Indexeringen av SAOB

Erik Bäckerud

We are creating a new digital version of the Swedish Academy dictionary, SAOB. We have previously scanned the printed originals to obtain a correct text in digital format and we are presently attempting to automatically identify important structures in the articles. The structures we identify include title, part of speech, etymology and the division of definitions among others. This paper will give details of our work and a review of how well we are succeeding and what difficulties we have encountered.

1. En digital SAOB

1.1. Bakgrund

Svenska Akademiens ordbok (SAOB) har digitaliserats en gång tidigare, den gången av en grupp knuten till dåvarande avdelningen Språkdata vid Institutionen för svenska språket, Göteborgs universitet (OSA 1996). Vi har tack vare nydigitaliseringen fått ned antalet fel i texten väsentligt, både vad gäller tecken och stilar.

De primära målen med indexeringsarbetet är:

- Att göra en webbpresentation av SAOB som uppfyller höga krav på läsbarhet och korrekthet och som riktar sig såväl till fackmän som till allmänheten.
- Att skapa indata till ett framtida redigeringssystem då ordboken skall uppdateras.

1.2. Grunderna

Skapandet av en ny digital version av SAOB företas i två steg. Det första steget, att skanna all text till en teckenrätt digital version, har jag beskrivit i

ett tidigare arbete (Bäckerud 2014). Detta arbete är nu avslutat och utfallet blev mycket gott.

I det andra steget, som beskrivs i denna artikel, utvecklar vi automatiska metoder för att identifiera information om strukturer som finns implicit i artikeltexterna och göra dessa explicita. Den första fasen av indexeringsarbetet är klar t.o.m. band 36. Arbetet fortsätter nu med kvalitetsförbättringar och indexering av nyare material.

De strukturer vi indexerat är (jfr Lundbladh 1992):

1) Uppslagsord med biformer
2) Sammansättningar
3) Avledningar
4) Särskilda förbindelser (dvs. förbindelser av verb och betonad partikel)
5) Ordklass
6) Etymologiparentes (i huvudet)
7) Formparentes i huvudet (dvs. stavningsformer m.m.)
8) Moment I, II osv., A, B osv. och 1, 2 osv.
9) Sekundära sammansättningar och avledningar
10) Översikt
11) Hänvisningsartiklar

Utöver ovan uppräknade punkter har vi också till punkt 8 indexerat alla undermoment ned till den lägsta nivån; α', β' osv. och till punkt 9 har vi även identifierat sekundära särskilda förbindelser vilka förekommer enstaka gånger i ordbokstexten.

1.2.1. Indata

Det material vi haft tillgång till under projektet är texten till SAOB band 1 t.o.m. 36 omfattande uppslagsorden A till UTSTÄDES. De olika banden finns tillgängliga i olika format eftersom redaktionen använt olika tekniska lösningar vid olika tider. Band 1–31 finns i tryckt form på papper och har skannats, medan de senare banden finns i olika ordbehandlarformat.

Totalt har vi därmed 36 band med sammanlagt 195 miljoner skrivtecken fördelade på 84.000 artiklar, varav 54.000 är definitionsartiklar. Totalt har vi identifierat cirka 567.000 uppslagsformer inklusive hänvisningar och biformer.

2. Indexeringsarbetet

2.1. Om artiklarnas uppbyggnad

Här nedan följer exempel på hur artiklar i SAOB är uppbyggda och hur de olika strukturer vi indexerar ser ut.

