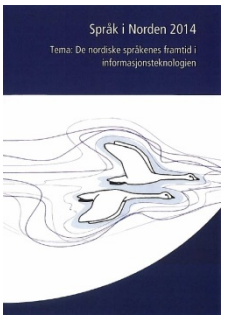
Sprog i Norden

Titel: FIN-CLARIN – en humanistisk forskningsinfrastruktur med betoning på språk

Forfatter: Krister Lindén

Kilde: Sprog i Norden, 2014, s. 126-132

URL: http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive



Betingelser for brug af denne artikel
Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Søgbarhed
Artiklerne i de ældre numre af Sprog i Norden (1970-2004) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

FIN-CLARIN – en humanistisk forskningsinfrastruktur med betoning på språk

Krister Lindén

Miljardvis med ord och tusentals timmar med audio och video behövs som material för humanistisk forskning och i synnerhet språkforskning. Dessutom behöver forskarna redskap för att förädla och jämföra sina egna datasamlingar med allmänna datasamlingar. När ett forskningsprojekt är slut behövs det lagrings- och spridningsplatser för att göra rådata, redskap och forskningsresultat tillgängliga och användbara. Data, redskap och gemensamma användningsmöjligheter bildar tillsammans en forskningsinfrastruktur, som gör det möjligt att verifiera tidigare resultat och effektivare göra nya rön, när alla inte behöver starta från noll med att samla data och bygga analysredskap.

FIN-CLARIN – en forskningsinfrastruktur

FIN-CLARIN är en forskningsinfrastruktur som tillhandahåller språkdata och språkredskap för humanistiska forskare. FIN-CLARIN är en nationell nod i Finland för CLARIN ERIC[1] (European Research Infrastructure Consortium) – den europeiska infrastrukturen för humanistisk forskning, som grundades den 29 februari 2012. Genom att samarbeta med andra inom CLARIN ERIC, kan FIN-CLARIN fokusera på de två officiella språken i Finland, dvs. finska och svenska, även om FIN-CLARIN också tillgängliggör andra språks resurser producerade av forskare i Finland.

FIN-CLARIN[2] är ett nationellt konsortium av universitet och forskningsinstitut i Finland. FIN-CLARIN är en distribuerad forskningsinfrastruktur, där de olika konsortiemedlemmarna bidrar med språkdatasamlingar och språkredskap från sina olika styrkeområden. FIN-CLARIN koordineras av Helsingfors universitet, som samlar in allmänt tillgängliga språkdataresurser och som i Finland har specialiserat sig på att bygga

1 www.clarin.eu
2 www.helsinki.fi/finclarin

språkteknologiska analysredskap. En viktig partner i samarbetet är CSC – Tieteen tietotekniikan keskus Oy[3], som är ett IT-center för vetenskap, ägt av staten och administrerat av undervisnings- och kulturministeriet. CSC tillhandahåller en gemensam plats för lagring och sökning i dataresurser och användning av språkredskap. En annan viktig partner är Institutet för de inhemska språken[4], som upprätthåller omfattande språkliga arkiv och samlingar.

FIN-CLARIN är en virtuell forskningsinfrastruktur som erbjuder två centraliserade tjänster: Språkbanken i Finland[5] (Kielipankki) och Vetenskapstermbanken i Finland[6] (Tieteen termipankki). För att rationalisera utvecklingen av språkresurser för svenska har Språkbanken i Finland speciellt värnat om samarbetet med Språkbanken i Göteborg, medan Vetenskapstermbanken i Finland samarbetar med Rikstermbanken i Sverige. Forskarna i samiska vid Universitetet i Tromsø använder redskap utvecklade av FIN-CLARIN och har börjat utveckla sin egen terminologiplattform med utgångspunkt i Vetenskapstermbanken i Finland. Förbundet för hörselskadade i Finland utvecklar en videokorpus och en videobaserad ordbok för teckenspråk där de kombinerar videosökredskap från Språkbanken i Finland och lexikografiska redskap från Vetenskapstermbanken.

