Ordböcker på Internet och Internet som ordbok

Lars Törnqvist

This paper gives an overview of present trends in Internet lexicography. Some methods of using web dictionaries and other websites for lexicographic purposes are presented. It ends up in a question: Are dictionaries really needed in the future?

Nyckelord: Internetlexikografi, elektroniska ordböcker.

Denna artikel inleds med en översikt över utvecklingen inom Internetlexikografin. Sedan följer en presentation av hur man kan använda webben för att sammanställa ordböcker eller som ersättning för ordböcker. Det hela utmynnar i frågan om ordböcker egentligen behövs i framtiden.

Bakgrund

År 2001 startade jag webbplatsen Thesaurus Lex, som är inriktad på ordböcker. Under åren har jag byggt upp en stor samling länkar till ordböcker av olika slag på webben. I samlingen ingår både språkliga ordböcker och encyklopedier samt hybrider mellan dessa. För närvarande omfattar samlingen ett par tusen länkar, sorterade efter ordbokstyp och språk. Detta ger goda möjligheter till överblick över ordboksfloran på Internet och dess utveckling under de senaste åren. En översikt över fackordböcker på webben presenterades i en tidigare artikel (Törnqvist 2007). I denna artikel ligger fokus på tvåspråkiga ordböcker, särskilt fackspråkliga.

Länksamlingen är delvis en biprodukt till min verksamhet som terminolog och lexikograf, genom att jag har lagt upp länkar till källor som jag har använt i arbetet. Flera av de ordlistor som jag har arbetat med finns också i länksamlingen, antingen på uppdragsgivarens webbplats eller på min egen. Erfarenheter från arbetet med dessa webbordlistor har gett impulser till de tankegångar som presenteras nedan.

Tendenser

Ordböckerna på Internet utvecklas i snabb takt till antal, storlek och form. Eftersom utvecklingen är i ett ganska tidigt stadium präglas den fortfarande av experimenterande. Det går dock att urskilja vissa allmänna tendenser.

Fritt tillgängliga – mer eller mindre

Numera förväntar sig folk att få tillgång till information utan kostnad. Här har Wikipedia och dess systerprojekt Wiktionary lagt ribban. När det nu finns ett uppslagsverk och en ordbok med stor omfattning och någorlunda anständig kvalitet som är helt gratis, så är det svårt för konkurrerande utgivare att ta betalt för motsvarande information. Det har också blivit vanligt att lexikala produkter som har finansierats med forskningsmedel publiceras gratis på webben i stället för att tryckas och säljas. Informationen får då mycket större spridning, vilket ofta är huvudsyftet. Förutom dessa moderna ordböcker har även en rad äldre ordböcker lagts upp på webben, bland annat i det ideella Projekt Runeberg. Det rör sig då om verk som inte längre är skyddade av upphovsrätt.

Konkurrensen med gratisprodukterna gör det svårt för ordboksutgivare som inte kan finansiera verksamheten genom ideellt arbete eller anslag från myndigheter eller forskningsstiftelser. Det har därför utvecklats olika sätt att sprida lexikal information gratis men ändå få viss finansiering av verksamheten. Ett sätt är att begränsa den kostnadsfria tillgången till materialet.

Den svensk-engelska ordboksdatabasen Tyda.se har begränsat den fria tillgängligheten för varje användare till 30 sökningar per vecka, medan större antal sökningar kräver abonnemang. Gränsen är satt så att en lekman som bara slår upp enstaka ord då och då klarar sig med den fria sökmängden, medan skolor och företag som utnyttjar ordboken i stor omfattning varje dag tvingas bli betalande abonnenter.

