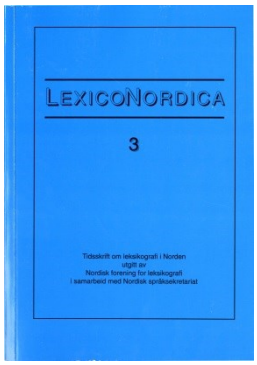
LexicoNordica

Titel:	Rätt och fel i korpusen
Forfatter:	Martin Gellerstam
Kilde:	LexicoNordica 3, 1996, s. 35-48
URL:	http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



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 LexicoNordica (1-16) 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.

Martin Gellerstam

Rätt och fel i korpusen

Today, language corpora are necessary requisites for the compiling of dictionaries. But to what extent should they have an impact on the lexical description? Opinions differ between lexicographers who use the corpus more or less as an inspiration for lexical description and those who rely heavily on corpus data (e.g. the wording of examples in Collins COBUILD). To what extent is a corpus reliable as a source for lexical data? This question is discussed against the background of a new Dutch spelling guide based on a 50 million word corpus.

1. Korpus och ordböcker

Att använda korpus när man gör ordböcker – över dagens eller gårdagens språk – är idag närmast en självklarhet. Historiska ordböcker har länge gjort det även om korpusarna kanske inte alltid varit så stora och de har härbärgerats på papperslappar snarare än i digital form. Men någon form av korpus har funnits vilket man i dagens korpusboom ibland glömmer bort. Enspråkiga ordböcker över det moderna språket har tidigare inte varit lika noga med att redovisa en korpus. Det viktiga har varit att man har fått med aktuella ord, ofta genom att konsultera andra ordböcker och nyordssamlingar. Textexemplen har företrädesvis konstruerats av redaktionen. Detsamma har tidigare gällt tvåspråkiga ordböcker.

Läget har på senare tid radikalt förändrats, bland annat genom korpuslingvistikens och datateknikens intåg i lexikografin. Men även om det idag är självklart att man skall ha en korpus så är det inte självklart hur långt man skall använda den. Till skillnad från de historiska ordböckerna så handskas nuspråkiga ordböcker lite mindre varsamt med exemplen: man talar vanligen inte om var de kommer ifrån (direkta beläggsuppgifter saknas) och man anpassar dem normalt för att de inte skall vara för långa eller komplicerade eller på annat sätt ägnade att förbrylla läsaren. I de få fall – jag tänker särskilt på Collins COBUILD – där man låter korpusen ha vitsord också ifråga om exempel, utsätter man sig för hård kritik från lexikografer som visserligen använder korpusar men inte går så långt i renlärighet. Man vill bygga på korpusen men – i likhet med Linje två i den svenska kärnkraftsomröstningen – göra det "med förnuft".

Men hur långt har egentligen korpusen vitsord? Man kan säga att det finns en skala från korpusskeptiker, som vet vad som är den goda normen och inte låter verkligheten störa, till korpusaktivister som är beredda att låta korpusen bestämma allt. Korpusskeptikerna finns framför allt i länder med starkt normerande tradition, möjligen kopplat till en rätt stor skillnad mellan tal och skrift. Särskilt den västeuropeiska traditionen har varit starkt pro-korpus, något som kanske inte bara har en ideologisk bakgrund utan också bottnar i bättre datautrustning än vad som finns i många andra länder. Vi som anser oss vara korpusanvändare med förnuft är kanske inte särskilt rädda för kritik från korpusskeptikernas sida. Snarare behöver vi värja oss mot föreställningen att korpusen skulle kunna användas mer än vad vi är beredda att göra. Låt oss materialisera en lite mer ortodox inställning ifråga om korpusanvändning och se hur en person som John Sinclair formulerar tankarna bakom COBUILD (Sinclair 1987). I förordet till ordboken tar Sinclair upp frågan om korpus ("usage") och auktoritet ("authority") och diskuterar förhållandet mellan dessa två begrepp:

> These concepts must support each other or noone will respect either of them. If /.../ authority is not backed up by usage, then noone will respect it. It will be seen as unrealistic, arbitrary, old-fashioned and a barrier to free expression. Similarly, noone will respect usage if it is merely an unedited record of what people say and write. Unless they have the support of authority, people will be unable to distinguish between good and bad usage; it will not be possible to teach or use the language with any confidence. /.../ Any successful record of the language such as a dictionary is itself a contribution to authority. People tend to believe that dictionaries tell them what is or is not allowed. Actually, the rules of a language are very flexible, but it is difficult to show all this in a dictionary. However much the editors may say they are only recording and following usage, there is no doubt whatsoever that they take thousands and thousands of decisions which contain an element of subjective judgement.

