Vi som går köksvägen

Språkteknologer och korpuslingvister i Litteraturbanken

Lars Borin

Drömmen om det digitala handbiblioteket

Mina första tankar om det som jag idag brukar kalla ”textteknologi för forskning och utbildning”[1] väcktes då jag hade förmånen att under ett par år samarbeta med Donald Broady i forskningsprogrammet *Wallenberg Global Learning Network* (WGLN), finansierat av Knut och Alice Wallenbergs stiftelse. Vårt dagliga samarbete bedrevs inom en nod av detta internationella nätverk, i Uppsala Learning Lab. Här introducerade Donald mig till sitt arbete med uppmärkning av den digitala versionen av nationalutgåvan av Strindbergs samlade verk.[2] Arbetet bestod bland annat i (manuell) uppmärkning av en rad informationskategorier – typografiska, dokumentstrukturella och innehållsliga – i Strindbergtexterna. Senare har Donald bl.a. initierat ett projekt för ”märkning av utbildningsinnehåll”[3], där bl.a. personreferenser och en del annan information i en lärobok försågs (igen manuellt) med passande uppmärkning. Båda typerna av uppmärkta digitala material hör hemma i vad Donald en gång har kallat det ”nya handbiblioteket”, som framförallt karakteriseras av *tillgänglighet*:

[1] Till exempel i Lars Borin, ”Textil. Textteknologi i forskning och lärande” *forhum* 1/2005, pp. 11–13. Se vidare nedan.

[2] Donald Broady, ”Hur kodar man Röda rummet?” I Mats Rolén (red.), *ABM, IT och forskningen. Rapport från en konferens på Kungliga Biblioteket den 17 november 1999*, Riksbankens Jubileumsfond, Stockholm 2000.

[3] Donald Broady, Märkning av utbildningsinnehåll. Projektbeskrivning 17 juni 2003, reviderad 19 sept. 2003. <http://www.skeptron.ilu. uu.se/broady/dl/mu-plan-030919.pdf> [nedladdad 14/11 2005].

> Dokumenten skall vara tillgängliga i flera hänseenden. De skall vara lätt åtkomliga, kanske lokalt lagrade på skivminne i den egna datorn eller i ett lokalt nätverk. De skall vid behov ofördröjligen kunna bearbetas med användarens egna favoritprogram (textbehandlingsprogram, bildbehandlingsprogram, kalkylprogram etc). De skall kunna kompletteras och förses med randanmärkningar och korshänvisningar (så kallade hypertextlänkar). Samlingen skall vara överblickbar, man skall utan stort besvär erhålla automatiskt genererade innehållsförteckningar eller kartor över dokumentbasens aktuella bestånd, och man skall på enkelt sätt kunna revidera inte blott de enskilda dokumenten utan även själva dokumentbasens struktur (ungefär som när man flyttar om böckerna på hyllorna eller sorterar pappren i pärmarna).[4]

Litteraturbanken

Litteraturbanken <http://litteraturbanken.se> är ett pilotprojekt som finansierats av Riksbankens Jubileumsfond 2004–2005 och som bedrivs i samarbete mellan Svenska Akademien, Kungliga biblioteket och Språkbanken vid Göteborgs universitet (där jag är föreståndare).

På initiativ av författaren Sven Lindqvist arrangerade Svenska Akademien 2002 ett seminarium för att dryfta frågor om elektronisk lagring av litterära verk. Som ett resultat av seminariet uppdrogs till Johan Svedjedal, professor i litteraturvetenskap vid Uppsala universitet, att utreda förutsättningarna och formerna för en svensk litteraturbank. Utredningen – som finns tillgänglig på Litteraturbankens webbplats[5] – resulterade våren 2003 i en ansökan till Riksbankens Jubileumsfond om stöd för ett pilotprojekt för ”inrättandet av en Litteraturbank för att bevara och tillgängliggöra digitala versioner av främst skönlitterära verk, dels redan föreliggande, dels sådana som åstadkoms genom digitalisering av tryckta böcker”[6]. Ansökan beviljades, och arbetet med att förverkliga Litteraturbanken satte igång i januari 2004. Språkbanken har i pilotprojektet ansvarat för den digitala infrastrukturen i den framväxande Litteraturbanken: textimport, materialformat, funktioner för sökning i och presentation av materialet, samt webbplatsens utformning.

Den design av Litteraturbanken som växte fram under början av pilotprojektet bär många drag av Donald Broadys handbibliotek. Litteraturbanken skulle innehålla sådana ”självklara” funktioner som: möjlighet att ordna materiallistor enligt olika kriterier (titel, författare, period), tillgång till hypertextuellt

---

[4] Donald Broady, ”Det nya handbiblioteket” i Lars Höglund (red.), *Biblioteken, kulturen och den sociala intelligensen. Aktuell forskning inom biblioteks- och informationsvetenskap.* Forskningsrådsnämnden / Valfrid, Göteborg 1995 pp. 83–107. Ur html-versionen på webben 15/11 2005 <http://www.skeptron.ilu.uu.se/broady/dl/p-broady-9304.htm>

[5] Se Johan Svedjedal, ”www.litteraturbanken.se. Planer och principer”, *Samlaren* 125, 2004a pp. 237–263.