2.1.1. Huvudet

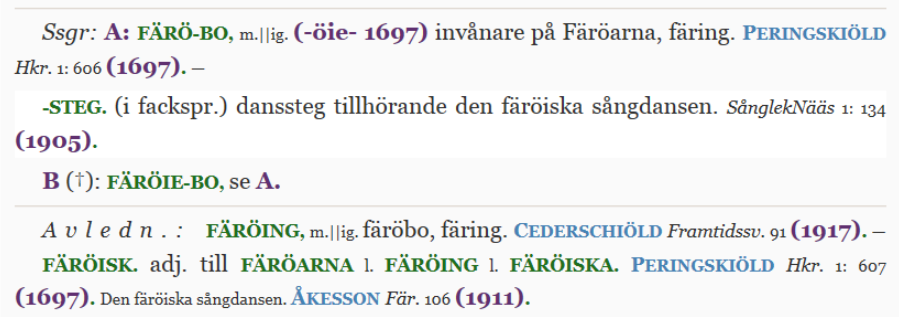
Flera av punkterna som listas i 1.2 ovan kan man få fram genom att analysera den inledande delen av artikeln, det så kallade huvudet. I nedanstående figur har jag markerat uppslagsordet med biformer (dvs. FLÅTE och FLOTTA), ordklass med homografsiffra, formparentes och etymologiparentes.

FLOTTE *flot*^{3} *e*^{2}, sbst.^{2}, l. (numera bl. starkt bygdemålsfärgat) **FLÅTE** *flå*^{3} *te*^{2}, r. l. m.; best. **-en** ((†) **-flåttn** RÅLAMB 10: 42 **(1691))**; pl. **-ar;** äv. (numera bl. bygdemålsfärgat samt i Finl.) **FLOTTA** *flot*^{3} *a*^{2}, sbst.^{2}, r. l. f.; best. **-an;** pl. **-or** ((†) **-er** (i vissa fall möjl. att hänföra till sg. **flotte**) *2Krön.* 2: 16 *Bib.* **(1541)**, HALLENBERG *Hist.* 5: 260 **(1796)). (flott-** (**flått-**) *c.* **1540** osv. **flot- 1538–1734. flåt(h)- 1588–1889. -e 1538** osv. **-a** *c.* **1540** (*oblik kasus*)**–1695** (*nom.*) **–1916 (:** flottor**))** [fsv. *floti,* motsv. d. *flaade,* isl. *floti,* feng. *flota,* eng. *float,* bildat på det svaga rotstadiet till **FLYTA;** jfr **FLOTA,** sbst.^{1}, **FLOTT,** sbst.$^{1-2}$, **FLOTT,** adj.$^{1-2}$, **FLOTTA,** sbst.^{1}, **FLOTTA,** v.$^{1-2}$]

Figur 1: Huvudet i artikeln FLOTTE sbst.^{2}.

2.1.2. Övriga uppslagsformer

I SAOB förekommer särskilda förbindelser (verb med partikel), sammansättningar och avledningar i så kallade ramsor. I exemplet nedan ser vi dels en sammansättningsramsa med två underordnade ramsor med olika förled, dels en avledningsramsa med ett par avledningar.



Ssgr: **A:** **FÄRÖ-BO,** m.||ig. **(-öie- 1697)** invånare på Färöarna, färing. PERINGSKIÖLD *Hkr.* 1: 606 **(1697).** –

-STEG. (i fackspr.) danssteg tillhörande den färöiska sångdansen. *SånglekNääs* 1: 134 **(1905).**

B (†): **FÄRÖIE-BO,** se **A.**

A v l e d n . : **FÄRÖING,** m.||ig. färöbo, färing. CEDERSCHIÖLD *Framtidssv.* 91 **(1917).** – **FÄRÖISK.** adj. till **FÄRÖARNA** l. **FÄRÖING** l. **FÄRÖISKA.** PERINGSKIÖLD *Hkr.* 1: 607 **(1697).** Den färöiska sångdansen. ÅKESSON *Fär.* 106 **(1911).**

Figur 2: Sammansättnings- och avledningsramsor.

I exemplet identifieras sammansättningarna FÄRÖ-BO, FÄRÖ-STEG och FÄRÖIE-BO samt avledningarna FÄRÖING och FÄRÖISK.

2.1.3. Sekundära konstruktioner och biformer

I SAOB kan både sammansättningar och avledningar i sin tur ha sammansättningar och avledningar. Dessa kallas då sekundära sammansättningar respektive avledningar. Här nedan visas ett exempel från artikeln TRO sbst., under sammansättningen TRO-LOVEN där det förekommer två sekundära sammansättningar trolovan-spillare och trolovan-öl.

