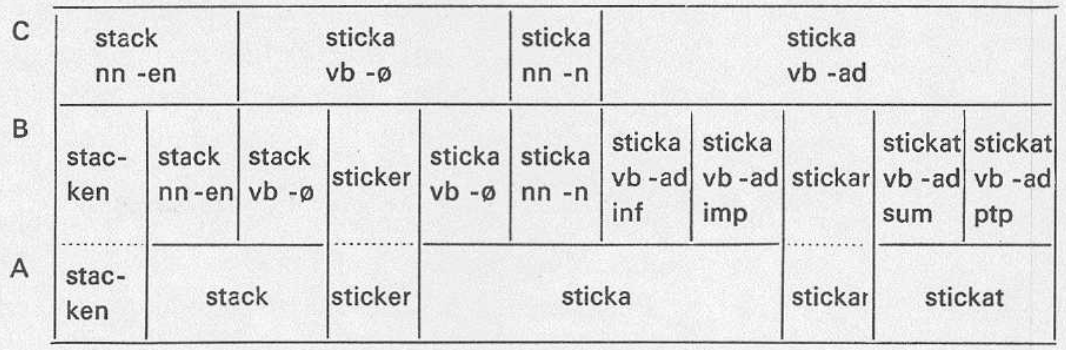
Som framgår av förkortningslistan nedan är antalet klasser 14. En speciell karaktär har klasserna abbreviationer (förkortningar) och utländska enheter, som naturligtvis inte är vad man traditionellt menar med ordklasser. Det kan vidare nämnas att klassen proprier (egennamn) definieras med hjälp av versalitet och att klassen pronomen är negativt bestämd såsom den restklass som erhålles när alla utnyttjade kriterier har applicerats.

Grafordet STICKA kan liksom STACK delas upp i två komponenter tillhörande olika ordklasser, substantiv och verb. Men inom ramen för verbklassen finns två olika böjningsserier, *sticka stack* osv. respektive *sticka*



Figur II. *Bestämning av flexionsseriers lemmastatus.*

stickade osv. En uppdelning på grundval av detta slags flexionsserier utgör det andra steget i homografsepareringen. Den innebär samtidigt att de berörda enheternas tillhörighet till *lemman* bestäms. Ett lemma kan kortfattat definieras som en grupp ordformer inom en ordklass vilka kan hänföras till antingen en och samma flexionsserie (i fråga om oböjliga ord endast omfattande grundformen) eller flera i tal och/eller skrift konvergerande serier vars divergenser visar rent fakultativ (fri) variation. Definitionens innebörd framträder kanske tydligare i figur II som åskådliggör bestämningen av flexionsseriers lemmastatus. Serierna i det första exemplet är utan vidare heterolemmatiska (konvergens saknas). I det andra exemplet finns konvergenser, men divergenserna visar inte fakultativ variation och vi har alltså att göra med två heterolemmatiska serier också här. Divergenserna i det tredje exemplet visar både variation och opposition (*nitar* kan ibland utbytas mot *niter*) — serierna är att betrakta som heterolemmatiska med överspridning. I det sista exemplet tillhör serierna ett och samma lemma (*svanar* kan utbytas mot *svanor* utan att den denotativa betydelsen ändras). Flexionsserierna hämtas i princip ur tredje upplagan 1964 av Illustrerad svensk ordbok utgiven av Bertil Molde (se



A grafordsnivå
B homografkomponentnivå
C lemmanivå
1, 3, 5 heterografi
2 extern homografi
4 extern och intern homografi
6 intern homografi

Figur III. *Exempel på homografseparering och lemmatisering.*

vidare nedan under Enhet). Det bör påpekas att s. k. suppletiv böjning inte accepteras i här föreliggande undersökning: formen *värre* ingår exempelvis i serien — *värre värst,* inte i en serie *ond värre värst.*

De två första momenten i homografsepareringen gäller *extern homografi,* homografi mellan lemman. I det tredje steget inriktas intresset på den *interna homografin,* homografin mellan former inom ett och samma lemma. Intern homografi råder sålunda exempelvis mellan infinitiv- och imperativformerna av verbet *sticka* (det som har preteritum *stickade*) och mellan supinum och en av participformerna av samma verb (t. ex. *hon har stickat ett täcke* och *ett stickat täcke*), dvs. i de fall då en och samma ordkropp utnyttjas i olika grammatiska funktioner inom lemmat. Se närmare under Enhet.

