Gammalt projekt och nya resurser – SAOB:s hantering av nytillkomna elektroniska textmängder

Bodil Rosqvist & Bo-A. Wendt

As all the big historical dictionary projects launched in the 19th century, *Svenska Akademiens ordbok* (SAOB) is based on paper excerpts manually selected from authentic texts. Today entirely new possibilities have emerged when it comes to anchoring lexical analysis and description in authentic language use. In this article we report on how the editorial staff of the SAOB handles the fact that nowadays our source material can so easily be complemented by electronic sources. We also discuss the upcoming revision of the SAOB and the question of what material(s) to use in the future.

1. SAOB och dess excerptmaterial

Ordbok över svenska språket utgiven av Svenska Akademien (SAOB) är en historisk ordbok som beskriver det svenska språket från 1521 fram till idag. Projektet inleddes redan på 1880-talet, och det material som ännu är ryggraden i projektet skapades utifrån sin samtids förutsättningar, i en tid när man knappast ens kunde drömma om de möjligheter som datorteknikens textdatabaser nu har givit oss. Materialet insamlades genom selektiv excerpering, där man valde ut i genomsnitt ett par, tre belägg per sida i de valda källorna och gjorde excerpter av dessa samman med det excerperade ordets närkontext.[1] Sådan excerpering genomfördes i ett mycket stort antal mycket mångskiftande slags verk. Resultatet har därmed blivit ett brett men inte så djupgående urval av språkbelägg. Relativt sett missgynnas språkets allra vanligaste ord i materialet, men å andra sidan fångas väldigt många olika ord upp, också ovanligare ord och sådana vars bruk inskränker sig till vissa genrer eller vissa ämnesområden.

Denna stora bredd, både genremässigt och inte minst viktigt för en historisk ordbok kronologiskt, är den verkligt stora fördelen med SAOB:s samlingar (SAOBArkiv), liksom omvänt den relativa hanterligheten i omfång för språkets mer frekventa ord vid det praktiska analysarbetet. Det blir tydligt om man jämför med nutidens allt flera och allt större elektroniska korpusar, som per definition är totalkonkordanser och därmed har ett djup som SAOBArkiv inte kommer ens i

[1] Lundbladh (1996:95) framhåller att denna excerpering oriktigt har beskrivits som slumpmässig. Tvärtom är urvalet högst medvetet från excerpistens sida. Ibland har detta helt visst gynnat det anmärkningsvärda eller rent av konstiga, men givna instruktioner har understrukit vikten av att också excerpera det typiska och vardagsgrå. Åtminstone idealt kan urvalet därför beskrivas som stratifierat.

närheten av men som föreligger till priset av antingen – och vanligen – att korpusen omfattar mycket färre olika källor eller att den har enastående stort och därigenom vid detaljanalys svårhanterligt omfång. Och inte ens de allra största korpusarna har samma bredd i genrer och tid som SAOBArkiv. En bredd av detta slag är förstås mycket viktig för en historisk ordbok där man på förhand inte vet vilka ord som över huvud taget funnits eller vilka användningar ett i och för sig känt ord en gång i tiden har kunnat ha. Å andra sidan är förstås det motsvarande mindre djupet en uppenbar nackdel när det kommer till att med bästa möjliga precision hitta det äldsta belägget på en viss betydelse eller användning. (Jämför med diskussionen hos Svensén (2004:56–57) om det han kallar beläggsmetoden kontra korpusmetoden.)

