SALDO 1.0
(Svenskt associationslexikon version 2)

Lars Borin
Markus Forsberg
Lennart Lönngren

Språkbanken, Göteborgs universitet, 2008-05-07

Innehåll

1 SALDO – en svensk språkteknologiresurs baserad på SAL

Detta dokument beskriver SALDO, en omfattande lexikonresurs för modernt svenskt skriftspråk som är avsedd för användning i språkteknologisk forskning och utveckling av språkteknologiska applikationer. Man kan betrakta SALDO som baslexikonresursen i en svensk BLARK (se avsnitt 2). SALDO bygger på Svenskt associationslexikon, ett semantiskt lexikon för svenska (se avsnitt 3).

SALDO är en elektronisk lexikonresurs avsedd för språkteknologiska tillämpningar. Det betyder att den är strukturerad i enlighet med detta och har ett innehåll som överensstämmer med detta mål, ett innehåll som av den anledningen på viktiga punkter avviker från vad man förväntar sig att finna i traditionella lexikon avsedda för mänskligt bruk (som ju också i allt större utsträckning görs tillgängliga i elektronisk form); se avsnitt 5. Slutligen distribueras den i ett format som är avsett och lämpar sig för programmatisk åtkomst, alltså användning som en komponent i datorprogram.

1.1 Villkor för användningen av SALDO

SALDO görs fritt tillgänglig under en licens, som tillåter all användning, inklusive modifiering och distribution av modifierade versioner, på villkor att upphovsmännen nämns på det sätt som anges samt att vidaredistribuerade versioner – vare sig de är modifierade eller ej – har licensvillkor som inte är mer restriktiva än de som gäller för denna version. Valfritt kan SALDO användas, modifieras och spridas under en av följande två licenser:

- *Creative Commons Erkännande–Dela lika (BY–SA) 2.5*; se appendix 1 samt <http://creativecommons.org/licenses/by-sa/2.5/se/>
- *GNU Lesser General Public License 3.0*; se appendix 2–3 samt <http://www.gnu.org/licenses/lgpl.html>

Språkbanken vid Göteborgs universitet och språkteknologiforskningsgruppen vid Chalmers tekniska högskola har båda en strävan att göra så mycket som möjligt fritt tillgängligt av de resurser och program som vår forskning resulterar i. Detta tror vi är bra för språkteknologiforskningen i allmänhet och den i Sverige i synnerhet. Samtidigt är det bra för våra forskargrupper; att göra allmänna resurser fritt tillgängliga är förmodligen ett bra sätt att på lång sikt få upp citeringsfrekvensen, medan krångliga licensvillkor och avgifter tenderar att ha motsatt effekt. I synnerhet licensavgifter på forskningsresurser underminerar förstås en av hörnstenarna i

akademisk forskning, nämligen reproducerbarheten, men även ett förbud mot modifiering är antitetiskt till forskningens anda.

Ett i det här sammanhanget mycket passande exempel är det ursprungliga amerikanska WordNet (Fellbaum 1998; <http://wordnet.princeton.edu/>). Det är tillgängligt med en licens som liknar den som vi har valt för SALDO; det är fritt användbart för alla ändamål, även kommersiella, och det får modifieras, men upphovsrätten kvarstår hos Princeton University.

Här kan man direkt jämföra WordNet med de olika europeiska ordnäten framtagna i EuroWordNet-projektet (Vossen 1999). Dessa distribueras genom ELRA <http://www.elra.info> och licenserna kostar pengar. Priset för en forskningslicens för en icke-ELDA-medlem är visserligen blygsamma 2 eurocent per synset, vilket gör t.ex. 880 euro för det nederländska ordnätet. En forskare eller forskargrupp som licensierar ett europeiskt ordnät får naturligt nog inte distribuera det vidare.

Det ursprungliga Princeton WordNet är den förmodligen mest använda och därmed mest citerade lexikaliska resursen i språkteknologiforskningen, medan man mycket sällan ser referenser i litteraturen till något europeiskt ordnät, och när man gör det är författaren nästan alltid en medlem i det ursprungliga EuroWordNet-konsortiet.[1]

Vidare är vi övertygade om att en fri resurs som SALDO inte utgör ett hot mot kommersiella intressen. Den konkurrerar inte med konventionella ordböcker, som har andra målsättningar och andra målgrupper och därmed annat innehåll och annan struktur (se avsnitt 5). Inget hindrar i själva verket en kommersiell ordboksutgivare att lägga till en produkt baserad på SALDO som ”grädde på moset” i en elektronisk ordbok, ungefär som WordNet används idag, t.ex. i (den kommersiella produkten) Visual Thesaurus <http://www.visualthesaurus.com>. Den främsta anledningen till att SALDO inte konkurrerar med kommersiella ordboksprodukter är just att SALDO inte är en ordboksprodukt. Det är en slags lexikalisk databas, som skulle kräva en stor och kvalificerad arbetsinsats om man skulle vilja bygga en produkt på grundval av databasen.[2]

1.2 Finansiering

SALDO har tagits fram med allmänna medel. Efter 2003 har Språkbanken vid Göteborgs universitet finansierat huvuddelen av arbetet på SALDO inom ordinarie budget tilldelad av Humanistiska fakulteten, Göteborgs

[1] Naturligtvis har det något med saken att göra att Princeton WordNet är för engelska, men det är åtminstone vårt intryck att detta inte förklarar mer än en del av skillnaderna i användning.

[2] Och en sådan produkt skulle inte utan omfattande bearbetning också av innehållet utgöra någon konkurrent till existerande kommersiella ordböcker.

universitet. Under 2008 har arbetet till en liten del finansierats med medel från VR/DISC-projektet *Framtidssäkring av Språkbanken/ Safeguarding the future of Språkbanken* (VR:s Dnr 2007-7430; projektledare Lars Borin, Språkbanken, Göteborgs universitet).

Lars Borins och Markus Forsbergs arbete på SALDO har delvis och tidvis under åren 2006–2008 finansierats av VR-projektet *Grammatiker som mjukvarubibliotek/ Library-Based Grammar Engineering* (VR:s Dnr 2005-4211; projektledare Aarne Ranta, Chalmers tekniska högskola).

2 En svensk språkteknologisk lexikalisk basresurs

Språkteknologi är ett samlingsnamn för sådan informations- och kommunikationsteknologi (IKT) som låter datorer hantera mänskligt språk i alla dess former – tal, skrift och teckenspråk. Språkteknologi är ett starkt tvärvetenskapligt forskningsområde som är relevant överallt där människor interagerar med datorer och faktiskt även vid interaktion människor emellan, i form av olika sorters kommunikationshjälpmedel.

I skönlitteratur och film pratar människor med intelligenta datorer som naturligtvis även förstår gester och kroppsspråk. Det är helt klart att språkanvändande datorer skulle kunna förändra vår vardag enormt, av den enkla anledningen att datorer och människor är bra på olika saker. Datorer byggs in i fler och fler apparater, och språkteknologi behövs för att möjliggöra interaktion med dessa allt mer avancerade tekniska produkter i våra hem, på vägarna och på arbetet. Den mesta informationen i våra IT-system uttrycks dessutom fortfarande i något mänskligt språk (och dessutom på allt fler språk). Den vanligaste användningen för datorer idag är förmodligen för informationshantering: att skapa, läsa, ordna eller söka information i form av text, ljud eller video. Vi behöver språkteknologi för att vi inte ska drunkna i all denna information och som hjälp för att skapa den, översätta den, läsa den och alltmer för att hämta relevanta delar av dokument (gärna i sammanfattad form) utan att läsa dem. Särskilt behöver vi hjälp för att hantera text på språk som vi inte kan så bra eller som vi inte har möjlighet att läsa (synskadade eller bilförare, t.ex.).

I den statliga utredningen om svenska språkets ställning (*Mål i mun*) understryks att språkteknologi har vittgående betydelse för svenskans framtid som fullödigt språk. Informationssamhället avancerar på bred front, och utan språkteknologi för ett språk kan man inte räkna med att upprätthålla önskvärd tillgång till digital information eller digitala tjänster på det språket. En satsning som 24-timmarsmyndigheten kan knappast förverkligas utan språkteknologi.

Språkteknologi har både språkoberoende och språkberoende aspekter. Detta betyder att resultat som kommer ur språkteknologisk forskning om svenska och andra språk i Sverige är högst relevanta för den internationella forskargemenskapen, men också att språkteknologi för svenska inte kommer till utan vidare; den måste skapas i Sverige.

Den språkteknologiska forskningen och utvecklingen av språkteknologisystem behöver en infrastruktur av allmänt tillgängliga och standardiserade basresurser, både data och program för att arbeta med dessa data (en grunduppsättning sådana resurser kallas med en engelsk förkortning för BLARK – Basic LAnguage Resource Kit). Sådana resurser måste skapas för varje språk för sig.

Tabell 1 visar översiktligt de önskade grundläggande språkresurskomponenterna i en svensk BLARK. Tabellen är hämtad ur VR-ansökan ”En infrastruktur för svensk språkteknologi”, 2006. Merparten av dessa resurser identifierades som antingen helt icke-existerande eller enbart delvis existerande i det planeringsprojekt som beviljades finansiering av VR på basis av ansökan.

BASIC LANGUAGE RESOURCES	
resource	size
Syntactically annotated text data (”treebank”)	10 MW out of which 1 MW manually checked
Speech database	1650 hours
Basic Swedish lexicon with (inflectional) morphological information	**50 000 lemmas**
Morph-based phonetically transcribed lexicon	50 000 morphs
Parallel corpus including multilingual lexicon	5 MW total (both directions for each language pair)
Swedish wordnet	>50 000 entries
Swedish framenet	>50 000 entries
Terminology resources	>5 subject fields
Noise database	–

Tabell 1: Grundläggande språkresurser i en svensk BLARK

Klart är exempelvis att flera grundläggande resurser inte finns allmänt tillgängliga för svenska och än mindre för många av Sveriges andra språk, t.ex. lexikonresurser med information om ords böjning och betydelser och god täckning (åtminstone 50.000 uppslagsord), stora databaser med talspråk och stora textdatabaser – textkorpusar – med en sammansättning som motsvarar det skrivna eller talade språkets genrefördelning och variation och som är försedda med rik språklig information om t.ex. ordklasser

och satsanalyser. Vi behöver både enspråkiga och flerspråkiga textkorpusar som avspeglar det faktum att svenskan är ett standardspråk med en lång skrift- och språkvårdstradition, men som lever och verkar i en vardag med dubbelriktad flerspråkighet: utåt mot de nordiska språken samt engelska och andra världsspråk, och inåt mot landets minoritets- och invandrarspråk.

För svenskans vidkommande har det saknats en allmänt tillgänglig lexikalisk basresurs för språkteknologiforskning och språkteknologiska tillämpningar, något som kan antas ha inverkat hämmande på utvecklingen av svensk språkteknologi. Ett svenskt lexikon med morfologisk information, innehållande minst 50.000 uppslagsord har angetts som en baskomponent i en svensk BLARK (se ovan). Det kartläggningsarbete som har gjorts i det VR-stödda planeringsprojektet ”En infrastruktur för svensk språkteknologi” har gett vid handen att en sådan allmänt tillgänglig lexikonresurs inte finns. Det är helt klart att det finns svenska morfologiska lexikon för språkteknologitillämpningar, som har utvecklats av kommersiella aktörer. Dessa är inte tillgängliga för andra.

Det finns också i allmänhet morfologisk och annan forminformation i lexikala databaser upplagda för arbete på lexikon för mänskligt bruk, både i pappersform och elektroniska. Dessa databaser är tillgängliga i olika grad, men i inget fall är dessa data helt fritt tillgängliga för alla slags ändamål. Dessutom ställer språkteknologianvändning i allmänhet delvis andra krav på lexikonresurser än vad mänsklig användning gör.

SALDO är en sådan resurs. Den innehåller ungefär 68.000 lemman (beroende på hur man definierar ett ”lemma”; se Borin 2008) med fullständig böjningsinformation, given i form av böjningsmönster (se Appendix 4).

Syftet med denna elektroniska ordbok är att vara en fri resurs för språkteknologi. Den är alltså avsedd som vad som ibland kallas maskinlexikon. Den är inte avsedd för direkt mänskligt bruk (åtminstone inte dess formella komponent; den semantiska strukturen har såvitt vi vet ingen motsvarighet någon annanstans och kan vara nog så intressant för en mänsklig brukare).

