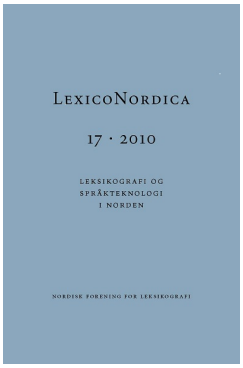
LexicoNordica

Titel: KTHs morfologiska och lexikografiska verktyg och resurser

Forfatter: Viggo Kann

Kilde: LexicoNordica 17, 2010, s.99-117

URL: http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



Søgbarhed
Artiklerne i de ældre LexicoNordica (1-16) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

KTHs morfologiska och lexikografiska verktyg och resurser

Viggo Kann

During the last 15 years the human language technology group at KTH has developed tools and resources that may be of interest to the lexicographical community. Several tools have been developed as part of the group's research on Swedish authoring tools: spelling error detection and correction, grammar checking, part-of-speech tagging, lemmatization, compound splitting, and an interactive learning environment called Grim. Most of the tools are open source and may be downloaded from <www.csc.kth.se/theory/humanlang>. We have also made several dictionaries available on the web: the Lexin series of dictionaries for 15 languages, the Scandinavian Dictionary, the Tvärslå dictionary collection, the Swedish Hyphenation Dictionary and the two crowdsourced resources The People's Dictionary of Synonyms and The People's English-Swedish Dictionary.

1. Språkteknologigruppen på KTH

Språkteknologigruppen är en tvärvetenskaplig forskargrupp inom avdelningarna för teoretisk datalogi och människa-datorinteraktion som hör till skolan för datavetenskap och kommunikation på KTH. Jag har lett gruppen från dess start för 15 år sedan. Vi arbetar med utbildning, forskning och utveckling i språkteknologi. Forskningen har finansierats av bland annat KTH, HSFR (Humanistisk-samhällsvetenskapliga forskningsrådet), Nutek, Vinnova, Vetenskapsrådet, Nordiska ministerrådet, Språkrådet och .SE-stiftelsen. Fem doktorander har doktorerat och disputerat helt inom gruppen. Vi arbetar i nära samarbete med språkteknologiforskarna på institutionen för data- och systemvetenskap vid Stockholms universitet och med talteknologigruppen vid KTH.

Våra huvudområden inom språkteknologi är svensk språkgranskning, informationssökning och informationsextraktion samt ordböcker. Vi vill att det vi arbetar med ska komma till användning och vara nyttigt för både allmänheten och andra forskare. Gruppens filosofi kan sammanfattas i en mening: Vi utvecklar effektiva och resurssnåla metoder för språkteknologiska system, i synnerhet för svensk text men också språkoberoende metoder, där resultaten, både algoritmer, program och resurser, görs fritt tillgängliga i möjligaste mån. Gruppens medlemmar och verksamhet inom forskning och utbildning beskrivs på <www.csc.kth.se/theory/humanlang> där också program och resurser finns för nedladdning.

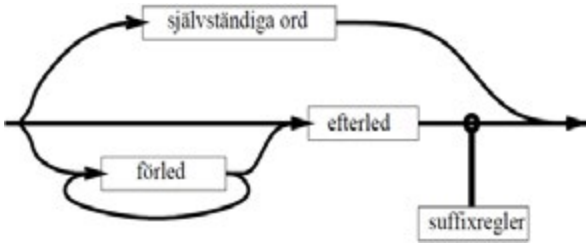
I denna artikel beskrivs och diskuteras verktyg och resurser med lexikografisk relevans som gruppen utvecklat under dessa femton år.

2. Stava – svensk stavningskontroll

Det första språkteknologiska problem vi studerade var svensk stavningskontroll. Vi konstruerade stavningskontrollprogrammet Stava, som likt de flesta sådana program är kontextlöst, det vill säga det kontrollerar varje ord separat, utan att ta hänsyn till vilka ord som står före och efter i texten. Detta gör att Stava aldrig kan upptäcka en felstavning som råkar sammanfalla med ett annat ord, till exempel om *spion* felstavas som *pion*. Eftersom vi vill bygga effektiva system som kan klara av realistiska indata har vi valt att använda SAOL (1986) som grundordlista. Ordlistan är bearbetad och lagrad så att Stava ska klara av att känna igen alla böjningsformer och sammansättningar på ett mycket effektivt sätt.