Möjligen blir man lite förvånad över hur självklart Sinclair avvisar en "unedited record of what people say and write" och hänvisar till det som i andra sammanhang (Zgusta 1996) kallats lexikografens "kreativitet". Och vilka är då närmare bestämt dessa "thousands and thousands of decisions"? Här kommer vi in på normativa resonemang och vi ska därför först titta lite på något som man skulle kunna beteckna som normerarens dilemma.

2. Språkvårdens Moment 22

Varför är normeraren – här kallad språkvårdaren – så defensiv, varför bestämmer man inte helt enkelt vad som är det "rätta bruket" och sedan ser till att bestämmelserna följs? Det finns exempel på att språkvårdsorgan följt en sådan maxim och spåren förskräcker. Ett ofta citerat exempel (se t.ex. Bengt Sigurds bidrag i Allén & Loman & Sigurd 1986) är stavningsreformen 1906. Då avvisade Svenska Akademien förenklingen av stavningen *-dt* i ord som *rödt, kastadt* osv. (till *rött, kastat* osv.). Man framhärdade långt efter det att svenska folket – och framför allt skolan – anammat den nya stavningen och tvingades till slut att i en senare upplaga av ordlistan ange sin egen form som ett förslag i andra hand.

Bilden av äldre, kloka språkvårdare som "bestämmer" hur språket skall se ut är naturligtvis falsk: dessa personers klokhet går i bästa fall ut på att konsolidera det som de flesta är eniga om och vid alternativa möjligheter förelå den lösning som man anser mest konsekvent och ändamålsenlig utifrån systemet. Detta är mycket nog och alldeles tillräckligt för språkvårdens existensberättigande. Man behöver inte måla upp ett språkligt Moment 22 där det som är språkligt korrekt är det som man i efterhand kunde konstatera var språkligt korrekt. Och i den mån som denna ovisshet ändå finns så är det något som man får lära sig att leva med. Språket är ett levande och föränderligt system. En av de första medlemmarna i Svenska Akademien, Anders von Höpken, uttryckte saken så här: "Det finns inget annat språk än det som folk i allmänhet talar och skriver". En korpus är ett utsnitt ur språket eller en del av språket som visar hur "folk i allmänhet talar och skriver". Är det inte lexikografens uppgift att återge just detta? Vari ligger de "tusentals beslut" som gör att ordboken skiljer sig från det språk som talas och skrivs? Var klämmer egentligen skon? Är det alla de tillfälliga ordbildningarna som aldrig finner vägen in i ordböcker (*Göteborgsläkare*), är det problemet med icke accepterade uttryck (*jag slog han*), är det icke accepterade uttal (*elektri'ker*), är det missuppfattningar som håller på att bli standard (*situationstecken* istället för *citationstecken*)?

Vi har hittills talat allmänt om korpusens användning för lexikografiska ändamål. Innan vi går vidare behöver vi precisera dessa ändamål. En korpus är bevismaterialet vid konstruktionen av en ordbok men inte verktygen att bygga ordboken. En aldrig så bra korpus hjälper dig inte att formulera definitionerna i ordboken eller den grammatiska beskrivningen. Korpusen kan enbart hjälpa dig att välja mellan olika frekventa uppslagsord, leverera autentiska och frekventa exempel på meningar, syntaktiska typer, kollokationer och avledningar till uppslagsordet. I en korpus kan saknas vissa betydelser hos ett ord, du

kan inte räkna med att hitta alla böjningsformer till ett ord osv. I fortsättningen skall vi huvudsakligen diskutera användningen av en korpus som underlag för en rättskrivningsordlista, alltså en ordlista över språket som ger många exempel på ord och visar hur de stavas och böjs, kanske också i vilka stilistiska sammanhang de förekommer. En sådan ordlista finns redan och vi skall se närmare på hur den ser ut.