3 SALDO: förhistoria

Svenskt associationslexikon (SAL), den lexikonresurs som ligger till grund för SALDO, är ett semantiskt lexikon, ett slags tesaurus. SAL är en relativt ny och relativt omfattande svensk tesaurus som dock är föga känd och därmed också lite använd. SAL skapades under åren 1987–1992 under ledning av Lennart Lönngren, då verksam vid Centrum för datorlingvistik

och Slaviska institutionen, båda vid Uppsala universitet.[3] Lexikonet har givits ut enbart i två små stencilupplagor i form av rapporter från Centrum för datorlingvistik, Uppsala universitet (Lönngren 1998), samt Institutionen för lingvistik, Uppsala universitet (Lönngren 1992) Dessutom har ända från början uppslagsorden och deras mest grundläggande semantiska relationer (se nedan) förelegat i elektronisk form, som rena textfiler.

SAL:s tillkomsthistoria dokumenteras i Lönngren 1989. De första källorna till ordförrådet var korpusmaterial, en lärobok i svenska för invandrare och ett populärvetenskapligt textmaterial. Dessutom innehåller SAL en relativt stor mängd – c:a 3000 – egennamn från olika källor, framförallt en liten encyklopedi. Så småningom utökades ordförrådet med hjälp av en lista av stickord ur *Svensk ordbok* (1986) som införskaffades från Språkdata, Göteborgs universitet. Den andra pappersversionen av SAL (Lönngren 1992) innehöll 71.750 lexikoningångar. SALDO innehåller 72.396 ingångar. Ökningen beror till en del på att nya ord har lagts till, men framför allt på att vissa ingångar tillhör mer än en ordklass eller uppvisar mer än ett böjningsmönster.[4]

Efter en lång viloperiod ”återuppväcktes” SAL i slutet av år 2003, när Lars Borin och Lennart Lönngren inledde ett samarbete syftande till att göra lexikonet tillgängligt online genom Språkbanken, Göteborgs universitet. En formell kontroll avslöjade vissa cirkeldefinitioner, vilka undanröjdes. År 2005 framställde en student i datalingvistik en grafisk prototyp till ett interface, benämnt SLV (Språkbanken Lexicon Visualization; Cabrera 2005). Tack vare detta interface har Lönngren kunnat förbättra bortåt ett tusen lexemanalyser, vilket gör SALDO till ett även i semantiskt avseende reviderat lexikon.

Snart insåg vi emellertid att SAL, för att bli en verkligt användbar språkteknologisk resurs, behövde kompletteras med åtminstone flektionell morfologisk information. Därmed startade arbetet på den nya versionen, SALDO.

4 SALDO: ett semantiskt lexikon

Som ett semantiskt lexikon kan SALDO beskrivas som ett lexikosemantiskt nätverk; det har viss yttre likhet med WordNet (Fellbaum 1998), men är i själva verket uppbyggt enligt helt andra principer.

[3] Lennart Lönngren var från 1993 fram till sin pensionering verksam som professor i ryska språket vid Universitetet i Tromsø (se <http://www.hum.uit.no/a/lonngren/>). Andra medverkande i arbetet med SAL var Gunilla Fredriksson som arbetade med lexikondefinitionerna samt Ágnes Kilár som svarade för programmeringen i det ursprungliga projektet.

[4] Således ska ”ingång” här förstås som en kombination av en semantisk och en morfosyntaktisk enhet, eller om man så vill ett lexem och ett lemma.

Strukturen i SALDO är baserad på två primära semantiska relationer. Varje uppslagsord är försett med en beskrivning (analys) bestående av en obligatorisk deskriptor, benämnd ”moder” och en fakultativ deskriptor, benämnd ”fader”. Som moder väljs det ord som bäst fyller följande krav: det skall vara nära semantiskt relaterat med nyckelordet och samtidigt mer centralt än detta. Med mer centralt menas bl.a. att det är mer frekvent och mer stilistiskt neutralt. Andra kriterier kan baseras på morfologisk eller semantisk asymmetri: t.ex. *sol* är mer centralt än *solig*, metall är mer centralt än *koppar*, *hus* är mer centralt än *fönster*, *häst* är mer centralt än *gnägga*. Ofta råder samstämmighet mellan dessa kriterier.

För att SALDO skall utgöra en sluten hierarkisk trädstruktur, behövs ett artificiellt ord, PRIM, vilket fungerar som moder till 50 oanalyserbara och sinsemellan orelaterade toppord.

En del uppslagsord har förutom den obligatoriska modern en fakultativ fader, vilken bl.a. tjänar till att differentiera ”syskongrupper” med gemensam moder.

SALDO (eller rättare sagt det underliggande SAL) har flera ovanliga egenskaper:

- det innehåller egennamn och fraser, vilka vanligen inte påträffas i konventionella lexika; även förkortningar och s.k. formord (*att*, *i*, *den*) är medtagna och försedda med deskriptor(er);
- det är strikt semantiskt baserat; alla uppslagsord är lexem, dvs ordbetydelser snarare än ord, och det finns ingen information om ordklass eller böjning;
- SALDO har i jämförelse med ett lexikalisk-semantiskt nätverk som WordNet mindre specificerade semantiska relationer, i jämförelse med konventionella thesauri, däremot, är relationerna i SALDO mer elaborerade.

Nedan ges några exempel på analysförsedda uppslagsord i SALDO, slumpvis hämtade under bokstaven ”B” i Lönngren 1992:

balkong : hus

bröd : mat + mjöl

brödföda : uppehälle

bröllop : gifta sig

Bulgakov : författare + rysk

De två primära lexiko-semantiska relationerna (moder och fader) ger upphov till ett stort antal deriverade relationer. Särskilt värdefulla är syskonrelationerna (gemensam moder, gemensam fader, gemensamma föräldrar), vilka, om deskriptorerna är välfunna, skall generera naturliga semantiska grupper.

Tack vare ovannämnda interface (SLV) har det blivit möjligt att upptäcka och eliminera en egenhet i strukturen hos SAL vilken tidigare accepterades som oundviklig. Det gäller multipla relationer mellan två lexem. T.ex. *segelbåt* ges naturligen beskrivningen *båt + segel*. Om *segel* återförs på *båt* blir resultatet att *segel* är samtidigt barn och make till båt. En korrigering av sådana fall visar sig oftast leda till bättre resultat även ”lokalt”.

En eventuell framtida bearbetning bör bl.a. inriktas på att utöka antalet väl lexikaliserade sammansättningar. Sådana ges i större ordböcker som exemplifieringar, men tyvärr ofta inte som egna uppslagsord. Att som skedde vid initialstadiet av arbetet med SAL hämta material från korpusar kan tyvärr leda till att många icke-lexikaliserade ”korpussammansättningar” införlivas. I SAL finns sålunda av denna anledning ordet *genomsnittsskatt*, men inte ”det bättre” *genomsnittsinkomst* (det senare finns som exempel, men inte som uppslagsord i *Svensk ordbok*).

5 SALDO: ett morfologiskt lexikon

Den mest avgörande skillnaden mot befintliga lexikon är förstås – som redan nämnts – ändamålet, nämligen användning i datorprogram för språkteknologi. Detta ändamål betingar framförallt distributionsformen, men av olika skäl, bland annat ändamålet, så skiljer sig också SALDO:s innehåll rent konkret på ett antal punkter från det man finner i befintliga lexikon för mänskligt bruk.

Den stora nyheten i SALDO jämfört med dess föregångare SAL är att varje lexikoningång nu är försedd med fullständig information om dess böjning. SAL innehöll enbart uppslagsformer utan någon som helst information om ordböjning eller ens ordklass.

Naturligtvis har vi utgått ifrån befintliga beskrivningar av svensk böjningsmorfologi när vi har tillfört denna information i SALDO. Vi har framför allt använt oss av NEO (Nationalencyklopedins ordbok; NEO 1995) – i form av den lexikaliska databas som legat till grund för arbetet med NEO – men också av dess föregångare SO och SAOL (Svenska Akademiens ordlista; SAOL 2006), samt Svenska Akademiens grammatik (Teleman, Hellberg & Andersson 1999) samt Hellberg 1978.

Av praktiska (och ibland teoretiska) skäl har vi ibland avvikit från dessa beskrivningar (som dessutom inte alltid överensstämmer sinsemellan). T.ex. har vi inte funnit Hellbergs ”tekniska stam” vara ett fruktbart be-

grepp.[5] Ordböjningsinformationen i SAL skiljer sig från den som man finner i konventionella ordböcker för mänskligt bruk. Här har vi styrts av tre delvis motstridiga metodologiska principer:

Generositet: SALDO-böjningsangivelserna är mycket generösa med vilka former som ska antas existera för ett givet lemma. Man kan säga att vi tar fasta på den idé om potential som finns i själva böjningsparadigmbegreppet. I allmänhet måste det finnas en klar(t formulerbar) grammatisk, semantisk eller pragmatisk anledning till att ett visst lexikonord inte ska anses kunna bilda alla de former som dess ordklasstillhörighet antyder. Här handlar det i praktiken om numerus (plural- och någon gång singularformer) hos substantiv, komparation hos adjektiv och vissa adverb samt passivt particip och s-former hos verb. Fler enheter är också i linje med detta angivna som böjliga i SALDO-morfologin än i normativa svenska ordböcker (t.ex. dna, som i SALDO dessutom har vacklande genus; både *dna:t* och *dna:n* godtas som bestämd form; se nedan). I allmänhet innebär detta att när det finns en skillnad mellan SALDO och moderna svenska ordböcker (i praktiken avses med de senare SAOL och NEO), så är SALDO generösare.

Precision: Samtidigt är det viktigt i en språkteknologiskt lexikonresurs att undvika övergenerering, åtminstone i de fall där detta leder till omotiverad ambiguitet i analyserna. Därför har en strävan i vårt arbete varit att varje ords böjningsmönster ska vara exakt ”så stort” som det behöver vara men inte större. För närvarande syns detta tydligast i förhållande till konventionella ordböcker i SALDO:s verbparadigm, där vissa verb inte antas kunna bilda perfektparticip enligt SALDO, medan en sån begränsning inte någonsin verkar finnas i de moderna ordböckerna. Sålunda skiljer SALDO-morfologin mellan tre verb "väga":

väga `vb_2a_viga` (d.v.s. *väger*, *vägde*, *vägt*, *vägd*)
(*Handlaren stod och vägde fisk.*)

väga `vb_2m_väga` (d.v.s. *väger*, *vägde*, *vägt*, –)
(*Fisken vägde sammanlagt nästan två ton.*)

väga `vb_1a_laga` (d.v.s. *vägar*, *vägade*, *vägat*, *vägad*)
(Mest i uttrycket *ovägat land*)

Variation: Ett språkteknologiskt lexikon ska kunna användas för att analysera fri text, t.ex. på webben, där man jämsides med standardböj-

[5] Hellberg själv betraktar också den tekniska stammen som just något som han tvingats ta till på grund av tekniska tillkortakommanden (Hellberg 1978: 15f), som knappast längre är förhanden.

ningsformerna hittar en hel del former som enligt normativa skriftspråksordböcker inte ska finnas. Vi har intagit ståndpunkten att vi i SALDO-morfologin erkänner en hel del belagda (men inte normativa) former (ex.-vis *datum* med genus utrum och tillhörande vacklande utral plural på *-ar/-er*, eller den alternativa pluralen *minutrar* av *minut*) samt vissa stavningsvarianter, men inte felstavningar. Gränsen mellan de två sistnämnda är förvisso långt ifrån knivskarp. Det finns förstås hur många klara felstavningar som helst, något som framstår med all önskvärd tydlighet när man googlar efter ordformer på webben. Samtidigt har vi även fall som ”microvågsugn”. Är det att betrakta som en felstavning eller en stavningsvariant? ”microvågsugn” ger 67.000 träffar på Google i början av december 2007, medan ”mikrovågsugn” ger 182.000 träffar. I SALDO betraktas det som en stavningsvariant. Analoga fall är legio.

Vi är medvetna om att den rika böjningsmorfologiska variation som man lätt belägger ex.-vis på internet har mer än en orsak. Delvis avspeglar den en existerande variation i språket som man brukar sträva efter att avskaffa med standardisering av skriftspråk. Delvis avspeglar den det faktum att svenska kanske bäst beskrivs som många olika, till stor del överlappande språk. Slutligen handlar det som sagt ofta om rena skrivfel.