Genom det sätt på vilket homografi här har preciserats får begreppet *polysemi* en entydig definition: den semantiska variationen inom ett lemma. De inledningsvis anförda fallen *mal* och *magasin* kan få tjäna som exempel.

Det skall tilläggas att man också har att räkna med *partiell homografi,* den som i normala texter råder mellan proprier och icke-proprier under förutsättning att den senare gruppens medlemmar står i början på meningar (eller i vissa andra ställningar) och alltså får versal: *Karl* : *karl; Sten* : *sten* osv. Ytterligare en typ av homografi kan exemplifieras med det tämligen frekventa grafordet *and* som i samtliga fall återgår på engelskans konjunktion. Andra exempel på vad som kan kallas *interlingval homografi* är svenskans *le* gentemot franskans *le,* svenskans *den* gentemot tyskans *den* osv.

Den tredje nivån i figur I omfattar lemman. En del ytterligare detaljproblem som aktualiseras vid lemmatiseringen kommer att tas upp i inledningen till andra delen av ordboken. Man kan lägga märke till att det mellan enheterna fr. o. m. homografkomponentnivån och uppåt råder

en s. k. taxonomisk relation. Övergången från grafordsnivån till homografkomponentnivån innebär däremot en uppdelning i komponenter. I den här föreliggande undersökningen indelas alltså vokabulären i 14 ordklasser, som på lemmanivå representeras av uppskattningsvis 70 000 enheter (lemmatiseringen är ännu inte slutförd), som på homografkomponentnivå representeras av i runt tal 112 000 enheter, som på grafordsnivå motsvaras av i runt tal 103 000 enheter.

Förhållandet mellan de tre grundläggande ordnivåerna sammanfattas i figur III.

Ordtyp

Man kan urskilja fyra olika ordtyper. De alfabetiska orden är uteslutande uppbyggda av alfabetiska grafem (bokstäver): *data, tandsprickningsbesvär, Boëthius* osv. I de hybrida orden är minst två av de tre olika slagen av grafem — alfabetiska, numeriska (siffror), junkturella (skiljetecken o. d.) — representerade: *11:e, 0,65, kr., schweizisk-fransk-belgisk-brittisk* osv. De numeriska orden är rena siffertal: *1970* osv. Den fjärde typen slutligen utgörs av logogrammen: %, & osv. De alfabetiska orden dominerar givetvis starkt i kvantitativt avseende (se tabellerna 1.3 och 2.6).

Enhet

På grafordsnivå utgörs en enhet av en *ordkropp*. På homografkomponentnivå utgörs en heterograf enhet likaså enbart av en ordkropp, medan en homograf enhet utgörs av en ordkropp plus en *klassbeteckning*. I det senare fallet fungerar alltså hela komplexet ordkropp plus klassbeteckning som sorteringsobjekt vid bearbetning. I numeriskt och initialalfabetiskt ordnade listor placeras klassbeteckningen efter ordkroppen, i finalalfabetiskt ordnade listor före ordkroppen.

I vissa fall är en notsiffra fogad till ordkroppen. Det är då regelmässigt fråga om en upplysning av icke-grammatiskt slag. Exempel är

kyskling[163]
lammorden[167]

där noterna meddelar att det i första fallet gäller en (medveten) förvanskning av *kyckling* och i andra fallet ordet *lamm-morden*.

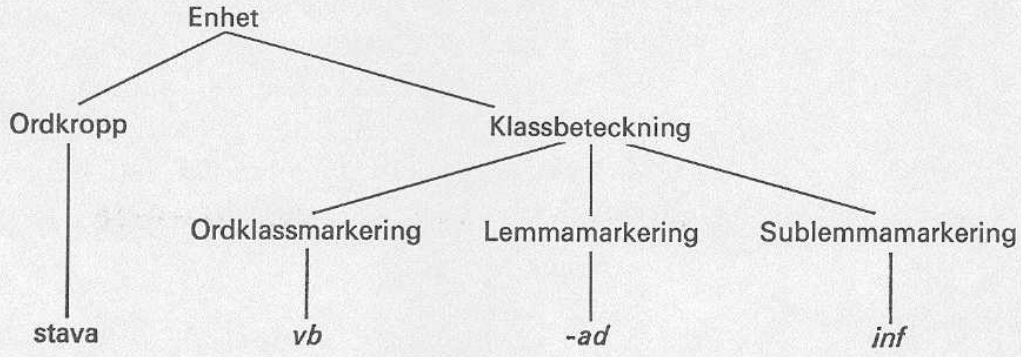
Klassbeteckningen har maximalt tre huvuddelar: uppgifter beträffande tillhörighet till ordklass, lemma och sublemma (den sistnämnda uppgiften kan i sin tur maximalt omfatta tre moment). Som enhet i ordboken har en homografkomponent alltså den principiella uppbyggnad som framgår av figur IV.