3. En korpusbaserad rättskrivningsordlista

I Holland utgavs i december 1995 en rättskrivningsordlista (*Woordenlijste Nederlandse taal*) baserad på en holländsk 50-miljonerordskorpus. Korpusen som ligger till grund för ordlistan heter Corpus of Present-Day Dutch och består av en enspråkig fulltextkorpus med 1600 texter, 50 milj. löpord och 700.000 olika ord från tiden 1979–1990.

Idén är alltså att göra en stavningsordlista med uppgifter om böjning utifrån en stor korpus. Korpusens storlek och fördelning gör det möjligt att ställa krav på frekvens och fördelning hos de ord som tas med i ordlistan. Ett tröskelvärde som använts i detta avseende är att ord(former) för att komma med i ordlistan skall ha minst två belägg i två deltexter. Korpusen används för tre saker: urval, stavning och böjning.

Skulle en sådan idé kunna anpassas till att gälla *Svenska Akademiens ordlista* (SAOL)? Frågan gäller hur långt man kan låta korpusen bestämma det lexikaliska innehållet och var ens egna normativa tankar kommer in och tar ifrån korpusen beslutanderätten. Vi skall nu se närmare på fördelarna och nackdelarna med att låta en korpus få en central roll vid produktionen av en ordbok.

4. Korpusens urval är ovedersägligt

Fördelen med att ha en korpus som grund för en ordbok är att urvalet i alla avseenden är ovedersägligt. Inte i den meningen att korpusen såsom stickprov ur en viss population skulle vara oomtvistad. Men om man väl accepterar det faktum att ordboken bygger på en viss korpus så är invändningar ifråga om urval, omfattning, språktyp etc. meningslösa. Om jag gör en ordbok över Strindbergs språk baserat på författarens samlade produktion så är det poänglöst att invända mot det ordförråd som blir resultatet. På samma sätt kan man stödja sig på en korpus över allmänspråket – även om ett sådant stickprov alltid kan utsättas för kritik. Invänder någon att ett vanligt ord inte finns med så kan man svar att det kanske finns i språket fast inte i just det här stickprovet. Gör man korpusen tillräckligt stor (så att den rentav omfattade allt som är skrivet

på ett språk under en viss tid, alltså hela populationen) så skulle man kunna säga att det som inte finns med knappast är värt att redogöra för i en ordbok. Och mellan ett litet stickprov (som t.ex. Brownkorpusen) och hela den skriftspråkliga populationen under en viss tid finns naturligtvis mellannivåer. En sådan är exempelvis den holländska 50-miljonerordskorpusen som vi just talade om.

Om korpusen inte alltid är tillräcklig så är den nödvändig för de flesta avgöranden om språkbruk. Ett bra exempel på korpusens oundgänglighet är förekomsten av variantformer och variantböjning. Hur fördelar sig former som *sjal* eller *schal*, *jogging* eller *joggning*, *chefläkare* eller *chefsläkare*? Det är lätt att exemplifiera vikten av korpusen som korrektiv genom att jämföra med existerande ordböcker. I Tabell 1 ges exempel på ett antal frekventa alternativformer som i olika lexikaliska källor (här SAOL11 och SOB) är mer eller mindre anpassade till fördelningen i korpusen (som dock representerar ett något senare språkbruk).

Uppgifterna i SAOL11 och SOB är någorlunda likartade men angivelserna om alternativens respektive vanlighet (där *eller* jämställer men *även* innnebär ett andrahandsalternativ) är olika anpassade till korpusdata. Skillnaderna är mindre en fråga om SAOL:s normerande inslag kontra SOB:s rent beskrivande utan snarare en olika långt gående anpassning till bruket.

Tabell 1. Hopskrivning eller särskrivning. Korpusfrekvenser i Press 95 och rekommendationer i SAOL11 och Svensk ordbok.

