Folkets användning av Lexin – en resurs

Viggo Kann
KTH Nada
Stockholm
viggo@nada.kth.se

26 augusti 2004

Sammanfattning

Tjänsten Lexin on-line är oerhört populär: i genomsnitt görs mer än tre uppslagningar per sekund, dygnet runt. Denna popularitet är en stor resurs som kan utnyttjas på följande sätt: Om alla människor som slår upp ett ord erbjuds att svara på en enkel fråga om huruvida två ord är synonyma så kan man på kort tid bygga upp ett stort synonymlexikon. I artikeln beskrivs hur man kan gå tillväga för att genomföra detta.

1 Introduktion

Att bygga ett nytt och omfångsrikt lexikon är ett enormt jobb och kostar därför normalt mycket pengar. Den som har bekostat framställningen av ett lexikon håller naturligtvis hårt i det, varför det inte kan spridas fritt och därigenom komma allmänheten tillgodo.

Internet har gjort det möjligt att skapa stora verk (som lexikon) genom att massor av människor drar var sitt strå till stacken. Ett verk som kräver en enorm arbetsinsats kan därmed skapas utan att varje enskilt bidrag behöver vara så stort. Ju fler personer som lämnar bidrag desto mindre behöver varje bidrag bli.

Det mest kända verket som byggts upp på detta sätt är encyklopedin Wikipedia[1], som grundades i januari 2001 av Internetentreprenören Jimmy Wales och filosofen Larry Sanger. Wikipedia har från starten varit ett internationellt projekt och finns idag på 41 olika språk. Den engelska Wikipedia är störst med omkring 330 000 artiklar. Den svenska Wikipedian[2] startades i maj 2001 och har just nu 37 800 artiklar. Vem som helst får bidra genom att skriva, komplettera eller korrigera en artikel.

Nyligen (i maj 2004) inleddes systerprojektet Wiktionary[3] som går ut på att alla intresserade gemensamt ska skapa en ordlista för alla språk. En stor fördel med verk skapade på detta sätt är att de kan göras fritt tillgängliga för alla. Detta ger också en extra motivation till att bidra till verket. Både Wikipedia och Wiktionary har öppet innehåll och använder sig av copyleftlicensen[4] GNU FDL.

Att starta ett nytt projekt som Wikipedia är inte lätt eftersom det kräver att mängder av personer som skulle vilja lämna bidrag måste hittas och kontaktas. Ett sådant projekt måste lanseras på en webbplats som har tillräckligt många besökare. Men det räcker

[1] http://en.wikipedia.org
[2] http://sv.wikipedia.org
[3] http://sv.wiktionary.org
[4] Copyleft infördes av Richard Stallman 1984, se http://www.gnu.org/copyleft/copyleft.html

inte att det är många besökare; det måste vara besökare som är intresserade av att hjälpa till.

Lexin on-line[5] är en webbplats med många besökare. I maj 2004 gjordes 8,7 miljoner uppslagningar, vilket innebär i genomsnitt 195 uppslagningar varje minut eller mer än tre uppslagningar per sekund, dygnet runt. 80 procent av uppslagningarna görs i det svensk-engelska lexikonet. Gemensamt för alla besökare är att de vill veta vad ett svenskt ord heter på ett annat språk eller vad ett ord på ett annat språk heter på svenska. De har kort sagt intresse för det svenska språket. Därför kanske många av dem skulle vara intresserade av att hjälpa till att bidra till ett svenskt synonymlexikon.

För att förverkliga denna idé utlyste jag ett programmeringsprojekt i en projektkurs vid KTH, där jag är professor. Kursen heter *Programutvecklingsprojekt med mjukvarukonstruktion* och en projektgrupp om åtta tredjeårsteknologer[6] valde att göra mitt projekt, som kallades Folkets synonymlexikon Synlex. Projektet genomfördes under våren 2004 och ledde fram till en prototyp som nästan kan sättas i drift direkt.