SALDO är alltså inte ett normativt lexikon utan strävar efter att vara deskriptivt. Samtidigt innebär tanken om böjningsparadigm som potentialer snarare än listor över belagda former naturligtvis att det finns ett normativt drag i SALDO, nämligen oundvikligen det som är språkvetenskapen eget, alltså formulerandet av lagbundenheter i våra språk. Det är också ett erkännande av det faktum att hur stor korpus man än samlar in, kommer man aldrig att se alla böjningsformer av alla lexikonord, inte ens i ett språk som svenska som ju i ett typologiskt perspektiv har en mycket fattig böjningsmorfologi.[6]

Samtidigt vet man som språkbrukare att vissa former av vissa ord inte bara är icke-belagda, utan faktiskt verkar vara icke-beläggbara, t.ex. de redan nämnda preteritumparticipformerna av vissa verb, komparativformer av (participliknande) adjektiv på *-ad* (alltså t.ex. *långfingrad*, men inte *glad*), etc. Som sagt kan anledningarna till att en eller flera former saknas vara av olika slag, semantiska eller formella (det senare verkar gälla för adjektiven på *-ad*), men även helt idiosynkratiska (Hetzron 1975). Naturligtvis har vi tagit hänsyn till detta. Vi har som redan sagts i allmänhet valt att hellre fria än fälla, vilket säkerligen leder till att SALDO:s morfologiska

[6]Men det är förstås inte konstigare än att man aldrig kommer att se ”alla språkets meningar” hur stor korpus man än samlar in.

beskrivning övergenererar. Detta är dock inget praktiskt problem, så länge som potentiella men icke möjliga former inte sammanfaller med faktiska former.

6 Flerordsenheter i SALDO

SALDO innehåller ungefär 2000 flerordslexem. Huvuddelen av dessa – drygt 1100 – är verb, mest partikelverb.

Teoretiskt och metodologiskt ser vi inget skäl att skilja mellan enords- och flerordslexem och -lemman. Av praktiska skäl hanterar ordböjningskomponenten i SALDO (se avsnitt 9 nedan) dock inte böjningsmönster eller enstaka medlemmar i böjningsmönster som består av mer än ett ord och där annat språkligt material kan infogas mellan orden. Praktiskt yttrar sig detta i att verbparadigmen inte innehåller de perifrastiska formerna (perfekt, pluskvamperfekt, futurum, etc.) och att verbflerordingar (huvudsakligen partikelverb) inte alls hanteras i denna version av SALDO:s ordböjningskomponent. Flerordslemman där annat språkligt material inte kan infogas hanteras redan nu (även i de fall där de ingående orden böjs, t.ex. *enarmad bandit*).

7 Encyklopediska drag i SALDO

SALDO innehåller många egennamn, en ordklass som inte brukar vara representerad i andra lexikon, utan snarare brukar sägas höra hemma i encyklopedier. Ursprungligen motiverades detta rent semantiskt, men vi kan bara konstatera att information om egennamns grammatiska uppträdande naturligtvis behövs i språkteknologiapplikationer lika mycket som för andra ordklasser. Morfologiskt och syntaktiskt finns inga tungt vägande skäl att dra en skiljelinje mellan egennamn och andra ordklasser. Den stora skillnaden finns på det semantiska planet och det är naturligtvis det som gör att egennamn inte brukar finnas med i konventionella ordböcker. De kan inte förses med en definition, parafras eller beskrivning av deras denotat (fast det skulle förvisso kunna sägas om en del grammatiska ord också). Däremot har det ansetts möjligt att förse dem med en association.

8 Tekniskt format

Grundlexikonet som utgör SALDO distribueras uppdelat i två textfiler, båda med en lexikoningång per rad. Varje rad är i sin tur uppdelad i ett antal fält åtskilda av tabbtecken.

Först i varje fil står en kort licenstext på svenska och engelska. Licenstextraderna inleds med en brädgård (#) först på varje rad. Detta tecken förekommer inte i själva lexikonlistan.

De båda lexikonlistorna är:

1. Lexemlistan, med fem fält:
   (a) lexem-id för uppslagslexemet
   (b) lexem-id för moderlexemet
   (c) lexem-id för faderlexemet
   (d) lemma-id för uppslagslexemets lemma
   (e) uppslagslexemets djup i lexikonets hierarkiska struktur (en siffra)
2. Lemmalistan, med fyra fält:
   (a) lemma-id för uppslagslemmat
   (b) grundform
   (c) ordklasskod
   (d) böjningsangivelse (se Appendix 4)

För närvarande (2008-05-07) innehåller lexemlistan 72.396 ingångar och lemmalistan 68.355 ingångar.

Lexem- och lemma-id är avsedda att kunna användas som XML-identifierare, eftersom vi planerar att skapa en OWL/XML-version av SALDO. De följer därför reglerna för XML-namn (se <http://www.w3.org/XML/>), vilket innebär att vissa tecken inte är tillåtna. I synnerhet används understreck för att återge mellanslag i flerordsenheter. Understreck står också först i identifieraren ifall grundformen börjar med ett tecken som inte får inleda ett XML-namn. I övrigt kodas förbjudna tecken om med en mekanism som liknar procentkodning av URL:er, men där tecknet ”·” (upphöjd punkt; Unicode/ISO 8849-1 #xB7) används som prefix istället för procenttecknet, som inte är tillåtet i XML-namn. Den upphöjda punkten följs av två hexadecimala siffror som ger kodpunkten för det tecken som avses. Hittills har det räckt med två positioner. Behöver man fler, föreslår vi att man efter ”·” använder gement ”u” plus fyra hexadecimala siffror eller versalt ”U” plus åtta hexadecimala siffror. På det viset kan godtycklig Unicode-kodpunkt återges.

SALDO:s grundordklasser är (ordklasskod inom parentes): substantiv (nn), egennamn (pm), adjektiv (av), pronomen (pn), räkneord (nl), verb (vb), adverb (ab), prepositioner (pp), konjunktioner (kn), subjunktioner

(sn), infinitivmärke (ie) och interjektioner (in). Flerordingar betecknas med grundordklassens kod + ”m” (nnm, vbm, etc.), förkortningar med grundordklassens kod + ”a” (pma, ava, etc.). Led i flerordingar betecknas med grundordklassens kod + ”f” (nnf, avf, etc.). Slutligen finns koden ”ssm” för flerordingar som utgör satser eller satsliknande uttryck.

9 Funktionell Morfologi

9.1 Bakgrund

Vi har använt oss av Funktionell Morfologi (FM) (Forsberg & Ranta 2004; Forsberg 2007) för att utveckla den morfologiska komponenten, ett program utvecklat på Chalmers Tekniska Högskola av Aarne Ranta och Markus Forsberg. FM är ett så kallat inbäddat språk för definition av morfologier i det funktionella programmeringsspråket Haskell (Jones 2003). Det inbäddade språket består av en språkspecifik och en språkoberende del, där den språkspecifika delen inkluderar den morfosyntaktiska beskrivningen och böjningsmaskineriet, och den språkoberoende delen sådant som är gemensamt för alla morfologier, sådant som analys och kompilering till andra format.

FM tog sin början 2000 då Aarne jobbade på Xerox, när Gérard Huet kom på besök och visade att det går utmärkt att skapa morfologier i CAML, ett annat funktionellt språk. Den rådande uppfattningen vid den tiden, som delades av Aarne, var att man skulle använda sig av finita tillståndsautomater för att skapa morfologiska komponenter. Gérard visade dock att det fanns mycket kvar att göra med avseende på *beskrivningen* av morfologier.[7]

I Paris, mars 2001, gjorde Aarne en fullständig implementation av de franska verbparadigmen i Grammatical Framework (GF), en grammatikformalism som bygger på typteori och funktionell programmering. Han blev uppmuntrad över hur lätt det gick, och i Siena juni 2001 gjorde han en liknande, nära fullständig implementation för italienska i Haskell, vilket samtidigt blev en början till Funktionell Morfologi. Hösten 2001 började han på en implementation av svenska, som har utgjort en grund för SALDO:s morfologiska komponent.

I slutet av 2001 anställdes Markus som projektassistent i forskningsprojektet ”Interaktiv Språkteknologi”. Han utvecklade FM vidare, vilket senare blev en del av hans avhandling *Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract*. Våren 2002 gjorde Aarne

[7] Däremot med avseende på vad beskrivningen *avbildar* är en finit tillståndsautomat många gånger att föredra, eftersom en automat är en kompakt struktur som ger snabb uppslagning.

och Markus ett försök att samla in svenska ord för att skapa ett fritt tillgängligt, morfologiskt lexikon via en hemsida,[8] men resultatet blev skralt.

Under åren 2003-2004 gjordes två examensarbeten som använde FM för att skapa en spansk (Andersson & Söderberg 2003) respektive en rysk (Bogavac 2004) morfologi. Dessa arbeten visade att det var möjligt att implementera huvuddelen av ett språks böjningssystem inom ramen för ett examensarbete, och dessutom, att det var möjligt för studenter som saknade tidigare erfarenhet av Haskell.

År 2007 trycktes Markus avhandling, och några månader innan dess en avhandling skriven av Otakar Smrz (Smrz 2007), vari han beskrev sin användning och utveckling av Funktionell Morfologi för arabiska. Arbetet med SALDO:s morfologiska komponent tog sin början detta år, men satte igång på allvar i slutet av året, då Markus blev anställd vid Språkbanken.

9.2 Introduktion

Det traditionella sättet att skapa en morfologi i FM är att först leta rätt på en lämplig grammatik för målspråket som beskriver böjningssystemet, för att sedan beskriva detta böjningssystem i FM. Därefter börjar man utveckla lexikonet för språket, möjligtvis med hjälp av automatiska metoder.

SALDO skiljer sig markant på denna punkt, eftersom i SALDO är ordmängden redan given och arbetet har varit att dels ge varje ord en paradigmangivelse, och dels att implementera paradigmangivelsen i FM:s böjningsmaskineri. Detta i sin tur innebär att vi har ett paradigmsystem som är avsevärt mycket större än tidigare morfologier, ungefär 10 gånger så många paradigm, eftersom denna tydliga distinktion mellan implementation och beskrivning omöjliggör att man hanterar enordsparadigmen genom uppräkning. Därtill har vi ett paradigmssystem som varit i ständig svängning, eftersom utvecklingen av paradigmsystemet snarare har varit en process än något på förhand givet. Allt detta har lett till att FM har utvecklats mycket den senaste tiden, speciellt med avseende på kvalitetsäkring (testning) och metodik. Denna utvecklingen kommer även de äldre morfologierna till gagn.

Vi kommer inte att gå in på några tekniska detaljer om utvecklingen av FM och implementationen av SALDO, utan vi kommer fokusera på hur slutprodukten används. Om läsaren är intresserad av mer tekniska detaljer, så hänvisas hon till Markus avhandling, och den framtida, för tillfället oskrivna, tekniska rapporten om FM och SALDO.

[8] http://www.cs.chalmers.se/~markus/svenska

9.3 Programvaran

9.3.1 Installation

Den morfologiska komponenten består av två delar, ett kommandoradsbaserat program och ett lexikon. Lexionet består av en listning av orden, annoterade med paradigmidentifierare och lemmaidentifierare.

Exempelvis har vi här ordet *hoppa* som anges böjas enligt paradigmet `vb_1a_laga` och med lemmaidentifieraren `hoppa..vb.1`. Paradigmidentifieraren har en mnemonisk kodning, som säger att det är ett verb i första konjugationen med perfekt particip som böjs enligt ordet *laga*.

```
vb_1a_laga "hoppa" {id("hoppa..vb.1")};
```

Nu över till installationen av programmet. Vi förutsätter att du använder dig av ett Unix-lik miljö (Linux, Mac OSX eller Cygwin för Windows). För installation av programmet krävs Haskellkompilatorn GHC[9] och en C-kompilator, exempelvis GCC. Installationen består av den traditionella konfigurera-bygg-installera-processen som ges här steg för steg. Resultatet av denna process är ett program `saldo` som har installerats i katalogen `\usr\local\bin`.

ladda ned källkoden:	`saldo_1.0.tgz`
packa upp:	`$ tar xvfz saldo_1.0.tgz`
byt katalog:	`$ cd saldo_1.0`
konfigurera:	`$ ./configure`
bygg:	`$ make`
installera:	`$ sudo make install`

9.3.2 Användning

Vi antar nu att ni lyckats installera programvaran och laddat ned medföljande lexikon. En första överblick av vad man kan göra med programmet fås genom kommandot `saldo -h`. Vi kommer nu att gå igenom de olika aspekterna av programmet.

```
$ saldo -h
FM 2.1 (c) M. Forsberg & A. Ranta, 2008, under GNU GPL
Usage: saldo [OPTION...] dictionary_file(s)...
  -i            --inflection           run inflection engine
  -s            --synthesiser          enter synthesizer mode
  -a            --analysis             pos tagging
  -t TOKENIZER  --tokenizer=TOKENIZER  select mode (default, words,
                                                    lines, norm)
  -m MODE       --mode=MODE            select mode (fail, lexfail,
```

[9] <http://www.haskell.org/ghc>

```
                                                    lexcomp)
  -p PRINTER    --printer=PRINTER       print using PRINTER
                                         (core, paradigms, tagset,
                                          words, lex, tables,
                                          extract, gf, latex, xml,
                                          sfst, sfstlex, sfstheader,
                                          lexc, xfst, sql, hundict,
                                          hunaffix, lmf)
  -e ENCODING   --encoding=ENCODING     select another morphosyntactic encoding (SUC)
  -q QUALITY    --quality=QUALITY       run tests (all, test, dup, undef,
                                                   pop, argc , dict)
  -h            --help                  display this message
  -v            --version               display version information
```

Till och börja med kan det vara värt att poängtera att FM först och främst är en utvecklingsmiljö, och att tanken är att när morfologin väl skall sättas i drift, så bör den exporteras till ett av exportformaten. Å andra sidan är FM:s analys inte särskilt långsam, bortsett från de tiotal sekunder det tar att räkna ut alla ordformer, så om användningen inte är ytterst tidskritisk, så kan man gott och väl använda FM direkt.

