

Avtryck från WGLN-projekten i forskningen

Lars Borin

Inledning

Minnet är kortare än man tror. Det känns som en evighet sen som jag var aktiv i ULL i WGLN-samarbetet, men det har faktiskt inte gått mer än ett drygt decennium sen jag inbjöds att bidra till en ansökan till Knut och Alice Wallenbergs stiftelse om medel för att sätta igång verksamheten i det nytillkomna Swedish Learning Lab med Uppsala Learning Lab som en av modernerna.¹⁸¹ Ansökan lämnades in den 31 oktober 1999, och mitt bidrag – ett delprojekt med rubriken *Digital resources in the humanities* – fanns i ett litet hörn av den breda verksamhet som planerades för SweLL i samarbete med Stanforduniversitetet i USA.

Ansökan var antagligen väl beredd i förväg, eftersom Wallenbergstiftelsen behövde mindre än en månad på sig för att godkänna den, så att SweLL-arbetet kunde komma igång i början av år 2000, och därmed även arbetet med de digitala resurserna, under förkortningen DRHum.

WGLN utlyste så småningom (under 2001) ytterligare projektmedel, och Uppsala/ULL och KTH tillsammans med universiteten i Hannover (Learning Lab Lower Saxony – L3S) och Stanford formulerade en sedermera beviljad ansökan med titeln *Personalized access to digital learning resources* (PADLR), där jag deltog i två delprojekt: *Personalized Access to Large Text Archives* (PALaTe) och *Personalized Learning Sequences* (PLeSe).

PADLR skulle starta den 1/9 2001, och det starkaste minnet från det projektet är något som inte hade det minsta med själva projektet att göra: Den sista dagen av projektavsparksmötet i Hannover råkade vara den 11 september 2001, vars händelser kastade en mörk skugga över både mötesavslutningen och hemresan som skedde via en nu extremt hårdbevakad flygplats.

Slutligen var jag med och formulerade innehållet i ett delprojekt i en ansökan om ytterligare ett projektår inom PADLR, med titeln *Personalized learning with text from humanities open repositories/archives* (PLeTHORA).

¹⁸¹ Mitt mest varaktiga bidrag till Swedish Learning Lab var kanske mitt förslag om att använda förkortningen SweLL (istället för Sw-LL). Då tänkte jag förstås närmast på den här betydelsen: ”*adj* 17. *informal* stylish or grand 18. *slang* excellent; first-class” (från *Collins English Dictionary*, 10th ed. 2009, via <<http://dictionary.reference.com/browse/swell>>)

Även denna ansökan beviljades, men jag hann lämna min anställning i Uppsala och flytta över till Göteborgs universitet innan arbetet i fortsättningsprojektet satte igång, och eftersom Göteborgs universitet inte ingick i WGLN kunde jag tyvärr inte fortsätta mitt engagemang där.

Visionen: en DRHum-värld

Vad hoppades jag kunna åstadkomma i DRHum-projektet och i de efterföljande PALaTe- och PLLeaSe-projekten? Mitt forskningsområde var och är språkteknologi, som grovt uttryckt handlar om att få datorer och datorsystem att hantera mänskligt språk ungefär som man tänker sig att människor gör det, men dessutom kombinerat med datorernas förmåga att lagra och i hög hastighet bearbeta enorma mängder data.

Mycken (kanske huvuddelen) humanistisk forskning – och därmed även utbildning inom humaniora – använder text som primära forskningsdata. I takt med att alltfler textkällor från alla tider görs tillgängliga i digitalt format, blir det en intressant forskningsfråga hur dessa stora textmängder på bästa sätt ska kunna komma forskning och utbildning till gagn. En del av detta nyttiggörande tror jag handlar om att skapa effektiva och intelligenta verktyg för att söka och navigera i stora textarkiv, verktyg som är anpassade till den aktuella användargruppens behov och önskemål. Att åstadkomma sådana verktyg var i ett nötskal ett viktigt mål för DRHum- och PALaTe-projekten.¹⁸² PLLeaSe-projektet däremot var helt inriktat på utbildning i språk, även om textmaterial spelade en viss roll också i det projektet, eftersom vi bland annat utforskade hur man med automatiska metoder skulle kunna välja ut autentiska texter av lämplig svårighetsgrad för språkinlärares extensiva läsning.

Jag var redan då, för tio år sedan, övertygad om att språkteknologi skulle kunna bidra substantiellt till att skapa sådana verktyg för att arbeta med text i forskning och utbildning, så min vision om vad vi skulle kunna åstadkomma i WGLN-projekten var just det – intelligent åtkomst till stora textarkiv med hjälp av språkteknologi, för studenter och forskare i humanistiska ämnen. Vi kom inte särskilt långt när det gällde just den aspekten av projekten (se nästa avsnitt), men för min egen del kom arbetet i WGLN-projekten och ULL att bli början till en forskningsinriktning som jag har fortsatt att ägna mig åt sedan dess och som i ökande omfattning bär rik frukt (se avsnitt 4).