Ordklassmarkering är obligatorisk för homografkomponenterna. Vid oböjliga ord är detta den enda information som ges:

men *kn*
vid *pp*.

Samma sak gäller i några fall även ord ur flekterande ordklasser. Formerna förekommer då huvudsakligen i stående förbindelser som i fallet

känn *nn*



Figur IV. *En homografkomponents uppbyggnad.*

där ordet hänför sig till frasen *ha på känn*. Restklassen pronomen markeras på samma sätt.

En ordklassmarkering (jämte lemma- och eventuell sublemmamarkering) utsätts i en del fall som ytligt sett är entydiga, t. ex.

odlingen *nn -en.*

Detta hänger samman med den speciella form av homografi, partiell homografi, som råder mellan proprier och icke-proprier (*Sten* : *sten*). Namnet *Odlingen* förekommer i en av de proprieförteckningar som har genomgåtts för forskningsgruppens homograflexikon. Förutom ortnamn är personnamn (för- och efternamn) beaktade i detta. Eftersom i princip alla ord i språket kan fungera som proprier, är noga taget varje ord en potentiell homograf. En film eller roman kan heta *Nu* eller *Fem* eller *Utsålt* osv. I detta läge har den utvägen valts, att klassbeteckning i hithörande fall — förutom i typen *odlingen* — endast sätts ut då en propriell homografkomponent är belagd i materialet. Ett sådant exempel är

Smycket *pm -s*
smycket *nn -t.*

Vid numeriska enheter utsätts inga klassbeteckningar. Den enda ordklass utom numeraler som skulle kunna komma i fråga vid en separering är proprier (boktitlarna 491 och 1984 är belagda i materialet). Proprier definieras emellertid med hjälp av versalitet, och sådan förekommer inte hos de numeriska orden. Inte heller logogrammen har några klassbeteckningar.

I de allra flesta fall är också en *lemmamarkering* utsatt. Den utgörs normalt av en böjningsangivelse. Som tidigare nämnts hämtas uppgifter om ordböjningen i princip från tredje upplagan av Illustrerad svensk ordbok, till vilken alltså generellt hänvisas (i de flesta fall återfinns uppgifterna också i nionde upplagan 1950 av Svenska Akademiens ordlista). Uppgifterna i denna ordbok har i ett avsevärt antal fall behövt modifieras och/eller kompletteras i anslutning till den på avgörande punkter formellt baserade beskrivningsmodellen, på grundval av textmaterialets vittnesbörd eller i sista hand på grundval av forskningsgruppens kännedom om enheternas uppträdande i språket. De båda sistnämnda källorna har också utnyttjats i sådana fall där ordet saknas i Illustrerad svensk ordbok. Här

skall vidare erinras om det tidigare nämnda förhållandet att beskrivningsmodellen inte innefattar suppletiv böjning.

Böjningsangivelserna avser för substantiv (vanligen) bestämd form singularis (*-en, -et* osv.), för adjektiv neutral form singularis i positiv (*-t* respektive *-ø,* se nedan) och för verb preteritumform (*-ad, -de* osv.). Exempel:

svalka *nn -n*
skär *av -t*
röra *vb -de.*

Uppgifterna om böjning är relaterade till ordens grundformer. Exempel:

fiskarna *nn -en*
fiskarna *nn -n.*

Här avses lemmat *fisk* respektive lemmat *fiskare.*

Vid sådana substantiv där bestämd form singularis inte är lemmaskiljande ansätts en böjningsuppgift för obestämd form pluralis (om denna är lemmaskiljande):

form *nn -ar*
form *nn -er.*

Böjningsangivelsen *-ø* avser att vederbörande böjningsform bildas utan tillägg av ändelse. Vid substantiv markerar den antingen bestämd form singularis eller obestämd form pluralis, vid adjektiv neutral form singularis i positiv och vid verb preteritumform (stark böjning):

ansökan *nn -ø obe*
fot *nn -ø plu*
stilla *av -ø plu*
springa *vb -ø.*

Som framgår av exemplen krävs också en sublemmamarkering vid hithörande substantiv och adjektiv.