Ordlistan är uppdelad i tre delar: självständiga ord, förled och efterled.

1. Ordlistan med **självständiga ord** består av de ord som inte kan ingå i sammansättningar, till exempel *ömsom*, *eller* (1400 ord).
2. I **förledsordlistan** ingår alla förled i sammansättningar, såsom *korv*, *medie*, *packnings* (23 000 ord).
3. **Efterledsordlistan** består av alla ord som kan vara efterled i sammansättningar, till exempel *kunskap*, *stort*, *låta*, *medium*. Orden förekommer i efterledsordlistan i de former som står i SAOL. För substantiv är det till exempel grundform, bestämd form och pluralform (100 000 ord).



Figur 1: Schema för hur ord kan bildas som hopslagningar av ord

Varje ord ur ordlistan med självständiga ord och ur efterledsordlistan är tillåtna ord. Sammansatta ord kan bildas med hopslagning (konkatenering) av ett eller flera ord ur förledsordlistan och ett ord ur efterledsordlistan, se figur 1. Exempel med ordleden ovan: *korv-packnings-kunskap*. SAOL är grunden till vilka ord som hamnat i vilken ordlista, men vi har gjort vissa avvikelser. Till exempel har alla substantiv lagts i efterledsordlistan även om de inte förekommer som efterled i några sammansättningar i SAOL. Huvudsaken för Stava är att det är teoretiskt möjligt för ordet att bilda ett efterled.

Böjningsformer som inte förekommer i efterledsordlistan genereras med suffixregler, omkring 1500 stycken. Exempel på en suffixregel:

-ornas ← -a, -an, -or

Denna regel säger att om ett ord förekommer med suffixen *-a*, *-an* och *-or* så ska även suffixet *-ornas* godkänns. Exempelvis finns *docka*, *dockan* och *dockor* i efterledsordlistan, varför *dockornas* godkänns. Det finns 1800 substantiv som passar in på denna regel.

Varje ordlista lagras som ett Bloomfilter (Bloom 1970) vilket är ett extremt kompakt och effektivt lagringssätt där varje ord representeras av ett antal till synes slumpmässigt utplacerade ettor i en binär datastruktur. Med denna lagring går det mycket snabbt att avgöra om ett ord finns med i ordlistan. Ytterst sällan, med en sannolikhet som kan väljas i förväg, gör algoritmen fel och godkänner ett ord som inte finns i ordlistan. Detta sker så sällan att det inte betyder något i praktiken, men det gör att det inte går att utvinna ursprungsordlistan, något ordlisteleverantörer i allmänhet kräver. Enda sättet att skydda ursprungsordlistan helt är faktiskt att stavningsalgoritmen svarar fel ibland (Kann, Domeij, Hollman och Tillenius 2001). Annars går det nämligen att utvinna ordlistan genom att ett program testar alla möjliga bokstavskombinationer mot stavningsalgoritmen.

Att dela upp SAOL i de tre ordlistorna och att skapa suffixreglerna var ett stort arbete. Sammansättningsgränserna i SAOL var inte uppmärkta på ett entydigt sätt och på många ställen var kodningen inkonsekvent. Under arbetet fann vi över 500 fel i SAOL, som alltså hade undgått mänskliga korrekturläsare. Vi skickade felen till SAOL-redaktionen men fick tyvärr ingen respons. I senare upplagor verkar dock felen vara åtgärdade.

Stavas kodade ordlistor och källkod finns fritt tillgängliga för nedladdning (STAVA). På samma webbsida finns också en webbversion av Stava som även kan stavningskontrollera webbsidor.

3. Utvidgningar av Stava

Stava har genom årens lopp vidareutvecklats till att bli ett allt mer komplett morfologiskt verktyg. Här beskrivs de viktigaste tilläggen.

3.1. Rättstavning

Första utvidgningen som gjordes var generering av rangordnade rättelseförslag. Rättelseförslag till ett felstavat ord tas fram genom att Stava går igenom alla tänkbara ord som skulle ha kunnat ge upphov till felstavningen och kontrollerar om orden existerar. Damereau (1964) har visat att 80 procent av alla mänskliga felstavningar beror på antingen omkastning av två intilliggande bokstäver eller insättning, borttagning eller utbyte av en bokstav. Om Stava inte hittar något korrekt ord med en tillämpning av Damereaus regler så provas två tillämpningar.

