Språkbanken

Årsrapport 2010

ÖVERSIKT

I DENNA ÅRSRAPPORT redovisas merparten av de språkteknologiverksamheter som bedrivs vid institutionen för svenska språket. I vårt identitetsarbete för vi samman dessa verksamheter under rubriken ”Språkbanken” (1) för att markera att de aktiviter som beskrivs här hör så nära ihop att de i praktiken utgör en forskningsmiljö, samt (2) därför att Språkbanken är en väl inarbetad benämning (som går tillbaka till 1970-talet) på dessa aktiviteter.

När vi talar om Språkbankens verksamhet menar vi således en språkteknologiforskningsverksamhet vid institutionen för svenska språket som finansieras av fakulteten under benämningarna *Språkbanken* och *språkvetenskaplig databehandling* (forsknings- och forskarutbildningsämne som deltar i grundutbildningen inom masterprogrammet i språkteknologi), samt ett antal externfinansierade forskningsprojekt inom språkteknologi.

KONFERENSER/PRESENTATIONER

Under denna rubrik har vi samlat den viktiga del av vår utåtriktade verksamhet som består i att vi presenterar vårt arbete i form av föredrag eller posterpresentationer vid konferenser och workshopar med öppen inbjudan att inkomma med bidrag (inklusive sådana som vi själva arrangerar eller är med om att arrangera). Inom området språkteknologi är sådana konferenser huvudpublikationskanalen. Bidrag skickas först in i fulltext (typiskt 4–8 sidor) och bedöms i normalfallet dubbelt anonymt (anonymiserade bidrag bedöms av anonyma granskare) av två till tre fackgranskare. Sådana presentationer återfinns i nästa avsnitt.

En annan viktig del av den utåtriktade verksamheten består i presentationer efter inbjudan eller egen anmälan vid seminarier, projektmöten o.dyl. Huvudskillnaden mot föregående kategori är avsaknaden av explicit fackgranskning. Naturligtvis avspeglar mängden och den geografiska spridningen av den här sortens presentationer ändå forskargruppens aktivitet och rykte. Dessa redovisas i avsnittet *Andra presentationer* nedan.

Konferens/workshop (plats)/månad (antal presentationer)

- NFL – Nordisk förening för lexikografis årliga symposium (Schæffergården, Köpenhamn)/januari (1),
- XML-Prague eXist-db workshop (Prag)/mars (1)
- LREC – International Conference on Language Resources and Evaluation (Malta)/maj (4),
- Developing multidimensional methods for vocabulary assessment (Stockholm; arrangör)/maj (1),
- FOT – Forum för textforskning (Lund)/juni (1)
- Second Louhi Workshop on Text and Data Mining of Health Documents, NAACL-HTL Workshop (Los Angeles)/juni (3)
- EURALEX – European Conference on Lexicography (Leeuwarden)/juli (1),
- SCL – Scandinavian Conference of Linguistics (Joensuu)/augusti (1),
- 2010 års nationella termkonferens: Professionen i språket -. språket i professionen (Skövde)/september (1),
- BioSEPLN Workshop on Language Technology applied to biomedical and health documents (Valencia)/september (1)
- IDS-WS (Göteborg; arrangör)/oktober (1),
- Computational approaches to synonymy (Helsingfors)/oktober (1),
- SLTC – Swedish Language Technology Conference (Linköping)/oktober (3),
- Readability and Multilingualism (i samband med SLTC, Linköping; arrangör)/oktober (1),
- Svenska Läkaresällskapets riksstämma (Göteborg)/december (1)

Andra presentationer (plats)/månad

- GSLT – Swedish National Graduate School of Language Technologys årliga retreat (Gullmarsstrand)/januari
- GU-online (Göteborg)/april
- Stockholms universitet, Lingvistik (Stockholm)/april
- Språkbanken för Litteraturbankens styrelse (Göteborg)/september
- Interedition thinktank och COST-A32 (München)/september
- CLARIN (Nijmegen)/september
- NEERI/D-SPIN Workshop (Wien)/oktober
- Glossa (Göteborg)/juni
- Portalens gymnasium (Göteborg)/december
- CLT – Centre for Language Technologys årliga workshop (Gullmarsstrand)/november
- CLT-seminarier (Göteborg)/återkommande under året

PROJEKT

Språkbankens arbete utförs inom en mängd olika projekt. Några av dessa projekt är interna, medan flertalet innebär samarbeten på nationell eller internationell nivå. Här redovisar vi de forskningsprojekt som vi deltog i under året.

