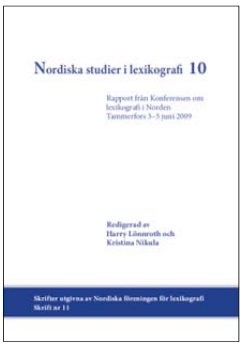
# NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Lexikografiska aspekter på Internet som källa till informellt språkbruk

Forfatter: Håkan Jansson

Kilde: Nordiska Studier i Lexikografi 10, 2010, s. 223-237
Rapport från Konferens om lexikografi i Norden, Tammerfors 3.-5. juni 2009

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






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

Håkan Jansson

# Lexikografiska aspekter på Internet som källa till informellt språkbruk

This paper treats Internet-based corpora in a lexicographical perspective, which first of all requires a quick look at the terms normative and descriptive lexicography. However the main part of the paper presents research on the compilation of Internet-based corpora, and compares that type of corpora with the traditional kind. This includes a discussion of the notion of the representativeness of the corpus, with reference to comparison between Internet-based corpora and traditional corpora such as the BNC. It is noted that Internet-based corpora offers the possibility to capture language from registers that hitherto has been unrepresented or at least underrepresented in traditional corpora. Reference is being made to some experiences of compilation of Swedish web corpora, notably how to evaluate the differences in word frequency, when different corpora are compared.

*Nyckelord*: Internet-baserad korpus, korpusbaserad lexikografi, informellt språk, balanserad korpus, SketchEngine

## 1. Inledning

Nyttan med att använda korpusar som grund för lexikografiskt arbete är det i dag knappast någon som ifrågasätter. Den som hävdar att en större korpus generellt sett ger ett bättre stöd än en mindre, och att en aktuell korpus är mer relevant än en äldre för beskrivning av samtidsspråket kommer troligen inte att bli emotsagd. Om man däremot hävdar att Internet kan vara en intressant källa för den som vill sammanställa en korpus, har man påstått något som kan ses som mer kontroversiellt. Likväl ska denna framställning behandla Internet-baserade korpusar, varför de är intressanta, och hur man ska bedöma deras innehåll.

En anledning till att man hittills förhållit sig skeptisk till Internet-baserade korpusar är möjligen att språket på Internet skulle vara mindre lämpligt som grundval för språkbeskrivning. Följande citat från *Nationalencyklopedin* (NE) kan ge en antydan om varför:

> Normalt är ordböcker *deskriptiva*, dvs. de beskriver (**det goda**) språkbruket. *Normativ* är däremot en ordlista som ”Svenska Akademiens ordlista över svenska språket” (SAOL). (artikeln *ordbok*, *Nationalencyklopedin*.)[1]

Citatet ur artikeln *ordbok* visar på en tydlig tendens i den lexikografiska traditionens praktik. Även när man säger sig arbeta deskriptivt, finns där en tydlig normativ underström. Det språkbruk man beskriver är det goda bruket. Naturligtvis är det möjligt att räkna upp exempel på motsatsen såsom *Norstedts svenska slangordbok* (Kotsinas 2000), *Stora fula ordboken* (Dagrin 2008) och andra, men det ändrar inte att huvudfåran i den lexikografiska traditionen är att beskriva det goda språkbruket. Det är inte helt utan skäl. Förutom det att allmänheten har en tendens att uppfatta ordböckernas lemmaurval som normativt, finns skälet att orden ska beläggas: ”Uppbyggandet av beläggsamlingar var tidigare det enda säkra sättet att från primärkällor samla in material till en ordbok.” Citatet från Bo Svensén (2004) antyder också den vanligaste metoden för beläggande fram till korpuslingvistikens tidevarv – att excerpera. Detta ger skriftspråket en särställning i förhållande till talspråket, och eftersom skriftspråk till sin natur blir mera formellt än talspråk, så får det formella språket en mycket mer framträdande roll än det informella språket. Man skulle kunna säga att det finns en obruten tradition av att lexikografi i första hand bygger på formellt (eller gott, om man så vill) skriftspråk. I dag har vi dock tillgång till verktyg och metoder vilka gör det möjligt att låta ett talspråksnära informellt språk få en större roll i den lexikografiska praktiken. Nu blir frågan om den lexikografiska traditionen inte bara är obruten utan om den ska visa sig obrytbar.