Mycket väsen för lite ULL?

Även om hoppet var att projekten skulle ge upphov till konkreta och varaktiga utbildningsapplikationer, så kom de i verkligheten snarare att bli

¹⁸² Ett annat, relaterat huvudmål handlade om digitala studentportföljer och hur sådana skulle kopplas till textarkiven.

explorativa och grundläggande för senare arbeten. Med tio års efterklokhet kan jag se att det nog inte kunde ha varit på något annat vis.

För det första: Det vi ville åstadkomma då kräver en omfattande infrastruktur – som bygger på standarder för dataformat, informationsstruktur och metadata – som vi visserligen diskuterade mycket redan i WGLN-projekten, men som vi inte hade några reella möjligheter att åstadkomma på egen hand. I själva verket är det först nu som en sådan infrastruktur börjar växa fram som ett resultat av internationella initiativ med oerhört mycket större resurser än våra tre små projekt i WGLN-samarbetet. Framför allt tänker jag här på ESFRI-infrastrukturprojektet CLARIN, som syftar till att bygga en europeisk IT-infrastruktur som ska möjliggöra just den sorts e-vetenskapstillämpningar för humaniora som jag drömde om i WGLN-projekten.¹⁸³ Därvid räknar man med att initialkostnaden för att åstadkomma en sådan grundläggande infrastruktur för ett enda språk kommer att ligga i storleksordningen 100–200 miljoner kronor, och då är inte ens kostnaderna för digitaliseringen av själva textarkiven inräknade.

För det andra: Deltagarna i WGLN-projekten hade förmodligen alltför olika utgångspunkter och därmed alltför olika tankar om vad som vore möjligt och önskvärt att åstadkomma. Våra diskussioner om tekniska standarder, till exempel, gick nog stundtals högt över huvudena på somliga av de medverkande företrädarna för de humanistiska disciplinerna. Det fanns också ett stort och – åtminstone enligt min uppfattning – aldrig överbryggt gap mellan dem med ett huvudsakligen tekniskt/vetenskapligt intresse av projekten (t.ex. jag själv) och dem som hade ett huvudsakligen pedagogiskt perspektiv på dem.

För det tredje: Den första WGLN-ansökan hade sprungit ur ett existerande samarbete som omfattade naturvetenskap, teknik och medicin. Det humanistiska inslaget kom in på sladden och begränsade sig till våra små projekt, som med några få procent av en inte överväldigande stor totalbudget inom WGLN skulle representera hela det humanistiska området. Till slut kom ”bara” en handfull humanistiska ämnen att bli representerade i projekten, men redan det lilla urvalet fragmenterade i praktiken de tillgängliga medlen så att varje inblandad individ därmed kunde avsätta enbart en liten del av sin arbetstid i ULL.

Det här betydde sammantaget att våra humanistiska WGLN-projekt huvudsakligen kom att röra sig inne i det befintligas fängelse, snarare än ute på visionernas vidder. Befintlig teknologi, snarare än nydanande tekniska lösningar och framväxande standarder – för vilka tekniska standardlösningar inte fanns än, och där projektbudgeten inte tillät egen nyutveckling – kom sålunda att karakterisera de (få) demonstrationsprototyper som togs fram inom projekten.

¹⁸³ Se <<http://www.clarin.eu>>.

På ett sätt kom dessa projekt alltså inte att motsvara mina visioner om vad de borde ha lett till. Det betyder dock alls inte att de var misslyckade. Dels gav de som sagt upphov till några demonstrationsprototyper, dels väckte de ett intresse hos flera av oss som medverkade i projekten för användningen av nydanande IKT i utbildning och forskning, som jag är övertygad har gett god utdelning om vi betraktar ett längre perspektiv än de få WGLN-projektåren.

Även under projekttiden åstadkom vi en del konkreta och solida resultat. I referensavsnittet nedan, under rubriken *Publikationer från WGLN-projekten*, återfinns de publikationer som DRHum-, PALaTe- och PLLeSe-projekten genererade under tiden de pågick. Tre internationella konferenspublikationer får ändå sägas vara ett gott utfall från en verksamhet med så pass blygsam budget som våra WGLN-projekt.