Skillnaderna mellan SAOBArkiv och moderna digitala korpusar kan tydliggöras med följande sifferförhållanden. SAOBArkiv omfattar numera uppemot 9 miljoner excerpter (där varje excerpt har ett stickord, som blir den enda lexikala ingången till excerpten).[2] SAOBArkiv källförteckning omfattar idag 21 460 poster. Man får då komma ihåg att många av dessa källor endast lämnat enstaka belägg, ibland bara ett enda; vi torde kunna räkna med att endast omkring 11 500 av källorna är genomexcerperade på det selektiva sätt som beskrivits ovan (Lundbladh 1996:94). Därmed skulle man alltså få ett genomsnitt på omkring 750 belägg per genomexcerperad källa. Detta kan så jämföras med Fornsvenska textbankens nysvenska annex som omfattar 786 397 ord fördelade på endast 19 källor[3], en 600-del av antalet källor i SAOBArkiv. Det ger ett genomsnitt på 41 389 ord/källa, mer än 50 gånger högre än snittet för SAOBArkiv. Ett material av innehållsligt helt annat slag är den av historiker i Uppsala sammanställda Gender and Work-databasen, som bygger på handskrivet material från 1500–1700-talet, främst domböcker. Denna omfattar 498 336 ord från 37 olika källor (ofta i urval, i form av nedslag för vissa år). Båda dessa textdatabaser har med nutida ögon sett ändå ett beskedligt omfång. Så mycket större är Korps historiska material, som med borträknande av fornsvenska källor omfattar hisnande 1,33 miljarder ord fördelade på 114 korpusar, i sig förstås överlag samlingskorpusar. Bland dessa samlingskorpusar återfinns Litteraturbanken, som idag i egen rätt består av 117 425 315 ord fördelade på 2 143 verk.

Så långt alltså bredd och djup översatta till nakna siffror, men intressantare blir det om vi ser på ett exempel på en följd av dessa förhållanden. Exemplet vi mer eller mindre godtyckligt har valt avser det alfabetiska spannet *yl–ymta* (främst för att det inte ger upphov till så mycken störande stavningsvariation vid korpussökningarna). Detta är ännu endast förberedande bearbetat vid SAOBArkiv redakti-

---

[2] Också det är ju för övrigt ett utslag av att SAOBArkiv är ett barn av en annan tid att vi faktiskt inte vet exakt hur många materialets excerpter är. Tidigare har ofta siffran 7,5 miljoner excerpter anförts (Hast 1983:160), men denna hänför sig till en tid då inte redaktionens egna kompletteringar av materialet, inte minst ur andra ordböcker, sparades efter redigering, vilket görs numera.

[3] Av dessa är 17 tillika SAOBArkiv-källor.

on. Av tabell 1 framgår vilka och hur många olika ord som SAOBArkiv och några elektroniska korpusar samt Dahlgrens Glossarium (som redovisar nysvenska ord som var föråldrade vid dess utgivning 1914) lyckas fånga upp i spannet. Som framgår härav återfinns i SAOBArkiv 16 olika ord, förutom ytterligare 10 hapax-ord som inte tagits upp i tabellen och rimligen inte heller skall beskrivas i ordboken. Av dessa 16 ord har Dahlgren med nio, vilket är rätt imponerande, i synnerhet om man tar i beaktande att han säkerligen uteslutit några av de andra orden för att de inte bedömts som föråldrade. Fornsvenska textbankens nysvenska annex har fem av orden, och samma antal återfinns i Gender and Work-databasen. Tillsammans får dessa båda databaser ihop sju ord i spannet. Lika många, men inte samma, är automatiskt identifierade i Korps historiska material. Poängen med denna jämförelse är å ena sidan att SAOBArkiv fångar upp betydligt fler olika ord än de två i huvudsak äldre nysvenska textdatabaserna tillsammans – och med belägg-stal för vardera ordet som någotsånär möjliggör en kvalificerad analys av dem. Å andra sidan har SAOBArkiv i jämförelse med Korp inte ohanterligt många belägg per ord. Så känns ju till exempel 15 000 belägg för *ymnig* lätt övermäktigt för en lexikograf som i sin historiska analys arbetar förutsättningslöst, nedifrån och upp.