## 2. Normativitet och deskriptivitet

Det inledande stycket kan ses som en antydan om att artikelförfattaren menar att i det närmaste all ordboksproduktion har ett mer eller mindre framträdande normativt sidospår. Frågan om relationen normativitet/deskriptivitet är dock inte den här presentationens huvudfråga, utan där hänvisas istället till en översikt i ett Euralexbidrag av Lars Trap-Jensen (2002). I den följande framställningen kommer det till en början att vara mer intressant att skilja mellan avsiktligt normativa ordböcker och oavsiktligt normativa, det vill säga deskriptiva, ordböcker. En sådan dikotomi kommer att ganska väl motsvara den skillnad som Lev V. Ščerba (1995) gör mellan å ena sidan standard-deskriptiv ordbok (där

1 Fetstil tillagd av mig.

han använder franska akademins ordbok som exempel, men lika gärna skulle ha kunnat välja SAOL), och det han kallar informationsordbok å andra sidan. Som exempel på den senare kategorin nämner Ščerba SAOB, men ODS eller *Norsk Ordbok* hade varit lika bra exempel. Eftersom Ščerbas kategorier delvis går in i varandra bör man också nämna kategorin översättningsordbok som exempel på en oavsiktligt normativ, eller deskriptiv ordbok.

Den avsiktligt normativa lexikografin kan naturligtvis vara av varierande slag, framförallt vad gäller syften, mål och rättslig status. Den svenska SAOL (2006) skulle då kunna sägas vara milt normerande, medan den isländska *Stafsetningarorðabókin* normerar strängare. Dansk sprognævns *Retskrivningsordbogen* skulle vara särskilt strängt normerande, enligt Henning Bergenholtz (2003). Det är också något ironiskt att SAOL:s kanske mest utnyttjade normativa funktion är helt oavsiktlig. De som spelar Scrabble/Alfapet använder SAOL som prövosten för om ord existerar i svenskan eller ej, men jag tror inte att någon av företrädarna för SAOL:s redaktion skulle säga att de ord som gallras inför varje ny utgåva därmed har utgått ur det svenska språket.

När det gäller de deskriptiva eller oavsiktligt normativa ordböckerna ska man lägga märke till att dessas informationsfunktion kan komma i konflikt med den oavsiktligt normativa funktionen. Om man har som mål att beskriva det goda bruket, kommer som en självklar konsekvens en del av språket att exkluderas från beskrivningen. Denna exklusion har kanske mindre betydelse i en del fall, särskilt om vi ser till ordböcker av typen SAOB, ODS och NO (dvs. den typ som Ščerba explicit nämner), men om vi betraktar några andra funktioner som den deskriptiva lexikografin förväntas uppfylla är situationen mera problematisk. Översättningsordböcker kommer till exempel att lämna sina brukare med olösta problem om de begränsar sin lemmalista till det goda bruket. På ett mer generaliserande plan skulle man kunna säga att de tidigare nämnda ordböckerna (SAOB, ODS och NO) i högre grad sysslar med språkbeskrivning av filologisk typ, och att det för deras del inte är något problem ur ett brukarperspektiv om informellt språk saknas, eller först i ett senare skede, efter att tydligare ha etablerat sig, vinner inträde i lemmalistan. En brukare som har ett behov av att få veta vad ett ord betyder vänder sig inte i första hand till filologiskt inriktade ordböcker, utan söker sig snarare till mer lättillgängliga definitionsordböcker, som till exempel *Bonniers Svenska ordbok* (2006), NOK:s *Stora svenska ordbok* (2006) eller *Norstedts Svenska ordbok* (2006), inlärningsordböcker, som *Lexin* (1999), eller L2→L1-ordböcker av olika slag. Om det sökta ordet då saknas i lemmalistan, har naturligtvis försöket att använda ordboken misslyckats, och ordboken har inte fyllt sin funktion.

Talar man om funktioner kan det också vara motiverat att byta metalexikografisk terminologi och istället för med Ščerba tala om informationsordbok

eller översättningsordbok använda Sven Tarps (2006) termer ordbokens kognitiva funktion respektive ordbokens kommunikativa funktion.

## 3. Korpuslingvistiska aspekter

Korpusars betydelse för lexikografiskt arbete har beskrivits av B. T. Sue Atkins och Michael Rundell (2008), och ägnas därför detta inte mer utrymme här, än vad som är nödvändigt för att framställningen ska bli någorlunda tydlig. Frågan om hur man kan urskilja att en korpus är välbalanserad kan vara lite mer kontroversiell, och ägnas därför lite mera utrymme. Slutligen kan det vara på sin plats att ta upp några konkreta jämförelser mellan arbete med traditionella och Internetbaserade korpusar, och några resultat av sådant arbete.

### *3.1 Traditionell eller Internetbaserad korpus?*

Ska nu alla som vill förbättra möjligheterna att undersöka alla delar av ordförrådet, även det informella och mycket moderna, skaffa sig tillgång till en Internetbaserad korpus? Svaret på den frågan är kanske inte självklart, men det finns trots allt mycket som talar för Internetbaserade korpusar. Den stora frågan för den som ska bygga upp en korpus för lexikografiskt bruk är naturligtvis dess representativitet. Med undantag för vissa begränsade historiska korpusar kommer varje korpus med nödvändighet att vara ett stickprov av den språkliga verklighet vi vill beskriva. Frågan om vår korpus kommer att ge en rättvis bild av vårt studieobjekt blir då central. Som jag ser det finns åtminstone fyra kriterier som är viktiga att undersöka för att avgöra en korpus representativitet: att den är väl dokumenterad, att den är balanserad, att den är stor och att den är aktuell.