Man kan också begränsa tillgängligheten genom att ge gratis åtkomst till vissa artiklar eller kortversioner av artiklar, medan betalande abonnenter har åtkomst till hela materialet. Denna metod används bland annat av Nationalencyklopedin. En variant är att tillhandahålla ett enkelt

sökgränssnitt gratis, medan avancerad sökning kostar pengar. Ett exempel på detta är SAOL på nätet. Man kan utan kostnad söka på uppslagsord i den senaste upplagan av Svenska Akademiens ordlista och får då fram en bild av den sida i den tryckta ordlistan där ordet står. Den kommersiella versionen SAOL Plus, för närvarande på CD, ger dessutom möjlighet till sökning på ordled och böjningsformer, användning av jokertecken, sökning på ord i ordförklaringar m.m. (Berg 2009). Ytterligare ett sätt är att utgivaren publicerar vissa av sina ordlistor gratis och tar betalt för andra. Så är det bland annat på webbplatsen Norstedts ord, som ger fri tillgång till Norstedts stora engelsk-svenska och svensk-engelska ordböcker – samt under en introduktionsperiod till motsvarande franska, spanska och tyska ordböcker – medan övriga ordböcker liksom tidigare tillhandahålls endast i betalversion som bok eller CD.

Det håller på att utvecklas olika sätt att finansiera ordboksutgivning på webben. Det handlar om abonnemang för tillgång till extratjänster, tillhandahållande av annonsplats, offentliga anslag och frivillig medverkan från brukarna. Ofta används flera av dessa finansieringssätt samtidigt, exempelvis annonsfinansiering och abonnemang. Det är dock svårt att få tillräckligt kommersiell finansiering för att utveckla och underhålla lexikala produkter med hög kvalitet. Återanvändning av gammalt och delvis undermåligt material är därför alltför vanligt.

Fler informationsyper i samma ordboksartikel

Det finns vissa tendenser till att inkludera fler informationstyper i ordboksartiklarna än vad som har varit vanligt förr. Man kan till exempel hitta synonymer och ordförklaringar i tvåspråkiga ordböcker. Ofta förekommer encyklopedisk information utöver de rena ordförklaringarna, särskilt i fackordböcker. Ett ordboksprojekt som har drivit mängden av informationstyper mycket långt är Wiktionary, där en och samma ordboksartikel kan innehålla fullständig uppsättning böjningsformer, definitioner av flera homonyma betydelser, synonymer, motsvarigheter på flera andra språk samt flera olika kategoriseringar.

Hyperlexikon

Det har börjat komma interna och externa hyperlänkar till andra ordbokstyper, exempelvis länkar mellan definitionsordbok, synonymordbok eller thesaurus och översättningsordbok.

(Lange & Törnqvist 2003). I Wikipedia och i viss mån Wiktionary har hyperlänkningen utvecklats långt (Törnqvist 2008).

Nya informationstyper

Det förekommer helt nya typer av information som aldrig har förekommit i tryckta ordböcker. Ett exempel på detta är uttalsordböcker där uttalet återges med ljud, antingen inspelat eller med talsyntes. Sådana uttalsangivelser finns för både svenska och engelska ord hos Tyda.se och Norstedts ord. Ett annat exempel är teckenspråkslexikon där tecknen återges med rörliga bilder, såsom Spreadthesign och Svenskt teckenspråkslexikon 2009.

Interaktivitet

Användarna engageras allt mer i ordboksarbetet. Brukarmedverkan varierar över ett brett spektrum. I ena änden av detta spektrum finns Wikipedia och Wiktionary som skapas helt och hållet av användarna, i mitten finns Folkets synonymlexikon Synlex där frågor om grad av synonymitet måste besvaras av användarna, och i andra änden finns ett stort antal ordboksprojekt där användarna kan skicka förslag och kommentarer via en e-postlänk.

Teknikdriven utveckling

Utvecklingen av nya ordboksprodukter drivs av språkteknologer och datorlingvister. Nya funktioner tillkommer därför på grund av att de är tekniskt möjliga, inte därför att de är efterfrågade av användarna. Detta innebär att fokus gärna ligger på de tekniska funktionerna, inte på innehållet. Typiska exempel på denna utveckling är korsordsordböcker och rimlexikon. Dessa baseras ofta på textkorpusar som används utan djupare språklig analys. Sålunda anger Den stora rimordlistan att *mage* rimmar på *bagage* och *tillkännage*. Homonymseparering görs inte heller. När man söker på *val* i Folkets synonymlexikon Synlex får man därför synonymer som *omröstning* och *kaskelot* i samma uppräkning. Ett problem med teknikfokuseringen är att man ofta använder gammalt ordboksmaterial för nya tekniska tillämpningar utan att uppdatera eller anpassa det för nya användningsområden.

