NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Nydigitalisering av SAOB

Forfatter: Erik Bäckerud

Kilde: Nordiske Studier i Leksikografi 12, 2013, s. 95-105
Rapport fra Konferanse om leksikografi i Norden, Oslo 13.-16. august 2013

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.

Nydigitalisering av SAOB

Erik Bäckerud

The Swedish Academy Dictionary, SAOB, has started a project with the aim to digitize all volumes of SAOB again to improve the quality of the text over the previous effort. The new version will first of all be correct as regards the actual characters in the text, also the typography will be identified correctly in most cases. In the second part of the project we attempt to automatically identify structures in the articles such as title, part of speech, etymologi, division of definitions and so on.

1. Syfte och bakgrund

1.1. Syfte

Digitaliseringen av SAOB är tänkt att tjäna två syften. Det första är för att kunna göra en version tillgänglig på internet som uppfyller högt ställda krav på korrekthet och läsbarhet och det andra syftet är att skapa en korrekt och strukturerad text som grund för det fortsatta arbetet med nästa upplaga av SAOB.

1.2. Bakgrund

Hittills har 36 band av SAOB tryckts omfattande A t.o.m. UTSUDDA. Det första bandet kom ut 1898 och band 36 trycktes 2012. Från och med band 32 (1993) skrivs ordbokstexten på dator så förlagan till denna text finns i digital form. De första 31 banden förelåg däremot i traditionell tryckt form och måste digitaliseras för att kunna användas i en elektronisk publicering.

Varje band av SAOB innehåller drygt fem miljoner tecken, vi har alltså omkring 200 miljoner tecken ännu så länge. Den färdiga SAOB beräknas bestå av 38 band.

1.3. Tidigare digitaliseringsprojekt

SAOB-redaktionen har tittat på några andra projekt innan vi satte igång att nydigitalisera SAOB. Främst har vi sett på *Deutsches Wörterbuch* von Jacob Grimm und Wilhelm Grimm och på *Ordbog over det danske Sprog*. Båda dessa verk har digitaliserats med hjälp av Kompetenzzentrum vid universitetet i Trier.

1.3.1. OSA-projektet

Texten till de första banden av SAOB har digitaliserats vid ett tidigare tillfälle i det så kallade OSA-projektet (OSA = Om svar anhålles) som genomfördes vid Göteborgs universitet 1982–1996 (OSA 1996). SAOB-redaktionen ansåg dock att det behövdes bättre noggrannhet på texten, både vad gäller tecken-

rätthet och typografi, för våra planer. Det bedömdes att det snabbaste och billigaste sättet att få en teckenrätt digital version av texten var genom nydigitalisering hellre än att försöka korrigera den gamla texten.

2. Digitaliseringen

I februari 2011 gjorde en delegation från SAOB ett besök hos Kompetenzzentrum i Trier. Det beslöts då att Kompetenzzentrum skulle få uppdraget att digitalisera de 31 första banden av SAOB. Det beräknades ta två år att genomföra hela projektet. Senare reviderades leveranstiden till slutet av november 2013 för de sista banden. I själva verket levererades de sista sex banden redan den 3 september, tre månader tidigare än beräknat. Således har vi nu tillgång till SAOBs alla 36 band i digitalt format.

2.1. Processen

En utmaning för den som digitaliserar en text som SAOB är de många olika stilarna och storlekarna på text som förekommer. Det finns ett tiotal olika som t.ex. stor och liten rubrikstil (fet), stor och liten antikva och text som är kursiverad, spärrad eller både och. Dessutom förekommer förutom de vanliga latinska

skrivtecknen även grekiska, runskrift och många andra speciella tecken.

Vi är därför glada över den stora noggrannhet som digitaliseringsprojektet uppnått. Uppskattningsvis ett felaktigt tecken per 30.000. Även vad gäller identifiering av de olika stilarna och graderna har god noggrannhet uppnåtts.

Här är ett exempel på hur en artikel kan se ut i den tryckta versionen av SAOB från 1903.