- AO – en akademisk ordlista för svenska (år 1); finansiär ISA
- CLARIN (år 3 av 4); EU-projekt med egen finansiering
- CLT:s textteknologilaboratorium (år 2 av 4); finansiär CLT (Göteborgs universitet)
- CONPLISIT (år 1 av 2); egen finansiering
- Digital areallingvistik (år 1 av 3); finansiär VR
- Effektiv informationsförädling i sjukvården (år 1 av 2); finansiär VGR regionala utvecklingsmedel
- Framtidssäkring av Språkbanken (år 3 av 3); finansiär VR
- GRUS – grundskoleelevers skrivande i ”en dator till en elev”-satsning (år 1); finansiär LUN/ISA
- IKT i lärarutbildningen, anknytning mellan forskning och undervisning, datorbaserad textanalys och ordförrådsbedömning (år 1); finansiär LUN
- KELLY (år 1 av 2); finansiär EU
- Kvalitetssäkring av SNOMED CT (år 2 av 2); finansiär Socialstyrelsen
- Litteraturbanken (permanent); finansiär Svenska Akademien, Kungl. Vitterhetsakademien
- MOLTO (år 1 av 3); finansiär EU
- Mätning av ordförrådet i andraspråket (år 2), finansiär RJ
- SPF (DB1800) (år 2 av 3); finansiär VR
- SweFN++ (år 1); finansiär CLT (Göteborgs universitet)

Ett antal projektansökningar har inlämnats för pågående och kommande projekt.

- RJ: SweFN++ (avslagen i första omgången)
- VR/DISC: SweFN++ (beviljad för 2011–2013)
- EU: META-NORD (partner) (beviljad för 2011–januari 2013)
- EU: CLARICLE (partner) (avslagen)
- VR: A person-centred communication and information intervention for patients undergoing colorectal cancer surgery (avslagen)
- Söderbergs stiftelser: Läkartidningens arkiv (avslagen)

INFRASTRUKTUR

Språkbanken arbetar aktivt för att utveckla en språkteknologisk infrastruktur. Detta arbete omfattar i dagsläget följande komponenter:

- lexikal infrastruktur,
- korpusinfrastruktur, samt
- metadata.

Uppbyggnaden av infrastruktur för lexikala resurser och korpora handlar för närvarande om att harmonisera och standardisera så många fria språkliga resurser som möjligt, samt att göra dem tillgängliga för forskningsvärlden. Vi skapar dessutom verktyg för att utforska dessa resurser, t ex SBLEX för lexikala resurser och Korp för korpora. Målet är vidare att alla dessa resurser ska vara väl beskrivna i ett metadata-repositorium som följer de standarder som finns inom området.

Öppenhet är ett av våra ledord, en filosofisk ståndpunkt vi försöker att tillämpa i största möjliga mån. Vi anser att forskning ska utföras öppet, för att tillåta granskning och samarbete. I denna öppenhet ingår att använda öppna standarder och licenser samt att använda och skapa verktyg med öppen källkod. Även om alla äldre resurser inte kan göras fritt tillgängliga, p.g.a. restriktiva licenser, strävar vi efter att samla fler och större fria resurser, för att främja språklig forskning och utveckling av språkteknologiska tillämpningar, i Sverige och världen.

SAMVERKAN

En central del i samverkan med det omgivande samhället består i att vi är representerade i relevanta externa organisationer. Här redovisas denna medverkan.

- SND:s vetenskapliga rådgivargrupp – Svensk Nationell Datatjänst
- Litteraturbankens styrelse
- Språkrådets rådgivargrupp för språkteknologi
- SIS TK115 – Swedish Standards Institute, Terminologi och språkliga resurser
- CLT:s styrgrupp – Centre for Language Technology
- GSLT:s ledningsgrupp – Graduate School of Language Technology

Utöver medverkan i dessa organisationer samarbetar vi, genom Språkbanken som helhet eller genom projekt, med följande organisationer, institutioner, och företag.

