

Minnesanteckningar - BLARK – Basic Language Resources Kit för svenska GSLT retreat workshop 2006-01-29

Organisatörer:

Anna Sågvall Hein

Eva Forsbom

En BLARK är en definition, inventering och instansiering av en uppsättning grundläggande språkresurser och -verktyg som kan användas inom forskning, industri och undervisning.

1. Resurser

Ingående komponenter (för såväl text som tal) är t.ex.

- korpusar (en- och flerspråkiga)
- terminologiresurser
- lexikala resurser (en- och flerspråkiga)
- grammatiska resurser
- grundläggande verktyg: moduler (t.ex. taggare, parsrar, grafem-till-fonem-omvandlare)
- riktmärkesresurser för utvärdering
- annotationsstandarder och -verktyg.

En viktig fråga är hur dessa resurser ska hänga ihop. Vi bör noggrant fundera på vilka annotationsstandarder vi bör ha och vid avvikelser specificera dem så att man på ett enkelt sätt kan göra konverteringar. Vi bör även fundera över om och hur resurserna ska paketeras till större sammanhängande komponenter. Det är också viktigt att beakta huruvida de enskilda komponenterna har klart avgränsade funktioner och väldefinierade gränssnitt, så att kommunikationen mellan de olika komponenterna underlättas.

2. Tillgänglighet

Den svenska BLARK:en bör vara tillgänglig för en bred publik (akademin, industrin och privatpersoner) och vara en sökbar resurs både på engelska och svenska.

Lämpliga verktyg för sökning och distribution kan vara t.ex. OLAC, som används av bl.a. LDC och Linguist List, eller dess europeiska motsvarighet IMDI (där Marcus Uneson, Lund, är med i styrelsen).

Vid inventering bör resurserna graderas efter tillgänglighet (baserat på exempelvis kostnad, insyn, användningsfrihet och standardisering). Resurser som saknas eller har låg tillgänglighet ska sedan utvecklas i prioritetsordning, med så hög tillgänglighet som möjligt. Vi bör se till att de ingående komponenterna är kvalitetssäkrade. Det bör också finnas kvalitetsgaranti för uppgraderingar.

För ett flertal språkliga resurser som används idag, t.ex. Stockholm-Umeå-korpusen och KTH News, finns särskilda licensavtal som gör att de inte kan släppas helt fria. I allmänhet är det extra svårt med skönlitterära texter och tidningstexter, lexikon samt radio- och tv-program med inspelat talmaterial. En möjlighet är att överväga att göra om dessa resurser för att göra dem mer tillgängliga. Det största problemet utgörs oftast av upphovsrättsavtal med förlag och liknande; ett arbete som är tidskrävande

och allmänt betungande. En gemensam satsning på en nationell språkresursuppsättning där den nya språkmyndigheten ansvarar för underhåll och liknande skulle underlätta förhandlingsläget med förlagen.

3. Tillhandahållande av resurser och verktyg

För att det skall uppstå intresse för att tillhandahålla resurser och verktyg till en svensk BLARK bör vi hitta vägar för att uppmuntra folk att göra det.

En idé är att vi släpper en BLARK-instansiering under en gemensam publikation. Den ursprungliga idén är att de som tillhandahåller resurser och verktyg gör det på sina egna villkor, och att villkoren anges i BLARK-specifikationen.

4. Arbetssätt

ELRA/ELDA har ett verktyg där man i form av matriser fyller i vilka språkliga resurser som bör finnas för språket i fråga. Ett förslag är att använda verktyget som utgångspunkt och anpassa det för svenska. Vi bör också ta del av erfarenheter från tidigare BLARK-projekt för nederländska/flamländska och arabiska. Liknande ansatser utförs också inom några LDC-projekt, t.ex. för språk som undervisas i mindre utsträckning, samt för något mer avancerade resurser för engelska, arabiska och kinesiska.

5. Finansiering och organisation

Det finns möjlighet att söka pengar för att utveckla en BLARK för svenska under Vetenskapsrådets satsning på infrastruktur, där Merle Horne (Lund) representerar humaniora. Ett planeringsbidrag kan sökas från Vetenskapsrådet för att utarbeta en gemensam ansökan till ett stort projekt. Ansökan om planeringsbidraget ska vara inne i början av april (19/4?).

På mötet bestämdes att vi ska bilda en grupp som utarbetar en gemensam ansökan om planeringsbidraget. Gruppen ska bestå av en representant per lärosäte inom GSLT (och representanter för industrin? och språkmyndigheten?).

Följande personer samlades för ett första kort strategimöte:

Lars Ahrenberg (Linköping)
Lars Borin (Språkdata/Göteborg)
Rolf Carlson (TMH/KTH)
Eva Forsbom (Uppsala)
Beata Bandmann Megyesi (Uppsala)
Joakim Nivre (Växjö)
Bengt Nordström (Chalmers)
Aarne Ranta (Chalmers)
Anna Sågvall Hein (Uppsala)
Marcus Uneson (Lund)

Anna Sågvall Hein tog på sig uppgiften att skriva ett utkast till en gemensam ansökan som gruppens medlemmar sedan utökar. I ansökan ska följande särskilt påpekas:

- Det finns en stor brist på språkteknologiska resurser i allmänhet, och för svenska i synnerhet.

- Det finns en upparbetad samarbetsorganisation för språkteknologi inom Sverige som är lämpad att utföra uppgiften.
- Återanvändningsbarhet för resurserna ska prioriteras.
- Gruppen har en stark vilja som manifesterats i ett flertal tidigare ansökningar, uppvaktningar och liknande (t.ex. uppvaktning av departement, Mål i mun-skrivelse, Bästa språket-proposition, och ansökan om trädbanksprojekt).

Referenser:

Bästa språket – en samlad svensk språkpolitik. Prop. 2005/06:2. URL

<http://www.regeringen.se/sb/d/108/a/50761>.

ELDA (Evaluations and Language resources Distribution Agency). URL <http://www.elda.org/blark/>.

E. Forsbom. Återanvändbarhet för språkvara, 2004. URL

<http://stp.lingfil.uu.se/~evafo/gslt/java/paper.ps.gz>. Literature review for GSLT Java course.

IMDI (ISLE Meta Data Initiative). URL <http://www.mpi.nl/IMDI/>.

S. Krauwer. ELSNET and ELRA: A common past and a common future. ELRA Newsletter, 3(2), 1998.

URL <http://www.elda.org/blark/fichiers/elsnet&elra.doc>.

S. Krauwer, B. Maegaard, K. Choukri, and L. Damsgaard Jørgensen. Report on BLARK for Arabic, 2004. URL <http://www.nemlar.org/Publications/index.htm>.

LDC (Linguistic Data Consortium). URL <http://www ldc.upenn.edu/>.

V. Mapelli and K. Choukri. Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps. ENABLER Deliverable 5.1, 2003. URL

<http://www.enabler-network.org/reports.htm>.

Mål i mun – Förslag till handlingsprogram för svenska språket. SOU 2002:27. URL

<http://www.regeringen.se/sb/d/108/a/1443>.

OLAC (Open Language Archives Community). URL <http://www.language-archives.org/>.

H. Strik, W. Daelemans, D. Binnenpoorte, J. Sturm, F. de Vriend, and C. Cucchiarini. Dutch HLT resources: From BLARK to priority lists. In Proceedings of ICSLP, Denver, USA, pp. 1549-1552, pages 1549 1552, Denver, USA, 2002. URL