Press 95		**SAOL 11 (rek.)**	**SOB (rek)**
efter hand 96	efterhand 159	efter hand äv. efterhand	efterhand äv. två ord
i dag >2000	idag 928	i dag el. idag	i dag äv. idag
i fred 68	ifred 22	i fred äv. ifred	i fred el. i fred
i gång 520	igång 631	i gång äv. igång	i gång el. igång
till buds 31	tillbuds 0	till buds äv. tillbuds	till buds äv. tillbuds
till synes 180	tillsynes 5	till synes äv. tillsynes	till synes äv. tillsynes
till godo 36	tillgodo 22	till godo äv. tillgodo	till godo äv. tillgodo
till rätta 106	tillrätta 57	till rätta el. tillrätta	till rätta äv. tillrätta

till känna 16	tillkänna 3	till känna el. tillkänna	till känna äv. tillkänna

Att bygga en ordbok på korpusgrund är för övrigt ingen nyhet för just den nämnda holländska ordlistan. Ett annat och mer känt exempel är *The American Heritage School Dictionary* (1972) som baseras på *The American Heritage Word Frequency Book* (1971), frukten av en textgenomgång av ett stort antal läroböcker i olika skolämnen i USA. I frekvensräkningen ges uppgift om frekvens, spridning över olika (skol)ämnen. Korpusen innehåller också ett stort urval exempel på ords användning.

Under senare år har som tidigare nämnts *Collins COBUILD* varit det mest kända exemplet på en ordbok som fast baseras på korpusmaterial. Eftersom korpusen här har använts betydligt mer konsekvent än i de flesta andra ordböcker i England och internationellt skall vi först undersöka de argument som talar mot en långtgående användning av en korpus för lexikografiska ändamål.

5. Argument mot en korpusbaserad ordbok

Den kritik som brukar riktas mot ett alltför troget utnyttjande av en korpus i ordbokssammanhang är följande.

För det första kan det invändas att korpusen är full av *tillfälliga bildningar*, framför allt en mängd sammansättningar som är "självklara". Skall en ordlista inkludera ord som *göteborgsläkare, huvudtalare, 3000-årig* och *tysk-fransk?* Tillfälligheter kan också gälla på satsplanet: många konstruktörer av ordböcker anser sig inte kunna använda exempel som de går och står i texten utan anpassar kontextberoende meningar efter vad läsaren anses kunna ha nytta av. Denna senare aspekt på tillfälligheter i korpusen skall här inte vidare diskuteras. Även om det är en viktig del i frågan om hur långt en korpus kan användas så har denna fråga för det första redan flitigt diskuterats (se t.ex. Malmgren 1994), för det andra berörs här huvudsakligen lexikografiska källor av typ rättskrivningsordlistor och slutligen handlar användningen av meningsexempel mer om pedagogik än om rätt och fel.

För det andra kan korpusen vara ett *snedvridet* eller *alltför litet stickprov* ur den språkliga populationen. Detta får återverkningar på alltifrån ordurvalet till fördelningen av grammatiska och stilistiska kategorier: det blir stora luckor i ordförrådet, semantiska fält är inte utfyllda (t.ex. alla månaders namn är med utom *november*), det saknas

information om böjningsformer för att inte tala om den morfologiska variationen och vissa stilregister kan saknas eller vara dåligt företrädda.

För det tredje kan man invända att *korpusen innehåller många fel*. Det kan röra sig om enkla saker som korrekturfel men också om s.k. normluckefel enligt Ulf Telemans förslag till terminologi (Teleman 1979), dvs. allt som strider mot språknormen (i stavning, ordbildning, ordböjning, syntax och semantik), stilfel uppkomna genom sammanblandning av tal- och skriftnormen och fel som innebär att den språkliga kompetensen helt enkelt strejkar (anakoluter etc.).

Låt oss se lite närmare på dessa synpunkter.

(a) *Tillfälliga ordbildningar*

En korpus har i språk som svenskan och tyskan många sammansättningar och avledningar som är tillfällighetsbildningar på olika sätt. Det kan röra sig om sammansättningar med proprier, avledningar av siffror, sammansättningar gjorda för stundens bruk (*ölkostnader, bilköp* och *slumskildring*). Problemet med tillfälliga ordbildningar löser man inte enbart med att utöka korpusen och sätta en lägsta gräns för hur vanliga orden skall vara för att tas med i lexikonet. Också tillfälliga bildningar kan vara frekventa i vissa texter. Om en misstänkt brottsling av tidningspressen allmänt kallas *bombmannen* så är detta knappast något skäl för att införa ordet *bombman* i en ordbok. Enbart frekvens är uppenbarligen ett otillräckligt kriterium som behöver kompletteras med någon form av spridningsuppgift. Ett sådant kan se ut på olika sätt, alltifrån ett enkelt krav att ett ord skall förekomma i minst två textkategorier (hur nu dessa ser ut) till ett spridningsindex som ger en statistisk sammanfattning av hur väl spritt ordet är i olika textkategorier.