Därtill kan nämnas att FM har stöd för sammansättningsanalys, men att vi i denna utgåva inte implementerat denna möjlighet. Däremot planerar vi att inkludera sammansättningsanalys i en senare utgåva av FM-SALDO.

9.3.3 Analys

Låt oss nu ta en titt på hur man kan använda sig av programmet för att analysera ordet *lexikon*. Vi startar programmet och ger lexikonet som argument, och matar in ordet *lexikon*.

Vi kan bland annat utläsa att den morfologiska komponenten består av ungefär 67.000 lemman, som expanderas till 760.000 ordformer, 825 paradigm, och att kompileringen tog 12 sekunder. När vi matade in ordet *lexikon* fick vi tre analyser. Analyserna är i så kallat JSON-format, ett enkelt dataformat som består av atomiska typer, sekvenser och records. Fördelen med JSON är att det är ett kompakt och entydigt dataformat som är enkelt att analysera.

```
$ saldo saldo.dict
FM 2.1 (c) M. Forsberg & A. Ranta, 2008, under GNU GPL

processing dictionary in file saldo.dict

computing ('.' = 20k word forms): ..................................

760k word forms (c: 760431, u: 559279)
compile time: 20.00 seconds
language id: SALDO v1.0
825 paradigms
67k entries (e: 66996, i: 0)

lexikon
{"lexikon":[
```

```
{"head":"lexikon","pos":"nn","param":"comp",
 "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon","e":"*"},
{"head":"lexikon","pos":"nn","param":"pl indef nom",
 "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon","e":"*"},
{"head":"lexikon","pos":"nn","param":"sg indef nom",
 "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon","e":"*"}
]}
```


Den vanliga användning är naturligtvis inte att mata in enskilda ord, utan att analysera hela texter, vilket görs på vanligt Unix-vis:

```
$ cat corpus.txt | saldo saldo.dict
```

Vi kan få ut annan information med `-m`-flaggan, en möjlighet som är speciellt värdefull vid själva utvecklingen av morfologin. Den enda aktuella inställningen är 'fail', eftersom vi ännu inte beskrivit någon sammansättningsanalys.

fail listar alla ord i analysen som FM-SALDO inte lyckats analysera.

lexfail skriver ut alla ord som inte finns i lexikonet (även om de har analyserats som sammansättning).

lexcomp skriver endast ut de ord som analyserats som sammansättningar.

Vidare kan vi även välja tokeniserare. För närvarande finns fyra tokeniserare, som beskrivs nedan. Om ingen tokeniserare anges, används default-tokeniseraren.

default blankteckenbaserad tokeniserare, som markerar ord med stor bokstav som tvetydiga. Nummer och symboler identifieras.

words blankteckenbaserad tokeniserare.

lines nyradbaserad tokeniserare.

norm nyradbaserad tokeniserare som annars beter sig som default.

9.3.4 Syntes

Syntesen i FM producerar alla böjningstabeller där ett givet ord är medlem. Om vi återgår till vårt exempelord *lexikon*, så får vi vid syntes av detta ord hela ordets böjningstabell. Om ordet skulle ha förekommit i någon annan böjningstabell, vilket det inte gör, så skulle även den böjningstabellen listas.

```
$ saldo -s saldo.dict
[...]

lexikon
{"lexikon":{
"lexikon..nn.1":[
{"word":"lexikon","head":"lexikon","pos":"nn","param":"sg indef nom",
                  "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikons","head":"lexikon","pos":"nn","param":"sg indef gen",
                  "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikonet","head":"lexikon","pos":"nn","param":"sg def nom",
                   "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikonets","head":"lexikon","pos":"nn","param":"sg def gen",
                    "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikon","head":"lexikon","pos":"nn","param":"pl indef nom",
                  "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexika","head":"lexikon","pos":"nn","param":"pl indef nom",
                 "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikons","head":"lexikon","pos":"nn","param":"pl indef gen",
                  "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikas","head":"lexikon","pos":"nn","param":"pl indef gen",
                  "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikonen","head":"lexikon","pos":"nn","param":"pl def nom",
                   "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikonens","head":"lexikon","pos":"nn","param":"pl def gen",
                    "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikon-","head":"lexikon","pos":"nn","param":"comp",
                  "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"},
{"word":"lexikon","head":"lexikon","pos":"nn","param":"comp",
                 "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon"}
]
}
```

9.3.5 Böjningsmaskineriet

Man kan använda sig av FM:s böjningsmaskineri direkt genom att ange flaggan `-i`. Vi kan till exempel böja vårt exempelord `lexikon`, genom att ange `nn_vn_lexikon "lexikon";`.

```
$ saldo -i

nn_vn_lexikon "lexikon" ;

{"nn_vn_lexikon \"lexikon\" ;":[
{"word":"lexikon","head":"lexikon","pos":"nn","param":"sg indef nom",
 "inhs":["n"],"id":"lexikon_nn_n","p":"nn_vn_lexikon","attr":"0"},
[...]
```

Notera att vi inte fick samma lemmaidentifierare som tidigare: `lexikon_nn_n` istället för `lexikon..vb.1`. Anledningen till detta är att när inte lemmaidentifieraren anges, så konstruerar FM en identifierare genom att sätta ihop exempelordet, ordklassen och de inherenta dragen. Det är dock bättre att ha en identifierare som entydigt identifierar ett lemma, vilket vi får genom att istället skriva: `nn_vn_lexikon "lexikon" {id("lexikon..nn.1")};`.

9.3.6 Ordklasstaggning

Detta alternativ används av `extract`, men kan naturligtvis vara användbart i andra sammanhang. Alla ord ges en ambiguitetsklass, d.v.s. en listning av dess analyser, och därtill delar en något naiv meningssegmenterare (den antar att '.', '!' eller '?' markerar en meningsgräns) upp intexten i meningssegment. Ett exempel klargör detta, här med meningen: *Bosse kör en röd bil.* Notera att meningen markeras med krullparanteser.

```
$ saldo -a saldo.dict
...

Bosse kör en röd bil.

{("Bosse",pm nom)
 ("kör",nn comp|nn sg indef nom|vb imper|vb pres ind aktiv)
 ("en",ab invar|al sg u|nl nom num|nn comp|nn sg indef nom)
 ("röd",av pos indef sg u nom)
 ("bil",nn comp|nn comp|nn sg indef nom)
 (".",spec)}
```

9.3.7 Exportformat

FM stödjer många exportformat, men det viktigaste är det som producerar fullformslexikonet: `-p lex`. Fullformslexikonet är, likt analyserna, JSON-formaterat, men endast radvis; hela lexikonet är dock inte ett enda JSON-objekt. Anledningen till detta är dels för att det skall vara enkelt att arbeta med delmängder av lexikonet, dels för att det är enklare att undvika minnesproblem om man inte försöker läsa in ett stort JSON-objekt på en gång. Vill du av någon anledning ha lexikonet som ett enda JSON-objekt, är detta dock enkelt att åtgärda: sätt hela lexikonet inom hakparentes och separera raderna med komma.

Låt oss återgå till vårt exempelord `lexikon`, men här i fullformsformat. Låt oss titta på en rad, och gå igenom vad de olika beteckningarna står för. Vi har åtta fält: `word` är ordformen, `head` är exempelordet, `pos` är ordklassen, `param` är den morfolosyntaktiska analysen, `inhs` är de inherenta drag, här genuset neutrum, `id` är lemmaidentifieraren, `p` är paradigmidentifieraren, och `attr` är ett attribut som beskriver hur ordet beter sig i sammansättningsanalys.

```
{"word":"lexikon","head":"lexikon","pos":"nn","param":"comp",
 "inhs":["n"],"id":"lexikon..nn.1","p":"nn_vn_lexikon","attr":"1"}
```

FM stödjer även många andra exportformat, där `core`, `paradigms`, `tagset` och `extract` ger information om böjningsmaskineriet, medan resten är olika manifestationer av fullformslexikonet.

core listar kärnlexikonet, d.v.s. alla paradigm i böjningsmaskineriet med deras exempelord.

paradigms listar alla böjningstabeller för alla paradigm i böjningsmaskineriet med deras exempelord.

tagset listar den aktuella morfosyntaktiska kodningen i JSON-format.

words listar alla ord (utan analys)

tables skriver ut alla ords böjningstabeller.

latex skriver ut alla ords böjningstabeller i ett LaTeX-formaterat format.

extract översätter paradigmen till `extract`:s lexikonextraktionsformat. Detta exportformat är experimentellt och ger endast tillfredsställande resultat för vissa av paradigmen.

gf skriver gf-regler. Denna översättning är partiell eftersom ett typsystem krävs.

xml skriver ut fullformslexikon i ett XML-format

sfst skriver ut fullformlexikon i ett SFST-format (använd hellre sfstlex)

sfstlex skriver ut fullformlexikon i ett SFST:s lexikonformat.

sfstheader skriver ut en SFST header för lexikonformatet.

lexc skriver ut fullformslexikonet i ett LEXC-format.

xfst skriver ut fullformslexikonet i ett XFST-format.

sql skriver ut fullformslexikonet i ett SQL-format.

hundict skriver ut fullformlexikonet i ett Hunspell-format (detta format skulle kunna göras bättre genom att flytta affixen till affixfilen.)

hunaffix skapar Hunspell:s affixfil.

lmf skriver ut fullformslexikonet i ett LMF-format (Lexical Markup Format)

9.3.8 Morfosyntaktisk kodning

En nyhet i FM-SALDO är samexisterande morfosyntaktiska kodningar. En av kodningarna har precedens, och denna kodning översätts till de andra kodningarna. Fördelen med att ha denna översättning i FM, snarare än att skriva ett eget översättningsprogram, är att kodningen får genomslag i FM:s alla aspekter, från analys till formatgenerering. För närvarande finns

endast en samexisterande kodning för FM-SALDO, och det är en partiell översättning av SALDO:s morfosyntaktiska kodning till SUC:s. Den är partiell eftersom de olika kodningarna inte kan översättas rakt av – exempelvis har inte SUC flerordningar, och SALDO saknar SUC:s uppdelning av olika pronomen.

```
$ saldo -e SUC saldo.dict
[...]
lexikon
{"lexikon":[
{"head":"lexikon","pos":"NN","param":"SMS",
 "inhs":["NEU"],"id":"lexikon..nn.1","p":"nn_vn_lexikon","e":"*"},
{"head":"lexikon","pos":"NN","param":"PLU IND NOM",
 "inhs":["NEU"],"id":"lexikon..nn.1","p":"nn_vn_lexikon","e":"*"},
{"head":"lexikon","pos":"NN","param":"SIN IND NOM",
 "inhs":["NEU"],"id":"lexikon..nn.1","p":"nn_vn_lexikon","e":"*"}
]}
```

9.3.9 Kvalitetsäkring

En viktig utveckling av FM är tillägget av testning. Man kan säga att testning ger oss en minsta nivå för kvaliteten på SALDO, där man för varje test kan säga: *åtminstone kan vi vara (hyfsat) säkra på att....*

Vid testningen upptäcks tre typer av fel: annotationsfel, implementationsfel och testfel. Annotationsfel är ord som har taggats med fel paradigm, implementationsfel är när pardigmet är felaktigt implementerat, och testfel när själva testet är fel.