Markeringen *-ø* förekommer också vid proprier, numeraler och abbreviationer. Den används i de fall där den belagda formen tillhör ett lemma vars stam slutar på *-s, -x* eller *-z.* Det specificeras genom en sublemmabeteckning om den belagda formen står i grundform eller genitiv. Exempel:

Agnes *pm -ø gru*
Agnes *pm -s.*

Det första fallet är alltså grundformen av lemmat *Agnes,* det andra genitiven av lemmat *Agne.* På motsvarande sätt:

sex *nl -ø gru*
SFS *an -ø gru.*

I vissa fall är de utnyttjade böjningselementen inte lemmaskiljande. Som lemmamarkering används då i första hand en uttalsanvisning (där tryckstyrka markeras med ' före den betonade vokalen och längd med :):

glatt *av 'a:*
glatt *av at: neu.*

Den första formen tillhör lemmat *glad,* den andra lemmat *glatt.* Några gånger är inte heller uttalet lemmaskiljande. I de fallen preciseras lemmat i en not (notsiffran fogas till ordklassbeteckningen):

glatt *ab*[41].

I noten anges att det gäller det lemma *glatt* som i komparativ heter *gladare* (till skillnad från det lemma *glatt* som i komparativ heter *glattare*).

Sublemmamarkeringen ger besked om vilken flexionsform inom lemmat som belägget hänför sig till. Det är här alltså fråga om intern homografi. Ett exempel är det i figur IV givna

stava *vb -ad inf*

som avser infinitivformen till skillnad från imperativformen av lemmat *stava.*

Separeringen genomförs principiellt så långt som det morfologiska systemet ger möjlighet till. Exempel:

läkare *nn -n sin*
läkare *nn -n plu*
dömde *vb -de prt*
dömde *vb -de ptp.*

De underliggande morfologiska mönstren representeras i dessa fall av böjningar som *patient* : *patienter* respektive *sköt* : *skjutne.* Vid tillämpningen av den nämnda principen för homografseparering beaktas inte enstaka fall (lexikaliskt eller textuellt sett) av typen *ord* : *(till) orda* bland substantiven (obsolet böjningsform) eller *en* : *ett* bland räkneorden (unik genusdistinktion). Så utsätts t. ex. inte genusbeteckning vid räkneordet *två.* Påträffas ett belägg låt oss säga på presens pluralis, på konjunktiv eller på plural *e*-form av adjektiv anges dock naturligtvis detta:

skildra *vb -ad prs*
ske *vb -dd kon*
svenske *av -t plu.*

I ett fall som

gammaldags *av -ø gru obe neu*

är sublemmamarkeringen maximal. Den form som avses är den som förekommer i uttryck som *ett gammaldags julbord,* dvs. den obestämda grundformen i neutrum (jfr *ett gott julbord*), till skillnad från bland annat den icke-neutrala form som förekommer i exempelvis *en gammaldags kontorslampa* (jfr *en god kontorslampa*) och som betecknas

gammaldags *av -ø gru obe utr.*

En viss ekonomi i notationen har eftersträvats. I själva verket står ju de båda nyss givna exemplen i opposition till kategorin pluralis (jfr *god/*

gott : *goda*). Orsaken till att *sin* inte behöver utsättas i sublemmamarkeringen är att genusdistinktionen *utr* : *neu* implicerar singularis, eftersom någon genusdistinktion inte realiseras i pluralis (det heter *goda* genomgående: *goda kontorslampor, goda julbord*).

I ett visst antal fall utsätts klassbeteckningar som innehåller sublemmamarkeringar vid enheter som från systemets synpunkt är möjliga homografer men som i praktiken knappast förekommer i mer än en komponent:

återfå *vb -ø inf*
vände *vb -de prt.*

Man kan emellertid konstruera kontexter av typen »Återfå talförmågan!» (sagt av regissör till skådespelare), där *återfå* skulle få beteckningen *vb -ø imp*. Någon alldeles strikt gräns mellan heterografi och intern homografi i sådana specialfall kan knappast dras.