2 Folkets synonymlexikon Synlex

Tanken är att varje gång en uppslagning görs i Lexin on-line så ska efter den vanliga lexikontexten ett möjligt synonympar presenteras, och användaren ska, helt frivilligt, få chansen att betygsätta hur pass nära synonymer dessa ord är på en skala från 0 till 5. Om användaren inte vet vad något av orden betyder kan han eller hon svara *vet inte*.

2.1 Vad är en synonym?

Det första problemet en synonymlexikonkonstruktör ställs för är att definiera vad som menas med att två ord är synonymer. Synonymer är ju ord som betyder samma sak. Men det är mycket ovanligt med ord som är helt utbytbara. Vissa ord har till exempel olika stilvärde (*flicka* och *tös*). Andra har överlappande men inte identisk betydelse (*god* och *smaklig*). Detta gör att språkvetare och lexikografer är försiktiga med att definiera exakt vilka ord som är synonymer.

I Synlex finns inte det problemet, för där är det folkets egen definition av synonym som används. Det betyder att om tillräckligt många personer anser att två ord är synonymer (till en viss grad) så kommer orden att vara det i lexikonet.

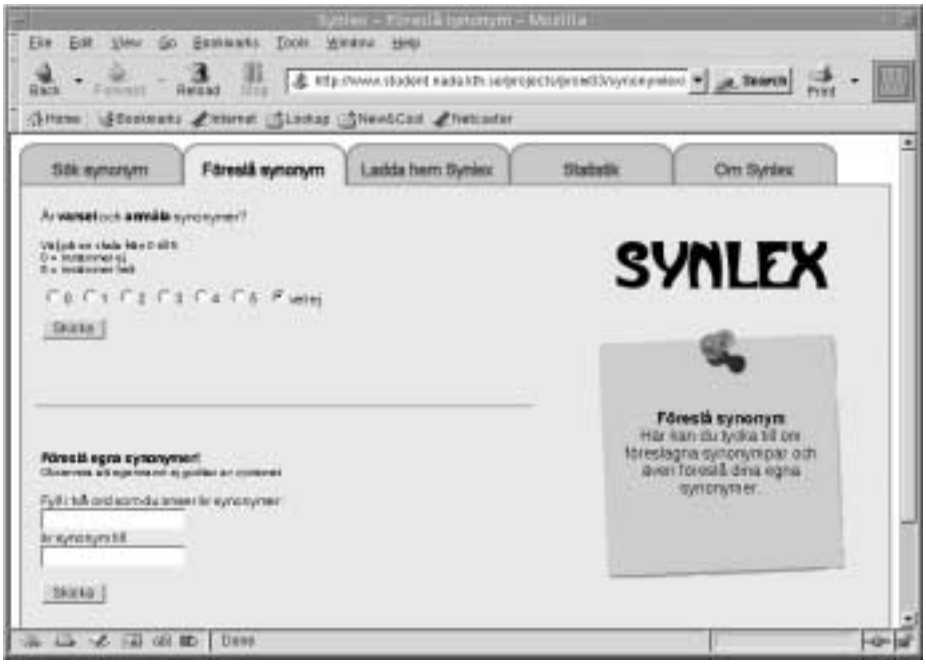
2.2 Folkets möjligheter

Den som betygsätter ett föreslaget synonympar ska få upp ett fönster med Synlex egen webbplats. Där kan man bedöma fler synonympar eller föreslå egna synonympar, se figur 1.

Under en annan flik kan man söka i synonymlexikonet Synlex. Innan ett synonympar läggs in i Synlex kommer det att ha testas och betygsatts av flera användare. Varje synonympar får därför ett medelbetyg som kan användas vid sökning i synonymlexikonet. En användare som vill ha synonymer till ett visst ord kan alltså välja att få synonymer som fått högt betyg (och därför är mycket nära i betydelse) eller synonymer som fått lägre betyg (och har liknande betydelse).

[5] `http://lexin.nada.kth.se`

[6] Gruppen bestod av Sara Björklund, Sofie Eriksson, Patrik Glas, Erik Haglund, Anna Hilding, Nicholas Montgomerie-Neilson, Helena Nützman, Carl Svärd.



Figur 1: Synlex webbplats.

