SAOLhist.se – för allmänt och vetenskapligt bruk

Louise Holmer, Sven-Göran Malmgren & Monica von Martens

The Swedish Academy Glossary (*Svenska Akademiens ordlista*, SAOL), a normative glossary with particular focus on spelling, inflexion, and pronunciation, has appeared since 1874. In 2006, the 13th edition appeared. The second, third, fourth and fifth editions were more or less reprints of the first edition. Thus, the 2006 edition may be regarded as the ninth (autonomous) SAOL edition. The most important parts of these nine editions, especially the complete sets of lemmas, have been transformed into a database. The database may be thought of as a matrix with more than 200 000 rows and eleven columns (the nine SAOL editions plus the Dalin dictionary, plus a column to the left with all lemmas that have occurred in at least one SAOL edition). The interface is very flexible, allowing the inclusion of other dictionaries. From each square of the matrix there is a link to a facsimile version of the full corresponding dictionary article. Several advanced searches are possible, e.g.: which words were added and omitted between any two consecutive editions, or: has the frequency of a particular suffix increased or decreased over time?

1. Inledning

Sedan 2013 finns databasen SAOLhist fritt tillgänglig på nätet på adressen saolhist.se. Databasen innehåller uppslagsorden (med ordklassangivelser) i nio av de 13 upplagor av *Svenska Akademiens ordlista* (SAOL) som hade kommit ut 2012. Det tidiga arbetet med bland annat digitaliseringen av de äldre upplagorna beskrivs närmare i Holmer (2012). Upplagorna 2–5 var i stort sett omtryck av den första upplagan och uteslöts. De upplagor som ingår i SAOLhist är sålunda SAOL 1 (1874), SAOL 6 (1889), SAOL 7 (1900), SAOL 8 (1923), SAOL 9 (1950), SAOL 10 (1973), SAOL 11 (1986), SAOL 12 (1998) och SAOL 13 (2006) (jfr t.ex. Malmgren 2014). Sedan april 2015 föreligger även SAOL 14, såväl i pappers- som databasform. Den ingår i en intern version av SAOLhist men ännu inte i den allmänt tillgängliga versionen.

Databasens struktur och det flexibla gränssnittet gör att man lätt kan komplettera den med data från andra ordböcker. Sålunda ingår även lemmauppsättningen i Dalins klassiska ordbok (1850–53) i SAOLhist. Ett samarbete har också inletts med Dansk sprognævn, vilket resulterat i att alla upplagor av den danska *Retskrivningsordbogen* (sedan 1955) nu föreligger i en ”systerdatabas” med ett modifierat gränssnitt (rohist.dsn.dk).

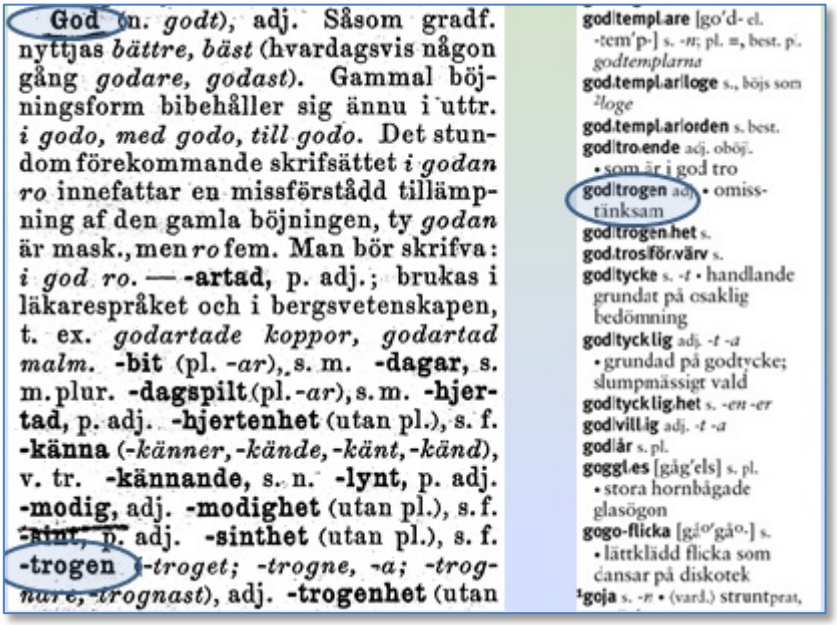
I det följande ger vi först en kort redogörelse för arbetet med att etablera SAOLhist, med fokus på de mest intressanta lexikografiska problemen. Sedan följer en diskussion av de datalogiska aspekterna på databasen. Slutligen ges några exempel på hur man kan använda SAOLhist – både som forskare och som allmänt språkintresserad.