Osynliga lexikon

Det finns en stor mängd mer eller mindre osynliga lexikon inbyggda i olika programvaror. Det handlar om synonym- och felstavningslexikon i sökmotorer, tvåspråkiga lexikon i program för maskinöversättning, uttalslexikon i program för talsyntes och taligenkänning (Törnqvist 2006), rättstavnings- och avstavningslexikon i ordbehandlings- och layoutprogram och liknande. Kanske är det inom det här området som den största mängden lexikografiskt arbete utförs i dag.

Nya sätt att använda webbordböcker

Internet innehåller alltså en stor mängd ordböcker som är utarbetade för olika ändamål. Men man kan också använda Internet för att sammanställa nya ordböcker för nya användningsområden. Det finns olika sätt att göra detta.

Samma ordbok i flera språkversioner

Definitionsordböcker för speciella områden ges ibland ut i flera språkversioner. Detta förekommer särskilt i länder med flera officiella språk och inom organisationer med internationell utbredning. De olika versionerna är vanligen helt fristående, var och en med sin egen alfabetiska sortering. Språkversionerna kan dock lätt länkas ihop, och vips har man en översättningsordbok. Ett par exempel är en flygordlista i engelsk och fransk version hos Canada Aviation Museum och en ordlista över mormonska uttryck i 23 språkversioner hos Jesu Kristi Kyrka av Sista Dagars Heliga.

Språklänkar och kategorier

Wikipedia har två kraftfulla funktioner som gör att den kan användas som termbank. Den ena funktionen är länkarna till motsvarande artiklar i andra språkversioner. Från den svenska artikeln *Talgoxe* kan man hoppa direkt till den norska artikeln *Kjøttmeis* eller den engelska artikeln *Great Tit*. Den andra funktionen är den hierarkiskt uppbyggda kategoriseringen, som i vissa fall har drivits mycket långt. Trots att Wikipedia egentligen är avsedd att användas som uppslagsverk för sakinformation har den på så sätt även blivit en flerspråkig termbank med systematisk översikt.

Flerspråkiga databaser

Inom det naturvetenskapliga området finns ett antal internationella databaser över bland annat djur- och växtarter. Ett par exempel är Avibase och FishBase. I dessa anges ofta arternas namn på ett stort antal språk, förutom det latinska vetenskapliga namnet. Dessutom finns ofta artbeskrivningar och annan faktainformation på ett eller flera språk samt bilder och kartor. Sådana databaser fungerar som både heltäckande och pålitliga översättningsordböcker för artnamn.

Internet som ordboksunderlag

Även andra texter på Internet kan användas som lexikon eller underlag för lexikon. Det levande språkbruket på Internet utgör ju en textkorpus. Visserligen är den obalanserad och dåligt uppmärkt, men den är ändå en oerhört rik källa till modernt språkbruk av alla slag. Att räkna Google-träffar har blivit det vanligaste sättet att snabbt få en ungefärlig uppfattning om vanligheten hos olika uttryck.

Många företag vänder sig till kunder i flera språkområden och publicerar därför sina webbsidor i flera språkversioner. Dessa webbsidor bildar en parallellkorpus som kan utnyttjas i tvåspråkig lexikografi, i synnerhet fackspråkslexikografi. Bland de parallella webbsidorna finns ofta produktkataloger, vilka kan innehålla mycket stora mängder facktermer. Dessa kan enkelt matchas mellan språken på samma sätt som man kan göra med ordböcker i flera språkversioner (se ovan). Katalogerna är ofta illustrerade, vilket underlättar matchningen. Som exempel kan nämnas IKEA, vars webbplats finns i 40 versioner för olika språk och

länder. Om man inte har tillgång till katalogdata på flera språk från samma företag kan man i många fall matcha webbplatser från olika företag, förutsatt att de täcker samma produktområde och innehåller bilder eller tydliga beskrivningar av produkterna.