-LOVEN l. **-LOVE** l. **-LOVAN**. [fsv. *trolovan*; formen **-lovan** snarast vbalsbst. till **-lova**] (†) (högtidligt l. edligt) löfte om trohet; äv. speciellare: trolovning. Ther om och alt annad wij eder formane, på samma hulskap och troloffue som j med oss hans nade haffue rethuiselige tilsagt, atj bliffue ther faste vtinnan och retther eder sielffue j alle motte. *G1R* 4: 434 **(1527)**. Ther trolofwan är skeedd medh lagha gåfwor och närwarande witne och sedan är sängelagh tilkommit, slijk handel moste man räkna gill och rätt in för Gudh. RUDBECKIUS *Kyrkiost.* 22 **(*c.* 1635)**. ALMQVIST *Amor.* 309 **(1822, 1839)**. Ssgr (†): **trolovan-** l. **troloves-** l. **trolovs-spillare**. person som omintetgjort trolovning. *VDAkt.* **1656**, nr 112. **-öl**. trolovningsöl. BUREUS *Suml.* 55 **(*c.* 1600)**. *HammarkDomb.* 16/9 **1623**. —

Figur 3: Exempel från artikeln TRO sbst.[1].

Vi ser här också flera exempel på sammansättningar med biformer. TRO-LOVEN har två biformer: TRO-LOVE och TRO-LOVAN och den sekundära sammansättningen trolovan-spillare har två biformer: troloves-spillare och trolovs-spillare.

2.1.4. Moment

Betydelsebeskrivningarna i SAOB är indelade i numrerade moment vilka förekommer hierarkiskt i inte mindre än fem olika nivåer.

1)	Romerska (över-)moment	I, II, III, ...
2)	Arabiska siffror (huvudmoment)	1, 2, 3, ...
3)	Undermoment	a, b, c, ...
4)	Grekiska undermoment	*α, β, γ,* ...
5)	Grekiska prim-undermoment	*α′, β′, γ′,* ...

Till ovanstående kommer att momenten på de två högsta nivåerna kan vara avskilda med grupperingsbokstäver som skrivs A, B, C osv. Dessa bryter dock inte numreringen utan finns endast för att förtydliga sambandet mellan olika intilliggande moment.

I nästa exempel visas början av en artikel som innehåller alla fem nivåerna av momentnumrering.

ANFORDRAN *an*³ *~fω*² *rdran* l. (hvard.) *~fω*² *d-* (**a`nfordran** WESTE), r. l. f. (m. LIND **(1749;** *jämte* f.**)**). vbalsbst. till **ANFORDRA.**

I. till **ANFORDRA I.**

1) framställande af fordran; fordran.

a) mer l. mindre abstr.: yrkande; kräfvande l. kraf.

α) föregånget af prep. *vid* l. *(up)på* i uttr. för tid l. sätt.

α‘) i fråga om penningfordran. *SFS* **1836,** nr 8, s. 2. Lån, att betalas vid anfordran,

Figur 4: Artikel med numrerade moment i fem nivåer.

2.2. Svårigheter

Arbetet med indexeringen har överlag gått bra men naturligtvis har vi stött på svårigheter. Ett återkommande problem har varit fel i den ursprungliga texten. Obalanserade parenteser och utelämnade bindestreck får obönhörligt indexeringsprogrammet att gå vilse.

Även när texten är korrekt finns det svåra fall där indexeringen kan gå fel. Några exempel ges nedan.

2.2.1. Exempel

Momentsiffror i SAOB skrivs som en siffra följd av en parentes, **1)**, eller som en bokstav följd av parentes, *a)*. En hänvisning till ett visst moment i en annan artikel som står inom parentes kan lätt förväxlas med en momentsiffra. I nedanstående exempel är **1)** inte början av ett moment.

— (jfr **AFDELA II 1) -SKÄRM**~ ². skärm, som begagnas till rums afdelande.

Figur 5: En hänvisning inom parentes.

För att undvika detta fel måste programmet söka bakåt i texten och se att det finns en oavslutad vänsterparentes 16 tecken tidigare. Ibland måste man se

ännu längre än så för att förvissa sig om att det inte handlar om en hänvisning.