Det är naturligtvis ett grundkrav att korpusen ska vara väl dokumenterad för att man ska kunna ta ställning till dess innehåll. Dokumentationen ska innehålla uppgifter dels om korpusens olika beståndsdelar med hänsyn till genrer med mera, dels också om varje texts ursprung. Atkins och Rundell (2008) framhåller *British National Corpus* (BNC) som särskilt väl dokumenterad. BNC:s dokumentation finns lätt tillgänglig på dess webbplats (Burnard 2007), och jag rekommenderar var och en att ta del av den som exempel på god dokumentation. Olika sätt att klassificera och beskriva förhållandet mellan olika texttyper i en korpus diskuteras förutom av Atkins och Rundell (2008) även av många andra, till exempel Serge Sharoff. I Sharoff (2006a) behandlas sammansättningen av innehållet i Internetkorpusar i jämförelse med BNC, och i Sharoff (2006b) jämförs två ryska korpusar (en traditionell och en Internetbaserad) med BNC i

detta avseende. En slutsats man kan dra av Sharoffs framställning är att det inte finns några generella skillnader mellan traditionella korpusar och Internetbaserade korpusar vad gäller dokumentationen av korpusarnas sammansättning.

När det gäller korpusens storlek kan man enkelt säga att tumregeln är ”ju större, dess bättre”. Atkins och Rundell (2008) argumenterar för att storleken har en väsentlig betydelse i det här fallet. Man anför bland annat Zipfs lag, som lite kortfattat säger att förhållandet mellan ”vanliga” och ”ovanliga” ord är någorlunda konstant i ett godtyckligt språk. Detta innebär att mindre frekventa ord, för att inte tala om ordkombinationer, av rent matematiska skäl behöver en korpus av viss storlek för att över huvudtaget dyka upp. För att sedan kunna belägga även den minsta variation i bruket krävs det en betydligt större korpus. Internetbaserade korpusar har bättre förutsättningar än traditionella korpusar att komma upp i storlek utan att resurskraven vad gäller personal och finansiering ska bli oöverstigliga. Mycket av arbetet med att sätta samman en Internetbaserad korpus kan med datorers hjälp göras automatiskt, och blir därmed relativt resurssnålt. Hur man konkret går till väga för att ställa samman en Internetbaserad korpus beskrivs bland annat av Marco Baroni et al. (2006, 2009), Adriano Ferraresi et al. (2008) och Serge Sharoff (2006a). Tabell 1 visar en översikt över ett antal Internetbaserade korpusar samt några engelska jämförelsekorpusar som alla nås via korpusvertyget *Sketch Engine*.

*Tabell 1. Översikt över Internetbaserade korpusar*

| Språk | Namn | Löpord |
|---|---|---|
| engelska | British Academic Spoken English Corpus (BASE) | 1 252 256 |
| engelska | British Academic Written English Corpus (BAWE) | 8 336 262 |
| engelska | British National Corpus (BNC) | 111 244 375 |
| engelska | ukWaC (*Internet-baserad*) | 1 526 599 198 |
| franska | French web corpus | 126 850 281 |
| grekiska | GkWaC (*Internet-baserad*) | 149 067 023 |
| italienska | itWaC (*Internet-baserad*) | 1 909 535 984 |
| japanska | JpWaC (*Internet-baserad*) | 409 384 405 |
| persiska | WBC-Per (*Internet-baserad*) | 6 375 735 |
| rumänska | Romanian web corpus | 53 457 522 |
| ryska | Russian Web Corpus | 187 965 822 |
| spanska | Spanish web corpus | 116 900 060 |
| svenska | Swedish web corpus | 18 080 394 |
| tyska | deWaC (*Internet-baserad*) | 1 644 785 836 |

Det är utan tvekan så att en Internetbaserad korpus har helt andra förutsättningar än en traditionell korpus att vara aktuell. Skillnaderna i förutsättningar blir än tydligare, om man tänker sig att någon del av materialet ska skannas in. I dag kan man dock tänka sig att den allra största delen (om inte allt) av en mer traditionell korpus ställs till förfogande i digital form. Trots det får man nog tänka sig en mer utdragen process, där många olika överenskommelser om tillgång till materialet ska träffas. För en Internetbaserad korpus är det bara det rena insamlings- och redigeringsarbetet som styr tidsåtgången. Av det som framgår av Ferraresi et al. (2008) om arbetet med *ukWaC*, kan man dra slutsatsen att det skulle vara fullt möjligt att sammanställa en så stor korpus (drygt 1,5 miljarder löpord) inom en tidsrymd av några månader.