De genererade förslagen rangordnas sedan i trolighetsordning med hjälp av ett felstavningsavstånd och ordfrekvenser. Felstavningsavståndet visar exempelvis att dubbelskrivning av en konsonant eller byte av en bokstav mot en intilliggande på tangentbordet ligger närmare det rättstavade ordet än andra fel. Att ordfrekvenser förbättrar rangordningen beror på att det är troligare att det är ett vanligt ord som råkat bli felstavat än ett ovanligt. På detta sätt genererar Stava rättelseförslag som är mycket bra. En undersökning har visat att 60 % av de felstavade orden i en text fick korrekt förstahandsrättelseförslag (Kann, Domeij, Hollman och Tillenius 2001). Vid en jämförelse mellan Stava, Microsoft Words rättstavningsmodul och Unixverktyget Ispell hade Stava klart bäst rättstavning (Bigert 2005).[1]

1 Korrekt rättelseförslag föreslogs i 97 % av fallen av Stava, 93 % av Ispell och 89 % av MS Word. Den korrekta rättelsen kom som för-

3.2. Ordklasstaggning och lemmatisering

Det är svårt att få tag i stora lexikon där orden är märkta med både ordklass och böjningsform. Det visade sig att vi med Stavas suffixregler kunde få ett sådant lexikon med minimalt arbete. Det enda som krävdes var att vi märkte varje suffixregel med vilken tagg den motsvarar. Exempelvis märktes exempelregeln i avsnitt 2 med *nn.utr.plu.def.gen*, vilket säger att alla 1800 ord som genereras med denna regel är substantiv i utrum, plural, bestämd form, genitiv.

På detta sätt blir Taggstava en ordklasstaggare för alla regelbundet böjda svenska ord, både enkla och sammansatta (TAGGSTAVA). Det är bara omkring 3 000 av de 100 000 orden i efterledsordlistan som inte blir taggade med denna metod, och dessa skulle det enkelt gå att ta hand om för hand. Taggstava tittar liksom Stava bara på enskilda ord utan kontext. Om man vill bestämma vilken av de tänkbara taggarna för ett ord som ordet har i sitt sammanhang så ska man använda en *disambiguerande ordklasstaggare*, såsom Granskatagger i avsnitt 4.

Med suffixreglernas hjälp kan även lemmaformen enkelt fås fram för alla ord som kan taggas, eftersom lemmaformens ändelse står först i regelns högerled. Lemmatisering kommer alltså på köpet.

3.3. Sammansättningsanalys

Med hjälp av förleds- och efterledsordlistorna kan Stava, som vi beskrivit i avsnitt 2, ta fram vilka ordled ett sammansatt ord består av. Många sammansatta ord får flera sammansättningsanalyser av Stava, och det vore värdefullt om ett system skulle kunna disambiguera bland dessa.

---

sta förslag i 88 % av fallen för Stava, 67 % för Ispell och 60 % för MS Word.

Vi har undersökt åtta metoder för disambiguering av flertydiga sammansättningar och kombinationer av dessa metoder (Sjöbergh och Kann 2006). Den bästa metoden var en statistisk kombinationsmetod som bygger på följande tre enskilda metoder:

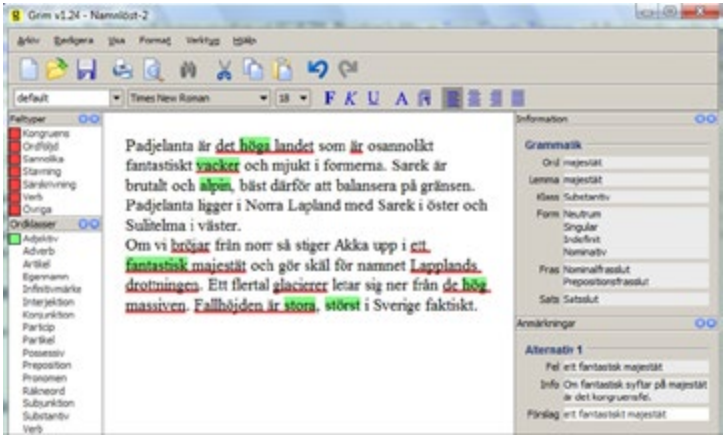
1. Välj sammansättningen med lägst antal ordled (*mun-vinklarna* bättre än *mun-vin-klarna* eftersom det bara har två ordled).
2. Välj sammansättningen med vanligast ordled (*upp-rättar* bättre än *upprätt-ar* eftersom *upp* och *rättar* är vanligare än *upprätt* och *ar*).
3. Välj sammansättningen med vanligaste kombinationen av ordledsordklasser (*upp-rättar* bättre än *upprätt-ar* eftersom preposition-verb är vanligare i sammansättningar än adverb-substantiv).