	SAOBArkiv	Dahlgren	NsvA	GaW	KorpH
YL s. ’ylande’	16	–	–	–	5 591
YLA, v.	143	x	1	–	1 996
YLLE, s.	885	–	1	6	21 854
YLLEN, adj.	143	x	1	1	307
YLONS, s. ’(slags) öl’	3	–	–	–	–
YLVA, s. ’varghona’	3	–	–	–	–
YLVA, v. ’yla?’	2	–	1	–	–
YMLA, v. ’mumla’	7	x	–	–	–
YMMEL, s. ’mummel’	8	x	–	1	–
YMN(E), s. ’ymnighet’	7	x	–	–	–
YMNA, v. ’göra fruktbar’	8	x	–	–	–
YMNELIGA, adv.	5	x	–	–	–
YMNIG, adj.	788	x	25	2	15 343
YMP, s.	75	–	–	2	178
YMPA, v.	803	–	–	–	1 339
YMTA, v. ’antyda’	9	(x)	–	–	–
Σ ord	**16 ord**	**9 ord**	**5 ord**	**5 ord**	**7 ord**
Σ belägg	2905		29	12	46 608

Tabell 1: Ord inom det alfabetiska spannet *yl–ymta* i SAOBArkiv, Dahlgrens G*lossarium*, Fornsvenska textbankens nysvenska annex (NsvA), Gender and Work-databasen (GaW) och Korps historiska material (ej fornsvenska källor, endast automatiskt identifierade ord; KorpH).

2. Elektroniska resurser i SAOB-arbetet idag

Digitala korpusar som Fornsvenska textbanken, Korps historiska material, Litteraturbanken och Gender and Work-databasen nämns ovan. Andra välkända elektroniska resurser är Projekt Runeberg och, inte minst, Google Books. Alla dessa resurser används ibland i arbetet med SAOB, exempelvis för att hitta äldre (eller yngre) belägg än de som finns i SAOBArkiv eller språkexempel som illustrerar en viss betydelse på ett bra sätt.

Databaser med äldre material är dock fortfarande förhållandevis små och/eller svåra och tidskrävande att söka i, ibland är de också osäkra till sitt innehåll. Det sistnämnda gäller i synnerhet Google Books. Procentuellt sett täcker SAOBArkiv också det äldre språket bättre än det nyare. Samtidigt som antalet publicerade texter har ökat lavinartat under 1900-talet har excerperingen på SAOB:s redaktion tvärtom minskat.

Vid arbetet med SAOB görs därför mest omfattande kompletterande sökningar i material från de senaste 150 åren, dvs. ”senare material” ur vårt historiska perspektiv. Vidare används främst tidningsdatabaser. Detta eftersom de är både väldefinierade till sitt innehåll och förhållandevis lättillgängliga. De tidningsdatabaser som används på redaktionen är framför allt:

- Mediearkivet
- KB dagstidningar (Svenska dagstidningar, Kungliga biblioteket)
- Finlandssvenska dagstidningar (Nationalbiblioteket i Finlands digitala samlingar)
- DN-arkivet (Dagens Nyheters arkiv)
- SvD-arkivet (Svenska Dagbladets arkiv)

Mediearkivet, KB dagstidningar och Finlandssvenska dagstidningar utgör samlingsarkiv för en stor del av den svenska dagspressen. Mediearkivet innehåller ca 70 olika tidningar från 1980-talet och framåt. Denna resurs används flitigt, för att hitta sena illustrativa belägg om sådana saknas i SAOBArkiv, för att bekräfta betydelseanalysen eller för att kontrollera brukligheten hos ord eller användningar.

KB dagstidningar, som tillhandahåller tidningar som publicerats från 1700-talet och framåt, och Finlandssvenska dagstidningar, som täcker perioden 1771–1920, används framför allt för att hitta äldre första belägg än de som finns i SAOBArkiv.