### *3.2 Balanserad korpus*

Frasen ”balanced corpus” ger cirka 5 650 träffar på Google. Bland dessa träffar finns korpusar som beskriver sig själva som balanserade och där finns forskare som påpekar vikten av en balanserad korpus. Tyvärr används begreppet oftast utan att det får en djupare definition. Många av dem som försöker definiera begreppet nöjer sig med att jämföra sin diskuterade korpus med BNC (British National Corpus, ca 100 miljoner löpord). Av denna praxis kan man dra slutsatsen att BNC av många anses som en standard för hur en väl balanserad korpus ska vara sammansatt. Även handböcker (se t.ex. McEnery et al. 2006: 13–21, Atkins & Rundell 2008: 61–76), hänvisar till BNC, men inte utan att först ha påpekat kopplingen till begreppet *representativitet*. En korpus är balanserad om proportionerna mellan olika delar av dess innehåll med avseende på olika relevanta parametrar (t.ex. domän, genre, ålder) är representativt för den språkliga verklighet man vill beskriva. Olika parametrars betydelse för en korpus’ representativitet finns väl analyserad i Douglas Biber (2007).

Med tanke på att en redaktionell process föregått publiceringen av huvuddelen av det språkliga material de flesta korpusar består av, kan man konstatera att korpusarnas textmaterial troligen är representativt för ”mer eller mindre vårdat” skriftspråk. Talspråk samt domäner och genrer som inte är vanligt förekommande i publicerade texter kommer däremot att vara underrepresenterade i de flesta korpusar – som därmed bara kan sägas vara balanserade i förhållande till ”publicerat skriftspråk”, och inte i förhållande till ett helhetsperspektiv på ett bestämt språk.

I en artikel som problematiserar idén om vad en balanserad korpus betyder i förhållande till korpusens faktiska innehåll har Ferraresi et al. (2008) jämfört BNC och den cirka 15 gånger mer omfattande webbkorpusen ukWaC. Man tar

*Tabell 2. Tyngdpunkter i ordförrådet i ukWaC och BNC*

| ukWaC | | |
|---|---|---|
| *Web and computers* | *Education* | *Public sphere issues* |
| website, email, link, software | students, skills, project, research | services, organizations, nhs, support |
| BNC | | |
| *Imaginative* | *Spoken* | *Politics and economy* |
| eyes, man, door, house | er, cos, sort, mhm | government, recession, plaintiff, party |

fram tyngdpunkter i ordförrådet genom att ställa samman frekvensbaserade ordlistor som sedan analyseras med hjälp av log-likelihood och andra metoder.[2] Tabell 2 visar hur tyngdpunkterna i ordförrådet skiljer sig åt mellan korpusarna, och ger exempel på ord som hör till respektive fält. Det finns i ukWaC en övervikt av ord med anknytning till *webb och datorer* (ordet *website* saknas i BNC), *undervisning* samt *samhällsfrågor*, medan BNC har övervikt i *skapande genrer*, *talat språk* samt *politik och ekonomi*. Det ska dock poängteras att ukWaC såsom varande en större korpus ändå i absoluta tal har fler ord inom de fält där BNC har sin tyngdpunkt. Fältet *talat språk* är där det enda undantaget, vilket är naturligt eftersom ukWaC saknar talat språk.

Ferraresi et al. (2008) visar att korpusarna visserligen är olika med hänsyn till tyngdpunkter i ordförrådet, men att det samtidigt inte är rimligt att med hänsyn till korpusarnas faktiska innehåll säga att den ena är bättre balanserad än den andra. En av anledningarna till skillnaderna i fältet webb och datorer kan vara en systematisk skillnad som består i att det står mer om webb och datorer på Internet än det gör i publikationer i allmänhet, en annan orsak kan vara korpusarnas ålder. Huvuddelen av BNC:s texter publicerades under åren runt 1990 (jfr Burnard 2007), medan texterna i ukWaC är insamlade 2005–2007 (Baroni et al. 2009).

2 En lättillgänglig introduktion till log-likelihood och andra metoder för korpusstatistik finns i McEnery et al. (2006: 52–58).

BNC:s övervikt inom fältet *imaginative* (*skapande genrer*) är rimligen ett resultat av en medveten principavvägning. Cirka en fjärdedel av materialet i BNC utgörs av skönlitteratur (Burnard 2007), medan Internet som helhet inte alls har den stora proportionen skönlitteratur.