Sortering

En rad olika typer av sorteringar förekommer i ordboken. De *numeriskt* ordnade bearbetningarna har som första sorteringsgrund fallande F, Fmod eller Ftext. Som andra sorteringsgrund används här *initialalfabetisk* ordning, dvs. i princip ordinär alfabetisk ordning med ordkroppens första grafem som första argument, dess andra grafem som andra argument osv.

Som första sorteringsgrund utnyttjas initialalfabetisk ordning i 2.2. Beteckningen initialalfabetisk ordning markerar en motsättning till *finalalfabetisk* ordning, där ordkroppens sista grafem är första argument, dess näst sista grafem andra argument osv. (jfr rimlexikonets princip). En ordbok efter denna sorteringsordning kallas ibland ett baklängeslexikon. Intresset inriktas i en bearbetning av detta slag på ordsluten, som är de viktigaste bärarna av grammatisk information. Listningarna under 2.3 utgör ett finalali etiskt lexikon. I de initial- och finalalfabetiska bearbetningarna sorte homografkomponenter i andra hand alfabetiskt efter klassbeteckningen.

Vad som närmare bestämt skall menas med alfabetisk ordning är långt ifrån självklart. Var skall exempelvis speciella tecken som ó och æ placeras, var skall numeriska grafem och logogram infogas, hur skall ordningsföljden mellan versaler och gemena bestämmas? I föreliggande fall har följande principer genomförts.

1) Jämförelsen mellan två ordkroppar sker teckenvis från vänster till höger vid initialalfabetisk ordning respektive från höger till vänster vid finalalfabetisk ordning; alla tecken beaktas.
2) Ordslut har högst prioritet (*Ada* före *Adam*).
3) De icke-alfabetiska tecknen kommer i ordningen » ' § : ; ( ) + ° ? = — ! * / 0 1 2 3 4 5 6 7 8 9 . , - % & × och har högre prioritet än de alfabetiska tecknen (*svensk-norsk* före *svenska*).
4) Versaler och gemena betraktas som likvärdiga; dock går versal före gemen i sådana fall där två ordkroppar endast skiljer sig från varandra genom distinktionen versal : gemen (*Sten* före *sten* före *Stens*).
5) Tecken som i sorteringshänseende kan anses vara varianter av samma siffra eller bokstav, exempelvis 1 ₁; E É e é è ê ë; V W v w; Y Ü y ü;

Ö Ø Œ ö ø œ, betraktas som likvärdiga, om de två jämförda ordkropparna i övrigt är olika; i annat fall sker prioritering förutom efter 4 ovan på följande sätt: (1) grundtecken (utan diakritikon), dvs. de icke-alfabetiska tecknen under 3 ovan och de 28 alfabetiska tecknen a–ö (ej w) (2) (tecken med) akut accent ´ (3) (tecken med) grav accent ` (4) (tecken med) cirkumflex ˆ (5) (tecken med) trema ¨ (6) tecknen ç ñ w ü ø; höjda och sänkta indexsiffror (7) ligaturerna æ œ (8) tecken som inte ingår i det latinska alfabetet (betraktas i sorteringshänseende som varianter av närmaste latinska motsvarigheter).

Eftersom de numeriska enheterna uppträder i initialalfabetisk respektive finalalfabetisk ordning i bearbetningarna 2.2 och 2.3.3, förtecknas de i 2.4 i *stigande talordning,* dvs. i den vanliga ordningen från lägsta till högsta talvärde.

Från ordtypernas synpunkt är listningarna av två slag. Normalt är enheterna *samsorterade,* men i bearbetningarna under 2.3 är de fyra ordtyperna — alfabetiska, hybrida och numeriska enheter respektive logogram — *särsorterade.*

Kvantitativa aspekter

Också de kvantitativa aspekterna är fyra: frekvens, spridning, rang och vokabulärsektion. De tas här upp i tur och ordning. Först skall emellertid ett generellt påpekande av kvantitativ art göras. Det gäller tabeller — både i inledningen och i själva ordboken — som innehåller procenttal. Vid summeringen av deluppgifterna kan en avvikelse på någon decimal från de angivna totalvärdena förekomma. Denna avvikelse är skenbar: deluppgifterna är avrundade, och totalvärdena har beräknats på basis av delvärden med ett betydligt större antal decimaler än det som ges i tabellerna.