Ibland begränsas visningen av texter i samlingsarkiven p.g.a. lagen om upphovsrätt. Därför används också några enskilda tidningars arkiv som är tillgängliga för prenumeranter, exempelvis DN:s och SvD:s arkiv.

2.1. Några specifika exempel från ordboksarbetet

Just nu är ord som börjar på *vre–vä* under redigering på SAOBArkiv redaktion, bl.a. sammansättningarna till ordet *värld*. För att visa vad tidningsdatabaserna kan

tillföra och vilket merarbete som hanteringen av dem medför i det enskilda fallet kommer några av dessa sammansättningar att diskuteras här.

I SAOBArkiv finns 9 belägg på ordet *världsartikel*, det äldsta från *Nordisk familjebok* (1920, 30:767). En tentativ definition, baserad på detta material lyder ”i sht *handel.* artikel (se d. o. III 3) som säljs i stora delar av världen”. En sökning i KB dagstidningar gör att årtalet för första belägg kan flyttas tillbaka 27 år, till 1893. Det tidigaste exemplet i Finlandssvenska dagstidningar är från samma år. (DN:s och SvD:s arkiv behöver inte undersökas eftersom de ingår i samlingsarkivet KB dagstidningar.) Under den aktuella perioden har ordet *värld* haft stavningsvarianter med *w-* och *-e-*. Således bör även dessa varianter beaktas. En sökning på strängen ”verldsartik*” visar sig ge det tidigaste belägget, från DN (19/11 1866, s. 2):

> Äfven vår stad har lemnat ett bidrag till den stora verldsexpositionen i Paris. Härifrån har neml. afsändts biskop Corts tofflor [...]. Som dessa tofflor ej synas ega någon annan märkvärdighet än den att hafva tillhört nämnde biskop föreställa vi oss, att de bland de öfriga verldsartiklarne komma att intaga en blygsam plats.

Som framgår av citatet ovan avser ordet *världsartikel* här inte en ’vara som säljs i stora delar av världen’. Artikeln i fråga är ett par tofflor som tillhört en biskop i Strängnäs och som ska ställas ut på världsutställningen i Paris. Den tidigare formulerade definitionen behöver alltså utökas för att inbegripa detta nytillkomna belägg. Den nya versionen lyder ”i sht *handel.* artikel (se d. o. III 3) som säljs i stora delar av världen; förr äv. om sak eller persedel av världsintresse”.

Av de ca 200 sammansättningar till *värld* som hittills är skrivna har tidigare belägg hittats för ca 25 %, bl.a. för orden *världsarena* (1906→1865), *världsatlas* (1924→1888), *världsberömmelse* (1929→1900), *världsess* (1963→1936) och *världsledande* (1974→1926). De kompletterande sökningarna i tidningsdatabaserna har alltså höjt kvaliteten på ordboksartiklarna, såväl vad gäller tidsfästning som definitioner.

Å andra sidan är kompletteringsarbetet tidskrävande. Även om tidningsdatabaserna kan ses som lättillgängliga kräver de ett omfattande merarbete. Äldre tidningar har ofta upp till sex spalter tättskriven text, och det är inte alltid som det aktuella textstället är markerat på sidan. Det kan ta lång tid att hitta fram till det eftersökta ordet. Vad gäller yngre tidningstexter, som skyddas av upphovsrättslagen, visas eftersökta textställen inte i sin helhet (KB dagstidningar begränsar exempelvis texter fr.o.m. 1900). Det kan därför vara omöjligt att citera texten utan att gå till originalkällan. Det kan också vara svårt eller omöjligt att med utgångspunkt i det lilla textutdrag som visas säkert avgöra det aktuella ordets betydelse.

2.2. Kvalitet kontra arbetsbörda och tidsåtgång

Det har framgått ovan att ett systematiskt och grundligt kompletteringsarbete kan

höja kvaliteten på ordboksartiklarna, och att kvalitetshöjningen har en prislapp i form av ökad arbetsbörda. Så som redaktionen arbetar för närvarande görs därför inte systematiska sökningar på varje enskilt ord. Det är upp till den aktuella redaktören att bedöma vad som är en ”rimlig insats” i det enskilda fallet.