Man kan nog slå fast att det inte kan vara representativt för något språk, att en fjärdedel av det totala språkanvändandet till sin helhet utgörs av skönlitteratur. Det kan å andra sidan finnas skäl att ha en stor proportion skönlitteratur i en korpus. Sålunda framgår det av Håkan Jansson (2006) att den så kallade reflexivkonstruktionen i PAROLE:s material användes mer kreativt i kåserier och skönlitteratur än i informativa och rapporterande genrer, och en mer kreativ språkanvändning kan antyda en större språklig variation, och därmed vara mer eftersträvansvärd i en korpus.[3]

## 4. Arbete med svenska webbkorpusar

Vi har i det föregående sett något om vad som går att säga om innehållet och strukturen i ukWaC jämfört med BNC. Vi har konstaterat att det finns skillnader mellan korpusarna, men att det inte går att säga vilken av dem som skulle vara mest ”representativ” för ett engelskt ordförråd i dag. Utan att ha kunnat undersöka ordförrådet på samma inträngande sätt som Ferrareasi et al. (2008), ska jag här ändå ge en antydan till jämförelse mellan olika korpusar som jag har undersökt. Tabell 3 visar frekvensen av två Internet-ankutna ord i tre svenska korpusar – PAROLE och GP04 från Språkbanken, Swedish web corpus sammanställd av mig och tillgänglig via Sketch Engine – som alla är ungefär lika stora (18–19 miljoner token). Vi kan notera att tidsperioden när materialet samlades in har stor betydelse – PAROLE började samlas in 1976 och har sin tyngdpunkt i språkmaterial från tidigt 1990-tal, GP04 innehåller material från *Göteborgs Posten* från 2004 och Swedish web corpus är insamlad under december 2008. Som påpekats ovan saknades ordet *website* i BNC, men den svenska motsvarigheten *webbplats* finns med i PAROLE som sammanställdes vid ungefär samma tid. Skillnaden i frekvens mellan PAROLE och GP04 kan nog till största delen förklaras med att det skiljer drygt 10 år i tillkomsttid mellan de båda. Däremot spelar nog tillkomsttiden mindre roll för skillnaden mellan GP04 och Swedish web corpus. Det är nog snarare så att webben är självrefererande, man skriver helt enkelt oftare om webbplatser och webbsidor på Internet än i andra media.

3 PAROLE är en ordklasstaggad korpus som sammanställts av Språkbanken vid Göteborgs universitet. Den kan nås via http://spraakbanken.gu.se/.

*Tabell 3. Frekvensen Internet-anknutna ord i tre svenska korpusar*

| | PAROLE (19 milj.) | GP04 (19 milj.) | Swedish web corpus (18 milj.) |
|---|---|---|---|
| webbplats | 16 | 94 | 661 |
| webplats | – | – | 4 |
| webbsida | 8 | 136 | 264 |
| websida | 8 | 6 | 30 |

Tabell 3 visar för övrigt också att normeringen av stavningen av *webb* – med dubbeltecknat b – har slagit igenom väl. I 1990-talets början (PAROLE) kunde man ofta se ordet stavat med enkeltecknat b <web>, men nu råder en tydlig övervikt för stavningen <webb>.

Språkbankens korpusar GP04 och PAROLE saknar i motsats till BNC inte ord som är så grundläggande för det moderna kommunikationssamhället som webbplats och webbsida. Men hur är det med det mer informella ordförrådet? Tabell 4 visar förekomsten av fyra informella ord i Språkbankens GP04 och PAROLE jämfört med Google.se och Swedish web corpus samt en av mig tillfälligt sammanställd AdHoc-korpus.

*Dagis* finns med i alla fyra korpusarna liksom i moderna ordböcker som SAOL XIII, NOK:s *Stora svenska ordbok* och *Lexin*. *Snackis*, *flummo* och *bloppis* saknas däremot i dessa ordböcker. Saknas de med hänvisning till en redaktionell avvägning av ordens vardaglighet eller stilnivå? Troligen inte. Andra minst lika vardagliga ord som *röv* – vardagligt eller anstötligt om 'stjärt' – brukar normalt finnas med i de flesta lexikon. Är det så att de har för låg frekvens för att få komma med? Nej! *Flummo* som med 4 040 instanser har lägst frekvens på Google.se av dessa ord, har ändå betydligt högre frekvens än många gamla fina ordboksord som *moltiga* 'tiga envist', som bara finns på 3 030 sidor i Google.se.

Saknas de på grund av att de skapats genom någorlunda regelmässig ordbildning: *Snackis* med *-is*-avledning till verbet *snacka* 'företeelse som det snackas om'; *flummo* med nedsättande *-o*-avledning till adjektivet *flummig* 'person som beter sig på ett icke-alert sätt'? Nej rimligen inte: *Bloppis*, bildat till *blogg* + *loppis*, är inte bildat på ett regelbundet sätt, medan ordbildningen i fallen *snackis* och *flummo* är så pass ogenomskinlig, att de normalt skulle tas in i lemmalistan om orden i övrigt anses meritera sig för en plats. Den troligaste anledningen till