Man kan alltså på denna punkt behöva precisera vilken roll en korpus bör spela för urvalet av uppslagsord. Det handlar inte om att kritiklöst redovisa allt som finns i korpusen (det kan man göra i en frekvensundersökning) utan om att hitta formella kriterier att välja ut de ord som kan ses som lexikaliserade enheter. För att återkoppla till en frekvensundersökning som ligger nära till hands (Nusvensk frekvensordbok) så är det den redovisning av korpusen som kallas basvokabulären – resultatet av ords frekvens och spridning – som skall gå vidare till ordboken.

Principen illustreras enkelt i följande översikt över förekomsten av ordet *rutin* i några olika källor (åtta olika korpusar i Språkbanken). Här rör det sig alltså inte om ordets fördelning på olika ämneskategorier utan helt enkelt om förekomsten i några olika korpusar. Exemplet

illustrerar enbart principen att ordet skall ha en viss spridning – frekvensen redovisas inte – för att få ingå i en viss ordbok. Av Språkbankens 130-tal belägg på ordet och dess böjningsformer visade sig följande ord (uppslagsord) förekomma i minst två källor:

Tabell 2. Förekomst av ordet *rutin* i mer än en källa i Språkbanken vid Göteborgs universitet

rutin	rutinarbete	rutinmässig
sjukhusrutin	rutinartad	rutinsak
vardagsrutin	rutinbesök	rutinundersökning
dagsrutin	rutinbetonad	rutinuppdrag
inköpsrutin	rutinerad	rutinåtgärd
orutin	rutinjobb	rutinärende
orutinerad	rutinkontroll	

Orden som uppfyller kraven på spridning är på det hela taget väl etablerade i svenskan och man slipper ifrån en rad tillfälliga bildningar i Språkbanksmaterialet som *brödjobbsrutin* och *daghemsrutin*. Men samtidigt kan man konstatera att flera av de ord som på grund av spridningen kvalificerat sig för ordboken inte kan anses mer lexikaliserade än de ord som avvisats av samma skäl. De kan kanske anses vara vanligare men är de mer "lexikaliserade"? Är *sjukhusrutin* mer lexikaliserat än *daghemsrutin*? Det finns inget enkelt sätt att ange graden av lexikalisering. Kanske är spridningen i kombination med frekvens det enda operationellt användbara sättet att ange om ett ord skall beskrivas i en ordbok.

Här finns också en avgörande skillnad mellan ordboken med sin semantiska beskrivning av ordförrådet och rättskrivnings-ordlistan med sin formella beskrivning av många ord. I det senare fallet behöver inte skiljelinjen mellan vad som är lexikaliserat eller inte (hur nu denna linje skall dras) bli avgörande för urvalet av ord.

Men om man använder frekvens och spridning i en korpus som kriterium på ordurval så hamnar man i ett annat dilemma: man riskerar att göra sig av med ord som har dålig spridning av det enkla skälet att de är ämnesbundna. Ett ord som *anställningsrutin* förekommer flera gånger men bara i en källa (Riksdagens snabbprotokoll 1978–79). Om korpusen är ämnesindelad bör denna typ av effekter bli ännu mer accentuerad: finns *avdragsgill* bara i texter om deklaration och skatt? Ändå skulle många anse ordet vara ett centralt ord i sitt ämne och en självklar kandidat till att ingå i en ordbok. Hur ser den procedur ut som automatiskt accepterar *avdragsgill* men avvisar *brödjobbsrutin*?

Det finns inget enkelt svar. Kanske kan man säga att alla ord som förekommer i fler än ett ämne eller en textkategori är starka kandidater

till urvalet i en ordbok. Men bland de ord som slås ut genom spridningskravet döljer sig väl lexikaliserade ord med ämnesspecifikt innehåll. Dessa ord kan inte lokaliseras på något enkelt automatiskt sätt. En väl planerad och stor korpus är därför i detta avseende ett nödvändigt men inte tillräckligt verktyg för urval till ordböcker.

(2) *Korpusen är otillräcklig eller snedfördelad*

Den andra invändningen handlar om att korpusen är otillräcklig: inte stor nog, inte balanserad nog. Den utgör därför ett otillräckligt underlag för ordurval men ger också otillräckliga besked på alla de punkter där en ordbok har att ge besked: stavning, ordböjning, ordbildning, syntaktiskt uppträdande, semantiska fält etc.