Något som talar för en ambitiös komplettering av materialet är att angivandet av äldsta belägg på ord- och betydelsenivå är en grundläggande uppgift för SAOB. Ett större material betyder också att betydelseanalyserna blir säkrare och mer representativa. På minussidan kommer dock att arbetet tar betydligt längre tid om SAOB:s uppgift utökas från att beskriva det material som finns i SAOBArkiv till att innefatta allt som, mer eller mindre enkelt, går att hitta ”någonstans i världen”. Dessutom skulle ett systematiskt utnyttjande av elektroniska resurser medföra en obalans i verket SAOB, eftersom det bara är under de senaste åren som dessa möjligheter till kompletteringar stått till buds. Det skulle också bli obalans i den kronologiska beskrivningen av språket, eftersom det i nuläget främst är den senare delen av den nysvenska perioden som kan förbättras. Såsom läget är just nu blir det snedfördelning även vad gäller genre; tidningstexter blir överrepresenterade p.g.a. sin lättillgänglighet. Det viktigaste argumentet mot en systematisk och grundlig komplettering är kanske ändå att det blir obalans mellan olika ordtyper. Sammansättningar och monosema ord gynnas medan homografer och polysema ord missgynnas.

3. SAOB:s material i framtiden

Färdigställandet av SAOB närmar sig, och inom några få år är det dags att börja uppdatera den tryckta upplagan. En stor fråga blir då vilket material som ska beaktas. En önskedröm skulle vara att ha en mycket stor elektroniskt sökbar, lemmatiserad och taggad korpus som grund för hela ordboksinnehållet, från A till Ö. Korpusen skulle vara balanserad med avseende på kronologi, ämnesområde, genre, regionalitet och genus och innehålla ett brett urval texter som föreligger i sin helhet och som täcker hela beskrivningsperioden. Helst skulle den också vara specialanpassad för ökad hanterlighet och innehålla genvägar som gör det möjligt för lexikografen att på ett enkelt sätt välja ut och citera textställen. Att åstadkomma en sådan korpus skulle kräva specialkunskaper på en rad olika områden och enorma insatser vad gäller arbete, tid och pengar. En representativ och i alla avseende balanserad korpus måste nog dessutom betraktas som en utopi. Det förhåller sig ju onekligen så att vi lexikografer, som Atkins & Rundell uttrycker det (2008:74), ”are sampling from a population whose nature is unknowable and whose extent is unlimited”. En möjlig lösning som vi ser framför oss just nu är en fortsatt kombination av olika resurser.

För det första har vi ”SAOBArkiv 2” till vårt förfogande, dvs. den del av SAOBArkiv som har kommit in ”för sent” för varje enskilt ordboksband. Det rör sig om ca 500 000 excerpter, insamlade från 1890-talet och framåt. För det andra

tror vi oss ha goda möjligheter att skapa en mindre egen digital korpus, anpassad efter våra behov och utformad med en medveten strävan mot balansering och representativitet. Grunden till en sådan korpus skulle vi kunna få via vetenskapligt noggranna projekt som t.ex. de tidigare nämnda Litteraturbanken och Gender and Work. För det tredje tänker vi oss att vi även fortsättningsvis kommer att använda fritt tillgängliga textdatabaser. De olika materialen skulle kunna användas på olika sätt för olika syften. Det är ju nämligen så att den kommande uppdateringen rör högst olika aspekter med delvis olika materialmässiga implikationer.