Numera kan man även söka efter bilder i Google och andra sökmotorer. Detta är ett snabbt sätt att söka efter betydelsen hos ord med konkret betydelse. Sådan sökning kan också ge mer korrekta uppgifter om betydelsen hos facktermer än traditionella ordböcker. Genom bildsökningen ser den fackkunnige användaren direkt vilken teknisk anordning som avses, medan allmänspråkliga ordböcker många gånger ger otydliga eller vilseledande uppgifter om facktermers ekvivalens inom ett givet fackområde. Om man exempelvis vill veta vilket verktyg som heter *pincers* på engelska ser man på bilderna att det är hovtång, medan vissa ordböcker ger den felaktiga ekvivalenten *kniptång*.

Några praktikfall

Kan de här möjligheterna att använda Internet som lexikografisk resurs vara till praktisk nytta? I så fall bör man kunna sammanställa ordböcker på ett mycket effektivt sätt. För att undersöka detta gjorde jag några små förstudier och tog sedan itu med uppbyggnaden av en stor termdatabas.

Förstudier: norsk-svenska specialordlistor

För att testa metoden gjorde jag små översättningsordlistor från norska till svenska inom fyra specialområden: sport, mat, kontor och djur. Ordlistorna gjordes som enkla tabeller med en kolumn för vardera språket samt två kolumner för ordklass och kort förklaring, vilka användes vid behov. Som källor använde jag uteslutande webbsidor.

Att använda språklänkningen i Wikipedia på svenska och norsk bokmål visade sig vara den snabbaste metoden att få ihop ett stort basförråd av ekvivalenta ordpar inom samtliga områden. Kategoriindelningen gav en god överblick över artiklarna och språklänkarna gav direkt åtkomst till ekvivalenterna. Från den norska artikeln *Breiflabb* var det bara att klicka sig direkt till den svenska artikeln *Marulk* och så vidare. Jag använde dock inte bara hyperlänkarna mellan artiklarna på de två språken, utan jag gick även igenom artikeltexterna

för att få större djup i ordförrådet. Vissa svårigheter visade sig ganska snart. Många artiklar saknades på det ena språket, oftast på norska, och innehållet i de motsvarande artiklarna kunde vara ganska olika till omfattning och karaktär. Det som gick att få ut av Wikipedia-länkningen var alltså ett relativt stort och brett ordförråd med stora luckor i täckningen av ämnesområdet och begränsad detaljeringsnivå. För att komplettera luckorna och öka detaljeringsnivån använde jag även andra källor. För matordlistan gick det att få ut ganska mycket från webbplatser med matrecept och liknande. Här fanns aldrig samma webbsidor i olika språkversioner, utan jag fick leta fram någorlunda motsvarande sidor med liknande innehåll. Resultatet blev god täckning när det gällde matvarorna, medan det var svårare att hitta motsvarigheter till maträtterna eftersom recepten sällan var helt ekvivalenta. För köksutrustning och liknande var det lätt att hitta ekvivalenter i illustrerade produktkataloger. IKEA:s webbplats var här till stor nytta. Samma metod användes med gott resultat för kontorsordlistan, medan det var svårt att hitta bra parallella texter till sportordlistan.

Djurordlistan visade sig vara enklast att bygga upp. Här hade Wikipedia ganska god täckning när det gällde fåglar, fiskar, däggdjur och många grupper av insekter. Dessutom finns det många flerspråkiga artdatabaser, särskilt för fiskar och fåglar. Det gick därför snabbt att ställa samman en ordlista med acceptabel täckning på artnivå. Jämförelser mellan de olika källorna avslöjade många fall av bristande överensstämmelse, oftast orsakade av dålig korrekturläsning. För att hitta ord för djur av olika ålder och kön, kroppsdelar och typiska beteenden var det nödvändigt att granska artikeltexterna, främst i Wikipedia. Även här var det ganska lätt att få acceptabel täckning av ordförrådet.

Resultatet av förstudien var att det gick mycket snabbt att samla ordpar upp till en viss detaljeringsnivå. När jag hade hittat bra källor var det inga problem att få ihop mellan 10 och 20 ordpar per timme med god kvalitet. För den som är van vid traditionellt terminologiarbete, där man ibland lägger ned flera timmar på varje termpost, är detta en förbluffande hastighet.