Testning körs med flaggan `-q` med något av argumenten som beskrivs nedan. Av dessa tester så är endast `test` språkberoende. För de språkberoende testerna skiljer man mellan positiva och negativa tester – i de positiva testerna söker man efter sådant som måste finnas, och i de negativa, efter sådant som inte får finnas. För närvarande består testningen av 18 positiva test och 9 negativa tester. Testerna är allt från att ett ord, som inte är en förkortning, skall innehålla en vokal, till att verb som är uppmärkt som saknande perfekt particip, verkligen saknar perfekt particip.

all kör alla tester.

dup sök efter lemma-id-dubletter.

undef sök efter paradigm som saknar definition, men refereras till i lexikonet.

pop sök efter paradigm som är definierade, men inte refererade till i lexikonet.

argc sök efter paradigm med felaktigt antal argument.

dict sök efter identiska ingångar, d.v.s. ingångar som formellt inte kan skiljas åt. Detta test tar längst tid, då alla ingångar jämförs med alla ingångar, vilket summerar till ungefär 2.2 miljarder jämförelser för SALDO.

test kör de språkspecifika testerna.

Referenser

Andersson, I. & T. Söderberg. 2003. "Spanish morphology – implemented in a functional programming language." Master's Thesis, Master's Thesis in Computational Linguistics, Göteborg University.

Bogavac, L. 2004. "Functional Morphology for Russian." Master's Thesis, Department of Computing Science, Chalmers University of Technology.

Borin, Lars. 2008. "Lemma, lexem eller mittemellan? Ontologisk ångest i den digitala domänen." In *Nog ordat? Festskrift till Sven-Göran Malmgren*, edited by Kristinn Jóhannesson, Hans Landqvist, Aina Lundqvist, Lena Rogström, Emma Sköldberg & Barbro Wallgren Hemlin, 59–67. Göteborg: Meijerbergs arkiv för svensk ordforskning.

Cabrera, Isabelle. 2005. "Språkbanken lexicon visualization. Rapport de stage." Projet réalisé au Département de Langue Suédoise, Université de Göteborg, Suède.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Forsberg, Markus. 2007. "Three tools for language processing: BNF converter, Functional Morphology, and Extract." Ph.D. diss., Göteborg University and Chalmers University of Technology.

Forsberg, Markus & Aarne Ranta. 2004. "Functional morphology." *ICFP'04. Proceedings of the ninth ACM SIGPLAN international conference of functional programming*. Snowbird, Utah: ACM.

Hellberg, Staffan. 1978. *The Morphology of Present-Day Swedish*. Data linguistica no. 13. Stockholm: Almqvist & Wiksell International.

Hetzron, Robert. 1975. "Where the grammar fails." *Language* 51: 859–872.

Jones, Simon P. 2003, May. *Haskell 98 Language and Libraries: The Revised Report*. Cambridge: Cambridge University Press.

Lönngren, Lennart. 1989. *Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi*. Centrum för datorlingvistik. Uppsala universitet. Rapport UCDL-R-89-1.

Lönngren, Lennart. 1992. *Svenskt associationslexikon. Del I-IV*. Institutionen för lingvistik. Uppsala universitet.

Lönngren, Lennart. 1998. "A Swedish associative thesaurus." *Euralex '98 proceedings, Vol. 2*. 467–474.
NEO. 1995. *Nationalencyklopedins ordbok*. Höganäs: Bra Böcker.
SAOL. 2006. *Svenska Akademiens ordlista över svenska språket*. Stockholm: Norstedts Akademiska Förlag.
Smrz, Otakar. 2007. "Functional arabic morphology. formal system and implementation." Ph.D. diss., Charles University in Prague.
Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska Akademiens grammatik, 1–4*. Stockholm: NorstedtsOrdbok.
Vossen, Piek, ed. 1999. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks for European Languages*. Dordrecht: Kluwer.

Appendix 1: Creative Commons BY–SA 2.5

Creative Commons Legal Code

Creative Commons
Erkännande-DelaLika 2.5

CREATIVE COMMONS CORPORATION ÄR INTE EN JURIDISK BYRÅ ELLER ADVOKATBYRÅ OCH ERBJUDER INTE JURIDISKA TJÄNSTER. CREATIVE COMMONS GER INGA GARANTIER FÖR INFORMATION SOM ERBJUDS OCH FRISKRIVER SIG HÄRMED FRÅN ALLT ANSVAR FÖR SKADOR SOM KAN UPPSTÅ TILL FÖLJD AV ANVÄNDNING AV INFORMATION.

Licens

ANNAN ANVÄNDNING ÄN SÅDAN SOM MEDGES UNDER DENNA CREATIVE COMMONS PUBLIKA LICENS (”CCPL” eller ”LICENS”) ELLER ANNARS ÄR TILLÅTEN ENLIGT TVINGANDE LAG ÄR FÖRBJUDEN. GENOM ATT NYTTJA DEN RÄTT SOM LICENSTAGAREN FÅR TILL VERKET ENLIGT DENNA LICENS ACCEPTERAR LICENSTAGAREN ATT VARA BUNDEN AV SAMTLIGA DE VILLKOR SOM ANGES NEDAN. OM LICENSTAGAREN INTE ACCEPTERAR SAMTLIGA VILLKOR ANGIVNA I DENNA LICENS HAR LICENSTAGAREN INTE NÅGON RÄTT ATT NYTTJA VERKET.

1. Definitioner

a. **”Verk”** betyder det upphovsrättsligt skyddade verk och/eller den närstående rättighet som erbjuds på de villkor som följer av denna Licens.

b. **”Samlingsverk”** är när flera oförändrade verk samlas till en enhet. Ett verk som utgör ett Samlingsverk kommer inte enligt dessa Licensvillkor att betraktas som ett Bearbetat Verk (enligt vad som anges nedan).

c. **”Bearbetat Verk”** betyder verk som gjorts om i annan form i vilken Verket kan bli omstöpt, omvandlat eller anpassat med undantag för att ett Samlingsverk inte enligt dessa licensvillkor skall betraktas som ett Bearbetat Verk. För det fall Verket är ett musikaliskt verk eller en ljudinspelning skall synkroniseringen med ett filmverk betraktas som ett Bearbetat Verk enligt denna Licens.

d. **”Licensgivare”** betyder den fysiska eller juridiska person som erbjuder Verket under denna Licens.

e. **”Upphovsman”** betyder den fysiska eller juridiska person som skapat Verket.

f. **"Licenstagaren"** betyder den fysiska eller juridiska person som nyttjar sina rättigheter under denna Licens som inte tidigare har brutit mot villkoren i Licensen avseende Verket eller den som har erhållit ett uttryckligt medgivande från Licensgivaren att använda den Licens som erbjuds enligt Dessa Licensvillkor trots tidigare brott mot Licensvillkoren.

g. **"Licenselement"** betyder de attribut som Licensgivaren valt och som ingår i denna Licens: Erkännande (Attribution), IckeKommersiell (Noncommercial), DelaLika (ShareAlike).

2. Inskränkningar. Dessa Licensvillkor skall inte på något sätt minska, begränsa eller annars inskränka några rättigheter som framgår av upphovsrättslagen eller annan tillämplig lag. Upphovsmannen ideella rättigheter påverkas inte av dessa Licensvillkor.

3. Licensupplåtelse. Enligt dessa Licensvillkor erhåller Licenstagaren en global, royalty-fri, icke-exklusiv, evig (för skyddstiden för ensamrätten enligt vad som följer av lag) licens att utnyttja de rättigheter som framgår i det följande.

a. att framställa exemplar av Verket, att infoga Verket i ett eller flera Samlingsverk och att framställa exemplar av sådana Samlingsverk;

b. att skapa och framställa exemplar av Bearbetat Verk;

c. att sprida exemplar eller upptagningar av Verket eller på annat sätt göra det tillgängligt för allmänheten, även såsom infogat i Samlingsverk;

d. att sprida exemplar eller upptagningar av Bearbetat Verk, eller på annat sätt göra det tillgängligt för allmänheten;

e. För det fall Verket är ett musikaliskt verk gäller att:

   i. **Royalties enligt förlagsavtal.** . Licensgivaren avsäger sig den exklusiva rätten att motta, antingen individuellt eller genom en intresseorganisation som företräder upphovsmän och artister (såsom STIM), royalties för nyttjande (enligt punkt 3) av Verket.

   ii. **Mekaniska rättigheter.**Om Verket är en inspelning avsäger sig Licenstagaren rätten att, antingen individuellt eller genom en intresseorganisation, motta royalties för nyttjande (enligt punkt 3) av Verket.

f. **Webcasting-rättigheter**. Om Verket är en inspelning avsäger sig Licenstagaren rätten att, antingen individuellt eller genom en intresseorganisation, motta royalties för nyttjande (enligt punkt 3) av Verket.

Ovanstående rättigheter får utövas i alla nuvarande och framtida media och format. Ovanstående rättigheter inkluderar rätten att utföra sådana ändringar som är tekniskt nödvändiga för att kunna utöva rättigheterna i andra media och format. Inga andra rättigheter än de som uttryckligen anges enligt ovan tillkommer Licenstagaren.

4. Inskränkningar.

a. Licenstagarens tillstånd som anges i punkten 3 ovan gäller endast under villkoren enligt denna Licens samt är förenat med följande inskränkningar:

   - Kopia av, eller Internet-adress (Uniform Resource Identifier) till, denna Licens skall bifogas med varje exemplar
   - Villkor i Licensen får inte ändras
   - Andrahandsupplåtelser av rättigheter till Verket är ej tillåtna
   - Alla hänvisningar till Licensen skall bibehållas
   - Tekniska åtgärder som begränsar rättigheter enligt denna Licens är ej tillåtna

   Ovanstående gäller även Verk som ingår i Samlingsverk, men det krävs inte att Samlingsverket förutom den del som härrör från Verket sprids och licensieras enligt villkoren i denna Licens. Om Licenstagaren skapar ett Samlings- eller Bearbetat Verk, måste Licenstagaren på anmodan från Licensgivaren, så långt det är praktiskt möjligt, ta bort sådan referens som anges i 4c.

b. Licenstagarens tillstånd till Bearbetat Verk som anges i punkten 3 ovan gäller endast under villkoren enligt denna Licens, eller senare version med samma Licenselement, eller Creative Commons Licens från annan jurisdiktion (”iCommons-licens”) innehållande samma Licenselement samt är förenat med följande inskränkningar:

   - Kopia av, eller Internet-adress (Uniform Resource Identifier) till, denna Licens skall bifogas med varje exemplar
   - Villkor i Licensen får inte ändras
   - Andrahandsupplåtelser av rättigheter till Verket är ej tillåtna
   - Alla hänvisningar till Licensen skall bibehållas
   - Tekniska åtgärder som begränsar rättigheter enligt denna Licens är ej tillåtna

   Ovanstående gäller även Verk som ingår i Samlingsverk, men det krävs inte att Samlingsverket förutom den del som härrör från Verket sprids och licensieras enligt villkoren i denna Licens.

c. Om Licenstagaren nyttjar rättigheter (enligt punkt 3) till Verket eller ett Bearbetat Verk eller Samlingsverk måste Licenstagaren tillse att alla hänvisningar till denna licens vidhålls, samt i relation till media eller framförandesätt:

- Upphovsmannen skall omnämnas i skälig omfattning. Detta sker genom att upphovsmannens namn (eller pseudonym), och/eller annan part som utses av Upphovsmannen och/eller Licensgivaren anges för omnämnande i Licensgivarens uppgift om upphovsrättsinnehav eller dylikt.
- Namnet eller titeln på Verket skall anges om uppgivet;
- Om praktiskt möjligt, skall den Internet-adress (Uniform Resource Identifier) som Licensgivaren uppger anges. Detta gäller endast om Internet-adressen refererar till uppgift om upphovsrättinnehav eller licensinformation för Verket.
- För Bearbetat Verk gäller dessutom, att man anger hur Verket används i Bearbetat Verk (till exempel ”fransk översättning av Verket av Upphovsmannen,” eller ”Filmmanus baserat på Verket av Upphovsmannen”). Sådant omnämnande skall införas skäligen. I fall av Bearbetat Verk eller Samlingsverk, skall all erkännande enligt denna punkt genomföras på sådant sätt som är jämförbart i status med annat angivande av upphovsmän.

5. Garantier och friskrivning

UTÖVER VAD SOM UTTRYCKLIGEN FÖRESKRIVS I DENNA LICENS ELLER SOM ANNARS SKRIFTLIGEN ÖVERENSKOMMITS ELLER KRÄVS ENLIGT LAG TILLHANDAHÅLLS VERKET I ”BEFINTLIG SKICK”, UTAN NÅGRA SOM HELST GARANTIER, VARKEN UTTRYCKLIGA ELLER IMPLICITA, UTAN NÅGRA BEGRÄNSNINGAR AVSEENDE GARANTIER AVSEENDE INNEHÅLLET ELLER KORREKTHETEN I VERKET.

