Inledning

Del 2 av Nusvensk frekvensordbok innehåller resultat av en fortsatt undersökning av det material som låg till grund för del 1. Bearbetningarna har skett i enlighet med den beskrivningsmodell som presenterades där (s. XVI–XXIV). Detta innebär att den ordnivå som fokuseras i del 2 är lemmanivån (grafordsnivån och homografkomponentnivån behandlades i del 1). Här bildar alltså exempelvis substantivet *stack* med sublemmaformerna *stack, stacken, stackar* osv. en enhet liksom verbet *sticka* med sublemmaformerna *sticka, stack, stuckit* osv. (liksom också verbet *sticka* med sublemmaformerna *sticka, stickade, stickat* osv.). Att etablera ett autentiskt materials lemman är emellertid långt ifrån så enkelt som exemplet antyder. Huvudproblemen och de principer som har följts behandlas längre fram i inledningen.

Antalet bearbetningar av textmaterialet i denna del av ordboken är femtiotvå. Av dessa gäller fyrtionio olika aspekter på orden, de tre återstående olika aspekter på grafemen. Fyrtiotvå av bearbetningarna rörande orden (1.1.1–1.3) har karaktären av ordlistor eller lexikon. De sju övriga (1.4–1.8.3) utgörs av tabeller. Detta gäller också de tre grafemlistningarna (2.1–2.3), som bildar ett appendix.

Fullständigt förtecknas ordmaterialet i listning 1.2.2. Detta är ett initialalfabetiskt ordnat lexikon som för varje lemma redovisar samtliga belagda sublemmaformer jämte kvantitetsuppgifter. I 1.3 ges en finalalfabetisk listning av lemmanas uppslagsformer. I den mån dessa har frekvensen 10 eller högre ingår de också i listningen 1.1.1, som är ordnad efter fallande frekvens.

Listningarna 1.1.2–1.1.6, 1.2.1 och 1.3 ger olika aspekter på basvokabulären på lemmanivå. Uppgifter som rör ordklasserna ges i 1.1.6.1–1.1.6.15, 1.1.7.1–1.1.7.3, 1.1.8.1–1.1.8.3 och 1.7. I listningarna 1.8.1–1.8.3 anges enheternas fördelning på vissa grammatiska kategorier. Data rörande sublemmaformerna kan man finna i 1.1.7.1–1.1.7.3, 1.1.8.1–1.1.8.3, 1.2.1, 1.2.2 och 1.6. Listningar som speciellt rör tidningsmaterialen och ämnessfärerna är 1.1.4, 1.2.3 och 1.2.5.1–1.2.5.5 respektive 1.1.5, 1.2.4 och 1.2.6.1–1.2.6.6. Enheternas fördelning på frekvenser respektive ordlängder visas i 1.4 respektive 1.5.

Material, språk, stil

Materialet utgörs av en artikelsamling som omfattar en miljon löpande ord ur fem ledande morgontidningar 1965. Hur detta har insamlats, hur det är sammansatt i politiskt och regionalt avseende, hur det fördelar sig på tidningar, genrer och ämnessfärer m. m. har beskrivits i inledningen till del 1 (s. XIV–XVI).

Här skall åter framhållas att materialet är att betrakta just som en samling artiklar. Från statistisk synpunkt är det fråga om ett sampel ur en viss population. Samplet kan närmare beskrivas som en slumpmässigt vald samling texter av en rad vana skribenter som behandlar växlande ämnen i ledande morgontidningars språkliga miljö. Antalet skribenter är 569 och antalet artiklar 1387.

Slumpmässigheten gäller inom ramen för de restriktioner som uppställdes och som avsåg att öka materialets språkliga homogenitet och därigenom göra resultaten mera gripbara. Förutom osignerade bidrag (inklusive nyhetsbyråtelegram, annonser och ledare) uteslöts sportartiklar, kåserier, artiklar om språk (metaspråkliga texter), artiklar som innehöll längre citat, insändare och artiklar av utländska författare. Vidare ansågs skribenter från en och samma tidning inte böra få dominera, varför endast varannan artikel av dem som kom från Svenska Dagbladet medtogs. Självfallet gällde också den tekniska restriktionen att artiklarna skulle vara satta med hålremsa — en praktisk-ekonomisk förutsättning för undersökningens genomförande.

I strikt mening bildas den teoretiska population som samplet är draget ur av alla texter av samma skribenter under samma förutsättningar. Vid bedömningen av denna populations relevans bör man bland annat beakta att en lång rad andra skribenter kan förväntas ansluta sig till eller eftersträva språkformen i materialet, att det rör sig om ett centralt område inom standardspråkligt skriftspråk och att antalet läsare som berörs är mycket stort.

Att materialet inte är avsett att i detalj återspegla de fem tidningarnas språkbruk hindrar inte att det kan ge intressanta upplysningar om relationen mellan de deltexter som producerats av skribenterna inom var och en av tidningarna. Härvid är det av betydelse att tidningsmaterialens fördelning på de tre genrerna (A allmänna reportage, K kultursidesartiklar, U utrikeskorrespondenters artiklar) i stora drag är likartad. De procentuella värdena visas i tabell I.