Det finns naturligtvis åtskilliga felkällor med den här metoden. Den viktigaste är att texterna på Internet håller mycket ojämn kvalitet, både innehållsmässigt och språkligt. Wikipedia sammanställs av allmänheten, och vissa artiklar förefaller vara skrivna av skolbarn med begränsad sakkunskap och outvecklad förmåga att uttrycka sig i skrift. Översättningar mellan olika språkversioner görs av amatörer som ibland har häpnadsväckande dåliga språkkunskaper. Länkningen mellan språken är inte heller felfri. Även webbkällor från företag och organisationer innehåller felaktigheter, ofta på grund av bristfällig översättning och dålig korrekturläsning. Det är därför viktigt att man i möjligaste mån använder flera oberoende källdokument.

Projektet Byggord

De norsk-svenska specialordlistorna var en förstudie för att testa nya metoder inför ett större projekt. Detta projekt, som har fått namnet Byggord, går ut på att bygga upp en flerspråkig databas över facktermer som används inom anläggning, husbygge, installation och fastighetsförvaltning. Databasen kommer att tas i drift i början av 2010 och är avsedd att fungera som ett kombinerat översättnings- och sökverktyg. Idén är att underlätta informationssökning på webben genom att tillhandahålla synonymer och andra närliggande termer samt ekvivalenter på andra språk. I det första skedet omfattar databasen tre språk: svenska, engelska och franska. Ordförklaringar ges bara i undantagsfall, särskilt när det behövs för disambiguering. Information om termernas innebörd ges i stället genom klassificering och externa länkar till information på webben.

Eftersom syftet är att hitta information på webben är det webbens språkbruk som ska återges i all sin vildvuxenhet. Både webbsidor och traditionella dokument – standarder, klassifikationssystem, fackordlistor och dylikt – har använts som källor. Dessa källor har visat sig komplettera varandra på ett bra sätt. Webbkällorna ger tillgång till ett mer varierat, verkligt språkbruk. Här hittar man också skillnaderna mellan brittisk och amerikansk terminologi, liksom franska termer som är speciella för Belgien eller Québec. De traditionella dokumenten syftar oftast till att standardisera språkbruket, och de innehåller därför mycket få synonymer och stavningsvarianter. Många termer i dessa källor visar sig vid närmare beskådan vara konstlade uttryck baserade på klassificering, som inte ger träffar vid webbsökning. De traditionella terminologikällorna ger dock en systematisk ingång som är värdefull när det gäller att täcka in fackområdets ordförråd.

Insamlingen av termer har gjorts efter de principer som har beskrivits ovan, och detta har visat sig fungera alldeles utmärkt. Termer från tryckta källor har vid behov kontrollerats genom webbsökning, vilket i hög grad har höjt tillförlitligheten hos materialet. Arbetet har gått ganska snabbt, och det har oftast gått att lägga in minst 100 trespråkiga poster per arbetsdag. Även det här projektet visar alltså att webben kan användas på ett effektivt sätt för ordboksarbete.

Kan webben i sig själv ersätta webbordböckerna?

Behövs det färdiga ordböcker sammanställda av lexikografer i framtiden, eller kommer användarna att själva utföra arbetet när de behöver det? På den frågan vill jag svara som den svenske politikern Yngve Holmberg: ”Frågan kan besvaras med både ja och nej, med reservation åt båda hållen.”

Ja: Alla användare har numera tillgång till lexikografiska verktyg. Hyperlänkning mellan språkversioner gör att Wikipedia och andra webbplatser på flera språk kan användas som översättningsordböcker.

Nej: Det går snabbare att söka i en färdig ordbok än att leta själv på webben och bedöma sökresultaten. Lexikografer gör också säkrare språkliga bedömningar än lekmän utan språkvetenskaplig skolning, vilket gör att en professionellt sammansntälld ordbok är mer pålitlig är resultatet från en webbsökning. Specialordböcker ger dessutom översikt över fackområden, vilket är värdefullt vid inlärning av terminologier på andra språk.