Frekvens

Den grundläggande uppgiften i samtliga bearbetningar är enheternas *observerade frekvens* F, antalet förekomster i materialet. Samma värde uppträder i vissa fall under beteckningen Abs *absolut frekvens* och då i motsättning till enheternas *relativa frekvens* Rel, i ordboken uttryckt i procent. Sett från en annan synpunkt kan värdet F anges som *textuell frekvens* Ftext, vilket i sin tur markerar motsatsförhållandet till *lexikalisk frekvens* Flex. Distinktionen görs exempelvis vid återgivningen av grafemfrekvenser — antalet förekomster av ett grafem i löpande text skiljer sig normalt kraftigt från antalet lexikonförekomster (där varje ord endast räknas en gång). I listningen över enheternas fördelning på frekvenser anges vidare den *kumulativa frekvensen,* vilken gör det möjligt att avläsa exempelvis hur stor textmassa ett givet antal av de vanligaste orden täcker.

De 24 mest frekventa homograferna (F > 4 000) har 64 belagda komponenter, och dessa åtföljs i listningarna på homografkomponentnivån av frekvensuppgifter försedda med en asterisk. Härmed avses *skattad frekvens.* Skattningen, som medgivit en stor arbetsbesparing vid separeringen av de sammanlagt 645 000 homografa beläggen, är baserad på stickprov

Tabell III. *Konfidensintervall för skattade frekvenser.*

			Konfidensintervall	
	Homograf-komponent	Skattad frekvens	Nedre gräns	Övre gräns
att	*ie*	12 534	12 054	13 023
	kn	11 232	10 743	11 712
av	*ab*	204	136	290
	pp	15 262	15 175	15 330
de	**	261	194	343
	al	5 695	5 459	5 932
	pn	4 997	4 761	5 233
den	*al*	9 290	8 981	9 593
	pn	5 378	5 075	5 687
det	**	8	1	39
	al	4 203	3 885	4 534
	pn	13 991	13 660	14 309
en	**	44	14	103
	ab	9	1	41
	al	18 184	17 910	18 436
	nl -s	2 101	1 854	2 369
	pn	53	20	116
ett	*al*	7 702	7 599	7 792
	nl -s	625	535	728
från	*pp*	4 042	4 035	4 042
för	*ab*	366	282	467
	kn	142	91	210
	pp	11 807	11 684	11 916
	vb -de prs	83	52	138
han	*pn*	6 446	6 435	6 446
har	*vb -de*	9 176	9 162	9 176
I	**	51	14	130
	an	26	3	93
	nl	13	1	61
i	*pp*	29 461	29 365	29 514
kan	*vb -de*	4 479	4 471	4 479
man	**	31	13	61
	nn[1]	183	136	243
	nn[2]	98	63	144
	pn	7 103	7 029	7 167
med	**	6	1	27
	ab	339	259	435
	an	6	1	27
	pp	11 700	11 604	11 782
men	**	6	2	20
	kn	5 677	5 657	5 685
	nn -et plu	3	1	15
	nn -et sin	3	1	15

Tabell III (*fortsättning*)

	Homograf-komponent	Skattad frekvens	Konfidensintervall Nedre gräns	Konfidensintervall Övre gräns
om	*ab*	376	305	459
	kn	2 283	2 125	2 448
	pp	5 199	5 028	5 366
på	*ab*	229	159	319
	pp	13 932	13 842	14 002
sig	*pn*	5 559	5 550	5 559
som	**	9	1	42
	kn	6 079	5 701	6 467
	pn	13 522	13 133	13 902
så	*ab*	3 979	3 927	4 024
	kn	122	91	160
	pn	125	93	164
	vb -dd imp	2	1	12
	vb -dd inf	5	2	16
till	*ab*	488	406	590
	pp	9 865	9 763	9 954
var	*ab*	131	100	171
	pn	166	129	209
	vb -ø imp	5	2	15
	vb -ø prt	3 839	3 789	3 888
är	*vb*	15 506	14 985	15 506