Tabell I. *Tidningsmaterialens procentuella fördelning på genrer.*

	GHT	SvD	ST	DN	SDS	*Hela materialet*
A	33,0	46,0	50,9	31,5	46,4	42,6
K	53,3	43,5	37,4	54,7	43,8	46,0
U	13,6	10,5	11,7	13,8	9,8	11,4
Totalt	99,9	100,0	100,0	100,0	100,0	100,0

Från språkstatistisk synpunkt är flera andra ting av betydelse för förståelsen och utnyttjandet av undersökningens resultat. Vad först datas säkerhet i förhållande till materialet beträffar, kan man räkna med en mycket hög tillförlitlighet. Upprepade genomgångar och korrigeringar borgar för detta. Den kombinerat manuella och maskinella konfrontationen mellan ordboken på grafordsnivå och de flera hundra

tusen beläggen i homografkonkordansen har spelat en stor roll också i detta avseende. De många olikartade bearbetningarna av materialet har möjliggjort en kontinuerlig kontroll och justering.

Säkerheten gentemot populationen har två huvudaspekter, den som gäller makrodata och den som gäller mikrodata. Som makrodata räknar jag (explicita respektive implicita) uppgifter rörande antalet enheter på de olika beskrivningsnivåerna, enheternas fördelning på frekvensområden, fördelningen på kategorier av typen ordklasser, deklinationsklasser etc., sammansättningars och avledningars typer och kvantiteter och mycket annat av liknande art. Medan makrodata alltså hänför sig till de mera generella fenomenen, avser mikrodata de mera speciella, i vårt fall främst de enskilda orden.

För både makro- och mikrodata gäller att slumpvariationens roll relativt sett är större vid låg frekvensnivå än vid hög. Normalt är frekvensnivån lägre i fråga om mikrodata än i fråga om makrodata. Ettfrekvensorden *taktegelsvis* och *korvhandlarhåll* kunde, för att ta ett par exempel, också ha hamnat utanför vårt material. Detta skulle däremot inte vara att vänta för exempelvis adverb på *-vis* såsom grupp (representerad i ordboken utöver av *taktegelsvis* av *uppskattningsvis, tallriksvis, samtalsvis, kvartalsvis, massvis, ingressvis, givetvis* och många andra).

Vid tolkningen av frekvensuppgifterna är det av betydelse att hålla ytterligare några faktorer i minnet. En är ordens spridning över deltexter som tidningsmaterialen och ämnessfärerna. Denna har behandlats i inledningen till del 1 (s. XXVII–XXX). En annan är fraseologins roll. Den höga frekvensen för *del* beror exempelvis på att förbindelser som *en del, en hel del, ta del av* m. fl. är vanliga. I del 3 av ordboken skall enligt planen denna beskrivningsnivå behandlas. Fler faktorer kunde nämnas. Satsstrukturen är en, synonymin (och över huvud taget semantiken) en annan. Studier av dessa är angelägna forskningsuppgifter.

Man kan ställa frågan om ords frekvensvärden har någon psykologisk realitet. Detta har testats på olika håll genom i huvudsak två typer av undersökningar. Den ena gäller vad man kan kalla tillgängligheten hos orden (på engelska availability) och den andra förtrogenheten med orden (familiarity). Med tillgänglighet menas (graden av) egenskapen hos ett ord att inställa sig hos språkbrukarna när ett tema ges (t. ex. »vapen», »på en resa»). Med förtrogenhet menas (graden av) egenskapen hos ett ord att förväntas bli hört, sett eller använt av språkbrukarna. Denna egenskap undersöks genom att man låter försökspersoner rangordna enheterna i listor av bjudna ord. På det hela taget har man funnit en god överensstämmelse mellan resultaten av sådana här test och frekvensundersökningar. En avvikande ordgrupp bildas av de konkreta substantiven (*tandborste* osv.), som tenderar att vara mer förtrogna än frekventa. Detta bör man också beakta när man utnyttjar frekvensundersökningar.

Hur skall man så uppfatta undersökningens resultat från lingvistisk synpunkt? Vad som kan studeras direkt är språkyttringar, resultat av språkproduktion. Genom att studera dessa kan man ytterst komma åt

två saker. Den ena är av kvalitativ natur: det språkliga systemet med dess regler och lexikaliska enheter, dvs. ungefär det som i generativ grammatik betecknas som den lingvistiska kompetensen. Den andra är av kvantitativ natur: den storhet som styr vår användning av den lingvistiska kompetensen, dvs. vad man (med en omdefiniering av en tidigare föreslagen term) kunde kalla den stilistiska kompetensen. I vidaste mening avgör den stilistiska kompetensen valen av presentationsform, kompositionssätt, grammatiska konstruktioner, lexikaliska enheter osv. Genom denna styrning sker den viktning av elementen i den lingvistiska kompetensen som vi kan avläsa i språkyttringens kvantiteter på olika plan. (Det kan nämnas att styrning också kan ske mot avvikelser från systemet. Ett exempel är ordbokens *horrogod* — jämför *herregud* — som ytterst återgår på Sandro Key-Åbergs scenprator »O».) Denna viktning får tänkas innefatta en förmåga att kalkylera de producerade formuleringarnas effekt.

I mycket enkel form åskådliggör figur I relationen mellan kompetenserna och språkyttringarna. Den streckade linjen markerar den stilistiska kompetensens styrning av den lingvistiska. Motsvarande heldragna linje markerar produktionen av språkyttringar. Den nedåtriktade pilen representerar produktionsriktningen eller syntesen, den uppåtriktade perceptionsriktningen eller analysen. Det skall understrykas att figuren är mycket förenklad och bara är avsedd att illustrera vissa grunddrag i teorin.

Figur I. *Förhållandet mellan stil och språk, kompetens och språkyttring.*

Studiet av den population som samplet representerar återspeglar alltså väsentliga drag i dels det språkliga systemet, dels det stilistiska systemet. Man får å ena sidan upplysningar om vilka enheterna och reglerna på olika plan är. Å andra sidan får man fram uppgifter om i vilken utsträckning de kommer till användning.