AUTOTYPOGRAFI *âu¹totyp¹ografi⁴* l. *-tå-* l. *-tω-*, l. *-ty¹p-*, äfv. *-på-* l. *-pω-*, äfv. *áω¹-*, äfv. *010104*, äfv. **AVTOTYPOGRAFI** *av¹-*, l. *01–*, r.; best. **-en**, äfv. **-n**. [jfr t. *autotypographie*, eng. *autotypography*, af gr. *ἀυτός* o. *τύπος* (se **AUTOTYPI**) samt *γράφειν* (jfr **AUTOGRAF**)] (förr) reproduktion af manuskript, handteckningar o. d. gm (det å ett gelatinskikt med bläck af visst slag, tusch osv. tecknade) originalets öfverförande på metallklisché o. anv. af tryckpress; jfr **ZINKOGRAFI**. *NF* 19 (1895).

Figur 1: Inskannad artikel från SAOB.

När texten digitaliserats och kontrollerats i Trier levereras den till redaktionen i ett format som kallas TUSTEP (TUebingen System of TExt Processing tools). TUSTEP är det format som används internt på Kompetenzzentrum. Artikeln ovan ser då ut så här när vi får den:

```
++<P> <A+1>AUTOTYPOGRAFI</A+1> <sup
char="#(GES)">au</sup>1totyp1ografi4 l. -t#;oa-
l. -t#P+^W#P--, l. -ty1p-, <A-1>äfv.</A-1> -p#;oa-
l. -p#P+^W#P--, <A-1>äfv.</A-1> <sup
char="#(GES)">#P+a^W#P-</sup>1-, <A-1>äfv.
010104, äfv. AVTOTYPOGRAFI</A-1> av1-, <A-1>l. 01--
, r.; best. -en,
äfv. -n. [jfr t. autotypographie, eng.
autotypography,
af gr. #G+%)ayt%/ow#G- o. #G+t%/ypow#G- (se
AUTOTYPI) samt #G+gr%/afein#G-
(jfr AUTOGRAF)]</A-1> (förr) reproduktion af
manuskript,
handteckningar o. d. gm (det #;oa ett gelatinskikt
med bläck af visst slag, tusch osv. tecknade)
originalets öfverförande p#;oa metallklisché o.
anv.
af tryckpress; jfr <A-1>ZINKOGRAFI. NF 19
(1895).</A-1></P>
```

Figur 2: Artikeln AUTOTYPOGRAFI i TUSTEP-format.

Texten i TUSTEP-format omvandlas därefter till XML enligt ett schema som beskriver all nödvändig typografisk information. Samma XML-schema används för alla band av SAOB. Nedan visas hur början av ovanstående artikel kan se ut i XML:

```
- <STYCKE art="AUTOTYPOGRAFI" hänvisning="false">
<StorRubrik id="A2573_146156">AUTOTYPOGRAFI</StorRubrik>
<b id="A2573_146157" />
<StorKursiv id="A2573_146158">au</StorKursiv>
<StorAntikva id="A2573_146159" rend="upphöjd">1</StorAntikva>
<StorKursiv id="A2573_146160">totyp</StorKursiv>
<StorAntikva id="A2573_146161" rend="upphöjd">1</StorAntikva>
<StorKursiv id="A2573_146162">ografi</StorKursiv>
<StorAntikva id="A2573_146163" rend="upphöjd">4</StorAntikva>
<b id="A2573_146164" />
<StorAntikva id="A2573_146165">l.</StorAntikva>
<b id="A2573_146166" />
<StorKursiv id="A2573_146167">-tå-</StorKursiv>
```

```
<b id="A2573_146168" />
<StorAntikva id="A2573_146169">l.</StorAntikva>
<b id="A2573_146170" />
<StorKursiv id="A2573_146171">-tω-,</StorKursiv>
<b id="A2573_146172" />
<StorAntikva id="A2573_146173">l.</StorAntikva>
<b id="A2573_146174" />
<StorKursiv id="A2573_146175">-ty</StorKursiv>
. . .
```

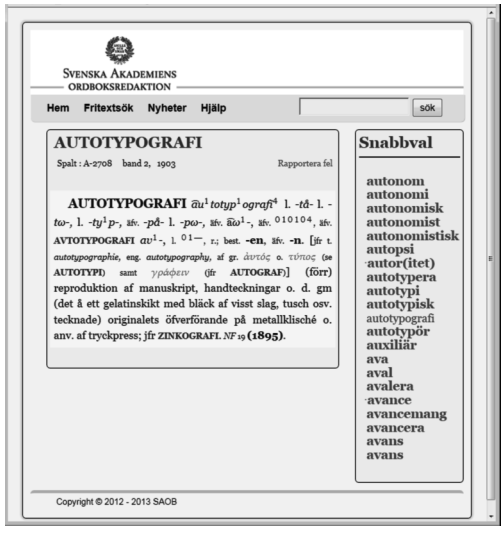
Figur 3: Artikeln som XML (avkortad).