2. Etableringen av SAOLhist: några lexikografiska problem

SAOLhist-projektet fokuserade från början på lemmauppsättningen, jämte informationen om ordklasstillhörighet. Den övriga informationen, t.ex. uttals- och böjningsinformationen och den i tidigare upplagor ytterst rudimentära betydelseinformationen, negligerades. Tre av de nio SAOL-upplagorna förelåg redan i digital form, nämligen upplagorna 11–13. De övriga skannades in och OCR-tolkades med fokus på att möjliggöra maskinell extrahering av feta teckensträngar (=uppslagsord) och (nästan) omedelbart följande ordklassangivelser, varefter normalisering och en första korrekturläsning genomfördes. Rättning och korrekturläsning utfördes till stor del av tredje- eller fjärdeterminsstuderande i nordiska språk. I ett senare skede gjordes mer sofistikerade korrekturläsningar, varvid bl.a. de garanterat korrekta uppslagsformerna i upplagorna 11–13 utnyttjades som jämförelsematerial.

De elva första SAOL-upplagorna tillämpade en nischalfabetisk modell med huvudlemman och sublemman (se Figur 1). Det innebar att de flesta uppslagsord inte återgavs explicit utan med hjälp av en divis som representerade ett grundord eller en förled i en sammansättning. Formalismen för att särskilja sammansättningar och avledningar utvecklades gradvis från en utgåva till nästa, vilket medförde att metoden för digitalisering och extrahering måste anpassas för varje separat utgåva. Efterlederna i de tidiga upplagorna

expanderades algoritmiskt till fullständiga uppslagsord för att göras jämförbara med uppslagsorden i de senare upplagorna, se t.ex. **godtrogen** i Figur 1.



Figur 1: Artiklarnas uppställning i SAOL 1 och SAOL 13.

Många av uppslagsorden behövde dock normaliseras, främst på grund av den stora stavningsreformen 1906 (alltså snart efter att upplaga 7 kom ut), men även till följd av senare ortografiska förändringar. I utdraget ur SAOL 1 i Figur 1 gäller det t.ex. **godhjertad** (>**godhjärtad**). Här liksom i andra avseenden är det SAOL 13 som utgör normen. För gammalstavade ord i de tidigare upplagorna som utmönstrats, t.ex. i upplaga 8, ansattes en tänkt nutida stavning, som i allmänhet gav sig naturligt. Exempelvis fick ordet *Afdånande* normalformen *avdånande*.

I de tryckta upplagorna 1–11 tillämpades en nischalfabetisk modell och i upplagorna 12–13 (och 14) en glattalfabetisk (se Figur 1). Dessa modeller återspeglade ett etymologiserande respektive ett rent formellt sätt att betrakta ordförrådet. I princip gjordes två likstavade ord med olika etymologier till två uppslagsord i de elva första upplagorna, och likstavade ord med samma etymologi presenterades som ett enda lemma, ofta även om de tillhörde olika

ordklasser. I upplaga 12 och senare hölls likstavade ord med samma böjning och uttal samman som ett uppslagsord, och likstavade ord med olika böjning eller uttal presenterades alltid som två uppslagsord. Den teoretiska grunden var den s.k. *lemma-lexemmodellen* (se t.ex. Allén 1999).