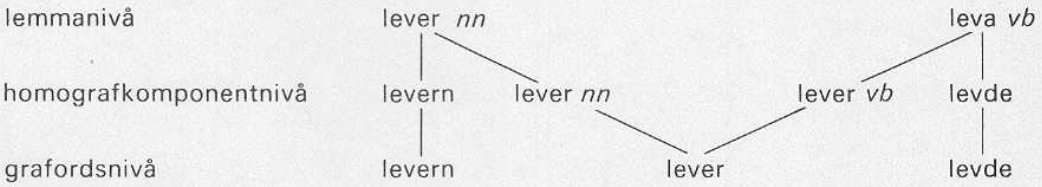
Kvalitativa aspekter

I detta avsnitt står givetvis lemmat och lemmatiseringen i centrum. Några ord om beskrivningsnivåerna inleder.

Beskrivningsnivå

I korthet skall de viktigaste begreppen och termerna belysas med hjälp av exemplet i figur II. Den första nivån bildas av *graforden,* som kan

vara *homografer* (*lever*) eller *heterografer* (*levern, levde*). Analysen av graforden leder närmast till den nivå där enheterna utgörs av *homografkomponenter* (substantivformen *lever,* betecknad *lever nn;* verbformen *lever,* betecknad *lever vb*) och samma heterografer som på grafordsnivån (*levern, levde*). Sedda från den tredje nivåns (lemmanivåns) synpunkt är enheterna på homografkomponentnivån *böjningsformer* (*lever nn, levern* respektive *lever vb, levde*) eller *variantformer* (t. ex. *karakteriserar, karaktäriserar*). Dessa former bildar *lemman* (*lever nn, leva vb, karakterisera/karaktärisera vb*). I anslutning härtill kan man som en sammanfattande benämning på böjningsformer och variantformer använda termen *sublemmaformer* (eller kortare *sublemman*).



Figur II. *Beskrivningsnivåer, homografseparering, lemmatisering.*

Vikten av att precisera vilken nivå man avser i ett visst sammanhang framstår tydligt, när man jämför de summauppgifter för antalet enheter på de olika nivåerna som undersökningen har givit. I runda tal är värdena följande. Materialet innehåller 71 000 lemman, som representeras av 112 000 sublemmaformer, som motsvaras av 103 000 graford. Skillnaden mellan antalet lemman och antalet sublemman är alltså drygt 40 000, mer än hälften av antalet lemman.

Ord kan definieras på flera andra sätt. Ett av dem aktualiseras i denna del av ordboken. Grafordet sådant det bestämts i undersökningen är normaliserat i två avseenden: dels är distinktionen mellan versal (stor bokstav) och gemen typ (liten bokstav) neutraliserad (graforden återges i del 1 med kapitäler), dels är junkturella grafem (skiljetecken o. d.) avskilda. Tabell 3.1 i del 1 gäller just grafemen i graforden definierade på detta sätt. Emellertid är naturligtvis också data om grafemen i de onormaliserade graforden av intresse. Sådana uppgifter ges i tabell 2.2 i denna ordboksdel. För att skillnaden gentemot de normaliserade graforden skall markeras har de onormaliserade graforden där benämnts *textord.*

I detta sammanhang skall också nämnas att bearbetningarna på grafemnivå innefattar en listning av digrammen (tvåkombinationerna) i sublemmaformerna. Det gäller här tabell 2.3. Lägg märke till att den tomma positionen före respektive efter varje ord räknas som en enhet. Ett fall som n + tomrum avser alltså n i ordslut.

Principer för lemmatisering

Med lemmatisering menas sammanföring av böjningsformer respektive variantformer i lemman. Den homografseparering som redovisas i del 1

av ordboken utgör ett väsentligt steg på vägen mot en lemmatisering. Genom den har nämligen homografkomponenternas lemmatillhörighet bestämts. Se s. XVII–XX med figurerna II och III i del 1.

Definitionen av ett *lemma* tar fasta just på böjning och variation (del 1, s. XVIII): ett lemma är 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.

De båda flexionsserierna *sticka, stack, stuckit* etc. och *sticka, stickade, stickat* etc. konvergerar bland annat i formen *sticka,* men divergenserna visar inte fakultativ variation (de åtföljs av en ändring i betydelsen) och serierna bildar därför två olika lemman. Serierna *simma, sam, summit* etc. och *simma, simmade, simmat* etc. konvergerar bland annat i formen *simma,* och eftersom divergenserna visar fakultativ variation har vi att göra med ett enda lemma.

Undersökningens formrikaste lemma är *ge* med 24 belagda sublemmaformer. Denna mängd förklaras till en del just av formvariation: imperativ *ge/giv,* infinitiv *ge/giva, s*-form i presens *ges/gives,* supinum *gett/givit* och *getts/givits.*

De båda oböjliga enheterna *också* och *även* återigen har ingen konvergens och bildar alltså var sitt lemma. Detsamma gäller exempelvis *inte, icke* och *ej* liksom *ned* och *ner.* Att oböjliga ord innefattas under beteckningen flexionsserier innebär bara att de betraktas som instanser av det specialfallet att en serie består av ett enda element. Ytterligare ett exempel är substantivet *gläntt* (i förbindelsen *på gläntt*).

Adjektivet *dålig* bildar ett lemma (med sublemmaformerna *dålig, dåligt* osv.) och adjektivet *värre* ett annat (med sublemmaformerna *värre, värst, värste* osv.). Likaså bildar exempelvis (det starka) verbet *vara* ett lemma och verbet *är* ett annat. Som flexionsserier accepteras med andra ord inte fall av s. k. suppletiv böjning (se s. XIX i del 1).