Från denna text i XML kan vi sedan producera artikeltext i många olika format. T.ex som PDF-filer som kan se ut så här:

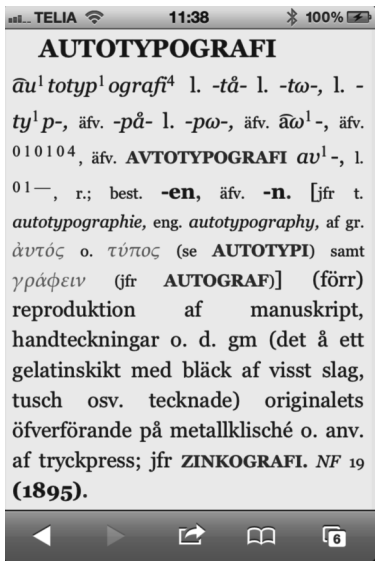
AUTOTYPOGRAFI *au*[1]*totyp*[1]*ografi*[4] l. *-tå-* l. *-tω-,* l. *-ty*[1]*p-,* äfv. *-på-* l. *-pω-,* äfv. *ω* [1]**-,** äfv. [010104], äfv. **AVTOTYPOGRAFI** *av*[1]**-,** l. [01—], r.; best. **-en,** äfv. **-n.** [jfr t. *autotypographie,* eng. *autotypography,* af gr. *ἀυτός* o. *τύπος* (se **AUTOTYPI**) samt *γράφειν* (jfr **AUTOGRAF**)] (förr) reproduktion af manuskript, handteckningar o. d. gm (det å ett gelatinskikt med bläck af visst slag, tusch osv. tecknade) originalets öfverförande på metallklisché o. anv. af tryckpress; jfr **ZINKOGRAFI.** *NF* 19 **(1895).**

Figur 4: Texten till artikeln AUTOTYPOGRAFI återskapad från det digitaliserade materialet.

Vi kan också från samma data göra webbsidor som kan visas på datorskärm eller mobila enheter. Så här kan texten se ut i redaktionens webbplats för korrekturläsning:



Figur 5: Artikeln AUTOTYPOGRAFI som webbsida.



Och så här ser samma artikel ut i en mobiltelefon:

Figur 6: Skärmbild från mobiltelefon.

2.2. Svårigheter

Även om hela processen att få SAOB digitaliserad har löpt på bättre än förväntat så har det naturligtvis varit några små bekymmer på vägen. Det första problemet vi ställdes inför var att tolka TUSTEP-kodningen som i vissa fall inte är helt konsekvent, och i ett fåtal fall rent felaktig. Tecknen % och # används t.ex. både för att inleda en sekvens som betecknar en symbol som saknas på tangentbordet och för att beteckna sig själva. Detta gör att det är litet svårt att avgöra om det är en speciell kodsekvens som börjar eller om det är ett enstaka tecken. Ett exempel på inkonsekvens är att ligaturen œ kan skrivas både som ”{oe}” och ”#.ö”.

När alla stilar är avkodade och alla specialtecken tolkats korrekt så översätts texten till Unicode. Detta går för det allra mesta bra även om det finns enstaka symboler och kombinationer av diakritiska tecken som inte återfinns i Unicode. Problemet med de saknade diakriterna har vi tills vidare löst genom att inte ta med dem utan endast grundsymbolen. Vi undersöker vidare om det finns bättre lösningar på detta problem.

Vissa tecken har man vid digitaliseringen inte lyckats tolka alls, t.ex. kan processen som används inte skilja på ligaturerna *æ* och *œ* (-ae-, -oe-) vilka blir mycket lika i kursiv stil. Här har SAOB-redaktionen fått hjälpa till med att identifiera vilket tecken som avses. Ett annat fall som vållat huvudbry är ord som är avstavade vid radslut. Här är det önskvärt att hålla samman orden så att typografin blir mer tilltalande, men också för att fulltextsökning skall fungera. I de allra flesta fall skall ord som bryts vid radslut hållas samman men ibland som t.ex. i vissa sammansättningar skall bindestrecket behållas. Även här har redaktionen fått rycka in för att kontrollera vilka ord som skall få behålla bindestrecket.