Men en del föll i god jord och bar frukt

Jag nämnde i föregående avsnitt att medverkan i WGLN-projekten väckte ett intresse hos bland annat mig för användningen av nydanande IKT i forskning och utbildning, framför allt i humanistiska ämnen. Sedan slutet av 2002 är jag föreståndare för Språkbanken vid Göteborgs universitet.¹⁸⁴ Språkbanken är en forsknings- och utvecklingsenhet med fokus på utveckling och tillhandahållande av så kallade språkresurser för i första hand svenska, i form av textkorpora och språkteknologiska lexikon samt språkteknologiska verktyg för att arbeta med resurserna. Språkbankens traditionella avnämargrupp är forskare i språkvetenskap och språkteknologi, men eftersom de flesta av våra resurser är fritt tillgängliga via webbgränssnitt har vi även många användare bland den språkintresserade allmänheten.

Sedan ett antal år tillbaka är vi alltmer inriktade även mot andra discipliner, sådana där text är en viktig primärkälla till forskningsdata. Här har absolut mitt tidigare engagemang i WGLN-projekten spelat en viktig roll. Det gav mig möjligheten att tänka koncentrerat på hur språkteknologi skulle kunna hjälpa forskare att använda stora textmängder mer effektivt som forskningsrådata. När jag därefter kom till Språkbanken hade jag alltså långt tänkta tankar om detta som jag sedan har försökt förverkliga och vidareutveckla inom de vidare ekonomiska och personella ramar som den miljön erbjuder.

Förverkligandet bygger i hög grad på en ständigt pågående utveckling av den grundläggande infrastrukturen i form av språkresurser och språkteknologiverktyg. Språkbanken är medlem av CLARIN (se ovan) och aktiv i den svenska standardiseringsorganisationen SIS spegelkommitté till ISO TC37, som svarar för internationell standardisering av format för språkresurser. Språkbanken utvecklar som sagt egna språkresurser, bland

¹⁸⁴ Se <<http://spraakbanken.gu.se>>.

annat språkteknologiska lexikon för olika stadier av äldre svenska, något som behövs för att skapa intelligenta verktyg för arbete med digitaliserade historiska textmaterial, t.ex. det stora äldre romanmaterialet i Litteraturbanken, vars tekniska infrastruktur har byggts upp och nu utvecklas och underhålls av Språkbanken.¹⁸⁵

Ett smakprov på den forskning som för min del har följt i spåren av mitt engagemang i WGLN-projekten ges under rubriken *Senare arbeten med rötter i WGLN-projekten* i referensavsnittet nedan. Där redovisar jag enbart publikationer där jag själv varit inblandad. Språkbankens verksamhet med språkteknologiskt stöd för forskning i olika discipliner är vidare än så, och den intresserade hänvisas till Språkbankens webbsidor för mer uttömmande information.¹⁸⁶

Slutord: Avtryck i forskningen – och i utbildningen

I tillägg till sitt traditionella samarbete med språkvetare av olika slag och språkteknologer har Språkbanken under senare år inlett samarbeten med forskare i bland annat historia, medicin och litteraturvetenskap om att utveckla e-vetenskapstillämpningar för att arbeta med digitala textarkiv i forskningen. Det är möjligt att detta skulle ha hänt även utan WGLN-projekten, men mindre troligt. Framför allt har WGLN-projekten på ett signifikant sätt berett vägen för dessa samarbeten.

Det är förvisso sant att fokus i vårt nuvarande arbete ligger på forskningsstöd, snarare än på stöd för högre utbildning, som ju var kärnan i WGLN-verksamheten. Samtidigt är vi helt övertygade om att en god högre utbildning bland annat kännetecknas av att forskningsmetodik och samma verktyg som används i forskningen även kommer in som ett naturligt inslag på alla nivåer i utbildningen. Av detta följer – och det är något som vi låter oss ledas av i vår forskning – att e-vetenskapstillämpningar bör utformas inte bara med tanke på att forskare ska kunna använda dem i sin forskning, utan även med tanke på att de ska kunna användas av studenter på olika nivåer i utbildningen. Därmed har våra WGLN-projekt inte bara lämnat avtryck i forskningen, utan faktiskt till och med kommit in i sin andra andning i Språkbanken.

¹⁸⁵ Se <<http://litteraturbanken.se>>.

¹⁸⁶ Se <http://spraakbanken.gu.se/swe/forskning>
<<http://spraakbanken.gu.se/swe/publikationer>>.