Denna kombinationsmetod delar upp 98 % av sammansatta ord korrekt. Sett över alla ord i den analyserade texten så gör sammansättningsuppdelaren bara fel på 0,1 %. Källkoden (skriven i programspråket C++) till sammansättningsuppdelaren är fritt tillgänglig.

4. Granska – svensk grammatikkontroll

Efter stavningskontrollprojektet angrep vår grupp grammatikgranskningsproblemet, något som är mycket svårt. Det finska företaget Lingsoft, som har gjort grammatikkontrollen i Microsoft Word, är den ledande producenten av svensk grammatikkontroll, men den är långt ifrån heltäckande, till exempel detekteras inte särskrivningar.

Vårt eget system Granska består till att börja med av en disambiguerande ordklasstaggare med lemmatiserare och ordböjare. Den kallas Granskatagger och finns fritt tillgänglig tillsammans med ett fritt lexikon.



Figur 2: Grim. Användaren, som just valt att alla adjektiv ska markeras, pekar på ordet *majestät* på rad 7 och får då upp analysen i fältet till höger.

Granska implementerar ett ganska avancerat grammatikregelspråk i vilket vi skrivit regler för bland annat särskrivningsdetektion. Det visade sig kunna användas också för att skriva en så kallad grund parser som analyserar och delar upp meningar i satser (Knutsson, Bigert och Kann 2003). Granska koncentrerar sig på svåra grammatiska fel och innehåller inte metoder för att hitta enklare fel såsom upprepade ord eller felaktiga skiljetecken, som till exempel Lingsofts grammatikkontroll hittar.

Vi har samlat flera av våra verktyg i ett gemensamt gränssnitt som kallas Grim (GRIM). Resultatet är en interaktiv lärmiljö som består av en ordbehandlare med stavningskontroll, regelbaserad och probabilistisk grammatikkontroll, presentation av ordklasser, sökning i lexikon med mera, se figur 2. Grim är avsett att användas i undervisning i svenska i olika stadier.

5. Nätordböcker

Gruppen har lagt upp en stort antal ordböcker på nätet. Några av dessa beskrivs här.

5.1. Lexin

Redan 1994 fick jag i uppdrag att lägga upp Skolverkets Lexin-lexikon på webben. I början av 1995 låg svensk-engelska och svensk-finska Lexin sökbara på webben, som de första svenska lexikonen på Internet. Lexins svenska ordbas består av 30 000 ord som valts för att utgöra ett lämpligt ordförråd för den som flyttar till Sverige. Den svenska ordbasen är numer översatt till 15 språk: albanska, arabiska, bosniska, engelska, finska, grekiska, kroatiska, nordkurdiska, persiska, ryska, serbiska, somaliska, spanska, sydkurdiska och turkiska. Alla dessa finns sökbara på webben (LEXIN). Stavas rättstavningsalgoritm finns inlagd i Lexin, varför felstavade sökningar på både källspråk och målspråk rättas automatiskt om de är entydiga. Om det finns flera möjliga rättelser får användaren möjlighet att välja mellan dessa. Dessutom har Lexins ordbas översatts till norska, danska och isländska och finns på webben i alla fyra länder.

Antalet uppslagningar i svenska Lexin har ökat år från år och är nu omkring 20 miljoner i månaden. Mängder av användarstatistik kan därmed samlas in. Denna statistik kan användas på flera sätt. När Skolverket skulle sätta ihop ett engelsk-svenskt lexikon tog vi fram en lista på de 40 000 vanligaste orden som användare slagit upp men som saknades i Lexin. Dessa ord översattes och blev därmed ett utmärkt komplement till Lexin, speciellt avpassat till det som användare brukar vilja slå upp. Statistiken kan också, tillsammans med användarenkäter, nyttjas för att se hur ordböcker används (Hult 2008).