Flexionsserierna hämtas i princip från Illustrerad svensk ordbok, utgiven av Bertil Molde, med de modifikationer som beskrivningsmodellen fordrar. Enligt kutym ger denna ordbok bara de former som är karakteristiska för respektive serier, men vid analysen utnyttjas de fullständiga serierna. Uppgifterna i ordboken kompletteras vidare med de upplysningar materialet ger och, då så krävs, den kännedom forskningsgruppens medlemmar har om språket.

Den variation som lemmadefinitionen förutser är inte bara böjningsvariation som i fallet *simma* eller *kemi* med de bestämda formerna *kemin* och *kemien* utan också uttrycksvariation som i fallet *karakterisera/karaktärisera* eller *essä/essay.* Den konvergens som alltid krävs för sammanföring finner man i de båda sista exemplen i uttalet. Korshänvisningar har inte införts; *camouflera* får sökas under *kamouflera* osv.

Omvänt är det skriftformen som ger konvergensen i exempelvis fallet *förut* 'tidigare'. De två uttalsformer som sammanhålls genom denna har betoningen på första stavelsen jämte grav accent respektive beto-

ningen på andra stavelsen. Ett annat exempel är *kilo* med tj-ljud eller k-ljud. Av speciella skäl har uttalskriteriet inte utnyttjats vid lemmatiseringen av proprier (se härom längre fram i detta avsnitt).

Ordet *sedan* kan i vissa sammanhang uttalas likadant som *sen* med samma funktion. Bestämmer man sig för att gå på *lexikaliskt uttal,* vilket förefaller rimligt i en vokabulärundersökning och därför har skett, får de båda orden emellertid ingen konvergens och bildar följaktligen var sitt lemma. På samma sätt blir det med *driftsäkerhet* gentemot *driftssäkerhet,* som alltså behandlas analogt med den morfologiska parallellen *ortnamn* gentemot *ortsnamn.* Likaså bildar *dumhuvud* ett lemma och *dumhuve* ett annat osv.

Genom lemmadefinitionens utformning blir gränslinjen mellan homografi och polysemi klar. *Homografi* avser den relation som råder mellan i skrift identiska ordformer tillhörande olika lemman (extern homografi) eller samma lemma (intern homografi). *Polysemi* avser betydelsevariationen inom ett lemma. Uppläggningen medger en konsekvent lexikologisk beskrivning. Det uppkommer skiljaktigheter gentemot traditionell lexikografi. I gängse ordböcker finner man exempelvis två olika enheter *skarv -en -ar* men en enhet *ackord -et -ø.* Enligt lemmadefinitionen får vi emellertid en (polysem) enhet i båda fallen. Semantiskt sett är det inte mer anmärkningsvärt att *skarv* kan avse en fog eller en fågel än att *ackord* kan avse ett beting eller en samklang av toner.

Varje lemma tilldelas en etikett, kallad *uppslagsform,* som principiellt utgörs av vad man brukar beteckna som grundformen plus en klassbeteckning (grammatisk specifikation). Den står alltså som representant för alla sublemmaformerna (inklusive grundformen). Lemmanas uppslagsformer behandlas i nästa avsnitt.

Efter dessa allmänna synpunkter skall några av de mera speciella problemen tas upp. Av ordtyperna har de rent numeriska orden (siffertalen) från början lämnats utanför den grammatiska bearbetningen. De har sålunda inte homografseparerats och lemmatiseras följaktligen inte heller. Att de fyra beläggen på 491 hänför sig till Görlings bok respektive Sjömans film och de två beläggen på 1984 till Orwells bok markeras alltså inte särskilt. Inte heller görs någon åtskillnad mellan grundtal och ordningstal, varför någon sammanföring med alfabetiskt skrivna numeraler inte utförts. Lemmatisering av logogram (§ etc.) kommer naturligtvis inte heller i fråga.

Det är sålunda de alfabetiska orden och hybridorden (t. ex. *30-årig*) som lemmatiseras. Dessa svarar för 99,2 % av enheterna på homografkomponentnivån. Lemmatisering över ordtypsgränsen sker i två typfall. Det ena är om en flexionsserie skär över denna, t. ex. *LP, LP:n, LP-n* eller *Demokritos, Demokritos'.* Det andra är om det föreligger alternativa former med konvergens i uttalet (sålunda en tillämpning av huvudregeln beträffande variation), t. ex. *sen se'n* eller *Genèveöverenskommelsen Genève-överenskommelsen.* Dock sammanförs inte alfabetiska ord med sådana hybridord som innehåller numeriska grafem: ett ord som *50-lapp* kunde lika gärna föras till lemmat *femtilapp* som lemmat *femtiolapp.* I linje med sådana fall ligger principen att inte heller sam-

manföra t. ex. *7:e* och *sjunde* (vilket dessutom hänger samman med att *7* enligt ovan inte har separerats i grundtal och ordningstal).

Abbreviationer (*an*) och utländska enheter (**) har i princip lämnats utan åtgärd vid homografsepareringen och lemmatiseras inte. Bland abbreviationerna utgör exempelvis *A.*, *a.*, *A* och *a* skilda enheter på lemmanivå. Enheterna är i allmänhet oböjliga, men undantagsvis förekommer genitivformer som *SACO:s* och dessa har då förts samman med grundformen.

Om en ursprunglig abbreviation är belagd i bestämd form och inte bara i genitiv, övergår den till klassen substantiv. Former som *LP, LP:n, LP-n, LP:ns* sammanförs således till lemmat *LP nn -n*. Detta får betydelse också för sammansättningar: *Parker-LP* betraktas som substantiv trots att böjningsformer saknas i materialet. *EP* kvarstår däremot som abbreviation på de böjningskriterier som angivits.

