# Slutrapport för projektet Folkets engelsk-svenska lexikon

Viggo Kann
viggo@algoritmica.se

Joachim Hollman
joachim@algoritmica.se

20 september 2011

## 1 Inledning

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. På engelska kallas detta för crowdsourcing. Det mest kända verket som byggts upp på detta sätt är encyklopedin Wikipedia.

I detta projekt har ett nytt sådant verk skapats, ett engelsk-svenskt och svensk-engelskt lexikon, kallat Folkets lexikon.

År 2008 beslutade Språkrådet, som ansvarar för den populära lexikontjänsten Lexin, att man avsåg att bara driva Lexin vidare med invandrarspråken. Därmed skulle man alltså sluta med svensk-engelska och engelsk-svenska Lexin. Folkets lexikon bygger i grunden på materialet i svensk-engelska och engelsk-svenska Lexin.

Detta projekt har gjort dessa lexikon fortsatt tillgängliga på webben och dessutom med möjlighet för användarna att utöka lexikonet genom att rösta på förslag till nya översättningar.

Folkets lexikon finns tillgängligt på `http://folkets-lexikon.csc.kth.se` och utvecklas vidare i fortsättningsprojektet Vidareutveckling av Folkets lexikon, som också finansierats av Internetfonden.

## 2 Mål och syfte

Projektet har tre målgrupper:

1. Internetanvändare som behöver slå upp i ett svensk-engelskt eller engelsk-svenskt lexikon.
2. Språkintresserade Internetanvändare som vill hjälpa till att utveckla lexikonet.
3. Utvecklare av fri programvara som behöver ett fritt tillgängligt engelsk-svenskt lexikon.

Syftet är att tillfredsställa alla tre målgrupperna genom att bygga ett utvidgningsbart engelsk-svenskt och svensk-engelskt lexikon.

Målsättningarna med projektet var tre:

1. En söktjänst med samma funktionalitet (till exempel rättstavning av felstavade sökord och uppslagning på både svenska och engelska) och lexikon som engelsk-svenska Lexin idag läggs upp hos KTH i samma ögonblick som Språkrådet stänger engelsk-svenska och svensk-engelska Lexin.
2. Varje användare av engelsk-svenska lexikonet får möjlighet att bedöma en översättning och erbjuds också möjligheten att utöka lexikonet på andra sätt.
3. Hela lexikonet görs tillgängligt i ett enkelt XML-format efter att det utvidgats och förbättrats tillräckligt.

## 3 Projektbeskrivning

Projektet har drivits av Viggo Kann och Joachim Hollman på Algoritmica Hollman & Kann HB och finansierats av IIS Internetfond.

Två servrar har köpts in, och dessa drivs av systemgruppen vid KTH Skolan för datavetenskap och kommunikation. Folkets lexikon har därmed sin hemvist på KTH. Språkrådet har överlåtit Lexins svensk-engelska och engelsk-svenska lexikon till KTH som startlexikon för Folkets lexikon.

Webbtjänsten är utvecklad med GWT. Joel Sjöstrand har gjort webbgränssnittets grafiska element. Översättningsförslagen är huvudsakligen framtagna med automatiska språkteknologiska metoder av Sara Stymne och Lars Ahrenberg, NLPLab, Institutionen för datavetenskap, Linköpings universitet.

Tiotusentals Internetanvändare har bidragit genom att rösta på översättningsförslag och själva bidragit med förslag.

## 4 Leverabler

1. Framtagning av data
   Avtal har skrivits med Språkrådet om övertagande av Lexins engelsk-svenska och svensk-engelska lexikon. Materialet har överförts till KTH och lagras i ett nytt XML-format som vi tagit fram speciellt för Folkets lexikon. Översättningsförslag har genererats av Linköpings universitet.

2. Söktjänst som ersätter Lexin
   Söktjänsten är byggd med GWT, Google Web Toolkit. Indexering av sökbara ord görs i förväg. Indexet och lexikonet läses in när servern startas. Servern skickar uppslagningsresultatet som XML till klienten som presenterar detta för användaren.

3. Bedömningsfråga
   Vid varje uppslagning i lexikonet får användaren en slumpvis vald översättningsbedömningsfråga. Svaret lagras på servern.