*Tabell 4. Frekvensen för fyra informella ord i fyra korpusar jämfört med Google.se*

| | PAROLE (19 milj.) | GP04 (19 milj.) | Swedish web corpus (18 milj.) | AdHoc ”slang” -korpus (3,6 milj.) | Google.se (ca 30 000 milj.)[1] |
|---|---|---|---|---|---|
| dagis | 410 | 418 | 370 | 84 | 192 000 |
| snackis | – | 5 | 12 | 10 | 108 000 |
| flummo | – | – | – | 14 | 4 040 |
| bloppis | – | – | – | 3 | 44 100 |

1 Storleken på Google.se beräknades genom att jämföra frekvenserna för 15 nyckelord i Swedish web corpus med samma ords frekvens på Google.se enligt nedanstående formel:

$$\frac{\Sigma\ \textbf{token}\ \text{SweWebC}}{f\ \textbf{nyckelord}\ \text{SweWebC}} \approx \frac{\Sigma\ \textbf{token}\ \text{Google.se}}{f\ \textbf{nyckelord}\ \text{Google.se}} \rightarrow \Sigma\ \textbf{token}\ \text{Google.se} \approx f\ \textbf{nyckelord}\ \text{Google.se} * \frac{\Sigma\ \textbf{token}\ \text{SweWebC}}{f\ \textbf{nyckelord}\ \text{SweWebC}} \approx 30\ \text{miljarder}\ \textbf{token}$$

De 15 nyckelorden indikerade en storlek på Google.se på mellan 15 och 40 miljarder token, med ett genomsnitt på 30 miljarder.

att de saknas, är att redaktörerna inte räknade med/kände till orden, eftersom de inte var representerade i den/de korpus(ar) som låg till grund för lemmalistan. Symptomatiskt nog saknas alla tre orden i PAROLE; *snackis* dyker upp för första gången i Språkbankens korpusar i GP01, men börjar nå något högre frekvens först i GP05 med 22 instanser. Såväl *flummo* som *bloppis* saknas i alla Språkbankens korpusar liksom i Swedish web corpus. För att hitta ord som dessa behöver man använda någon typ av AdHoc-korpus, det vill säga en korpus som skapas särskilt för att hitta vissa typer av ord. Den ”slang”-korpus som filtrerade fram *flummo* och *bloppis* skapades med hjälp av WebBootCaT (för närmare beskrivning av verktyget se Baroni et al. 2006). Genom att använda så kallade *frö-ord* (eng. *seed words*) som liknar de ord man hoppas finna, kan man söka sig fram till de segment av Internet som har potential att innehålla stora portioner av det ordförråd man vill beforska, och sammanställa det till en korpus man kan arbeta vidare med. Även om *flummo* och *bloppis* ändå skulle ha sorterats ut från listan över nyordskandidater efter en redaktionell process, är det viktigt att det är det redaktionella avgörandet som är orsaken, och inte brister i det underliggande korpusmaterialet.

### *4.1 Ordförrådets struktur i olika korpusar*

När Swedish web corpus skapades var ett av syftena att också få en Internetbaserad korpus som skulle ha texter som var jämförbara med huvuddelen av de texter som finns i Språkbankens korpusar, det vill säga som har en tyngdpunkt i journalistisk prosa och har blivit lästa av någon annan än författaren före publiceringen. För att uppnå detta syfte specificerades domäner (t.ex. expressen.se och konsumentverket.se) som hörde till tidningar, tidskrifter och offentliga organisationer, där korpusverktyget skulle göra sin insamling. Det visade sig emellertid att det inom dess domäner fanns till exempel diskussionsforum som kännetecknades av ett avsevärt mer informellt språkbruk än avsikt var att samla in. Det mer informella språkbruket avspeglas också i ordfrekvenserna för de vanligaste orden, vilket vi ser i tabell 5.

*Tabell 5. De 25 vanligaste orden i P97 jämfört med 3 Internet-baserade korpusar*

| Rang | Facktids | P97 | SweWC | Sve I |
|---|---|---|---|---|
| 1 | i | i | att | och |
| 2 | och | och | och | att |
| 3 | att | att | i | i |
| 4 | det | det | det | det |
| 5 | som | som | som | som |
| 6 | är | en | är | är |
| 7 | för | på | en | en |
| 8 | av | är | av | på |
| 9 | på | av | på | av |
| 10 | en | för | för | för |
| 11 | med | med | den | med |
| 12 | har | till | till | jag |
| 13 | till | den | med | den |

| Rang | Facktids | P97 | SweWC | Sve I |
|---|---|---|---|---|
| 14 | ett | har | inte | till |
| 15 | de | de | jag | inte |
| 16 | den | inte | om | om |
| 17 | om | om | har | har |
| 18 | inte | ett | de | de |
| 19 | vi | han | ett | ett |
| 20 | kan | men | så | så |
| 21 | ska | var | vi | men |
| 22 | från | jag | men | du |
| 23 | men | sig | man | kan |
| 24 | 2008 | från | kan | man |
| 25 | man | vi | sig | vi |
| saknas | jag/sig | kan/man | | sig |