[6] Ur Svedjedals utredning.

kringmaterial om författare och författarskap, fritextsökning i verk och urval av verk, valbar visning av kommentarer och ordförklaringar i vetenskapliga utgåvor, möjlighet att spara de egna gränssnittsinställningarna, samt tillgång till verken i text-, pdf- och digitalt faksimilformat. Den digitala handbiblioteksnaturen hos projektet – liksom den uttryckta ambitionen att Litteraturbanken ska vara en högkvalitativ resurs för forskning och undervisning – syns tydligare i andra planerade Litteraturbanksfunktioner: excerpering ur texter med automatiskt åtföljande metadata, möjlighet för användaren att göra ”marginalanteckningar” i texterna, en antologifunktion för lärare, och slutligen det som egentligen är huvudtemat för denna uppsats och som jag återkommer till nedan, nämligen inbyggt textteknologistöd.

Litteraturbanken invigdes officiellt på Bok- och Biblioteksmässan i Göteborg i september 2005. Många av de planerade funktionerna fanns ännu inte vid invigningen och som all komplex mjukvara i början uppvisar Litteraturbanken en hel del buggar. Tillräckligt mycket var dock färdigt för att man skulle få en god bild av visionen som hade fött Litteraturbanken, och förmodligen av den anledningen togs den emot med entusiasm. Litteraturbanken permanentas från och med 2006 med initial finansiering huvudsakligen från Svenska Akademien. Det betyder att arbetet med att utveckla den kan fortsätta utan avbrott.

Språkbanken: Språkteknologisk korpuslingvistik

Språkbanken <http://spraakbanken.gu.se> kommer även fortsättningsvis att svara för den digitala infrastrukturen i Litteraturbanken. Språkbanken är formellt knuten till Institutionen för svenska språket vid Göteborgs universitet men har ursprungligen haft ett nationellt uppdrag att samla in, bearbeta och lagra texter samt erbjuda språkliga data till intresserade användare, alltså i första hand språkforskare men med tiden också i allt större utsträckning allmänheten. Genom Språkbankens försorg har dessa användare sedan 1970-talet kunnat få systematisk tillgång till språkliga och i viss mån statistiska data ur svenska texter av olika slag. Utnyttjandet av Språkbankens resurser har varit kostnadsfritt. Språkbanken uppfattas fortfarande i mångt och mycket som en nationell resurs, exempelvis i betänkandet *Mål i mun. Förslag till handlingsprogram för svenska språket*[7].

Språkbanken besitter idag en i Sverige unik kombination av kompetens inom områdena svenska textkorpusar, parallella (tvåspråkiga) textkorpusar, svenska maskinlexikon, samt språkteknologiska verktyg för korpushantering, korpusannotation och korpuspresentation, inom en stabil organisation med resurser att kontinuerligt underhålla och tillhandahålla språkliga material i elektronisk form.

[7] SOU 2002:27.

Idag fokuseras Språkbankens arbete främst på *förädling* av materialen, i och med att införskaffandet av stora mängder råtext numera i många fall är en tekniskt trivial uppgift. Det nya problemet är att göra jättematerialen hanterbara för (språk)forskning. Texterna måste vara lingvistiskt bearbetade och uppmärkta och göras sökbara i så många intressanta dimensioner som möjligt. Till detta behövs standardiserade lagringsformat, ett heltäckande system för sökning och en uppsättning språkteknologiska verktyg för att utföra automatisk uppmärkning av materialen eller åtminstone maximalt rationalisera manuell uppmärkning av material. För grundläggande annotering förfogar vi nu över en ”språkteknologisk verktygslåda” för morfosyntaktisk taggning (ordklass och böjningsform), lemmatisering (återförande av textord till deras grundform), ”grund” syntaktisk analys och så kallad namnigenkänning, samt för parallelltexterna meningslänkning (mellan en eller flera meningar i originalet och motsvarande mening[ar] i översättningen).

Den primära uppgiften för verktygen ovan är att tillföra information till textmaterialen för att öka deras vetenskapliga användbarhet; man talar generellt om annoterade texter (i motsats till obehandlad text). Ett fruktbart sätt att tänka på detta är i termer av *förädling*; den ”råa” texten förädlas (genom att man tillför den grammatisk och lexikalisk kunskap) och blir därmed användbarare för språkvetenskapliga ändamål. Man får möjlighet att söka i texter efter lingvistiskt relevanta enheter såsom t.ex. lemman (alltså lexikonformer) eller ordklasser/ böjningsformer/ satsdelar/ namn, snarare än efter (delar av) ortografiska ordformer. Den aktuella målsättningen för Språkbanken är att enspråkiga (åtminstone moderna svenska) texter som standard skall vara försedda med morfosyntaktisk uppmärkning, namnuppmärkning och i många fall också syntaktisk uppmärkning, medan parallelltexter skall vara meningslänkade. Systemutvecklingsverksamheten i Språkbanken inriktas därför på att jämka ihop material- och annotationslagringsformat med beaktande av internationellt standardiseringsarbete inom språkteknologiområdet, och att anpassa verktygens in- och utdataformat på motsvarande sätt. Visionen är av en Språkbank utan gränser, där alla resurser potentiellt kan kopplas ihop genom standardiserade, väldokumenterade format och gränssnitt för att möjliggöra många olika slags språkteknologisk och språkvetenskaplig forskning, men även för att ge den språkintresserade allmänheten en intressantare och mer mångfacetterad resurs.