6. Ansvarsbegräsning. UTÖVER VAD SOM FÖLJER AV TILLÄMPLIG LAG OCH UTÖVER ERSÄTTNINGSSKYLDIGHET TILL OBEROENDE PART TILL FÖLJD AV BROTT MOT GARANTIERNA I PUNKTEN 5 SKALL LICENSGIVAREN INTE I NÅGOT FALL BLI ERSÄTTNINGSSKYLDIG TILL LICENSTAGAREN FÖR SKADA SOM FÖLJER AV DENNA LICENS ELLER ANVÄNDNING AV VERKET, ÄVEN OM LICENSGIVAREN HAR UPPLYSTS OM MÖJLIGHETEN AV SÅDAN ERSÄTTNINGSSKYLDIGHET.

7. Avtalets upphörande

a. Denna Licens och de rättigheter som är förenade därmed kommer automatiskt att upphöra om Licenstagaren bryter mot något villkor

i denna Licens. De fysiska eller juridiska personer som har erhållit Bearbetat Verk eller Samlingsverk från Licenstagaren under denna Licens kommer emellertid inte att få sin Licens avbruten förutsatt att dessa fysiska eller juridiska personer fortsatt uppfyller villkoren i denna Licens. Punkterna 1, 2, 5, 6, 7 och 8 skall äga fortsatt giltighet efter denna Licens upphörande.

b. Licensgivaren behåller rätten att påbörja eller upphöra spridning av Verket, under förutsättning att en sådan förändring inte innebär att denna Licens dras tillbaka (eller någon annan licens som har erbjudits eller skall erbjudas enligt villkoren som följer av denna Licens) och att denna Licens fortsätter gälla om den inte upphört enligt ovan.

8. Övrigt

a. Varje gång Licenstagaren nyttjar (enligt punkt 3) ett Verk eller ett Samlingsverk, erbjuder Licensgivaren mottagaren av Verket samma Licens till Verket som Licenstagaren har erhållit och som följer av dessa villkor.

b. Varje gång Licenstagaren sprider, offentligen framför eller på annat sätt gör ett Bearbetat Verk tillgängligt för allmänheten, erbjuder Licensgivaren mottagaren av det ursprungliga Verket samma Licens till verket som Licenstagaren har erhållit och som följer av dessa villkor.

c. Om någon del av Licensen skulle befinnas vara ogiltig, otillåten eller overkställbar skall detta inte påverka giltigheten av övriga bestämmelser som skall fortsätta att äga giltighet. Villkor som befinns vara ogiltiga, otillåtna eller overkställbara skall, i den mån så är möjligt, jämkas så att de blir giltiga, tillåtna respektive verkställbara och därvid i så hög utsträckning som möjligt tolkas i enlighet med parternas ursprungliga intentioner.

d. Part skall inte anses ha avstått från att göra villkor gällande eller tillåtit brott mot villkor om detta ej skett skriftligen.

e. Licensen skall utgöra parternas fullständiga reglering av allt som det berör, och alla skriftliga och muntliga åtaganden och utfästelser som föregått Licensen är utan verkan. Ändringar av Avtalet skall ske skriftligen och undertecknas av Licensgivaren och Licenstagaren för att vara gällande.

Creative Commons är inte en part till detta Avtal och Licens och ger inga garantier eller andra utfästelser i samband med Verket. Creative Commons kommer inte att vara ersättningsskyldigt till Licenstagaren eller någon annan part på något sätt för skada som uppkommer i samband med

denna Licens. Oavsett vad som stadgats i de tidigare två (2) meningarna i detta stycke skall Creative Commons ha alla rättigheter och skyldigheter enligt detta Avtal om Creative Commons uttryckligen har angett sig som Licensgivare. Förutom för det begränsade syftet att visa för allmänheten att Verket är licensierat under en Creative Commons-licens har ingen part rätt att utan skriftligt godkännande använda varumärket ”Creative Commons” eller något relaterat kännetecken eller logotyp som tillhör Creative Commons. All tillåten användning skall utföras i enlighet med från tid till annan gällande regelverk för varumärkesanvändning som utfördas av Creative Commons och publiceras på Creative Commons webbplats eller annars kan erhållas på begäran.

Creative Commons kontaktas via http://creativecommons.org/.

Appendix 2: GNU Lesser General Public License 3.0

This version of the GNU Lesser General Public License incorporates the terms and conditions of version 3 of the GNU General Public License, supplemented by the additional permissions listed below.

. Additional Definitions.

As used herein, "this License" refers to version 3 of the GNU Lesser General Public License, and the "GNU GPL" refers to version 3 of the GNU General Public License.

"The Library" refers to a covered work governed by this License, other than an Application or a Combined Work as defined below.

An "Application" is any work that makes use of an interface provided by the Library, but which is not otherwise based on the Library. Defining a subclass of a class defined by the Library is deemed a mode of using an interface provided by the Library.

A "Combined Work" is a work produced by combining or linking an Application with the Library. The particular version of the Library with which the Combined Work was made is also called the "Linked Version".

The "Minimal Corresponding Source" for a Combined Work means the Corresponding Source for the Combined Work, excluding any source code for portions of the Combined Work that, considered in isolation, are based on the Application, and not on the Linked Version.

The "Corresponding Application Code" for a Combined Work means the object code and/or source code for the Application, including any data and utility programs needed for reproducing the Combined Work from the Application, but excluding the System Libraries of the Combined Work.

a. Exception to Section 3 of the GNU GPL.

You may convey a covered work under sections 3 and 4 of this License without being bound by section 3 of the GNU GPL.

b. Conveying Modified Versions.

If you modify a copy of the Library, and, in your modifications, a facility refers to a function or data to be supplied by an Application that uses the facility (other than as an argument passed when the facility is invoked), then you may convey a copy of the modified version:

a) under this License, provided that you make a good faith effort to ensure that, in the event an Application does not supply the function or data, the facility still operates, and performs whatever part of its purpose remains meaningful, or

b) under the GNU GPL, with none of the additional permissions of this License applicable to that copy.

c. Object Code Incorporating Material from Library Header Files.

The object code form of an Application may incorporate material from a header file that is part of the Library. You may convey such object code under terms of your choice, provided that, if the incorporated material is not limited to numerical parameters, data structure layouts and accessors, or small macros, inline functions and templates (ten or fewer lines in length), you do both of the following:

a) Give prominent notice with each copy of the object code that the Library is used in it and that the Library and its use are covered by this License.

b) Accompany the object code with a copy of the GNU GPL and this license document.

d. Combined Works.

You may convey a Combined Work under terms of your choice that, taken together, effectively do not restrict modification of the portions of the Library contained in the Combined Work and reverse engineering for debugging such modifications, if you also do each of the following:

a) Give prominent notice with each copy of the Combined Work that the Library is used in it and that the Library and its use are covered by this License.

b) Accompany the Combined Work with a copy of the GNU GPL and this license document.

c) For a Combined Work that displays copyright notices during execution, include the copyright notice for the Library among these notices, as well as a reference directing the user to the copies of the GNU GPL and this license document.

d) Do one of the following:

0) Convey the Minimal Corresponding Source under the terms of this License, and the Corresponding Application Code in a form suitable for, and under terms that permit, the user to recombine or relink the Application with a modified version of the Linked Version to produce a modified Combined Work, in the manner specified by section 6 of the GNU GPL for conveying Corresponding Source.

1) Use a suitable shared library mechanism for linking with the Library. A suitable mechanism is one that (a) uses at run time a copy of the Library already present on the user's computer system, and (b) will operate properly with a modified version of the Library that is interface-compatible with the Linked Version.

e) Provide Installation Information, but only if you would otherwise be required to provide such information under section 6 of the GNU GPL, and only to the extent that such information is necessary to install and execute a modified version of the Combined Work produced by recombining or relinking the Application with a modified version of the Linked Version. (If you use option 4d0, the Installation Information must accompany the Minimal Corresponding Source and Corresponding Application Code. If you use option 4d1, you must provide the Installation Information in the manner specified by section 6 of the GNU GPL for conveying Corresponding Source.)

e. Combined Libraries.

You may place library facilities that are a work based on the Library side by side in a single library together with other library facilities that are not Applications and are not covered by this License, and convey such a combined library under terms of your choice, if you do both of the following:

a) Accompany the combined library with a copy of the same work based on the Library, uncombined with any other library facilities, conveyed under the terms of this License.

b) Give prominent notice with the combined library that part of it is a work based on the Library, and explaining where to find the accompanying uncombined form of the same work.

f. Revised Versions of the GNU Lesser General Public License.

The Free Software Foundation may publish revised and/or new versions of the GNU Lesser General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Library as you received it specifies that a certain numbered version of the GNU Lesser General Public License "or any later version" applies to it, you have the option of following the terms and conditions either of that published version or of any later version published by the Free Software Foundation. If the Library as you received it does not specify a version number of the GNU Lesser General Public License, you may choose any version of the GNU Lesser General Public License ever published by the Free Software Foundation.

If the Library as you received it specifies that a proxy can decide whether future versions of the GNU Lesser General Public License shall apply, that proxy's public statement of acceptance of any version is permanent authorization for you to choose that version for the Library.

Appendix 3: GNU General Public License 3.0

Preamble

The GNU General Public License is a free, copyleft license for software and other kinds of works.

The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program–to make sure it remains free software for all its users. We, the Free Software Foundation, use the GNU General Public License for most of our software; it applies also to any other work released this way by its authors. You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for them if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs, and that you know you can do these things.

To protect your rights, we need to prevent others from denying you these rights or asking you to surrender the rights. Therefore, you have certain responsibilities if you distribute copies of the software, or if you modify it: responsibilities to respect the freedom of others.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must pass on to the recipients the same freedoms that you received. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

Developers that use the GNU GPL protect your rights with two steps: (1) assert copyright on the software, and (2) offer you this License giving you legal permission to copy, distribute and/or modify it.

For the developers' and authors' protection, the GPL clearly explains that there is no warranty for this free software. For both users' and authors' sake, the GPL requires that modified versions be marked as changed, so that their problems will not be attributed erroneously to authors of previous versions.

Some devices are designed to deny users access to install or run modified versions of the software inside them, although the manufacturer can do so. This is fundamentally incompatible with the aim

of protecting users' freedom to change the software. The systematic pattern of such abuse occurs in the area of products for individuals to use, which is precisely where it is most unacceptable. Therefore, we have designed this version of the GPL to prohibit the practice for those products. If such problems arise substantially in other domains, we stand ready to extend this provision to those domains in future versions of the GPL, as needed to protect the freedom of users.

Finally, every program is threatened constantly by software patents. States should not allow patents to restrict development and use of software on general-purpose computers, but in those that do, we wish to avoid the special danger that patents applied to a free program could make it effectively proprietary. To prevent this, the GPL assures that patents cannot be used to render the program non-free.

The precise terms and conditions for copying, distribution and modification follow.

TERMS AND CONDITIONS

. Definitions.

"This License" refers to version 3 of the GNU General Public License.

"Copyright" also means copyright-like laws that apply to other kinds of works, such as semiconductor masks.

"The Program" refers to any copyrightable work licensed under this License. Each licensee is addressed as "you". "Licensees" and "recipients" may be individuals or organizations.

To "modify" a work means to copy from or adapt all or part of the work in a fashion requiring copyright permission, other than the making of an exact copy. The resulting work is called a "modified version" of the earlier work or a work "based on" the earlier work.

A "covered work" means either the unmodified Program or a work based on the Program.

To "propagate" a work means to do anything with it that, without permission, would make you directly or secondarily liable for infringement under applicable copyright law, except executing it on a computer or modifying a private copy. Propagation includes copying, distribution (with or without modification), making available to the public, and in some countries other activities as well.

To "convey" a work means any kind of propagation that enables other parties to make or receive copies. Mere interaction with a user through a computer network, with no transfer of a copy, is not conveying.

An interactive user interface displays “Appropriate Legal Notices” to the extent that it includes a convenient and prominently visible feature that (1) displays an appropriate copyright notice, and (2) tells the user that there is no warranty for the work (except to the extent that warranties are provided), that licensees may convey the work under this License, and how to view a copy of this License. If the interface presents a list of user commands or options, such as a menu, a prominent item in the list meets this criterion.

a. Source Code.

The “source code” for a work means the preferred form of the work for making modifications to it. “Object code” means any non-source form of a work.

A “Standard Interface” means an interface that either is an official standard defined by a recognized standards body, or, in the case of interfaces specified for a particular programming language, one that is widely used among developers working in that language.