Skillnaden mellan de båda modellerna kan åskådliggöras med hjälp av ett enkelt exempel, ordet *ljus*. Ordet är både substantiv och adjektiv och behandlas därför som två lemman i SAOL 12–14. Men oavsett ordklass är det förstås fråga om samma etymologi, och *ljus* tas därför upp som *ett* lemma i tidigare upplagor av SAOL (se Figur 2).

Figur 2: Artikeln **ljus** i SAOL 9 och SAOL 13.

Därmed uppstår ett problem: ska *ljus* tas upp som ett eller två uppslagsord i SAOLhist, som ska täcka alla upplagor? Även här bestämde vi oss för att låta SAOL 13 utgöra normen, och därmed står *ljus* som två lemman i SAOLhist. Se Figur 3.

Normaliserad ordform	SAOL 8 (1923)	SAOL 9 (1950)	SAOL 10 (1973)	SAOL 11 (1986)	SAOL 12 (1998)	SAOL 13 (2006)
ljus, adj	ljus2 (a.)	ljus2 (adj.)	ljus2 (adj.)	ljus2 (adj.)	²ljus (adj.)	²ljus (adj.)
ljus, subst	ljus1 (s.)	ljus1 (s.)	ljus1 (s.)	ljus1 (s.)	¹ljus (s.)	¹ljus (s.)

Figur 3: Behandlingen av ordet *ljus* i SAOLhist.

SAOLhist-databasen kan presenteras som en matris med över 200 000 rader (en rad för varje lemma som förekommit i minst en SAOL-upplaga) och elva kolumner (nio SAOL-upplagor, Dalins ordbok och vänsterkolumnen med de normaliserade uppslagsformerna). Ett valfritt urval av kolumnerna, med undantag för vänsterkolumnen, kan undertryckas. Varje ruta är länkad till

motsvarande faksimilsida i den aktuella ordboken, där hela artikeln går att läsa. Tack vare särskilda sökverktyg kan man lätt få fram när ett ord första eller sista gången togs med i SAOL, vilka ord som tillkommit respektive försvunnit mellan två upplagor etc. Man kan söka såväl baklänges som framlänges, vilket ger möjlighet till intressanta morfologiska studier.

3. Datalogiska aspekter

Uppbyggnaden av SAOLhist-systemet finansierades med projektmedel från Svenska Akademien, utan några åtaganden om fortsatt finansiering av underhåll och drift. Vid designen av databasen och gränssnittet var det bl.a. av den anledningen nödvändigt att inta en minimalistisk hållning och välja lösningar som inte medförde licenskostnader eller andra fasta kostnader.

Målet var att med hjälp av ett självförklarande och flexibelt webbgränssnitt, och med respekt för användarens integritet, göra det historiska materialet tillgängligt för både lekmän och forskare. Tack vare samarbetet med Dansk Sprognævn under år 2014 kunde vi vidareutveckla funktionaliteten och konfigurationsmöjligheterna så att samma programpaket kan användas mot olika databaser.

Som databasmotor används MySQL/MariaDB och programmen är skrivna i scriptspråket php. Databasstrukturen är väldigt enkel, för varje ordlista finns en innehållsförteckning med sidreferenser och en tabell med grundinformation. Denna tabell innehåller kolumnerna ”normaliserad ordform”, ”ursprunglig ordform”, ”ordklass”, ”sidhänvisning” och ”ordklassgrupp”. Istället för ordklass och ordklassgrupp kan valfritt annat klassificeringsbegrepp användas – i vår interna miljö använder vi även böjningsinformation i klassificeringen för att åstadkomma homografseparering. Utifrån dessa tabeller genereras tre aggregerade tabeller:

- en med alla ordformer som ska vara sökbara (med kolumnerna: ordform, normaliserad form, bok)
- en med länkar och länktext per normaliserad form och bok (hur länkar och länktext ska genereras specificeras separat för varje boktabell)

- en med länkar och länktext per normaliserad form, ordklassgrupp och bok (hur länkar och länktext ska genereras specificeras separat för varje boktabell)

Användargränssnittet består av en välkomstsida där man väljer vilka böcker som ska ingå i bearbetningen och en huvudsida för sökning och tabellvisning. Detaljinformation presenteras huvudsakligen i pop-upfönster. Avancerade sökparametrar kan visas eller döljas allt efter behov. Vid en enkel sökning behöver man endast fylla i ett ord, med eller utan jokertecken (wildcards).