Nationella samarbeten

- Gothia Forum för klinisk forskning
- GPCC – Centrum för personcentrerad vård, Göteborgs universitet
- Institutet för svenska som andraspråk, Göteborgs universitet
- Institutionen för historiska studier, Göteborgs universitet
- Lexikaliska institutet, Göteborgs universitet

- Sahlgrenska universitetssjukhuset
- Socialstyrelsen
- Uppsala universitet

Internationella samarbeten

- Berkeley FrameNet
- Centrum för Internationalisering och Parallellspråklighet, Köpenhamns universitet
- CLARIN – Common Language Resources and Technology Infrastructure
- Kelly – Keywords for Language Learning for Young and adults alike
- Max Planck-institutet för evolutionär antropologi
- META-NORD – The Multilingual Europe Technology Alliance, specifically the Baltic and Nordic parts
- MOLTO – Multilingual Online Translation
- Tekstlaboratoriet, UiO – Universitetet i Oslo

BESÖK

I kategorin besök återfinns två typer av aktiviteter: (1) besök inom ramen för institutionella informationsutbyten och (2) besök av gästforskare eller seminariegäster. Dessa två kategorier redovisas under separata rubriker nedan.

Institutionella informationsutbyten

- Nasjonalbiblioteket/Norsk språkbank (juni)
- SLS/Finlandssvensk korpus (juni)

Gästforskare/seminariegäster

- Bernard Comrie, Max Planck-institutet för evolutionär antropologi/California Santa Barbara-universitet
- Kalervo Järvelin, Tammerfors universitet
- Adam Kilgarriff, Lexical Computing Ltd
- Mats Lundälv, DART
- Paul Rayson, Lancasters universitetet
- Josef Ruppenhofer, Saarlands universitetet
- Srikant Sarangi, Cardiffs universitetet
- Stefan Schulz, Albert-Ludwigs-universitetet i Freiburg
- Mike Scott, WordSmith Tools
- Serge Sharoff, Leeds universitet

BEDÖMNINGSUPPDRAG

Utöver deltagande med egna presentationer vid konferenser och workshopar deltar vi även som fackgranskare av andra bidrag. Här redovisas sådana bedömningsuppdrag.

- Second Louhi Workshop on Text and Data Mining of Health Documents (NAACL-HLT Workshop), Los Angeles
- SLTC 2010 – Swedish Language Technology Conference, Linköping
- LREC – The 7th Language Resources and Evaluation Conference, Malta
- SemEval-2010 – The 5th International Workshop on Semantic Evaluation (ACL Workshop), Uppsala
- IceTAL, Reykjavík

NY PERSONAL

Under året har Språkbanken anställt flera nya medarbetare.

- Elena Volodina (forskningsingenjör; vikariat 6+6 mån)
- Martin Hammarstedt (systemutvecklare; vikariat 6 mån)
- Jonatan Uppström (systemutvecklare; vikariat 6 mån)
- Ivanka Ivanova (projektassistent; 2 mån)

PUBLIKATIONER

UTBILDNING

Grundläggande och avancerad nivå

- Inlärarkorpusar i forskning och undervisning (masternivå) ht 2010 (kurstillfälle)
- Introduction for Computer Scientists, MLT (masternivå) ht 2010
- Introduction for Linguists, MLT (masternivå) ht 2010
- Korpuslingvistiska metoder och verktyg (master- och forskarnivå) vt 2010 (kursansvar och examination)
- Overview lecture: Information retrieval, MLT (masternivå) ht 2010 (kurstillfälle)
- Overview lecture: Language technology resources, MLT (masternivå) ht 2010 (kurstillfälle)
- Workshop on WordSmith Tools endagskurs/tutorial i Göteborg, ca 30 deltagare, November (arrangör)
- Handledning av kandidat-, magister- och masterarbeten

Forskarutbildning

- Kurser:
  - Datamaskinell lexikologi (forskarnivå) ht 2010 (kursansvar och examination)
  - Korpuslingvistiska metoder och verktyg (master- och forskarnivå) vt 2010 (kursansvar och examination)
- Handledning:
  - Språkvetenskaplig databehandling (5 doktorander)
  - Svenska som andraspråk (1 licentiand)