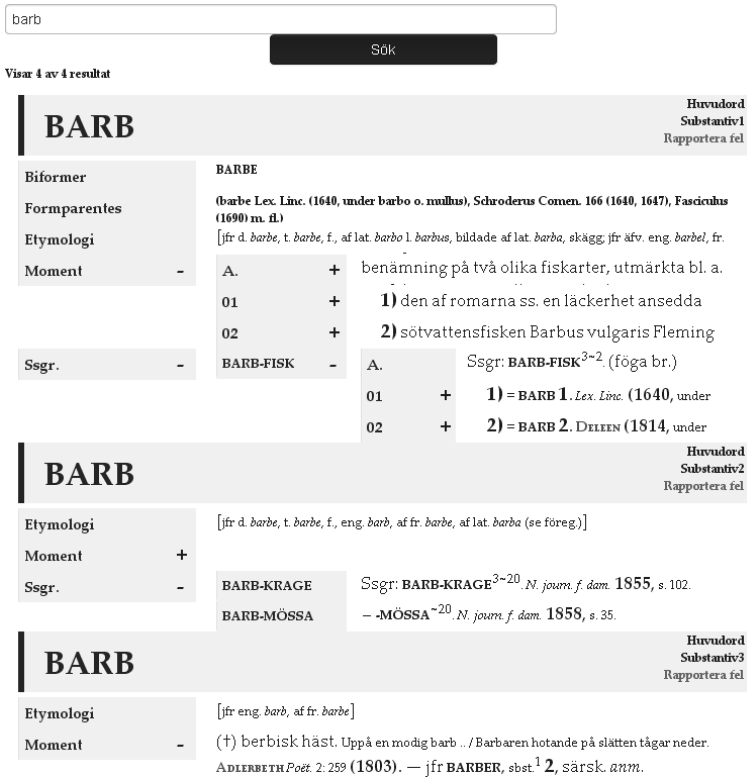
3. Indexeringen

Svenska Akademien har även startat ett projekt för indexering av SAOB. Detta projekt går ut på att med datorns hjälp identifiera flera viktiga element i varje artikel. T.ex. identifierar vi uppslagsord med stavningsvarianter, ordklass, etymologi och momentindelning. Till de uppslagsord som identifieras hör även avledningar, sammansättningar och särskilda förbindelser. I detta projekt använder vi den ovan beskrivna nydigitaliserade texten som indata.

Vi har medvetet valt att begränsa indexeringen till ett relativt litet antal strukturer som identifieras i texten. Saker som uttal, hänvisningar inom artiklar m.m. blir inte identifierade i denna första fas. Eftersom vi har en bestämd mängd resurser till vårt förfogande har vi valt att prioritera det nödvändigaste.

Denna struktur kommer att göra det möjligt för oss att göra en betydligt mer läsvänlig webbpresentation än vad som varit möjligt utan strukturering. Den kommer också att utgöra en god grund för ett redigeringssystem att användas vid en uppdatering då första upplagan av SAOB är färdigställd, vilket beräknas ske år 2017.

Arbetet med indexeringen gör redaktionen tillsammans med externa konsulter. Detta arbete är ännu inte avslutat men en del av vad vi åstadkommit hittills syns på nedanstående bild. Här har jag valt att visa uppslagsorden BARB i stället för AUTOTYPOGRAFI eftersom det finns mer intressant struktur här:



Figur 7: Exempel på strukturer som identifierats i BARB[1-3].

Litteratur

Cederholm, Yvonne, Mickel Grönroos, Susanne Manker & Lena Rogström (2000): *SAOB – en bok för hela folket.*

Rapport 2 från projektet OSA. GU-ISS-00-2. Research reports from the departement of Swedish, Göteborgs universitet.

DWB = *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm*. <http://woerterbuchnetz.de/DWB/> (oktober 2013).

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

ODS (1918–2005) = Dahlerup, Verner m.fl.: *Ordbog over det danske sprog inkl. supplement*. København: Det Danske Sprog- og Litteraturselskab. <http://ordnet.dk/ods> (oktober 2013).

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.

SAOB (1893–) = *Ordbok över svenska språket utgiven av Svenska Akademien (Svenska Akademiens ordbok)*, Lund. <http://g3.spraakdata.gu.se/saob/> (november 2013).

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