Kritik brukar riktas från lexikografiskt håll mot "små" standard-korpusar av typ Brown-korpusen och LOB-korpusen. De är visserligen användbara för en hel del syften och har betytt åtskilligt för korpusforskningen men de är helt otillräckliga när det gäller att få grepp om mindre vanliga ords språkliga uppträdande. Mot denna bakgrund skall man se uppbyggandet av dagens mycket stora textkorpusar på hundra miljoner löpande ord (*British National Corpus*) eller ännu mer (*Collins COBUILD*).

Det finns därför bara ett sätt att svara på den invändningen: man måste ha en stor och väl balanserad korpus för att överhuvudtaget med någon säkerhet uttala sig om ordförrådet och det språkliga bruket överhuvudtaget. Detta gäller oavsett om man är ute efter att beskriva bruket eller normera bruket: den språkliga verkligheten måste speglas i verkliga texter.

Här finns dessvärre en paradox: det man kanske mest vill ha svar på – t.ex. ovanliga ords böjning – ger korpusen ofta knapphändiga svar på. Så här kan det se ut om man letar i en korpus som Press 95 (som är under uppbyggnad men för närvarande består av ca 6 miljoner löpande ord). Målet är att skaffa fram autentiska exempel på böjningen av latinska ord på *-us*. I Tabell 3 (se nästa sida) ges också de mycket varierande böjningklasser (här icke närmare redovisade) som knutits till varje ord i SAOL.

Lägg märke till att ord som för språkvetare är självklara (*genus, korpus, numerus, kasus*) inte kan beläggas böjningsmässigt. Detta är kanske inte så förvånande – för att få besked på dessa punkter krävs en korpus som också innehåller språkvetenskapliga texter.

Tabell 3. Böjningsuppgifter för vissa ord på *-us* i Press 95.

Ord	Data i Press 95
thymus	
dekanus	en dekanus
manus	-et -en
genus	
bonus	-en -ar
opus	ett opus
korpus	
numerus	
virus	-et -ar -en (pl.)
korus	i korus
tesaurus	
kasus	
lapsus	
passus	en passus, passusar

Men svaret är inte alltid fler och bättre fördelade texter. De latinska orden på *-us* är allmänt svårhanterliga och språkbrukare tenderar att undvika böjning. Detta är ett välkänt fenomen. Det kan gälla engelska ord av typ *scanner* där språkbrukaren ställs inför en försvenskad pluralis och en engelsk, det kan gälla pluralformer till vanliga svenska ord som *baby* och *taxi*. Men korpusen svarar inte bara på hur ord böjs (om den nu överhuvudtaget ger något svar), den säger också något om vilka former som faktiskt realiseras, något som konstruktören av en ordlista inte alltid är medveten om. Böjningen i korpusen kan slumpmässigt (?) saknas på någon punkt i böjningssystemet: *bok-boks-boken-bokens-böcker-böckers-böckerna-böckernas.* Skall man i så fall fylla ut systemet och säga att det är komplett? Hur gör man då om man hittar böjningen *kärlekar* (som i filmen *Tre kärlekar*)? Och vad skall man inte fylla ut utan lita på korpusen (t.ex. avsaknaden av**glädjor* av *glädje*)? Hur gör man med komparation hos adjektiven: skall man fylla ut där ett adjektiv kan ha komparation (även om inget exempel finns i korpusen) men inte fylla ut där man tycker att komparation är omöjlig (*död-dödare-dödast*)?

Bristen på sammanhang i semantiska fält är en vanligt förekommande kritik mot korpusbaserade ordböcker. Man kan lätt peka ut enskilda ord inom sådana fält som slumpmässigt kan saknas. Det kan röra sig om namn på månader, gradbeteckningar, verktyg eller överhuvudtaget sådant som kan utgöra en del av ett större begreppsfält. Man kan med rätta fråga sig varför man skall ha med *april* men inte *maj*,

generalmajor men inte*generallöjtnant, lövsåg* men inte *fogsvans* etc. Samtidigt skall man komma ihåg att alla fält inte nödvändigtvis skall utfyllas: det heter *pojkklippt* men inte**flickklippt*, *pojkvalp* men inte **flickvalp*. osv.