[1] *Lemmat* man *-en män.*

[2] *Plu av* man (*utan sin bes*) *man.*

om lägst 1 700 förekomster av homografen i fråga. Konfidensintervallen för de skattade frekvenserna ges i tabell III. Tabellen skall läsas så, att de faktiska frekvenserna (frekvenserna i populationen) med 95 % säkerhet ligger inom de angivna gränserna. Vidare skall nämnas att 18 icke-propriella komponenter, som har den skattade frekvensen 0 och därför inte förtecknas i ordboken, har följande konfidensintervall: 0–7 från *ab,* så *nn -n,* var *av -t,* var *nn -et plu,* var *nn -et sin;* 0–8 kan *nn -en;* 0–12 man *ab,* man *nn -ar*; 0–14 har *nn -ø plu,* har *nn -ø sin*; 0–17 med *nn -en*; 0–18 för *av -t,* för *nn -en,* för *vb -de imp*; 0–21 är *nn -en*; 0–25 det *kn*; 0–27 en *nn -en*; 0–38 i *ab.*

Två identiska frekvenssiffror kan gälla helt olika verkligheter: i det ena fallet kan beläggen hänföra sig till en enda artikel, medan beläggen i det andra fallet kan vara väl spridda över materialet. Propriet *Krupp* med frekvensen 59 (fördelad på endast två artiklar av en författare i en tidning) är givetvis inte lika karakteristiskt för materialet som det lika frekventa verbet *fungera* (väl spritt över samtliga tidningar). Det framstår därför som önskvärt att utöver den observerade frekvensen ge ett värde som är modifierat med hänsyn till ordens spridning. Det värde som

utgör en sammanvägning (genom multiplikation) av frekvensen och det valda spridningsmåttet (se nedan) benämns *modifierad frekvens* Fmod. Det bör poängteras att det är fråga om ett teoretiskt värde (väntevärde), ägnat att diskriminera ord med dålig spridning.

Spridning

Det spridningsindex som valts är en generalisering av ett index som används av Alphonse Juilland och E. Chang-Rodriguez i Frequency dictionary of Spanish words (1964, s. LIII). Formeln blir

$$D = 1 - \frac{s}{m\sqrt{n-1}}$$

där m är medelvärdet, n antalet deltexter och s standardavvikelsen, beräknad enligt formeln

$$s = \sqrt{\frac{\sum(x-m)^2}{n}}.$$

Värdet på D *dispersionen* varierar mellan 1 och 0. En perfekt fördelning över deltexterna ger värdet 1 (vilket inget ord i materialet uppnår), medan förekomst i endast en deltext ger värdet 0. Indexet har fördelen att vara oberoende av frekvensernas storlek i deltexterna; det är proportionen mellan deltexternas frekvenser som återspeglas. När spridningsvärdet genom multiplikation sammanvägs med frekvensen F kommer produkten att ligga obetydligt under F hos ord med god spridning (t. ex. *fungera*), medan ord som uppträder i endast en deltext (t. ex. *Krupp*) och alltså får dispersionsvärdet 0 följaktligen också får Fmod-värdet 0. De deltexter som avses vid beräkningen av Fmod-värdet är de fem tidningarnas bidrag. Värdena på F i deltexterna har justerats med hänsyn till dessas olika storlek. Vid sidan av *dispersionen över tidningarna* DT ger ordboken uppgifter om *dispersionen över ämnessfärerna* DÄ.

Jämte dispersionsvärdena förekommer i ordboken ett grövre spridningsmått, som här kallas *kontribution*. Detta ger upplysning om antingen vilka och därmed hur många deltexter som lämnar bidrag till en viss enhets frekvens eller också — om det avsedda antalet deltexter är stort — enbart hur många deltexter som bidrar. Det förra gäller *tidningarna* Tidn och *ämnessfärerna* Ämn, det senare *författarna* För. Dispersions- och kontributionsvärdena, som alltså belyser olika drag i enheternas fördelning över materialet, återspeglar i komprimerad form de synnerligen utrymmeskrävande delfrekvenserna.