I nästa exempel har redaktören inte skrivit ut de fulla formerna i avledningen FÖRSTÄRKING. Det som avses är ju att de tre orden FÖRSTÄRKING, FÖRSTÄRKNING och FÖRSTÄRKUNG alla hänvisar till artikeln FÖRSTÄRKNING.

FÖRSTÄRKING, -NING, -UNG, se FÖRSTÄRKNING.

Figur 6: Svårtolkad hänvisning.

Svårigheten här består i att programmet inte kan urskilja vilken del av ordet FÖRSTÄRKING som skall föras ihop med respektive efterled. Det blir därmed omöjligt att återskapa de korrekta formerna automatiskt.

2.3. Utfall

Under hösten 2014 genomförde jag och Erik Norberg, en timanställd student, ett större test där vi korrekturläste en procent av alla artiklar i ordboken, slumpmässigt utvalda.

Enligt detta test var mer än 98 % av de kontrollerade artiklarna korrekt indexerade. De två vanligaste felen visade sig vara dels att en eller flera momentsiffror ej blivit identifierade i betydelsebeskrivningen, dels att en uppslagsform tolkats som biform till det närmast föregående ordet trots att den faktiskt utgör ett nytt uppslagsord.

Det förekom också en del fel i artiklarnas huvud, biformer hittas inte alltid och inte heller ordklass. Blott i några enstaka fall hade indexeringen spårat ur helt och missat en större mängd moment eller uppslagsformer och det berodde då oftast på fel i indatat som t.ex. obalanserade parenteser. Vi räknar med att snart ha en korrekthetsgrad på över 99 %.

Vi har även programmerat ett antal automatiska tester som bland annat letar efter ensamma moment, moment som inte kommer i ordning eller som bryter hierarkin bland momentnivåer. Vi söker även efter sammansättningar som ej kommer i alfabetisk ordning, vilket oftast är ett tecken på att något är fel i indexeringen.

3. Framtiden

Vi fortsätter arbetet med digitaliseringen av ordboken. Närmast på tur står att indexera häfte 37:1 som kom ut i december 2014 och som omfattar uppslagsorden UTSUG till VEDERSYN. Vi utför också återkommande stickprovskontroller för att hitta och åtgärda fel i indexeringen.

Nästa stora projekt blir att färdigställa en webbversion av ordboken som kan publiceras på nätet. En prototyp till en sådan webbsida finns redan idag och arbetet pågår med att utveckla denna till en färdig produkt.

Figur 7: Skärmklipp från indexeringsprojektets prototypsida.

Längre fram kommer vi, förutom att indexera nya häften efterhand som de kommer ut, att försöka fördjupa indexeringen. Det finns mycket som är intressant att identifiera, t.ex: hänvisningar inom och mellan artiklar, källor, brukligheter och fackområde, etymologier med mera.

Litteratur

Ordböcker

SAOB = *Ordbok över svenska språket utgiven av Svenska Akademien (Svenska Akademiens ordbok)*, 1893 ff. Lund.

SAOB i digitaliserad form (band 1–35) = <g3.spraakdata.gu.se/saob/> (november 2014).

Annan litteratur

Bäckerud, Erik (2014): Nydigitaliseringen av SAOB. I: Ruth Vatvedt Fjeld & Marit Hovdenak (red.): *Nordiske studier i leksikografi 12. Rapport fra Konferanse om leksikografi i Norden Oslo 13.-16. august 2013*. 95-105.

Lundbladh, Carl-Erik (1992): *Handledning till Svenska Akademiens Ordbok*. Stockholm: Norstedts.

OSA (1996) = Sture Allén, Yvonne Cederholm, Sofie Kokkinakis Johansson, Lena Rogström, Rudolf Rydstedt & Lars Svensson (1996): *Om svar anhålles. Rapport från projektet OSA*. GU-ISS-96-4 Research reports from the departement of Swedish, Göteborgs Universitet.

Erik Bäckerud
systemansvarig
Svenska Akademiens ordboksredaktion
Dalbyvägen 3
SE-224 60 Lund
erik.backerud@svenskaakademien.se