(3) *Korpusen innehåller fel*

Den vanligaste invändningen mot att basera lexikaliska källor på korpusmaterial gäller dock inte vare sig tillfälliga bildningar eller korpusens otillräcklighet utan dess bristande tillförlitlighet. Den innehåller fel. Kritiken vänder sig framför allt mot två typer av fel. Den första typen är tämligen harmlös och består helt enkelt av sådana fel som alla är överens om är fel: korrekturfel, stavfel, omkastningar, konstiga avstavningar etc. och annat som visserligen kan vara irriterande för läsaren men som inte heller försvaras av någon utan på sin höjd urskuldas.

Hit hör kanske en typ av fel som är medvetna förvrängningar av författaren till en text. I Press 65 citeras ett programblad där publiken hälsas välkommen till teatern med enbart o-ljud (*Horrogod so rologt ott no kom*). Och i en annan text i samma korpus uppträder – i något semesterparadis – maträtten *kyskling med behäkning* (dvs. kyckling plus något som är helt obegripligt). Det behöver knappast sägas att dessa förvrängningar inte hör hemma i en ordbok.

Den andra typen av "fel" är viktigare: här gäller det tänjningar i den språkliga normen som gärna betraktas som fel tills de slutligen accepteras. Typen finns på alla nivåer. På stavningens område kan vi se på sådan variation som *cigarrett:cigarett* där variantformen igår var fel men idag är accepterad, om än motvilligt. Det finns många liknande exempel som just gäller anpassningen av främmande ord. Ordet *juice* gavs tidigt en försvenskad variant *jos* som av korpusdata att döma inte har slagit igenom och på senare tid kommit att exemplifiera välmenande omsorg om språket som är dömd att misslyckas.

På uttalssidan finns många standarduttal som vän av ordning vill utesluta trots att en korpus visar att "felen" är rikligt företrädda. I fall som *acceptera* och *elektriker* (ofta uttalade /assepte'ra/ och /elektri'ker/) låtsar lexikaliska källor inte om uttalens existens. Samtidigt erbjuds språkbrukare alternativ som enligt korpusdata avvisas (*sambo* uttalas mot ordböckernas önskan med akut accent).

Fortsätter man till ordbildningen så accepterade länge ordböcker enbart bildningen *ortnamn* trots att korpusdata visade att *ortsnamn* var

det normala – men ordboksredaktörerna visste genom studier i nordiska språk att det skulle heta *ortnamn* och så blev det.

Ordböjningen uppvisar en liknande skillnad mellan korpusdata och ordboksrekommendationer. Standardexemplet är de relativt framgångsrika ansträngningarna från språkvårdshåll att undgå de negativa verkningarna i svensk ordböjning av engelskt plural-s (se Söderberg 1983). Resultatet blir att framför allt engelska lånord på *-er* ibland något hårdhänt tilldelas svensk plural på -ar (*containrar* etc). Korpusdata är i detta fallet mindre följsamma gentemot språkvårdarens rekommendationer.

Rena syntaktiska felaktigheter förekommer givetvis också i korpusen. En vanlig typ är brott mot valensregler (heter det *värna om x* eller bara *värna x* och finns det någon betydelseskillnad?). Hit hör också prepositionsbruket där språkbrukare är osäkra och normen vacklar (heter det *intresserad av* eller *i* eller *för*?). Och hur är det med satsflätor (*ett brev som jag inte minns om jag har besvarat*) som tidigare ofta fungerade som språkvårdens måltavla? Och brott mot subjektsregeln (*efter att ha ätit en god lunch avgick tåget*)?

På den semantiska sidan är det inte heller svårt att finna korpusdata som bryter mot, eller åtminstone tänjer, normen. Ett av många exempel är den bildliga användningen av ordet *andas* (här exemplifierat med korpusen Riksdagsspråket 1978–79). Vad kan något (*ett svar, en motion, ett uttalande* etc) *andas*? Ett svar kan andas *optimism, misstro, missnöje, oro, tillförsikt* m.m. Men kan en proposition andas "ett mycket stort ansvar"? Eller kan ett svar andas "samma slags tongångar som när vi 1972 tog upp asbest"?