Eftersom utländska enheter inte har lemmatiserats, har böjningsformer av dem inte presenterats under de uppslagsformer man skulle kunna ansätta i vederbörande språk. En sådan lemmatisering inom främmande språksystem har inte ansetts ligga inom undersökningens ram. Engelskans *thing* och *things* blir skilda uppslagsformer. Den etruskiska genitiven *zilacal* har inte förts till grundformen *zilac* osv. Principen kan någon gång leda till att enheter från olika språk strålar samman i en uppslagsform: *die* kan vara dels tyskans artikel, dels engelskans verb. Ett främmande ord som i materialet visar svensk böjning klassificeras i enlighet med svenskans system. Det från spanskan lånade *plaza* förekommer exempelvis i formen *plazan* och kvalificerar sig därmed för klassen *nn* (substantiv), inte ** (utländska enheter).

Klassen proprier (*pm*) kräver en något utförligare kommentar. Dess medlemmar bestäms på grundval av versalitet (förekomst av stor bokstav). Alla ord med initial versal är visserligen inte proprier, men inget ord utan initial versal är å andra sidan proprium. Grundläggande är att den räjong som ett propriums versal hänför sig till sammanfaller med ordet. Härigenom bortfaller alla de enheter som har versal uteslutande därför att de inleder en mening och sålunda normaliseras i enlighet med vad som sades i avsnittet Beskrivningsnivå ovan. En meningsbörjande versal kan emellertid ha den dubbla funktionen att också hänföra sig till det inledande ordet, som då blir proprium. Som meningar behandlas också syntagmer som bildar proprier på frasnivå. I *Röda havet* urskils alltså ett adjektiv (*röd*) och ett substantiv (*hav*), i *Nya Älvsborg* ett adjektiv (*ny*) och ett proprium (*Älvsborg*). Varje ord i språket kan i själva verket göras till proprium. Vi får därför inte bara proprier som *Älvsborg* utan också proprier som *Utvandrarna, Utsålt* (en revymusical), *Naturvårdsnämnden* etc.

I enlighet med huvudprincipen blir fall av typen *Göteborgskonstnärerna* däremot inte proprium, eftersom versalen endast hänför sig till en del av ordet. En annan typ av ord med versal som inte blir proprium kan exemplifieras med *Kr.* (för *Kristus*). Enheter av detta slag klassificeras i flödesgrammatiken som abbreviationer. Ytterligare en typ utgörs av utländska enheter sådana som tyskans *Liebe* och engelskans *Swedish*.

Det fonematiska kriteriet (konvergens i uttalet) har inte utnyttjats vid lemmatisering av proprier. Skälet är att proprier i stor utsträckning utgör en internationell ordskatt med de uttalskomplikationer som detta medför. *Tanzania, Paris, Kosygin* och *Stockholm* tillhör flera olika vokabulärsystem. Principen har följts också vid fall av varierande translitterering som *Krusjtjev, Chrusjtjov* osv. Man kan inte generellt vara säker på att inte olika namn avses.

Utöver grundform och genitivform förekommer hos proprier i enstaka fall andra former. Lemmat *Saab pm -s* omfattar t. ex. formerna *Saab, Saaben* och *Saabs* (formen *SAAB* tillhör klassen abbreviationer). Några gånger är den bestämda formen lemmats uppslagsform och den obestämda sublemmatisk, så i fallen *Kanslihuset* och *Stadshuset*. Exemplet lyder: »Man var ju — utom i Kanslihus och Stadshus — ytterst obenägen att över huvud taget se något positivt i ett kommunalt engagemang.»

I överensstämmelse med att polysemin, betydelsevariationerna inom lemmana, inte redovisas på den beskrivningsnivå som ordboken avser har inte proprierna separerats med avseende på referenter (individer). *Nilsson* kan bland annat syfta på sångerskan *Birgit Nilsson* och politikern *Torsten Nilsson*. *Chaplin* kan syfta på skådespelaren eller filmtidskriften osv.

I del 1 av ordboken förekom fjorton ordklasser eller om man så vill klasser av ord. I denna del tillkommer en femtonde. Det gäller här en skara enheter som från grammatisk synpunkt utgör fragment och konstellationer av speciella slag. Exempel är *teater-* och *konsert-* (som i »teater-, konsert- och biografsalong»), *Strömfjord–Los* (som i »Strömfjord–Los Angeles») och *Stockholm–Malmö*. Ordgruppen ges blank klassbeteckning. Behovet att presentera den förelåg inte i del 1, eftersom medlemmarna i klassen aldrig är homografa.

Lemmatiseringen liksom ansättandet av uppslagsformer (inklusive klassbeteckningar) har utförts med hjälp av ett programsystem som baserats på svenskans morfologi. Se härom Staffan Hellbergs uppsats Automatisk lemmatisering (1971; dupl.). Körningarna resulterade i en till cirka 95% korrekt avgränsning av lemman. Cirka 85% av de korrekt bestämda lemmana fick helt riktig uppslagsform. Efter granskning av resultatet genom två av varandra oberoende arbetslag utfördes maskinell korrigering i flera omgångar.

Lemmanas uppslagsformer

Varje lemma förses som nämnts med en *uppslagsform* som i princip består av en *ordkropp* och en *klassbeteckning*. Det som sägs i avsnittet Enhet på s. XX–XXIV i del 1 gäller i tillämpliga delar också här. Klassbeteckningen omfattar normalt två delar, en *ordklassmarkering* och en *lemmamarkering* (som i de flesta fall är en böjningsangivelse). Någon *sublemmamarkering* blir det självfallet inte tal om när det gäller lemmats uppslagsform. En sådan tillkommer i stället varje sublemmaform som är internt homograf. Ett lemma med de belagda sublemmaformerna kan alltså se ut så här:

hus *nn -et*
hus *gru plu*
hus *gru sin*
husen
husens
huset
husets

Sublemmamarkeringarna anger att det är fråga om grundformen i pluralis respektive grundformen i singularis. Ordningen mellan sublemmaformerna är alfabetisk enligt den specifikation som givits på s. XXIV–XXV i inledningen till del 1.