4. Exempel på sökningar

För att sökresultat i SAOLhist ska vara intressanta från svenska språkets (och inte bara från SAOL:s) synpunkt, krävs förstås att urvalet ur det samtida ordförrådet i var och en av SAOL-upplagorna är någorlunda representativt. Man kan räkna med att i det närmaste alla viktiga enkla (icke-sammansatta och icke-avledda) ord är med, åtminstone från och med upplaga 7.[1] Man kan också räkna med att urvalet av sammansättningar och avledningar till stor del styrs av frekvenskriterier, särskilt i upplagorna 11–13 som är korpusbaserade, men troligen också i tidigare upplagor. I synnerhet de sju nyaste SAOL-upplagorna i SAOLhist-databasen bör därför kunna säga något intressant om det svenska ordförrådets utveckling under mer än 100 år.

Vi ska se på två olika sökningar, en morfologisk och en semantisk. Vid samtliga sökningar inkluderar vi även den internt tillgängliga SAOL 14. Av utrymmesskäl tas i allmänhet bara en delmängd av sökresultaten med. Den första morfologiska sökningen är tydligt relaterad till den aktuella samhällsdebatten. Eftersom klimatfrågorna har kommit i centrum under de senaste decennierna, kan man vänta sig en ökning av antalet sammansättningar med *klimat-* i de senaste SAOL-upplagorna. Så är också i hög grad fallet (se Figur 4). Intressant nog får man inte alls samma utslag om man i stället söker på efterledssammansättningar med *-klimat* (se Figur 5).

[1] De sex första upplagorna hade en viss puristisk prägel, varför många lånord som var fullt brukliga inte kom med. Se t.ex. Malmgren (2002).

SAOL 12 (1998)	SAOL 13 (2006)	SAOL 14 (2015)
klimat (s.)	klimat (s.)	klim·at (s. *+et; pl. +*)
klimatanläggning (s.)	klimatanläggning (s.)	klim·at|an·lägg·ning (s. *+en +ar*)
		klim·at|an·passa (v. *+de +t*)
		klim·at|an·pass·ning (s. *+en +ar*)
		klim·at|av·tal (s. *+et; pl. +*)
		klim·at|be·red·ning (s. *+en +ar*)
		klim·at|bov (s. *+en +ar*)
klimatbälte (s.)	klimatbälte (s.)	klim·at|bälte (s. *+t +n*)
		klim·at|de·batt (s. *+en +er*)

Figur 4: Sökresultat för söksträngen *klimat%*.

SAOL 12 (1998)	SAOL 13 (2006)	SAOL 14 (2015)
	affärsklimat	af·färs|klim·at
	arbetsklimat	arbets|klim·at
bioklimat	bioklimat	bio|klim·at
	börsklimat	
debattklimat	debattklimat	de·batt|klim·at
		driv·hus|klim·at
ekvatorialklimat	ekvatorialklimat	ekv·at·ori·al|klim·at
fastlandsklimat	fastlandsklimat	fast·lands|klim·at
	företagarklimat	före·tag·ar|klim·at
företagsklimat	företagsklimat	före·tags|klim·at
förhandlingsklimat	förhandlingsklimat	för·handl·ings|klim·at

Fig. 5. Sökresultat för söksträngen *%klimat*.

Trots att betydelsebeskrivningarna i SAOL ofta är mycket kortfattade, går det faktiskt också att göra vissa semantiska studier med hjälp av SAOLhist.

Ett exempel erbjuder halvsynonymerna *doft* och *lukt*. För många språkbrukare står *doft* alltid för en angenäm och *lukt* för en oangenäm förnimmelse. Denna intuition kan kontrolleras mot efterledssammansättningarna med *-doft* respektive *-lukt* i SAOLhist. Här tänker vi oss närmast en synkron undersökning och begränsar oss till de senaste upplagorna (se Figur 6 och 7).