Slutligen kan nämnas en typ av korpusdata som visserligen är autentiska men behäftade med ett systematiskt fel: texterna består av översättningar som – det vet vi – alltid av påverkade av förlagan (Gellerstam 1996). Om det engelska *conversation* ("samtal") översätts till svenska med *konversation* ("artigt samtal (utan djupare innehåll)") så är översättningen inte lämplig att använda som korpusunderlag för en ordbok.

6. En korpusstyrd rättskrivningsordlista?

Det finns uppenbarligen en hel del invändningar mot att låta en korpus få direkt genomslag i en ordbok. Det finns fog för Sinclair's "thousands of decisions" innan verkligheten omsätts till en idealiserad lexikalisk beskrivning. Men kanske är svårigheterna större ju djupare beskrivningen är, syntaktiskt och semantiskt. Stannar man på den formella nivån där ord stavas, böjs och varieras – dvs. handlar det om en

rättskrivningordlista – så är goda korpusdata så nödvändiga att man undrar på vilka punkter man egentligen frångår dessa källor för att ta redaktionella beslut. Ett sådant avsteg är medvetna rekommendationer om "lämpligt" språkbruk i fall där korpusen kanske talar i annan riktning. Det rör sig främst om alternativformer där man från språksystemets synpunkt borde föredra en inhemsk form framför en mer frekvent men oönskad nyhet. Som exempel skulle kunna nämnas försök att introducera den svenska pluralformen i engelska lånord på *-er* (*scanner* etc.) trots att orden åtminstone i början är motspänstiga mot en sådan pluralform.

Om man bortser från sådana medvetna strävanden så finns ingen anledning att frångå de besked som ges i korpusen. Dock måste frågan om vad som är lexikaliserade ordbildningar och "tillfälliga" förmodligen hänskjutas till en lexikograf. I övrigt skall korpusen ha vitsord under förutsättning att följande krav uppfylls, liknande de som ställs för den holländska rättskrivningsordboken:

1. Korpusen skall vara stor, avspegla ett aktuellt språkbruk och vara väl fördelad över olika ämnen, genrer och stilarter.
2. Urvalet av ord skall baseras på korpusen – se dock inskränkningen ovan – liksom uppgifter om vilka böjningsformer som aktualiseras i korpusen.
3. Informationen om ett ord (urval, stavning, böjning) skall stödjas av minst två källor.

Litteratur

Allén, Sture 1971: *Nusvensk frekvensordbok. 2.* Stockholm: Almqvist & Wiksell.

Allén, Sture & Loman, Bengt & Sigurd, Bengt 1986: *Svenska Akademien och svenska språket*. Stockholm: Norstedts.

Gellerstam, Martin 1978: Att välja sina ord. Om ordurval till enspråkiga ordböcker. *Rapporter från Språkdata.* 9. Göteborgs universitet, Institutionen för språklig databehandling.

Gellerstam, Martin 1996: Translations as a source for cross-lingustic studies. I: Karin Aijmer & Bengt Altenberg & Mats Johansson (eds), *Languages in Contrast*. Lund: Lund University Press.

Malmgren, Sven-Göran 1994: Språkprovens form och funktion i svenska betydelseordböcker från Östergrens Nusvensk ordbok till Svensk ordbok. I: *LexicoNordica* 1, 107–117.

Collins COBUILD English Language Dictionary. Editor in Chief: John Sinclair. London/Glasgow:Collins 1987.

SAOL = *Svenska Akademins ordlista över svenska språket*. 11 upplagan 1986 Stockholm: Norstedts.

SOB = *Svensk ordbok.* Utarbetad vid Språkdata, Göteborgs universitet. 1986. Stockholm: Esselte Studium

Söderberg, Barbro 1983: *Från Rytters och Cowboys tillTjuvstrykers. S-pluralen i svenskan*. Stockholm: Almqvist & Wiksell International.

Teleman, Ulf 1979: *Språkrätt: om skolans språknormer och samhällets.* Lund: LiberLäromedel.

The American Heritage School Dictionary. 1972.

The American Heritage Word Frequency Book. 1971.

Woordenlijst Nederlandse taal. 1995. Haag: SDU.

Zgusta, Ladislav 1996. The Lexicographer´s Creativity. I. Martin Gellerstam et al (eds), *Euralex '96. Proceedings I–II*. Göteborg: Göteborg University, department of Swedish, 323–336.