4. Lexikonutvidgning
   Svaren på översättningsbedömningsfrågorna analyseras, dåliga översättningsförslag kastas bort och förslag som bedömts tillräckligt bra görs om till lexikonposter och läggs till Folkets lexikon.

5. Användarnas egna förslag
   Översättningsbedömningsfrågorna utökades med ett fält där användaren kan ange eget förslag till översättning. Dessa förslag automatgranskas (se nedan) och blir sedan nya översättningsförslag som presenteras för andra användare.

## 5 Resultat

De fem leverablerna ovan har levererats. Folkets lexikon las ut på webben i sin första version i mars 2009 och har sedan utvecklats steg för steg. En mycket användbar funktion som vi lagt till i uppslagningen är smart automatkomplettering, så att de ord som användaren troligast vill slå upp presenteras i en kompletteringsmeny. Vi använder statistik för vilka uppslagningar som är vanligast när vi väljer ut vilka fem kompletteringsförslag som ska ges.

I november 2010 la Språkrådet ner svensk-engelska och engelsk-svenska Lexin. Då ansåg Språkrådet att Folkets lexikon var moget att ta över.

### Hur missbruk förhindras

Det finns tyvärr personer som försöker förstöra på Internet. Därför har vi byggt hinder i Folkets lexikon för missbruk. För det första krävs det många bedömningar av samma ordpar för att det ska komma med och bli sökbart i lexikonet. När användaren får ett ordpar presenterat för

bedömning har systemet valt det slumpmässigt från listan med över 50 000 förslag. Det är alltså inte möjligt för en användare att tycka till om samma förslag många gånger. För att ett ordpar ska bli felaktigt bedömt krävs alltså att en stor del av användarna har svarat fel. Eftersom det finns ett alternativ *vet inte* så bör den största felkällan vara personer som medvetet svarar felaktigt. Den stora majoriteten av användare är inte ute efter att förstöra.

Det är också tänkbart att användare missbrukar möjligheten att föreslå egna översättningar. Därför stavningskontrollerar systemet alla föreslagna ord (på svenska och engelska) och kan därmed sålla bort dåliga ord liksom svordomar och fula ord. Dessutom är det så att när en användare föreslår en ny översättning så läggs ordparet bara till den tidigare listan av översättningsförslag, och det måste därmed bedömas på samma sätt som övriga ordpar innan det tas med i Folkets engelsk-svenska lexikon.

# 6 Utvärdering och analys

## 6.1 Utvärdering av resultat

Folkets lexikon är mycket välanvänt och tycks uppskattat av användarna. Användningsstatistik finns på webbsidan (under Om Folkets lexikon).

Bedömningarna som görs håller däremot inte genomgående hög kvalitet. Många användare (eller robotar) svarar utan eftertanke, vilket gör att brusnivån är ganska hög. Vi har därför varit tvungna att justera värdena för hur många och vilka bedömningar som krävs för att en översättning ska godkännas.

GWT, som vi använt för att utveckla tjänsten, var i början buggigt och ofullständigt, varför vi har behövt lägga ner mer tid än planerat för att gå runt problem och bygga egna lösningar. Dessutom har GWT kommit i flera nya versioner som inte varit helt bakåtkompatibla. I skrivande stund verkar GWT ha stabiliserats.

## 6.2 Förslag på förbättringar

- Utvidga Folkets lexikon med andra fria lexikonresurser, såsom Folkets synonymlexikon och Saldo, Svenskt associationslexikon.
- Utveckla stöd till forskare som vill studera hur Folkets lexikon används. Samla statistik för både uppslagningar och inställningar.
- Inför klickbara länkar inom lexikonet.
- Inför fritextsökning.
- Konvertera huvudorden till grundform (i Lexin används verbens presensform som uppslagsord).
- Utveckla möjligheter för användarna att utvidga och korrigera lexikonets samtliga fält.
- Utveckla en webbservice så att andra tillämpningar kan anropa Folkets lexikon.

# 7 Framtida arbeten

Vi har startat ett fortsättningsprojekt, Vidareutveckling av Folkets lexikon, där förbättringsförslagen ovan ska genomföras.