När det gäller de homografkomponenter som har skattad frekvens (asteriskförsedd frekvensuppgift) erhålles vid beräkningen självfallet inga värden på frekvenser i olika deltexter. Detta innebär att dispersions- och kontributionsvärdenas faktiska storlek inte kan kalkyleras. Vederbörande enheter har här tilldelats dispersionsvärden DT och DÄ som utgör medelvärdena i rangklasserna i fråga enligt tabell IV. En rangklass omfattar 100 enheter med icke skattad frekvens, varvid dock alla enheter med rangklassens lägsta frekvenstal inkluderas i klassen. Fmod för enheter

Tabell IV. *Medelvärden av DT och DÄ.*

	Rangklass			
	Nedre gräns	Övre gräns	DT	DÄ
Alfabetiska enheter	1	131	,946	,902
	132	235	,920	,884
	236	337	,896	,850
	338	440	,893	,823
	441	544	,884	,811
	545	646	,858	,783
	647	756	,863	,769
	757	861	,847	,750
	862	972	,838	,749
	973	1 087	,844	,718
	1 088	1 187	,839	,737
	1 188	1 304	,831	,721
	1 305	1 416	,807	,710
	1 417	1 527	,792	,673
	1 528	1 656	,794	,686
	1 657	1 779	,799	,703
	1 780	1 891	,792	,680
	1 892	2 031	,783	,678
	2 032	2 193	,778	,674
	2 194	2 311	,766	,669
	2 312	2 445	,758	,612
	2 446	2 602	,766	,636
	2 603	2 801	,748	,649
	2 802	3 002	,743	,624
	3 003	3 128	,722	,627
	3 129	3 233	,741	,665
	3 234	3 370	,694	,584
	3 371	3 521	,732	,596
	3 522	3 660	,735	,642
	3 661	3 821	,734	,624
	3 822	3 990	,696	,602
	3 991	4 206	,689	,598
	4 207	4 420	,683	,580
	4 421	4 668	,680	,584
	4 669	4 920	,676	,595
	4 921	5 167	,671	,560
	5 168	5 497	,647	,563
	5 498	5 843	,652	,551
	5 844	6 256	,634	,536
	6 257	6 772	,624	,524
	6 773	7 360	,603	,524
	7 361	8 010	,581	,497
	8 011	8 777	,575	,485
	8 778	9 701	,565	,470
	9 702	10 929	,539	,464
	10 930	12 467	,514	,434
	12 468	14 363	,469	,401
	14 364	17 125	,434	,371
	17 126	21 418	,382	,327
	21 419	28 540	,309	,264

Tabell IV (*fortsättning*)

	Rangklass			
	Nedre gräns	Övre gräns	DT	DÄ
	28 541	44 009	,197	,173
	44 010	106 612	,000	,000
Hybrida enheter	1	108	,376	,525
	109	211	,335	,327
	212	685	,237	,225
	686	963	,150	,141
	964	4 821	,000	,000
Numeriska enheter	1	100	,770	,712
	101	209	,614	,560
	210	359	,385	,369
	360	463	,209	,181
	464	918	,001	,001
Logogram	1	6	,215	,326

med skattad frekvens grundar sig alltså på de tilldelade DT-värdena. I kolumnerna för kontributionsvärdena Tidn, Ämn och För sättes i dessa fall —.

Rang

Två av de numeriskt ordnade listorna (1.1 och 2.1.1) ger ordens *rangnummer* Rnr: det frekventaste ordet har rangnummer 1, det närmast följande 2 osv. Det är här fråga om statistiskt korrekta rangnummer, vilket innebär att ord som ligger på samma frekvens också har samma rangnummer. Vid beräkningen har rangnummer som faller mellan två heltal avrundats neråt. Antalet olika rangnummer sammanfaller med antalet olika frekvenser i materialet.

Vokabulärsektion

Av praktiska skäl är hela ordmaterialet ($F \geqslant 1$) inte representerat i mer än två listningar, 2.2 och 2.3. Det är uppenbart att ord med enstaka belägg är av underordnat intresse exempelvis i listor sorterade efter fallande F. Dessa omfattar därför endast ord med frekvensen $F \geqslant 10$. Med utgångspunkt i den modifierade frekvensen Fmod har vidare framtagits en *basvokabulär,* som innehåller ord med Fmod $\geqslant 4{,}125$. Det ansatta minimivärdet är det som erhålles om en enhet har ett belägg i varje tidning. I anslutning till basvokabulären förekommer en lista över ord med $F \geqslant 10$ som inte har kvalificerat sig till basvokabulären. Denna sista vokabulärsektion innehåller extremt snedfördelade ord som trots relativt hög frekvens alltså inte kommer upp till Fmod-värdet 4,125.