Ägarskapet till Lexin övergick från Skolverket först till Myndigheten för skolutveckling och därefter till Språkrådet.

5.2. Skandinavisk ordbok

Skandinavisk ordbok utvecklades 1994 av Nordiska språksekretariatet inom Nordiska ministerrådet och består av 10 000 lemman som skiljer mellan de skandinaviska språken svenska, danska och norska. Vi gjorde en webbversion av detta lexikon för över tio år sedan (SKANDORD). Ordboken innehåller bara orden, inga definitioner eller annan information.

5.3. Tvärslå

Det nordiska samarbetsprojektet Nordisk nätordbok finansierades av Nordiska ministerrådet 2005–2007. I projektet skapades ett bibliotek av flerspråkiga nätordböcker mellan minst två av de nordiska språken och engelska. Dels skapades några nya flerspråkiga ordböcker med automatiska eller halvautomatiska metoder, dels samlades existerande ordböcker ihop: svensk-engelska Lexin, engelsk-svenska Lexin, svensk-finska Lexin och Skandinavisk ordbok var med bland dessa ordböcker. Ordböckerna har lagrats i ett för projektet speciellt framtaget XML-format, så att de skulle kunna användas på ett enkelt och enhetligt sätt i bland annat forskningssammanhang. Inom projektet tog vi också fram söktjänsten Tvärslå, i vilken man kan slå upp i alla ingående ordböcker samtidigt. Antingen kan man slå upp på alla språk samtidigt eller på ett specificerat språk (TVÄRSLÅ).

5.4. Avstavningslexikon

Boken Svenskt avstavningslexikon (Klingspor 1985) är utgången från förlaget, och Språkrådet som har rättigheterna till materialet

har beslutat att göra lexikonet tillgängligt på webben. Det handlar om 33 000 lemman och ungefär 100 000 avstavade ordformer. Materialet har skannats in, vilket gör att många fel införs. Med ekonomiskt stöd av Ebba Danelius stiftelse har vi automatgranskat och rättat det inskannade materialet från boken och tagit fram en rudimentär söktjänst som kommer att utvecklas vidare (AVSTAV).

6. Utveckling av fria nätordböcker

Ett problem inom språkteknologisk forskning och utveckling är att det inte finns så många tillgängliga språkteknologiska resurser, såsom ordböcker, korpusar och trädbanker. En anledning är att det kräver mycket manuellt arbete att sätta ihop en stor resurs, och att den som gjort det vill ha den för sig själv eller ta betalt för resursen. Detta gäller ofta även resurser som tagits fram i projekt finansierade med statliga medel. Även för resurser som är tillgängliga gratis gäller ofta att det finns begränsningar av hur de får användas och ändras.

Inom programutvecklingsvärlden finns det en stark rörelse för fri programvara. Den härstammar ur hackerkulturen, men allt fler företag och individer har insett att fri programvara är något som alla har glädje av och som kan snabba på utvecklingen av nya och existerande program. Kända exempel på fri programvara är GNU-projektets verktyg och operativsystemet Linux.

På samma sätt är fria språkresurser något som skulle gynna utvecklingen av nya språkteknologiska system. En fri språkresurs skulle kunna definieras på följande sätt, med en lätt modifiering av GNU-projektets grundare Richard Stallmans definition av fri programvara (Stallman 1996):

1. Frihet att använda resursen i en tillämpning.
2. Frihet att studera resursen och modifiera den.

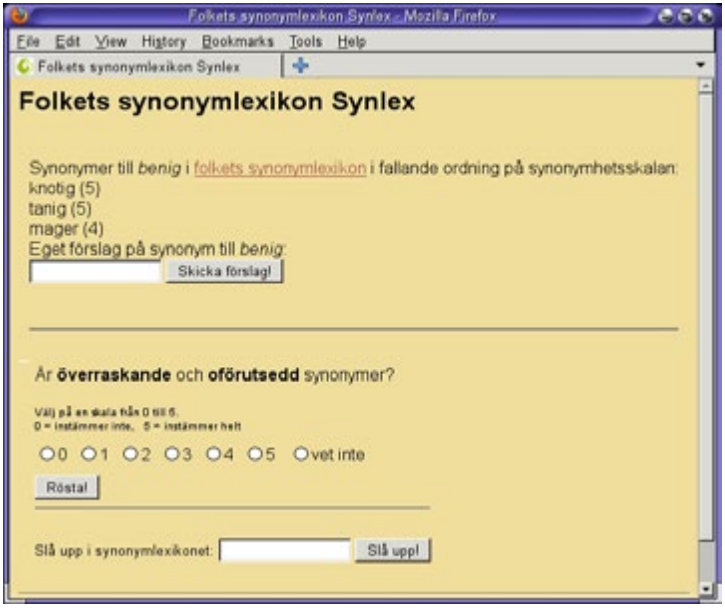
3. Frihet att ta en kopia av resursen.
4. Frihet att förbättra resursen och låta andra få kopior av den förbättrade resursen.