En snabb blick räcker för att konstatera att språkbrukarnas intuitioner i det närmaste stämmer. Det enda som egentligen stör bilden är ordet *spritdoft*, som dock möjligen uppfattas som positivt laddat av vissa språkbrukare.

skunk||lukt		
skur||lukt	skur||lukt	
skvattram⌈s⌉||lukt		
snusk||lukt	snusk||lukt	snusk||lukt
sprit||lukt		
späck||lukt		
stall||lukt	stall||lukt	
stek||lukt	stek||lukt	

Figur 6: Sökresultat med exempel från upplaga 9, 10 och 11 för söksträngen *%lukt*.

rökelse||doft	rökelse||doft	rökelse||doft
schersmin||doft	schersmin||doft	
skur||doft		
skvattram⌈s⌉||doft	skvattram⌈s⌉||doft	skvattram||⌈s⌉||doft
smultron||doft		
sprit||doft		sprit||doft
stek||doft	stek||doft	
syren||doft	syren||doft	syren||doft

Figur 7: Sökresultat med exempel från upplaga 9, 10 och 11 för söksträngen *%doft*.

5. Slutord

I artikeln har vi velat visa upp SAOLhist som en resurs för olika typer av användare, från forskare till allmänt språkintresserade. De båda exemplen på sökningar är i och för sig av relativt avancerat slag, och betydligt enklare sökningar är givetvis också möjliga. Möjligheterna till trunkerade sökningar gör också SAOLhist användbar för exempelvis korsordslösare och Word-Feud-spelare. En av de vanligaste frågorna som ställs till SAOL-redaktionen handlar om i vilken upplaga ett visst ord kom in i ordlistan, och sådana frågor kan man lätt få svar på i SAOLhist. Det ska slutligen påpekas att den indexering av SAOLhist som är tillgänglig på nätet fortfarande inte är hundraprocentigt korrekt; påpekanden om kvarvarande felaktigheter tas tacksamt emot. Tack vare de lätt åtkomliga faksimilsidorna finns dock alltid originaltexterna till hands.

Litteratur

Ordböcker

Dalin, A.F. (1850–55): *Ordbok öfver svenska språket.* Stockholm.

RO = *Retskrivningsordbogen*, udg. af Dansk Sprognævn. Udg. 1–4. 1955–2012. København.

SAOL 1, 6, 7, ..., 14 = *Svenska Akademiens ordlista*, uppl. 1, 6, 7, ..., 14. Stockholm 1874–2015.

Övrig litteratur

Allén, Sture (1981): The lemma-lexeme model of the Swedish Lexical Database. I: Ralph, Bo (red.). *Modersmålet i fäderneslandet. Ett urval uppsatser under fyrtio år av Sture Allén.* Meijerbergs arkiv för svensk ordforskning 25. Göteborg 1999.

Holmer, Louise (2012): SAOLHist – alla upplagor av SAOL i en och samma databas. I: Eaker, B., L. Larsson & A. Mattisson (red.), *Nordiska studier i lexikografi 11. Rapport från Konferensen om lexikografi i Norden, Lund 24–27 maj 2011*. Skrifter utgivna av Nordiska föreningen för lexikografi 12. Oslo, 287-295.

Malmgren, Sven-Göran (2002): Normering i Svenska Akademiens ordlista 1874–1950: principer och resultat. I: *LexicoNordica* 9, 5-20.

Malmgren, Sven-Göran (2014): Svenska Akademiens ordlista genom 140 år: mot fjortonde upplagan. I: *LexicoNordica* 21, 81-98.

Louise Holmer
bitr. forsk., doktorand
Inst. för svenska språket
Göteborgs universitet
Box 200
SE-405 30 Göteborg
louise.holmer@svenska.gu.se

Sven-Göran Malmgren
professor emeritus
Inst. för svenska språket
Göteborgs universitet
Box 200
SE-405 30 Göteborg
malmgren@svenska.gu.se

Monica von Martens
systemutvecklare
Inst. för svenska språket
Göteborgs universitet
Box 200
SE-405 30 Göteborg
monica.von.martens@gu.se