För utrymmes vinnande har sublemmaformerna undertryckts i bearbetning 1.2.2 (lexikon med sublemmaformer), om tre villkor är uppfyllda. För det första skall endast en sublemmaform vara belagd inom vederbörande lemma. För det andra skall den belagda formen vara identisk med ordkroppen i lemmats uppslagsform. För det tredje får den belagda formen inte ha sublemmamarkering. Ingen information går på detta sätt förlorad, och det vunna utrymmet har kunnat användas till att öka antalet medtagna bearbetningar.

Ordkroppen i uppslagsformen överensstämmer med lemmats grundform (om sådan finns) enligt gängse grammatisk och lexikologisk beskrivning: obestämd form singularis för substantiv, infinitiv aktivum för verb osv. Många gånger är grundformen inte belagd i materialet.

En del lemman saknar vad man brukar mena med grundform. Ett exempel är *kläder,* ett annat *tretiden.* I sådana fall ansätts den form som utgör stickord i Illustrerad svensk ordbok, om denna har ordet. Om inte, får som i andra sammanhang materialet och forskningsgruppens språkkännedom fälla avgörandet.

Som man kan vänta saknas i Illustrerad svensk ordbok särskilt uppgifter om en mängd sammansättningar, karakteristiska för ett autentiskt material. I dessa fall ansätts en uppslagsform som bygger på den som gäller för (respektive skulle ha gällt för) motsvarande simplex. Ett belägg som *Adenby-flaskorna* ger alltså uppslagsformen *Adenby--flaska.*

Vid variation hos ett lemmas grundmorfem — fall som *essä/essay, dra/draga* — väljs den variant som är vanligast i materialet till uppslagsform. I fråga om verben fälls därvid avgörandet av i första hand infinitiven, i andra hand infinitivens *s*-form och i tredje hand presensformen. Substantiven behandlas på motsvarande sätt. Paradigmets former gås igenom i tur och ordning. Kan inget avgörande träffas på basis av frekvens, gäller att försvenskad stavning går före utländsk (*rondör* före *rondeur*), hybridord går före alfabetiskt ord (*Stockholms--press* före *Stockholmspress*) och form med versal går före form utan (*Väststat* före *väststat*).

Också inom de böjningsmorfem som valts till lemmamarkering förekommer variation, t. ex. i fall som *varvsindustrin/varvsindustrien.* I den rätt stora grupp det här gäller bestäms lemmamarkeringen inte av alternativens frekvens i det enskilda fallet utan av frekvensen inom

ordgruppen som helhet. Liksom uppslagsformen i det citerade fallet blir *varvsindustri nn -n* blir den därför i ett fall som *konfektionsindustrien* (med endast denna form belagd) på samma sätt *konfektionsindustri nn -n*. Vid neutrala ord av typen *trä, bageri* är däremot den längre formen vanligast. Uppslagsformerna inom denna grupp får därför utformningen *trä nn -et*, även om som i fallet *knä* formen med *-t* är vanligare än den med *-et*.

Ingen lemmamarkering förekommer inom vissa grupper av ord. Hit hör de utländska enheterna, klassen pronomen, de numeriska graforden inom klassen numeraler och naturligtvis den speciella gruppen med blank klassbeteckning. Vidare gäller detta oböjliga enheter, t. ex. *ifall kn*. Inom flekterande ordklasser förekommer enheter med inskränkt böjning som enbart får ordklassmarkering, t. ex. *gläntt nn, värre av* och *är vb*. Slutligen gäller det klassen abbreviationer utom de fall som slutar på *-s, -x* eller *-z* (t. ex. *SFS*), där klassbeteckningen är *an -ø* (vid sublemmaformen anges samtidigt om den representerar grundformen eller genitivformen).

I vissa fall slutligen innehåller uppslagsformen en notsiffra. Den kan avse en förklaring till ordkroppen och sätts då med rak stil eller till klassbeteckningen och sätts då med kursiv stil. Noternas text är placerad längst bak i boken.

Kvantitativa aspekter

De kvantitativa aspekterna på materialet behandlades tämligen utförligt i inledningen till del 1. Här skall huvudsakligen sådana frågor tas upp som är speciella för del 2.

Frekvens

De frekvensuppgifter för lemman och sublemmaformer som baseras på skattningen av frekvenserna för de 24 mest frekventa homografernas komponenter markeras som tidigare med asterisk. (Däremot utsätts ingen asterisk vid motsvarande Fmod-värden i de bearbetningar där sådana anförs.) Konfidensintervallen för dessa homografkomponenters frekvenser gavs i tabell III i inledningen till del 1 (s. XXVI–XXVII). I tabell II på nästa sida ges nu på motsvarande sätt konfidensintervallen för frekvenserna hos de lemman som innehåller en eller flera skattade delfrekvenser.

Bestämningen av konfidensgränserna för de aktuella lemmanas frekvenser har skett så, att de nedre gränsvärdena för sublemmaformernas frekvenser har adderats för att ge den nedre gränsen för ett lemmas frekvens och de övre gränsvärdena har adderats för att ge den övre gränsen. Detta enkla förfarande utgör en approximation men torde inte ge något fel av betydelse. Approximationen ligger egentligen däri att delfrekvensernas konfidensgränser inte är beräknade på ett symmetriskt sätt och att man därför adderar vad som faktiskt är asymmetriska konfidensgränser. Detta hänger i sin tur samman med att vissa skattningar på grund av språkets natur är grundade på ett mycket litet antal observationer.