Är en ordbok som man kan slå upp i på nätet en fri resurs? Nej, enligt ovanstående definition är det inte nödvändigtvis så. För att den ska vara en fri resurs måste det också gå att ladda ner den, studera den och ändra den. Hur kan man då lösa upphovsrättsproblemet och skapa en fri resurs utan att några få personer behöver lägga ner enorma mängder ideellt arbete? Jo, det finns språkteknologiska metoder som automatiskt kan skapa stora resurser som visserligen till delar är skräp, men där mycket av innehållet är korrekt. Om det omfattande granskningsarbetet av en sådan resurs kan fördelas på många människor så behöver varje persons bidrag inte vara så stort. Detta arbetssätt att utnyttja folket till att gemensamt konstruera något kallas med ett nytt engelskt ord för *crowdsourcing*.

6.1. Folkets synonymlexikon

Vi har använt den ovan nämnda metoden för att skapa ett svenskt synonymlexikon (Kann och Rosell 2005). Vi började med att automatiskt konstruera en massa ordpar som skulle kunna vara synonymer. Detta gjorde vi genom att med hjälp av elektroniska ordböcker översätta svenska ord till engelska och sedan tillbaka till svenska igen. Därigenom hittas många äkta synonymer men också mängder av falska synonymer på grund av homonymi. Dessa ordpar rensade vi sedan med hjälp av Random indexing, en statistisk metod som grundar sig på i vilka kontexter orden förekommer; ord som förekommer i liknande kontext är förmodligen lika varandra, och ord som är lika varandra skulle kunna vara synonymer.

Därefter lät vi tiotusentals människor bidra genom att kontrollera synonymiteten hos ordpar. Till detta fick vi tillstånd att an-



Figur 3: Användargränssnitt för Folkets synonymlexikon

vända Lexins webbplats under ett par månader. Under denna tid fick alla som gjorde en uppslagning i Lexin också upp en fråga om ett ordpar, se nedre delen av figur 3 för ett exempel. Användarnas ordparsbedömningar analyserades och de par som fick goda omdömen lades in i synonymlexikonet.

På kort tid skapades härigenom ett svenskt synonymlexikon (SYNLEX). Därefter togs frågorna bort från Lexins webbplats, men en länk till Folkets synonymlexikon behölls. Den som vill slå upp i lexikonet behöver göra minst en ordparsbedömning först. Det går också bra att föreslå egna synonymer som sedan bedöms av andra användare. På detta sätt växer synonymlexikonet hela tiden i omfattning och kvalitet. I januari 2010 hade 3,6 miljoner bedömningar gjorts. Det fanns då 80 000 ordpar (som i medeltal bedömts med minst 2 på den femgradiga skalan). Totalt hade

123 000 ordpar föreslagits av användare, varav 74 000 olika. 24 000 av dem hade dittills accepterats av andra användare.

En finess med Folkets synonymlexikon jämfört med vanliga synonymlexikon är att synonymerna är graderade mellan 2 och 5, genomsnittet av alla användares bedömningar. Det är alltså varje användares egen intuitiva uppfattning om begreppet synonymitet som bestämmer hur ett ordpar graderas, se till exempel figur 3. Det skapade lexikonet bygger således på folkets egen definition av synonymitet, vilket förhoppningsvis är precis vad folket vill! Eftersom lexikonet är skapat av hela folket så är det en fri resurs som kan laddas ner i sin helhet.