Referenser

Publikationer från WGLN-projekten

- Babić, Sanja, Camilla Bengtsson och Mattias Lingdell, "DIDAX – a system for online testing: technical documentation", *Reports from Uppsala Learning Lab* 4.2002. <<http://www.ull.uu.se/images/stories/rapporter/rull04.pdf>>
- Borin, Lars, "Where will the standards for Intelligent Computer-Assisted Language Learning come from?", Ingår i: *LREC 2002. Third International Conference on Language Resources and Evaluation. Workshop Proceedings. International standards of terminology and language resources management*, Las Palmas: ELRA, 2002. (Även som *Reports from Uppsala Learning Lab* 5.2002. <<http://www.ull.uu.se/images/stories/rapporter/rull05.pdf>>)
- Borin, Lars, Karine Åkerman Sarkisian och Camilla Bengtsson, "A stitch in time: Enhancing university language education with web-based diagnostic testing", Ingår i: *20th World Conference on Open Learning and Distance Education. The Future of Learning – Learning for the Future: Shaping the Transition. Düsseldorf, Germany*, 2001. (Även som *Reports from Uppsala Learning Lab* 1.2002 <<http://www.ull.uu.se/images/stories/rapporter/rull01.pdf>>)
- Larsson, Esbjörn, György Nováky och John Rogers, "DRHum in History – a status report", *Reports from Uppsala Learning Lab* 2.2002. <<http://www.ull.uu.se/images/stories/rapporter/rull02.pdf>>
- Nilsson, Kristina och Lars Borin, "Living off the land: The Web as a source of practice texts for learners of less prevalent languages", Ingår i: *Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation*, Las Palmas: ELRA, 2002.
- Sjunnesson, Jan, "Digital learning portfolios: inventory and proposal for Swedish teacher education", *Reports from Uppsala Learning Lab* 3.2002. <<http://www.ull.uu.se/images/stories/rapporter/rull03.pdf>>

Senare arbeten med rötter i WGLN-projekten

- Andréasson, Maia, Lars Borin, Markus Forsberg, Jonas Beskow, Rolf Carlson, Jens Edlund, Kjell Elenius, Kahl Hellmer, David House, Magnus Merkel, Eva Forsbom, Beáta Megyesi, Anders Eriksson och Sven Strömqvist, "Swedish CLARIN activities", Ingår i: *Proceedings of the Nodalida 2009 workshop on CLARIN activities in the Nordic countries*, Odense: NEALT, 2009.
- Borin, Lars, "Sparv i tranedansen eller fisken i vattnet? Språkteknologi och språklärande", Ingår i: Patrik Svensson (red.), *Från vision till praktik: Språkutbildning och informationsteknik* (Nätuniversitetet, Rapport 1 2006), Härnösand: NSHU, 2006.
- Borin, Lars, "Vi som går köksvägen: Språkteknologer och korpuslingvister i Litteraturbanken", Ingår i: Mikael Börjesson, Ingrid Heyman, Monica Langerth Zetterman, Esbjörn Larsson, Ida Lidegran och Mikael Palme (red.), *Fältanteckningar: Utbildnings- och kultursociologiska texter tillägnade Donald Broady*, Forskningsgruppen för utbildnings- och kultursociologi, ILU, Uppsala universitet, 2006.
- Borin, Lars och Markus Forsberg, "Something old, something new: A computational morphological description of Old Swedish", Ingår i: *LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, Marrakech: ELRA, 2008.

- Borin, Lars, Markus Forsberg och Dimitrios Kokkinakis, "Diabase: Towards a diachronic BLARK in support of historical studies", Ingår i: *Proceedings of LREC 2010*, Valletta: ELRA, 2010.
- Borin, Lars och Dimitrios Kokkinakis, "Literary onomastics and language technology", Ingår i: Willie van Peer, Sonia Zyngier och Vander Viana (red.), *Literary education and digital learning. Methods and technologies for humanities studies*, Hershey • New York: Information Science Reference, 2010.
- Borin, Lars, Dimitrios Kokkinakis och Leif-Jöran Olsson, "Naming the past: Named entity and animacy recognition in 19th century Swedish literature", Ingår i: *ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, 2007.
- Borin, Lars och Klas Prütz, "New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language", Ingår i: Guy Aston, Silvia Bernardini och Dominic Stewart (red.), *Corpora and language learners*, Amsterdam: John Benjamins, 2004.
- Borin, Lars och Anju Saxena, "Grammar, incorporated", Ingår i: Peter Juel Henriksen (red.), *CALL for the Nordic languages* (Copenhagen Studies in Language 30), København: Samfundslitteratur, 2004.
- Wittenburg, Peter, Nuria Bel, Lars Borin, Gerhard Budin, Nicoletta Calzolari, Eva Hajicova, Kimmo Koskenniemi, Lothar Lemnitzer, Bente Mægaard, Maciej Piasecki, Jean-Marie Pierrel, Stelios Piperidis, Inguna Skadina, Dan Tufis, Remco van Veenendal, Tamás Váradi, Martin Wynne, "Resource and service centres as the backbone for a sustainable service infrastructure", Ingår i: *Proceedings of LREC 2010*, Valletta: ELRA, 2010.