The “System Libraries” of an executable work include anything, other than the work as a whole, that (a) is included in the normal form of packaging a Major Component, but which is not part of that Major Component, and (b) serves only to enable use of the work with that Major Component, or to implement a Standard Interface for which an implementation is available to the public in source code form. A “Major Component”, in this context, means a major essential component (kernel, window system, and so on) of the specific operating system (if any) on which the executable work runs, or a compiler used to produce the work, or an object code interpreter used to run it.

The “Corresponding Source” for a work in object code form means all the source code needed to generate, install, and (for an executable work) run the object code and to modify the work, including scripts to control those activities. However, it does not include the work’s System Libraries, or general-purpose tools or generally available free programs which are used unmodified in performing those activities but which are not part of the work. For example, Corresponding Source includes interface definition files associated with source files for the work, and the source code for shared libraries and dynamically linked subprograms that the work is specifically designed to require, such as by intimate data communication or control flow between those subprograms and other parts of the work.

The Corresponding Source need not include anything that users can regenerate automatically from other parts of the Corresponding

Source.

The Corresponding Source for a work in source code form is that same work.

b. Basic Permissions.

All rights granted under this License are granted for the term of copyright on the Program, and are irrevocable provided the stated conditions are met. This License explicitly affirms your unlimited permission to run the unmodified Program. The output from running a covered work is covered by this License only if the output, given its content, constitutes a covered work. This License acknowledges your rights of fair use or other equivalent, as provided by copyright law.

You may make, run and propagate covered works that you do not convey, without conditions so long as your license otherwise remains in force. You may convey covered works to others for the sole purpose of having them make modifications exclusively for you, or provide you with facilities for running those works, provided that you comply with the terms of this License in conveying all material for which you do not control copyright. Those thus making or running the covered works for you must do so exclusively on your behalf, under your direction and control, on terms that prohibit them from making any copies of your copyrighted material outside their relationship with you.

Conveying under any other circumstances is permitted solely under the conditions stated below. Sublicensing is not allowed; section 10 makes it unnecessary.

c. Protecting Users' Legal Rights From Anti-Circumvention Law.

No covered work shall be deemed part of an effective technological measure under any applicable law fulfilling obligations under article 11 of the WIPO copyright treaty adopted on 20 December 1996, or similar laws prohibiting or restricting circumvention of such measures.

When you convey a covered work, you waive any legal power to forbid circumvention of technological measures to the extent such circumvention is effected by exercising rights under this License with respect to the covered work, and you disclaim any intention to limit operation or modification of the work as a means of enforcing, against the work's users, your or third parties' legal rights to forbid circumvention of technological measures.

d. Conveying Verbatim Copies.

You may convey verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice; keep intact all notices stating that this License and any non-permissive terms added in accord with section 7 apply to the code; keep intact all notices of the absence of any warranty; and give all recipients a copy of this License along with the Program.

You may charge any price or no price for each copy that you convey, and you may offer support or warranty protection for a fee.

e. Conveying Modified Source Versions.

You may convey a work based on the Program, or the modifications to produce it from the Program, in the form of source code under the terms of section 4, provided that you also meet all of these conditions:

a) The work must carry prominent notices stating that you modified it, and giving a relevant date.

b) The work must carry prominent notices stating that it is released under this License and any conditions added under section 7. This requirement modifies the requirement in section 4 to "keep intact all notices".

c) You must license the entire work, as a whole, under this License to anyone who comes into possession of a copy. This License will therefore apply, along with any applicable section 7 additional terms, to the whole of the work, and all its parts, regardless of how they are packaged. This License gives no permission to license the work in any other way, but it does not invalidate such permission if you have separately received it.

d) If the work has interactive user interfaces, each must display Appropriate Legal Notices; however, if the Program has interactive interfaces that do not display Appropriate Legal Notices, your work need not make them do so.

A compilation of a covered work with other separate and independent works, which are not by their nature extensions of the covered work, and which are not combined with it such as to form a larger program, in or on a volume of a storage or distribution medium, is called an "aggregate" if the compilation and its resulting copyright are not used to limit the access or legal rights of the compilation's

users beyond what the individual works permit. Inclusion of a covered work in an aggregate does not cause this License to apply to the other parts of the aggregate.

f. Conveying Non-Source Forms.

You may convey a covered work in object code form under the terms of sections 4 and 5, provided that you also convey the machine-readable Corresponding Source under the terms of this License, in one of these ways:

a) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by the Corresponding Source fixed on a durable physical medium customarily used for software interchange.

b) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by a written offer, valid for at least three years and valid for as long as you offer spare parts or customer support for that product model, to give anyone who possesses the object code either (1) a copy of the Corresponding Source for all the software in the product that is covered by this License, on a durable physical medium customarily used for software interchange, for a price no more than your reasonable cost of physically performing this conveying of source, or (2) access to copy the Corresponding Source from a network server at no charge.

c) Convey individual copies of the object code with a copy of the written offer to provide the Corresponding Source. This alternative is allowed only occasionally and noncommercially, and only if you received the object code with such an offer, in accord with subsection 6b.

d) Convey the object code by offering access from a designated place (gratis or for a charge), and offer equivalent access to the Corresponding Source in the same way through the same place at no further charge. You need not require recipients to copy the Corresponding Source along with the object code. If the place to copy the object code is a network server, the Corresponding Source may be on a different server (operated by you or a third party) that supports equivalent copying facilities, provided you maintain clear directions next to the object code saying where to find the Corresponding Source. Regardless of what server hosts the Corresponding Source, you remain obligated to ensure that it is available for as long as needed to satisfy these requirements.

e) Convey the object code using peer-to-peer transmission, provided you inform other peers where the object code and Corresponding Source of the work are being offered to the general public at no charge under subsection 6d.

A separable portion of the object code, whose source code is excluded from the Corresponding Source as a System Library, need not be included in conveying the object code work.

A "User Product" is either (1) a "consumer product", which means any tangible personal property which is normally used for personal, family, or household purposes, or (2) anything designed or sold for incorporation into a dwelling. In determining whether a product is a consumer product, doubtful cases shall be resolved in favor of coverage. For a particular product received by a particular user, "normally used" refers to a typical or common use of that class of product, regardless of the status of the particular user or of the way in which the particular user actually uses, or expects or is expected to use, the product. A product is a consumer product regardless of whether the product has substantial commercial, industrial or non-consumer uses, unless such uses represent the only significant mode of use of the product.

"Installation Information" for a User Product means any methods, procedures, authorization keys, or other information required to install and execute modified versions of a covered work in that User Product from a modified version of its Corresponding Source. The information must suffice to ensure that the continued functioning of the modified object code is in no case prevented or interfered with solely because modification has been made.

If you convey an object code work under this section in, or with, or specifically for use in, a User Product, and the conveying occurs as part of a transaction in which the right of possession and use of the User Product is transferred to the recipient in perpetuity or for a fixed term (regardless of how the transaction is characterized), the Corresponding Source conveyed under this section must be accompanied by the Installation Information. But this requirement does not apply if neither you nor any third party retains the ability to install modified object code on the User Product (for example, the work has been installed in ROM).

The requirement to provide Installation Information does not include a requirement to continue to provide support service, warranty, or updates for a work that has been modified or installed by the recipient, or for the User Product in which it has been modified or instal-

led. Access to a network may be denied when the modification itself materially and adversely affects the operation of the network or violates the rules and protocols for communication across the network.

Corresponding Source conveyed, and Installation Information provided, in accord with this section must be in a format that is publicly documented (and with an implementation available to the public in source code form), and must require no special password or key for unpacking, reading or copying.

g. Additional Terms.

"Additional permissions" are terms that supplement the terms of this License by making exceptions from one or more of its conditions. Additional permissions that are applicable to the entire Program shall be treated as though they were included in this License, to the extent that they are valid under applicable law. If additional permissions apply only to part of the Program, that part may be used separately under those permissions, but the entire Program remains governed by this License without regard to the additional permissions.

When you convey a copy of a covered work, you may at your option remove any additional permissions from that copy, or from any part of it. (Additional permissions may be written to require their own removal in certain cases when you modify the work.) You may place additional permissions on material, added by you to a covered work, for which you have or can give appropriate copyright permission.

Notwithstanding any other provision of this License, for material you add to a covered work, you may (if authorized by the copyright holders of that material) supplement the terms of this License with terms:

a) Disclaiming warranty or limiting liability differently from the terms of sections 15 and 16 of this License; or

b) Requiring preservation of specified reasonable legal notices or author attributions in that material or in the Appropriate Legal Notices displayed by works containing it; or

c) Prohibiting misrepresentation of the origin of that material, or requiring that modified versions of such material be marked in reasonable ways as different from the original version; or

d) Limiting the use for publicity purposes of names of licensors or authors of the material; or

e) Declining to grant rights under trademark law for use of some trade names, trademarks, or service marks; or

f) Requiring indemnification of licensors and authors of that material by anyone who conveys the material (or modified versions of it) with contractual assumptions of liability to the recipient, for any liability that these contractual assumptions directly impose on those licensors and authors.

All other non-permissive additional terms are considered "further restrictions" within the meaning of section 10. If the Program as you received it, or any part of it, contains a notice stating that it is governed by this License along with a term that is a further restriction, you may remove that term. If a license document contains a further restriction but permits relicensing or conveying under this License, you may add to a covered work material governed by the terms of that license document, provided that the further restriction does not survive such relicensing or conveying.

If you add terms to a covered work in accord with this section, you must place, in the relevant source files, a statement of the additional terms that apply to those files, or a notice indicating where to find the applicable terms.

Additional terms, permissive or non-permissive, may be stated in the form of a separately written license, or stated as exceptions; the above requirements apply either way.

h. Termination.

You may not propagate or modify a covered work except as expressly provided under this License. Any attempt otherwise to propagate or modify it is void, and will automatically terminate your rights under this License (including any patent licenses granted under the third paragraph of section 11).

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright

holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, you do not qualify to receive new licenses for the same material under section 10.

i. Acceptance Not Required for Having Copies.

You are not required to accept this License in order to receive or run a copy of the Program. Ancillary propagation of a covered work occurring solely as a consequence of using peer-to-peer transmission to receive a copy likewise does not require acceptance. However, nothing other than this License grants you permission to propagate or modify any covered work. These actions infringe copyright if you do not accept this License. Therefore, by modifying or propagating a covered work, you indicate your acceptance of this License to do so.

j. Automatic Licensing of Downstream Recipients.

Each time you convey a covered work, the recipient automatically receives a license from the original licensors, to run, modify and propagate that work, subject to this License. You are not responsible for enforcing compliance by third parties with this License.

An "entity transaction" is a transaction transferring control of an organization, or substantially all assets of one, or subdividing an organization, or merging organizations. If propagation of a covered work results from an entity transaction, each party to that transaction who receives a copy of the work also receives whatever licenses to the work the party's predecessor in interest had or could give under the previous paragraph, plus a right to possession of the Corresponding Source of the work from the predecessor in interest, if the predecessor has it or can get it with reasonable efforts.

You may not impose any further restrictions on the exercise of the rights granted or affirmed under this License. For example, you may not impose a license fee, royalty, or other charge for exercise of rights granted under this License, and you may not initiate litigation (including a cross-claim or counterclaim in a lawsuit) alleging that any patent claim is infringed by making, using, selling, offering for sale, or importing the Program or any portion of it.

k. Patents.

A “contributor” is a copyright holder who authorizes use under this License of the Program or a work on which the Program is based. The work thus licensed is called the contributor’s “contributor version”.

A contributor’s “essential patent claims” are all patent claims owned or controlled by the contributor, whether already acquired or hereafter acquired, that would be infringed by some manner, permitted by this License, of making, using, or selling its contributor version, but do not include claims that would be infringed only as a consequence of further modification of the contributor version. For purposes of this definition, “control” includes the right to grant patent sublicenses in a manner consistent with the requirements of this License.

Each contributor grants you a non-exclusive, worldwide, royalty-free patent license under the contributor’s essential patent claims, to make, use, sell, offer for sale, import and otherwise run, modify and propagate the contents of its contributor version.

In the following three paragraphs, a “patent license” is any express agreement or commitment, however denominated, not to enforce a patent (such as an express permission to practice a patent or covenant not to sue for patent infringement). To “grant” such a patent license to a party means to make such an agreement or commitment not to enforce a patent against the party.

If you convey a covered work, knowingly relying on a patent license, and the Corresponding Source of the work is not available for anyone to copy, free of charge and under the terms of this License, through a publicly available network server or other readily accessible means, then you must either (1) cause the Corresponding Source to be so available, or (2) arrange to deprive yourself of the benefit of the patent license for this particular work, or (3) arrange, in a manner consistent with the requirements of this License, to extend the patent license to downstream recipients. “Knowingly relying” means you have actual knowledge that, but for the patent license, your conveying the covered work in a country, or your recipient’s use of the covered work in a country, would infringe one or more identifiable patents in that country that you have reason to believe are valid.