Man kan också ladda hem hela den aktuella versionen av Synlex, närmare bestämt alla synonympar som har betyg 3 eller högre, i en komprimerad XML-fil. Eftersom Synlex är skapat av folket och inte har några upphovsrättsproblem så kan användaren göra vad han vill med synonymlexikonet – det är verkligen helt fritt.

På Synlex webbplats går det också att få diverse statistik över synonyminsamlingen och den aktuella versionen av synonymlexikonet.

2.3 Konstruktion av synonymparsförslag

Det enda språkliga arbete som krävs för att ovanstående ska kunna genomföras är att en tillräckligt stor lista med förslag till synonympar kan skapas. Jag har skapat en lista med en kvarts miljon synonympar genom att använda ett svensk-engelskt lexikon framlänges och baklänges. Alla svenska ord som har samma engelska ord som översättning föreslås som synonympar.

Det finns andra hel- och halvautomatiska sätt att skapa förslag till synonympar. Ett välkänt och bra, men datorresurskrävande, sätt är LSA[7] (Latent Semantic Analysis) som bygger på en matematisk matrisreduktionsmetod som heter singulärvärdesfaktorisering. Till den lista med en kvarts miljon synonympar som hittills framställts har inte LSA använts, men om man använder LSA skulle man antingen kunna sålla bort dåliga synonymparsförslag eller få fram nya förslag.

[7]Se Landauer och Dumais: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge, Psychological Review, 104(2), 211-240.

2.4 Hur missbruk förhindras

Det finns tyvärr personer som försöker förstöra på Internet. Det finns inbyggda hinder i Synlex för missbruk. För det första krävs det många bedömningar av samma synonympar för att det ska komma med och bli sökbart i Synlex. När användaren får ett synonympar presenterat för bedömning har Synlex valt det slumpmässigt från listan med en kvarts miljon förslag. Det är alltså inte möjligt för en användare att tycka till om samma förslag många gånger. För att synonympar ska bli felaktigt betygsatt krävs det alltså att en stor del av användarna som betygsatt just det paret har svarat fel. Eftersom det finns ett alternativ *vet inte* så kan vi anta att den största felkällan är personer som medvetet svarar felaktigt. Gissningsvis är den stora majoriteten av användare inte ute efter att förstöra, så då bör det inte vara någon större risk för att denna typ av missbruk ska påverka lexikonets kvalitet.

Det är också tänkbart att användare missbrukar möjligheten att föreslå egna synonympar. Därför stavningskontrollerar Synlex alla föreslagna ord och kan därmed sålla bort dåliga ord liksom svordomar och fula ord. Dessutom är det så att när en användare föreslår ett nytt synonympar så läggs synonymparet bara till den tidigare listan av synonymparsförslag, och det måste därmed bedömas på samma sätt som övriga synonympar innan det tas med i Synlex.

Slutligen finns det ett administratörsprogram till Synlex där administratören kan gå in och manuellt ta bort oönskade synonymer och förslag.

3 Slutord

Tillsammans med teknologerna Sara Björklund, Sofie Eriksson, Patrik Glas, Erik Haglund, Anna Hilding, Nicholas Montgomerie-Neilson, Helena Nützman och Carl Svärd har jag konstruerat skelettet till ett synonymlexikon. Det enda som återstår är att låta Lexinanvändarna betygsätta synonymparsförslagen.

Hur lång tid tar det innan alla synonymparsförslag fått tillräckligt många betygsättningar för att svaret ska gå att lita på? Säg att det behövs fem betygsättningar av varje förslag och att bara var tionde användare av Lexin svarar på synonymparsfrågan. Då tar det ändå bara två månader innan synonymlexikonet är klart – så stor är användningen av Lexin on-line. Eftersom det är nästan omöjligt för användare att manipulera resultaten så borde lexikonet hålla god kvalitet, och kvaliteten blir bara bättre ju fler bedömningar som görs.

När synonymlexikonet är klart kan man förstås gå vidare och låta folket konstruera andra språkresurser, såsom ordlistor mellan nya språk, ordlistor med ordförklaringar etc.