Tabellen visar de 25 vanligaste orden i Språkbankens P97 jämfört med tre Internet-baserade korpusar – Swedish web corpus, Facktidskrifter-bygg, en relativt liten korpus (ca 100 000 ord) som skapades för att utveckla metoder och Svenska I som skapades för att söka efter källor till informellt språk).[4] När vi jämför korpusarna ser vi att det finns ganska stora likheter i frekvenslistan mellan P97 och Facktids liksom å andra sidan mellan SweWC och Sve I. Vi kan notera att P97 och Facktids har en helt identisk fördelning av frekvenserna för de fem första orden, men ännu mer slående är likheten mellan SweWC och Sve I. Bland de 25 första orden är det bara *sig*/*du* som skiljer mellan de båda. Såväl förekomsten av *du* som det faktum att *jag* är något vanligare i Sve I än i SweWC går helt i linje med att Sve I förväntades innehålla vad som kunde kallas ”skrivet tal”; både *du* och *jag* är ju karaktäristiska för den situationsbundna referens som är typisk för talspråk.

I en jämförelse av de vanligaste orden i en korpus är det oundvikligt att formorden dominerar, men som vi sett ovan kan man ändå dra några slutsatser av just pronomenfördelningen. Eftersom jämförelsen gäller relativa frekvenser, innehåller tabellen inga siffror, men som exempel kan nämnas att frekvensen för pronomenet *jag* är 83 procent högre i SweWC än i P97 ($f$ = 0,0081 jämfört med $f$ = 0,0044).

Vi kan således notera att de iakttagelser vi kunnat göra angående fördelningen av de vanligaste orden, styrker de intryck som gavs av att Swedish web corpus kommit att innehålla informellt språk i en grad som från början inte alls var avsikten. Det innebär att den blir mindre intressant att jämföra med Språkbankens korpusar, men å andra sidan har den då kommit att innehålla mer av det informella språk som av allt att döma hittills har varit underrepresenterat i Språkbankens korpusar.

## 5. Slutsatser

Det är naturligtvis så att lexikografins möjligheter att ge en god och allsidig bild av det ordförråd man avser att beskriva alltid är beroende av vad det finns för källmaterial att tillgå. Under de senaste åren har möjligheterna att ta del av den stora språkliga variation och rikedom som Internet bjuder ökat betydligt. Språk som engelska, italienska och tyska har i dag möjlighet att ge lexikograferna till-

4 P97, Presstext från 1997, valdes till den här jämförelsen, eftersom den är yngst av de korpusar som finns tillgängliga med färdig frekvenstabell på Språkbankens hemsida.

gång till korpusar som närmar sig två miljarder token i storlek. Det är självklart inte så att korpusstorleken är den enda faktor som har betydelse när det gäller att hitta detaljer värda att beskriva i ett ordförråd. I vissa fall är det dock avgörande, till exempel när man vill beskriva kollokationer och andra typer av relationer mellan ord; om man till exempel vill undersöka relationer där de engelska orden *service* och *game* figurerar, kan man gå till BNC med 54 511 träffar på *service* och 14 träffar på kombinationen, eller så kan man söka i ukWaC med 1 252 389 *service* och 689 träffar på kombinationen. BNC:s 14 exempel är, inte oväntat, ganska triviala, men bland de 689 träffarna på ukWaC finns betydligt intressantare träffar som till exempel ”by racing away with a love service game” (genom att sätta fart med ett blankt servgame).

Förutom att korpusar i storleksklassen över 100 miljoner token av praktiska skäl i stort sett alltid är Internet-baserade, kan det finnas andra skäl att arbeta med den typen av korpusar. Vi har här ovan sett att vissa delar av ordförrådet kan vara svåra att fånga i en korpus som till största delen bygger på tidningstext eller annat publicerat material. Jämförelse mellan korpusar med olika typer av huvudsakligt innehåll pekar också mot att Internet-baserade korpusar troligen skulle ha en struktur som ligger talspråk närmare.

Om vi bara ser till möjligheten att använda mycket stora korpusar och riktigt aktuella korpusar, så har Internet-baserade korpusar otvetydigen ett försteg. Om vi dessutom betraktar möjligheten att ringa in delar av ordförrådet som i de flesta korpusar hittills haft en låg eller närmast obefintlig representation, så blir försteget för Internet-baserade korpusar ännu större.

## Litteratur

### *Ordböcker och Internetbaserade hjälpmedel*

Sketch Engine. http://www.sketchengine.co.uk/

Språkbanken vid Göteborgs universitet. http://spraakbanken.gu.se/

Bonniers svenska ordbok = Malmström, Sten, Iréne Györki & Peter A. Sjögren, 2006. Bonniers svenska ordbok. Stockholm.