Reservation: De allmänt tillgängliga verktygen är trubbiga och kräver expertkunskap för att kunna användas effektivt. Därför kommer allmänheten att ha behov av ordböcker även i framtiden. Men även om ordböcker behövs, så kommer de knappast att kunna fylla alla behov. De kommer aldrig att täcka in hela ordförrådet med alla betydelsenyanser i alla fackspråkliga tillämpningar. Egen sökning kommer därför att vara nödvändig, särskilt för facköversättare och andra professionella användare.

Det stora problemet för lexikografin i framtiden är nog att ordböcker förväntas vara gratis. Detta ger finansieringsproblem, vilka kan medföra att det inte tas fram något nytt lexikaliskt material av god kvalitet. Det är därför stor risk att man i hög grad kommer att återanvända material från redan existerande ordböcker i nya skepnader, trots att detta material åldras i snabb takt. Men även om det inte skulle göras några bra ordböcker i framtiden är det ingen katastrof för den kvalificerade användaren, för det kommer samtidigt allt bättre möjligheter att själv hitta den information som inte längre finns i ordböcker.

Källor

Tryckta källor

Berg, Sture (2009): Om ordböjning och SAOL Plus. I: Martin Gellerstam (red.) *SAOL och tidens flykt: Några nedslag i ordlistans historia.* Stockholm: Norstedts. S. 139–165.
Lange, Sven & Törnqvist, Lars (2003): Thesaurus Lex: En ny svensk elektroniskt länkad thesaurus. I: Zakaris Svabo Hansen & Arnfinnur Johansen (red.) *Nordiske studier i leksikografi 6: Rapport fra Konference om leksikografi i Norden Tórshavn 21.-25. august 2001*. Tórshavn: Nordisk forening for leksikografi. S. 181–190.
Törnqvist, Lars (2006): Uttalslexikon för talsyntes. I: Henrik Lorentzen & Lars Trap-Jensen (red.) *Nordiske Studier i Leksikografi 8: Rapport fra Konference om Leksikografi i Norden Sønderborg 24.–28. maj 2005.* København: Nordisk forening for leksikografi. S. 373–381.
Törnqvist, Lars (2007): Språklig och encyklopedisk information i fackordlistor på Internet. *LexicoNordica* 14. S. 35–47.
Törnqvist, Lars (2008): Hyperlexikon – finns de? I: *Nog ordat?: festskrift till Sven-Göran Malmgren den 25 april 2008*. Göteborg: Meijerbergs institut för etymologisk forskning. Meijerbergs arkiv för svensk ordforskning 34. S. 382–389.

Webbsidor

Avibase. <http://avibase.bsc-eoc.org/avibase.jsp>.
Byggord. <http://byggord.thesauruslex.se>.
Canada Aviation Museum. <http://www.aviation.technomuses.ca/collections/glossary/>.
Den stora rimordlistan. <http://www.gameelite.se/rimma/>.
FishBase. <http://www.fishbase.se/search.php?lang=Swedish>.
Folkets synonymlexikon Synlex. <http://lexin2.nada.kth.se/synlex.html>.
Google. <http://www.google.com>
IKEA. <http://www.ikea.com/se/>.
Jesu Kristi Kyrka av Sista Dagars Heliga. <http://www.mormon.org/glossary/ >.
Nationalencyklopedin. <http://www.ne.se>.
Norska specialordlistor. <http://www.thesauruslex.se/ordlista/nospecial.htm>.
Norstedts ord. <http://www.norstedtsord.se>.
Projekt Runeberg. <http://runeberg.org>.
SAOL på nätet. <http://www.svenskaakademien.se/web/Ordlista.aspx>.
Spreadthesign. <http://www.spreadthesign.com/country/se/>.

Svenskt teckenspråkslexikon 2009.
<http://www.ling.su.se/pub/jsp/polopoly.jsp?d=10567&a=56692>.
Thesaurus Lex. <http://www.thesauruslex.com>.
Wikipedia, svenska. <http://sv.wikipedia.org>.
Wikipedia, norska (bokmål). <http://no.wikipedia.org>.
Wikipedia, engelska. <http://en.wikipedia.org>.
Wiktionary, svenska. <http://sv.wiktionary.org>.

URL-adresserna är kontrollerade 2009-11-30.