Textteknologi för humanistisk forskning och utbildning

Vad har då detta med Litteraturbanken att göra? Jag vill hävda att Litteraturbankens verksamhet kan sägas utgöra ett specialfall av *användning av text* i humanistisk forskning och utbildning. Jag tror att den sortens textanvändning kan

berikas av de möjligheter som språkteknologin erbjuder för att skapa nya verktyg för forskning och undervisning i humaniora.

Textteknologi använder jag ofta som en sammanfattande benämning på (1) den språkteknologi där man främst arbetar med skrivet språk (vilket innefattar såväl skriftspråk som transkriberat talspråk), och där framträdande roller spelas av s*pråkteknologisk korpuslingvistik* (ett skilt forskningsområde från språkvetenskaplig korpuslingvistik) och (2) språkteknologisk *informationsförädling* (bl.a. informationssökning, informationsextraktion och dokumentsammanfattning).

Tematiken här kan sammanfattande karakteriseras som *texttillgänglighet*, alltså samma grundtanke som i Donald Broadys digitala handbibliotek. Vi kan se textteknologi som den senaste i en rad ”hjälpteknologier” som har uppstått vid sidan av de huvudteknologier som har hittats på för att göra text tillgänglig på nya sätt, t.ex. katalogiseringssystem i bibliotek och arkiv samt index av olika slag i tryckta böcker. När digital produktion av nya texter och digitalisering av äldre texter blir allt mer allmänt förekommande och det föredragna sättet att sprida text är via elektroniska media, kommer hjälpmedel för att finna, korrelera och presentera textinformation – alltså textteknologi – att ha en naturlig plats i alla sammanhang där text är viktig, t.ex. sådan vetenskaplig forskning där text (både dess innehåll och dess form) är objekt, såsom i de språkliga, litterära och historiska vetenskaperna. Textteknologi möjliggör ”dynamisk interaktion med en text”[8], något som man i viss utsträckning redan har tagit fasta på inom humanistisk forskning, så att man som jag redan nämnt ovan har börjat att manuellt föra in information (”annotationer”) i texter som ska ge användaren många samtidiga perspektiv på ett och samma textmaterial och även underlätta sökning och navigering i textmaterialet.

Vad textteknologin tillför är möjligheten att skapa denna information *automatiskt*, vilken inte ska underskattas, eftersom man då inte begränsar sig till de texter man redan av någon anledning har haft möjligheten att märka upp, utan kan arbeta med textmaterial i samma ögonblick som de blir digitalt tillgängliga. Det var den insikten som slog mig, språkteknologen, i samarbetet med Donald Broady, humanisten.

I det perspektivet finns det stora synergieffekter i att se till att Litteraturbanken och Språkbanken delar så mycket som möjligt av den digitala infrastrukturen. För det första finns det inget som hindrar att lagringsformaten för Litteraturbankens texter är ett som tillåter att man automatiskt tillför dessa texter språkvetenskapliga annotationer, nämligen således i stort sett samma format som Språkbankens korpusar. Det blir sedan en fråga om presentation ifall ett textmaterial ska vara att betrakta som ett litterärt verk i ”salongen” Litteraturbanken eller en (del av en) korpus i Språkbanken, åtkomlig via de underlig-

[8] Raymond G. Siemens, “A new computer-assisted literary criticism?”, *Computers and the Humanities* 36, 2002, pp. 259–267 (citat från p. 261).

gande ”köksregionerna”. Man har då vunnit enhetlighet i materialhanteringen och gett språkteknologer och språkvetare ett mer mångfacetterat textmaterial.

Men inte nog med det: Man har också samtidigt gett dem som huvudsakligen vill studera texterna ur en litterär synvinkel nya analysverktyg. Exempelvis borde den som är intresserad av *litterär onomastik*[9] vara betjänt av att automatiskt kunna få namn av en viss typ utmärkta i sin litterära text och den som vill utforska hur begrepp används i litteraturen – eller i historien som den speglas i litteraturen – skulle kunna ta sig in i texterna via Språkbankens lexikonresurser för att på en enda gång och automatiskt komma åt alla relevanta ord för en begreppslig utredning, säg hur djurmetaforer används i ett visst författarskap eller en viss tid. På det viset är tanken om Litteraturbanken och Språkbanken i förening ett steg i riktning mot förverkligandet av Donald Broadys vision om det digitala handbiblioteket.

[9] Johan Svedjedal, ”Almqvist och namnen. En studie i litterär onomastik”, *Samlaren* 125, 2004, pp. 52–77; Karina van Dalen-Oskam & Joris van Zundert, ”Modelling features of characters. Some digital ways to look at names in literary texts”, *Literary and Linguistic Computing* 19, 2004, pp. 289–301.