Dagrin, Bengt G., 2008: Stora fula ordboken: baktalade och försummade ord i full frihet. Stockholm.

Kotsinas, Ulla-Britt, 2000: Norstedts svenska slangordbok. Stockholm.

Lexin = Martin Gellerstam, Kerstin Norén, Myndigheten för skolutveckling, Svenska datatermgruppen & Göteborgs universitet. Institutionen för språkvetenskaplig databehandling: Svenska ord: med uttal och förklaringar. Stockholm 1999.

NE = Åström, Kenneth, Christer Engström, Kari Marklund & Statens kulturråd: Nationalencyklopedin: ett uppslagsverk på vetenskaplig grund utarbetat på initiativ av

Statens kulturråd. Höganäs 1989. http://www.bib.miun.se/php/go.php?url=www.ne.se/
NO = Hellevik, Alf, 1966: Norsk ordbok: Ordbok over det norske folkemålet og det nynorske skriftmålet. Oslo.
NOK:s Stora svenska ordbok = Köhler, Per Olof, Ulla Messelius & Christian Mattsson, 2006: Natur och kulturs stora svenska ordbok. Stockholm.
Norstedts svenska ordbok = Norstedts svenska ordbok: en ordbok för alla. Stockholm 2006.
ODS = Ordbog over det danske Sprog. København 1919–1956. http://ordnet.dk/ods/
Retskrivningsordbogen = Dansk Sprognævn. Retskrivningsordbogen. København 2001.
SAOB = Ordbok över svenska språket. Lund 1893–.
SAOL XIII = Svenska Akademiens ordlista över svenska språket. 13 uppl. Stockholm 2006.
Stafsetningarorðabókin = Íslensk málnefnd & Dóra Hafsteinsdóttir. Stafsetningarorðabókin. Reykjavík 2006.

*Annan anförd litteratur*

Atkins, B. T. Sue & Rundell, Michael, 2008: The Oxford guide to practical lexicography, Oxford.
Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta, 2009: The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. I: Language Resources and Evaluation 43. http://dx.doi.org/10.1007/s10579-009-9081-4/
Baroni, Marco, Adam Kilgarriff, Jan Pomikálek & Pavel Rychlý, 2006: WebBootCaT: a web tool for instant corpora. I: XII EURALEX International Congress. Turin.
Bergenholtz, Henning, 2003: Bryder Dansk Sprognævn den danske sproglov? Sprogpolitik i teori og praksis. I: Från Närpesdialekt till EU-svenska. Festskrift till Kristina Nikula, red. av Harry Lönnroth. Tampere. S. 17–31. http://www.idiomordbogen.dk/Lit/Nikula.pdf/
Biber, Douglas, 2007: Representativness in corpus design. I: Corpus linguistics: Critical concepts in linguistics, red. av Wolfgang Teubert & Ramesh Krishnamurthy. London/New York.
Burnard, Lou, 2007: Reference Guide for the British National Corpus (XML Edition) British National Corpus Consortium: Research Technologies Service at Oxford University Computing Services. http://www.natcorp.ox.ac.uk/XMLedition/URG/
Ferraresi, Adriano, Marco Baroni, Silvia Bernardini & Eros Zanchetta, 2008: Introducing and evaluating ukWaC, a very large web-derived corpus of English. I: 4th Web as Corpus Workshop (WAC-4). Can we beat Google? Marrakech. http://webascorpus.sourceforge.net/download/WAC4_2008_Proceedings.pdf/
Jansson, Håkan, 2006: Har du ölat dig odödlig? En undersökning av resultativkonstruktioner i svenskan. Göteborg.
McEnery, Tony, Richard Xiao & Yukio Tono, 2006: Corpus-based language studies: an advanced resource book. London.

Ščerba, Lev V., 1995: Towards a General Theory of Lexicography. I: International Journal of Lexicography 8. S. 314–350.
Sharoff, Serge, 2006a: Creating general-purpose corpora using automated search engine queries. I: Wacky! Working papers on the Web as Corpus, red. av Marco Baroni & Silvia Bernardini. Bologna. http://wackybook.sslmit.unibo.it/pdfs/sharoff.pdf/
Sharoff, Serge, 2006b: Methods and tools for development of the Russian Reference Corpus. I: Corpus linguistics around the world, red. av Andrew Wilson, Dawn Archer & Paul Rayson. Amsterdam.
Svensén, Bo, 2004: Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik. 2 uppl. Stockholm.
Tarp, Sven, 2006: Leksikografien i grænselandet mellem viden og ikke-viden. Århus.
Trap-Jensen, Lars, 2002: Descriptive and Normative Aspects of Lexicographic Decision-Making: The Borderline Cases. I: Proceedings of the Tenth EURALEX International Congress EURALEX 2002 Copenhagen, Denmark, August 13–17, 2002, red. av Anna Braasch & Claus Povlsen. Köpenhamn.