Tabell II. *Konfidensintervall för skattade frekvenser.*

Lemma		Skattad frekvens	Konfidensintervall Nedre gräns	Konfidensintervall Övre gräns
att	*ie*	12 534	12 054	13 023
att	*kn*	11 233	10 744	11 713
av	*ab*	204	136	290
av	*pp*	15 283	15 196	15 351
de	**	261	194	343
den	*al*	19 188	18 325	20 059
den	*pn*	26 879	26 009	27 742
det	**	8	1	39
en	**	44	14	103
en	*ab*	9	1	41
en	*al*	25 887	25 510	26 229
en	*nl -s*	2 726	2 389	3 097
en	*pn*	85	52	148
från	*pp*	4 042	4 035	4 042
för	*ab*	366	282	467
för	*kn*	142	91	210
för	*pp*	11 807	11 684	11 916
föra	*vb -de*	345	314	400
ha	*vb -de*	12 924	12 910	12 924
han	*pn*	9 083	9 072	9 083
I	**	51	14	130
I	*an*	26	3	93
I	*nl*	15	3	63
i	*pp*	29 461	29 365	29 514
kunna	*vb -de*	6 468	6 460	6 468
man	**	31	13	61
man	*nn*[1]	593	546	653
man	*nn*[2]	99	64	145
man	*pn*	7 103	7 029	7 167
med	**	6	1	27
med	*ab*	339	259	435
med	*an*	6	1	27
med	*pp*	11 700	11 604	11 782
men	**	6	2	20
men	*kn*	5 677	5 657	5 685
men	*nn -et*	6	2	30
om	*ab*	376	305	459
om	*kn*	2 283	2 125	2 448
om	*pp*	5 199	5 028	5 366
på	*ab*	229	159	319
på	*pp*	13 932	13 842	14 002
sig	*pn*	5 559	5 550	5 559
som	**	9	1	42
som	*kn*	6 079	5 701	6 467
som	*pn*	13 522	13 133	13 902
så	*ab*	3 979	3 927	4 024
så	*kn*	122	91	160
så	*pn*	125	93	164
så	*vb -dd*	10	6	31
till	*ab*	488	406	590
till	*pp*	9 865	9 763	9 954
var	*ab*	131	100	171
var	*pn*	457	420	500
vara	*vb -ø*	7 110	7 057	7 169
är	*vb*	15 509	14 988	15 509

[1] *Lemmat* man *-en män.*

[2] *Lemmat* man (*utan sin bes*) *man.*

I tabellerna 1.8.1–1.8.3 anges enheternas fördelning på vissa grammatiska kategorier (komparationsklasser, deklinationsklasser, genusklasser, konjugationsklasser). Värdena i dessa tabeller bygger på en genomgång av materialet enligt följande förutsättningar. De belagda ordformerna ger långt ifrån alltid upplysningar om lemmanas klasstillhörighet. I den utsträckning som uppgifterna har kunnat kompletteras med hjälp av Illustrerad svensk ordbok har detta skett. Anges i denna ordbok flera alternativ, har det första valts. I en lång rad fall, främst när det gäller sammansättningar, har uppgifterna emellertid inte kunnat kompletteras på detta sätt. Forskningsgruppens kännedom om språket har då fått fälla utslag. Självfallet har det i viss utsträckning blivit fråga om rätt svåra avväganden.

Spridning

Samma spridningsvärden som tidigare, dispersion och kontribution, har använts (se del 1, s. XXVIII–XXX). Likaså har enheter med skattad frekvens tilldelats dispersionsvärden enligt samma princip som förut.

De har alltså fått de värden på DT och DÄ som utgör medelvärdena i de rangklasser de tillhör enligt tabell III. En rangklass bildas av 100 enheter med icke skattad frekvens. Härvid är dock att märka att alla enheter med samma frekvens som den lägsta i en rangklass förs till denna rangklass.

Tabell III. *Medelvärden av DT och DÄ.*

Rangklass				Rangklass			
Nedre gräns	Övre gräns	DT	DÄ	Nedre gräns	Övre gräns	DT	DÄ
1	126	0,947	0,902	3 139	3 323	0,704	0,586
127	229	0,916	0,863	3 324	3 437	0,700	0,597
230	333	0,900	0,815	3 438	3 537	0,692	0,607
334	435	0,881	0,794	3 538	3 649	0,662	0,561
436	537	0,860	0,780	3 650	3 781	0,640	0,541
538	641	0,864	0,751	3 782	3 917	0,643	0,559
642	742	0,870	0,781	3 918	4 082	0,669	0,557
743	850	0,838	0,742	4 083	4 245	0,624	0,526
851	958	0,830	0,712	4 246	4 406	0,643	0,559
959	1 064	0,816	0,705	4 407	4 597	0,611	0,543
1 065	1 172	0,805	0,687	4 598	4 804	0,633	0,503
1 173	1 273	0,826	0,742	4 805	5 061	0,608	0,514
1 274	1 382	0,832	0,706	5 062	5 331	0,602	0,506
1 383	1 486	0,802	0,670	5 332	5 625	0,587	0,499
1 487	1 600	0,807	0,729	5 626	5 974	0,577	0,489
1 601	1 701	0,796	0,692	5 975	6 412	0,559	0,461
1 702	1 826	0,791	0,692	6 413	6 868	0,544	0,452
1 827	1 949	0,796	0,683	6 869	7 434	0,527	0,444
1 950	2 053	0,773	0,639	7 435	8 124	0,527	0,428
2 054	2 154	0,769	0,662	8 125	8 933	0,486	0,417
2 155	2 281	0,738	0,625	8 934	10 054	0,475	0,398
2 282	2 392	0,755	0,635	10 055	11 436	0,436	0,361
2 393	2 542	0,723	0,620	11 437	13 427	0,388	0,332
2 543	2 674	0,754	0,620	13 428	16 254	0,345	0,290
2 675	2 806	0,723	0,583	16 255	21 158	0,277	0,239
2 807	2 972	0,710	0,601	21 159	31 197	0,176	0,151
2 973	3 138	0,690	0,585	31 198	71 178	0,000	0,000