Baksidan med att använda allmänheten som arbetskraft är att det förekommer individer som försöker förstöra. Alla projekt som bygger på crowdsourcing måste hantera detta på något sätt. Vi har på följande tre sätt försäkrat oss mot att missbruk av lexikonet ska påverka kvaliteten:

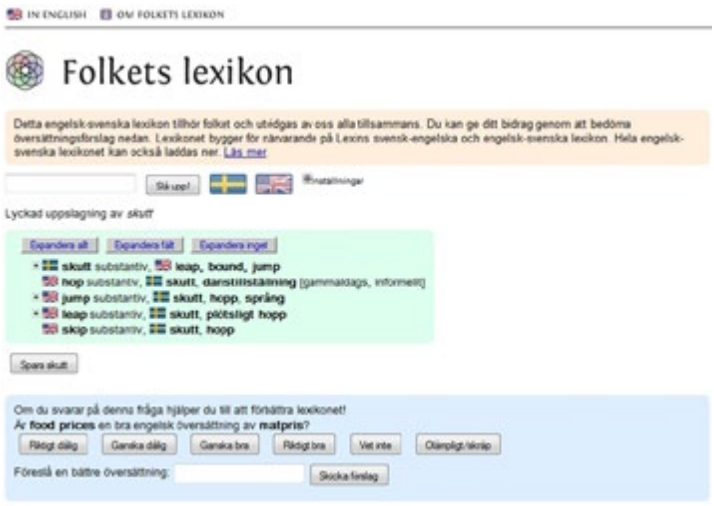
1. Många bedömningar krävs innan ett ordpar anses vara bra.
2. Vilket ordpar som ska bedömas väljs slumpmässigt från en lista som består av nästan 100 000 par.
3. Ordpar som föreslås av användare stavningskontrolleras innan de läggs till den enorma listan.

Det tycks som att dessa försäkringar är starka nog för att synonymlexikonets kvalitet ska hållas tillräckligt hög. En stickprovsundersökning där Folkets synonymlexikon jämfördes med en tryckt synonymordbok visade att alla synonymer som graderats 3 eller högre var angivna som synonymer i den tryckta ordboken eller var uppenbarliga synonymer som av någon anledning missats i den tryckta ordboken. I stort sett alla synonymer i den tryckta boken var med i Folkets synonymlexikon. De som saknades var mestadels omoderna eller mycket ovanliga.

6.2. Folkets lexikon

Det Lexinlexikon som flest använder och som antagligen därför genererar flest frågor till Språkrådet, som numera äger Lexin, är det svensk-engelska/engelsk-svenska. Det tillhör inte Språkrådets arbetsuppgifter att driva ett svensk-engelskt lexikon. Därför har Språkrådet nyligen överlåtit svensk-engelska/engelsk-svenska Lexin till KTH, som gör det tillgängligt i form av Folkets lexikon (FOLKETS). Projektet stöds ekonomiskt av .SE-stiftelsens Internetfond.

Tanken är att Folkets lexikon ska bli en fri resurs som ständigt utvecklas av folket. Forskare på Linköpings universitet har tagit fram en lång lista med översättningsförslag som bedöms av användarna av Folkets lexikon, se figur 4. På samma sätt som i Folkets synonymlexikon så kommer översättningsförslag, som användarna bedömer är bra, att läggas till lexikonet. Användare får också ge egna förslag som kommer att bedömas av andra användare och så småningom komma in i lexikonet. Även andra tillägg och ändringar kommer att bli möjliga att föreslå.



Figur 4: Användargränssnitt för Folkets lexikon

För att Folkets lexikon ska kunna växa ännu snabbare lägger vi in andra fria resurser i det, bland annat Folkets synonymlexikon och Svenskt associationslexikon (SALDO; se även Borins uppsats i denna volym).

7. Sammanfattning

Delar av femton års produktion från språkteknologigruppen på KTH har beskrivits. De flesta verktyg och resurser som tagits fram av gruppen är fritt tillgängliga och kan laddas ner från gruppens webbsida (HLTWEBB). En stor fördel med fritt tillgängliga verktyg är att såväl forskare vid högskolorna som utvecklare i industrin och privatpersoner som programmerar för nöjes skull kan få tillgång till moderna och fullskaliga verktyg och med hjälp av dem bygga nya språkteknologiska tillämpningar. Ett annat sätt att skapa stora fria resurser är med hjälp av crowdsourcing. Folkets synonymlexikon och Folkets lexikon blir hela tiden större och bättre, i och med att användare i tusental föreslår tillägg och bedömer varandras förslag. Det är ett alternativ till det traditionella mycket arbetskrävande sättet att ta fram lexikon. Min uppfattning är att crowdsourcing kommer att användas i många sammanhang i framtiden, men att det kommer att fungera väl och bli resultera i välfyllda lexikon endast i vissa gynnsamma fall. Popularitеten hos den webbsida där resursen byggs är kritisk, liksom att sidan måste besökas av användare som är villiga att hjälpa till att bygga just denna resurs.