If, pursuant to or in connection with a single transaction or arrangement, you convey, or propagate by procuring conveyance of, a covered work, and grant a patent license to some of the parties receiving the covered work authorizing them to use, propagate, modify or convey a specific copy of the covered work, then the patent license you grant is automatically extended to all recipients of the covered work and works based on it.

A patent license is "discriminatory" if it does not include within the scope of its coverage, prohibits the exercise of, or is conditioned on the non-exercise of one or more of the rights that are specifically granted under this License. You may not convey a covered work if you are a party to an arrangement with a third party that is in the business of distributing software, under which you make payment to the third party based on the extent of your activity of conveying the work, and under which the third party grants, to any of the parties who would receive the covered work from you, a discriminatory patent license (a) in connection with copies of the covered work conveyed by you (or copies made from those copies), or (b) primarily for and in connection with specific products or compilations that contain the covered work, unless you entered into that arrangement, or that patent license was granted, prior to 28 March 2007.

Nothing in this License shall be construed as excluding or limiting any implied license or other defenses to infringement that may otherwise be available to you under applicable patent law.

l. No Surrender of Others' Freedom.

If conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot convey a covered work so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not convey it at all. For example, if you agree to terms that obligate you to collect a royalty for further conveying from those to whom you convey the Program, the only way you could satisfy both those terms and this License would be to refrain entirely from conveying the Program.

m. Use with the GNU Affero General Public License.

Notwithstanding any other provision of this License, you have permission to link or combine any covered work with a work licensed under version 3 of the GNU Affero General Public License into a single combined work, and to convey the resulting work. The terms of this License will continue to apply to the part which is the covered work, but the special requirements of the GNU Affero General Public License, section 13, concerning interaction through a network will apply to the combination as such.

n. Revised Versions of this License.

The Free Software Foundation may publish revised and/or new versions of the GNU General Public License from time to time. Such

new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies that a certain numbered version of the GNU General Public License "or any later version" applies to it, you have the option of following the terms and conditions either of that numbered version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of the GNU General Public License, you may choose any version ever published by the Free Software Foundation.

If the Program specifies that a proxy can decide which future versions of the GNU General Public License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Program.

Later license versions may give you additional or different permissions. However, no additional obligations are imposed on any author or copyright holder as a result of your choosing to follow a later version.

o. Disclaimer of Warranty.

THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

p. Limitation of Liability.

IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT

LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

q. Interpretation of Sections 15 and 16.

If the disclaimer of warranty and limitation of liability provided above cannot be given local legal effect according to their terms, reviewing courts shall apply local law that most closely approximates an absolute waiver of all civil liability in connection with the Program, unless a warranty or assumption of liability accompanies a copy of the Program in return for a fee.

END OF TERMS AND CONDITIONS

How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively state the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

```
<one line to give the program's name and a brief idea of what it does.>

Copyright (C) <textyear>  <name of author>

This program is free software: you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation, either version 3 of the License, or
(at your option) any later version.

This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
GNU General Public License for more details.

You should have received a copy of the GNU General Public License
along with this program.  If not, see <http://www.gnu.org/licenses/>.
```

Also add information on how to contact you by electronic and paper mail.

If the program does terminal interaction, make it output a short notice like this when it starts in an interactive mode:

```
<program>  Copyright (C) <year>  <name of author>

This program comes with ABSOLUTELY NO WARRANTY; for details type ‘show w’.
This is free software, and you are welcome to redistribute it
under certain conditions; type ‘show c’ for details.
```

The hypothetical commands `show w` and `show c` should show the appropriate parts of the General Public License. Of course, your program's commands might be different; for a GUI interface, you would use an "about box".

You should also get your employer (if you work as a programmer) or school, if any, to sign a "copyright disclaimer" for the program, if necessary. For more information on this, and how to apply and follow the GNU GPL, see `http://www.gnu.org/licenses/`.

The GNU General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Lesser General Public License instead of this License. But first, please read `http://www.gnu.org/philosophy/why-not-lgpl.html`.

Appendix 4: Paradigmidentifierarnas uppbyggnad

```
Böjningangivelser:
ordklass_böjning(_exempelord)

enordingar:

substantiv (och substantiviska förkortningar; nna):
nn_{0-7|o|v|i|d|r|n}{n|u|p|v}(_exempelord)
0 = ingen plural; 1-7 = deklination enl. SAG;
o = oregelbunden; v = vacklande böjning;
i = oböjlig; d = bestämd form
r = best. plural -na; n = best. plural -en
n = neutrum; u = utrum; p = plural; v = vacklande genus;
d = bestämd form pl


egennamn (och egennamnsförkortningar; pma):
pm_{n|u|v|f|m|h|p|w}{l|p|o|e|w|a|t}(x)(_exempelord)
n|u|v|f|m|h|p|w neutrum/utrum/vacklande/femininum/maskulinum/
                mänskligt/plural/vacklande neutrum-plural

(Namntaxonomin är baserad på den som Dimitrios Kokkinakis
 har utarbetat)
l|p|o|e|w|a|t location/person/organization/event/work and art/
              artefacts, products/taxonomy
x beror på vilken namnkategori det handlar om:

för l: x = a|g|p|f|s
       astronomical/geographical/political/facilities/streets
för p: x = h|m|a|c
       human/mythological/animals/collectives
för o: x = f|s|c|p|m|e|a|g
       financial/sports/cultural/political/media/educational/
       air industry/corporations, governmental
för e: x = h|n|c|s|r
       historical/natural/cultural/sports/religious
för w: x = b|m|a|p|n|c
       books/tv, radio/art/projects/newspapers, etc/operas, plays
för a: x = m|e|c|w|g|f|p|a
       medical/food/computer/water transport/ground transport/
       air transport/prizes/general artefacts
för t: x = b|z|a|m
       botany/zoology/anatomy/medicine


verb:
vb_{0-4|o|v|i}{a|s|m|d|k}(_exempelord)
1-4 är konjugation enl. SAG
0 = går inte att ange konjugation
```

```
o = halvsvaga och oregelbundna verb i SAG
v = vacklande böjning
i = oböjligt
s = deponens (eller annars obligatorisk s-form)
a = icke-deponens, bildar perfektparticip
m = icke-deponens, bildar inte perfektparticip
d = bara enstaka former finns
k = konjunktiv


adjektiv:
av_{1|2|0|o|i|v}({d|n|k|s})(_exempelord)
1-2 = deklination enl. SAG
0 = ingen syntetisk komparation
i = oböjligt
o = oregelbundet
v = vacklande böjning
d = (endast) bestämd form
n = neutrum
k = komparativ
s = superlativ


adjektiv, förkortning:
ava_i_kungl

adverb tar komparationsdeklination
ab_{1|2|i}({k|s})_exempelord
k = komparativ
s = superlativ

annars oböjliga (i) eller oregelbundna/idiosynkratiska (o)


pronomen:
pn_o_ord
("o" kunde eventuellt utvecklas till olika typer av pronomen)


artiklar:
al_o_den
al_o_en


räkneord:
nl_{g|i|n}_exempelord
g = grundtal
i = oböjligt räkneord
n = normalt räkneord med både grundtal och ordningstal
    i paradigmet
```

```
oböjliga klasser:

ordklass_i_exempelord
(t.ex. "pr_i_i", "ab_i_inte", "kn_i_och", "sn_i_om", etc.)



flerordingar:

substantiv:
nnm_{0-7|v|o|i|d|g|s}{n|u|p}{#|a|c}(_exempelord)
0 = ingen plural; 1-7 = deklination enl. SAG ;
o = oregelbunden; v = vacklande böjning;
i = oböjlig; d = bestämd form; s = har genitivattribut
n = neutrum; u = utrum; p = plural; v = vacklande genus;
"#" anger vilket ord i ordningen i uppslagsformen
    som tar böjningen
    ("0" = sista ordet)
a = substantivfras med kongruerande adjektivattribut
c = koordination

egennamn:
pmm_{n|u|v|f|m|h|p|w}{n|u|v|f|m|h|p}
    {#|a|c}{l|p|o|e|w|a|t}(x (y))(_exempelord)
n|u|v|f|m|h|p|w neutrum/utrum/vacklande/femininum/maskulinum/
                mänskligt/plural/vacklande neutrum-plural

# = vilket ord i ordningen i uppslagsformen som (som alternativ
    till sista ordet ibland) tar (genitiv)böjningen
(0 = bara sista ordet; i = oböjligt)
a = substantivfras med kongruerande adjektivattribut
c = koordination

(Namntaxonomin är baserad på den som Dimitrios Kokkinakis
 har utarbetat)
l|p|o|e|w|a|t location/person/organization/event/work and art/
             artefacts, products/taxonomy
x beror på vilken namnkategori det handlar om:

för l: x = a|g|p|f|s
       astronomical/geographical/political/facilities/streets
för p: x = h|m|a|c
       human/mythological/animals/collectives
för o: x = f|s|c|p|m|e|a|g
       financial/sports/cultural/political/media/educational/
       air industry/corporations, governmental
för e: x = h|n|c|s|r
       historical/natural/cultural/sports/religious
för w: x = b|m|a|p|n|c
       books/tv, radio/art/projects/newspapers, etc/operas, plays
```

```
för a: x = m|e|c|w|g|f|p|a
       medical/food/computer/water transport/ground transport/
       air transport/prizes/general artefacts
för t: x = b|z|a|m
       botany/zoology/anatomy/medicine


adjektiv:
avm_{1|2|0|o|i|v}{p|o|a|x}#_typadjektiv
1-2 = deklination enl. SAG
0 = ingen syntetisk komparation
i = oböjligt
o = oregelbundet
v = vacklande böjning
p = endast predikativt
o = endast opersonligt predikativt
a = endast attributivt
x = både predikativt och attributivt
# talar om var i uttrycket (vilken position) det böjda ordet står
  ("0" betyder sist eller att frasen inte böjs)
typadjektivet visar böjningen (samma som enordingarna)


verb:
vbm_{0-4|o|v}{a|s|m|d}"komponenter"{#}(_typverb)
1-4 är konjugation enl. SAG
0 = går inte att ange konjugation
o = halvsvaga och oregelbundna verb i SAG
v = vacklande böjning

s = deponens (eller annars obligatorisk s-form)
a = icke-deponens, bildar perfektparticip
m = icke-deponens, bildar inte perfektparticip
d = bara enstaka former finns

"komponenter" är en eller flera koder för olika komponenter i
flerordsverb, i samma ordning som i det beskrivna verbet
antingen bara:
i = "idiom"

eller en eller flera av:
(partiklar:)
p = (lätt adverbiell) partikel
l = substantiv, etc. (tung partikel)
d = tvåordsfras (t.ex. prepositionsfras)
zd = tvåordsfras inledd av reflexivt possessivpronomen
t = treordsfras
zt = treordsfras inledd av reflexivt possessivpronomen
q = fyrordsfras
zq = fyrordsfras inledd av reflexivt possessivpronomen
```

```
s = reflexivt objektspronomen
v = valensledspreposition (i SAL)

# anger vilket ord i ordningen i uppslagsformen som tar
  böjningen ("c" betyder koordination [tillsammans med "i"])

typverbet visar böjningen
(samma som enordingarna, men eventuellt med annan valens)

"bryta ut" -> vbm_4ap1_bryta_upp
"brås på" -> vbm_3sv1
"bry sig om" -> vbm_3aq1


konjunktioner/subjunktioner:
knm_i_vare_sig / snm_i_efter_det_att
= icke-uppbrytbar
knm_x1_ju_ju
= med inskjutet material i mellanrum 1


pronomen:
pnm_{i|o|x#}_exempel
i = oböjlig icke-uppbrytbar
o = böjs på något sätt (exemplet ger böjningen)
x# = med möjlighet att skjuta in material i mellanrum #


adverb:
abm_{i|x#}_exempel
i = oböjlig (som sig bör)
x# = med möjlighet att skjuta in material i mellanrum #


prepositioner
ppm_i_a_la
= icke-uppbrytbar
ppm_x#_exempel
= med NP:n i mellanrum #


satser:
ssm_{i|d}#(_exempel)
i = oböjlig
d = böjs (med restriktioner)
# = verbets position i uppslagsformen

ssm_d2_saken_är_biff
ssm_d2_svinhugg_går_igen
ssm_i1_märk_väl
```

```
xxf = ett textord av ordklassen xx som ingår i ett
      flerordsuttryck som också finns i lexikonet.
      Ska inte ge upphov till en ingång i fullformslexikonet.
```