En första viktig del av denna uppdatering är givetvis en komplettering med de ord som nu saknas i SAOB, det vill säga framförallt under 1900-talet i språket inkomna ord. Här kan man sannolikt komma rätt långt med den enklare lösning vi skisserade ovan. En annan del gäller komplettering av nya, återigen i huvudsak 1900-talsnya, användningar hos redan beskrivna ord. Detta innebär förstås lexikografiskt en större utmaning, eftersom det nu handlar om att gå in och ändra i befintliga artiklar. Materialmässigt bör det emellertid kunna klaras med samma eller liknande material som för kompletteringen av helt nya ord. Ett tredje slags komplettering gäller nya äldsta belägg hos redan beskrivna ord och användningar, ibland för den delen också yngsta belägg för mindre brukliga sådana. Just här, där det snarast handlar om material äldre än 1900-talet, kan man faktiskt tänka sig att korpusbalansering trots allt inte är en så tvingande nödvändighet. Eftersom man här kan utgå från SAOB:s befintliga beskrivning kan man i detta hänseende föreställa sig punktvisa sökningar i varjehanda äldre material, små som stora, genrespecifika som mer allmänna, utifrån devisen att varje nytt äldstabelägg (eller yngstabelägg) är en vinst i sig. Malmgren (1999:201) framhåller att äldstabelägg måste ses som approximationer, särskilt för äldre tid, men att kravet på exakthet samtidigt ökar med tillgången till stora korpusar.

Ett gott material är naturligtvis en grundförutsättning för en god ordbok, och också om vi här skisserat möjliga vägar för SAOB:s materialmässiga framtid, är det tveklöst så att mycket ännu återstår att begrunda och överväga i detta hänseende, innan arbetet med SAOB:s andra upplaga kan inledas på allvar.

Litteratur

Ordböcker, textsamlingar och textdatabaser

Dahlgren, F. A. (1914–1916): *Glossarium öfver föråldrade eller ovanliga ord och talesätt i svenska språket från och med 1500-talets andra årtionde*. Lund: C.W.K. Gleerups förlag.

DN-arkivet: <arkivet.dn.se> (maj 2017).

Finlandssvenska dagstidningar: <digi.kansalliskirjasto.fi/sanomalehti> (maj 2017).

Fornsvenska textbankens nysvenska annex: <project2.sol.lu.se/fornsvenska> (maj 2017).
Gender and Work-databasen (GaW): <gaw.hist.uu.se> (maj 2017).
KB dagstidningar: <tidningar.kb.se> (maj 2017).
Korps historiska material: <spraakbanken.gu.se/korp/?mode=all_hist> (maj 2017).
Litteraturbanken: <litteraturbanken.se> (maj 2017).
Mediearkivet: <retriever.se> (maj 2017).
SAOB = *Ordbok över svenska språket*, utgiven *av Svenska Akademien* (1893–). Lund.
SAOBArkiv = Samlingar til SAOB. Lund.
SvD-arkivet: <svd.se/arkiv> (maj 2017).

Annan litteratur

Atkins, Sue & Michael Rundell (2008): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
Hast, Sture (1983): Om SAOB:s material. I: *Nysvenska studier* 63, 159–192.
Lundbladh, Carl-Erik (1996): Värdering av SAOB:s korpus och språkprovssamling. I: *LexicoNordica* 3, 91–103.
Malmgren, Sven-Göran (2000): Korpusar i ordboksarbete och grammatikforskning. I: Gunilla Byrman, Hans Lindqvist & Magnus Levin: *Korpusar i forskning och undervisning. Rapport från ASLA:s höstsymposium, Växjö, 11-12 november 1999*. Uppsala: Association Suedoise de Linguistique Apliquée, 192–208.
Svensén, Bo (2004): *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik.* Stockholm: Norstedts Akademiska Förlag.

Bodil Rosqvist
Biträdande huvudredaktör för SAOB
Bodil.Rosqvist@svenskaakademien.se

Bo-A. Wendt
Huvudredaktör för SAOB
Bo.Wendt@svenskaakademien.se

Svenska Akademiens ordboksredaktion
Dalbyvägen 3
S-224 60 Lund