Litteratur

Ordböcker

Klingspor, Richard: *Svenskt avstavningslexikon*. Tryckeriförlaget, Stockholm 1985.

SAOL 1986 = *Svenska Akademiens ordlista*, upplaga 11. Norstedts.

Annan litteratur

Bigert, Johnny 2005: Unsupervised evaluation of Swedish spell checker correction suggestions. I: *Nordiska datalingvistikdagarna 2005*, Joensuu, Finland.

Bloom, Burton H. 1970: Space/time trade-offs in hash coding with allowable errors. I: *Communications of the ACM*, volym 13, 422–426.

Damereau, Fred J. 1964: A technique for computer detection and correction of spelling errors. I: *Communications of the ACM*, volym 7, 649–664.

Hult, Ann-Kristin 2008: Användarna bakom loggfilerna. Redovisning av en webbenkät i Lexin online Svenska ord. I: *LexicoNordica*, volym 15, 73–91.

Kann, Viggo, Domeij, Rickard, Hollman, Joachim och Tillenius, Mikael 2001: Implementation aspects and applications of a spelling correction algorithm. I: L. Uhlirova, G. Wimmer, G. Altmann, R. Koehler: *Text as a Linguistic Paradigm: Levels, Constituents, Constructs*. Festschrift in honour of Ludek Hrebicek. Quantitative Linguistics, vol. 60, WVT, 108–123.

Kann, Viggo och Rosell, Magnus 2005: Free construction of a free Swedish dictionary of synonyms. I: S. Werner: *Nordiska datalingvistikdagarna 2005*, Joensuu, Finland, 105–110. <http://www.csc.kth.se/tcs/projects/infomat/rapporter/kann-rosell05.pdf>

Knutsson, Ola, Bigert, Johnny och Kann, Viggo 2003: A robust shallow parser for Swedish. I: *Nordiska datalingvistikdagarna 2003*, Reykjavik, Island. <http://www.nada.kth.se/theory/projects/xcheck/rapporter/gta03.pdf>

Sjöbergh, Jonas och Kann, Viggo 2006: Vad kan statistik avslöja om svenska sammansättningar? I: *Språk och stil*, volym 16, 199–214.

Internethänvisningar

AVSTAV = Språkrådet: *Avstavningslexikon*. http://avstava.appspot.com

FOLKETS = Hollman, Joachim och Kann, Viggo: *Folkets lexikon*. <http://folkets-lexikon.csc.kth.se>

GRIM = Westlund, Knutsson och Sjöbergh med flera: *Grim – en interaktiv miljö med fokus på det svenska språket*. <http://skrutten.nada.kth.se/grim/>

HLTWEBB = Språkteknologi vid KTH. <http://www.csc.kth.se/theory/humanlang/>

LEXIN = Språkrådet: *Lexin*. <http://lexin.nada.kth.se>

SALDO = Språkbanken: Saldo*, svenskt associationslexikon*. <http://spraakbanken.gu.se/saldo/>

SKANDORD = Nordiska språksekretariatet: *Skandinavisk ordbok*. <http://www.nada.kth.se/skandlexikon/>

Stallman, Richard 1996: *The Free Software Definition*. <http://www.gnu.org/philosophy/free-sw.html>

STAVA = Kann, Viggo: *Stavningskontrollprogrammet Stava*. KTH. <http://www.nada.kth.se/stava>

SYNLEX = Kann, Viggo och Rosell, Magnus: *Folkets synonymlexikon*. <http://lexin.nada.kth.se/synlex.html>

TAGGSTAVA = Kann, Viggo: *Ordklasstaggaren Taggstava*. <http://www.csc.kth.se/tcs/projects/granska/taggstava.html>

TVÄRSLÅ = Projektet Nordisk nätordbok: *Tvärslå*.
<http://ordbok.nada.kth.se>

Viggo Kann
professor
KTH Skolan för datavetenskap
och kommunikation
SE-100 44 Stockholm
viggo@kth.se