Som ett slags grammatisk kontributionsuppgift kan man betrakta kolumnen Sub i listorna 1.1.7 och 1.2.1. Där anges för varje lemma hur många olika sublemmaformer som är belagda.

Rangnummer och ordningsnummer

I lista 1.1.1, som förtecknar lemmanas uppslagsformer efter fallande frekvens, ges enheternas rangnummer. Dessa har liksom tidigare beräknats så att högsta frekvens ger lägsta rangnummer, näst högsta frekvens näst lägsta rangnummer osv. Alla enheter med samma frekvens får alltså samma rangnummer. Observera att rangnumren inte bildar en följd av jämnt stigande tal. Om ett framräknat värde ligger mellan två heltal, avrundas det till närmast lägre heltal. Rangnummer meddelas också i tabell 1.4, som anger enheternas fördelning på frekvenser.

Enheterna i den numeriskt ordnade listan över basvokabulären 1.1.2 har försetts med ordningsnummer. Här får enheten med högst Fmod

nummer 1, den därpå följande nummer 2 osv. i jämnt stigande talordning. Utan att vara ett faktiskt rangnummer — som bör baseras på observerad frekvens — ger ordningsnumret en upplysning om relativ kvantitet. Så fungerar det i den initialalfabetiskt ordnade listan över basvokabulären 1.2.1, dit det har fått följa med ordkroppen från listan 1.1.2. Man kan också se ordningsnumret som ett identitetsnummer med vars hjälp man snabbt kan återfinna en enhet från den initialalfabetiska listan i den numeriskt ordnade.

Korrelation

För att underlätta en jämförelse mellan de vanligaste lemmana i hela materialet och dessa enheter i de olika tidningsmaterialen respektive ämnessfärerna ges i listningarna 1.2.3 och 1.2.4 korrelationsvärden grundade på de 1 000 vanligaste enheterna i hela materialet. Siffrorna anger en enhets rangnummer i vederbörande deltext minus rangnumret i hela materialet. Högre rangnummer i deltexten än i hela materialet ger alltså ett positivt tal i listningen, lägre rangnummer ger ett negativt tal och samma rangnummer en nolla. Är frekvensen 0 i en deltext, sättes ett streck i kolumnen för denna deltext. Detta inträffar i undantagsfall som abbreviationen *bl* med rangnummer 266, som har frekvensen 0 i SvD (där i stället formen *bl.* används).

Om totalfrekvensen är skattad, sättes asterisker på korrelationsvärdenas platser. I detta fall är nämligen frekvenserna i deltexterna och därmed rangnumren inte kända. Vid den allmänna beräkningen av rangnumren i deltexterna tilldelas enheter med skattad totalfrekvens samma rangnummer i alla deltexter som de har i hela materialet. Härigenom undviks de systematiska förskjutningar i rangnumren i deltexterna som annars skulle ha inträffat.

Vokabulärsektion

Som tidigare nämnts tar två listningar upp hela antalet enheter, nämligen den initialalfabetiska 1.2.2 (som också förtecknar sublemmaformerna) och den finalalfabetiska 1.3. Övriga listningar omfattar sådana delar av vokabulären som har bedömts vara relevanta från den synpunkt som anlagts i varje särskilt fall.

Åtskilliga av listningarna rör basvokabulären. Lägg märke till att dess enheter har markerats med halvfet stil i den finalalfabetiska listningen 1.3. Vid beräkningen av DT (respektive DÄ) och därmed Fmod har noggrannheten ökats något (avrundningsfelen vid beräkningarna motverkats) i jämförelse med del 1 av ordboken. Härigenom kan i enstaka fall sista siffran i ett värde avvika en enhet från motsvarande värde i den delen. Detta är i och för sig utan betydelse men nämns för ordningens skull. I samband härmed har villkoret för tillhörighet till basvokabulären avrundats till Fmod >4 (gentemot Fmod $\geqslant 4{,}125$ i del 1).

Listningarna 1.2.3–1.2.6.6 behandlar de »1 000 vanligaste enheterna» i hela materialet respektive de enskilda tidningsmaterialen och ämnessfärerna. Denna avgränsning skall förstås på följande sätt. Drar man

gränsen exakt vid den tusende enheten, hamnar man normalt i en alfabetiskt ordnad följd av enheter med samma frekvens. För att inte behöva skära i en sådan ordgrupp kan man som här har skett specificera såsom nödvändigt och tillräckligt villkor för tillhörighet till en viss lista att enheternas rangnummer skall vara högst 1 000. Detta innebär att de faktiska antalen enheter i de aktuella listorna varierar något, i vårt fall mellan 961 och 1 025.

Särskilda villkor har uppställts för listorna 1.1.7.1–1.1.7.3 och 1.1.8.1–1.1.8.3. Avsikten har varit att ta fram de adjektiv, substantiv och verb som har en rik flora av sublemmaformer respektive är hårt knutna till en enda sublemmaform.