FIN-CLARINs förberedelser inför CLARIN ERIC

FIN-CLARIN har förberett sig för Finlands inträde i CLARIN ERIC genom följande åtgärder:

1. Det autentiserings- och auktoriseringssystem som krävs för att få tillgång till språkdata inom CLARIN ERIC har implementerats av CSC med delfinansiering från Helsingfors universitet och det har testats av Finland, Tyskland och Holland i form av en tjänstefederation som kallas CLARIN SPF för att alla CLARIN-center ska få tillgång till varandras tjänster.
2. FIN-CLARIN erbjuder forskningsmaterial till hundratals forskare och tusentals studerande i Finland och många fler i hela Europa. CLARIN

3 www.csc.fi
4 www.sprakinstitutet.fi
5 www.kielipankki.fi
6 www.tieteentermipankki.fi

och FIN-CLARIN erbjuder öppna sökportaler[7] för språkdata i Europa och för att informera användare om vilka språkresurser som finns, var de finns och på vilka villkor de kan användas. En del material kan användas genom sökgränssnitt som Korp[8] för textmaterial och LAT[9] för audio- and videomaterial.

3. Licenstemplat för deponering och användning av språkresurser har skapats av Helsingfors universitet och de kan användas inom hela CLARIN (Oksanen & al., 2010).
4. CLARIN skapade en pilot för gemensam användning av språkredskap och språkdata som befinner sig på olika CLARIN-center genom en molntjänst kallad Weblicht. Språkredskap placerade i Finland deltog också pilotprojektet[10].
5. FIN-CLARIN har skapat nya multifunktionella språkresurser så som den syntaktiskt analyserade finska textsamlingen FinnTreeBank (Voutilainen & al., 2012) och den finska synonymordboken FinnWordNet (Lindén & Carlson, 2009; Lindén & Niemi 2013) samt skapat redskap med HFST – Helsinki Finite-State Technology (Lindén & al., 2009, 2011, 2013) för att bearbeta språkdata, vidareutveckla existerande och integrera nya språkmaterial i Språkbanken i Finland.
6. För att utvidga språkmaterialet i Språkbanken i Finland har FIN-CLARIN ingått ett licensavtal med Kopiosto beträffande digitaliserade material i Finlands nationalbibliotek. Materialet omfattar nu cirka 5 miljarder ord finskspråkiga tidskrifter från åren 1820–1940 och cirka 3 miljarder ord svenskspråkiga tidskrifter från åren 1750–1940.
7. FIN-CLARINs personal vid Språkbanken i Finland inom Helsingfors universitet och CSC har rest runt till språkresurskonsortiet FIN-CLARINs medlemsorganisationer och gett råd om hur man kan använda språkredskap och standarder i forskningsprojekt för att skapa resurser som kan återanvändas inom CLARIN.
8. Språkbanken i Finland har integrerats med Scientist's User Interface vid CSC[11] för att skapa en plattform för forskare att bearbeta och distribuera språkdata.
9. FIN-CLARIN ordnar möten med sin styrgrupp ungefär var tredje må-

7 www.clarin.eu/vlo/ och metashare.csc.fi
8 korp.csc.fi
9 http://lat.csc.fi
10 https://weblicht.sfs.uni-tuebingen.de/
11 https://sui.csc.fi

nad för att informera om den senaste utvecklingen och samla feedback och utvecklingsförslag. Dessutom har FIN-CLARIN årligen organiserat kurser vid CSC med avsikt att träna nya forskare i att använda forskningsinfrastrukturen och få konkreta exempel på hur infrastukturen används i praktiken.

Målsättningen för CLARIN och FIN-CLARIN

CLARIN är en virtuell infrastruktur som erbjuder en ny plattform för humanistisk forskning, där en forskare kan arbeta vid sin egen arbetsstation, hitta språkmaterial i enorma datasamlingar, friktionsfritt få de nödvändiga tillstånden och påbörja sin forskning utan fördröjning.

CLARIN grundar sig på nationella initiativ som producerar språkmaterial för lagring i pålitliga CLARIN-center. Språkresurserna består av språkmaterial såsom maskinellt läsbara texter, lexikon, terminologier, digitaliserade inspelningar av talspråk och redskap för att bearbeta, söka och förädla sådant språkmaterial. Målet med CLARIN är att lösa tre problem som för tillfället hindrar forskare att till fullo utnyttja språkmaterial:

1. Digitalt språkmaterial existerar vanligen, men användarna kan inte hitta det. Detta kan lösas med genom att man erbjuder metadata för språkresurserna som kan samlas in i en gemensam databas. Federerad sökning, dvs. en innehållssökning i samlingarna på alla datacenter inom CLARIN, förbättrar ytterligare möjligheterna att hitta lämpliga språkresurser.
2. När relevant material hittas, är det inte alltid lätt att veta hur man ska få tillstånd att använda det. Detta kan lösas genom att kategorisera och standardisera användarlicenserna och erbjuda potentiella användare en standardiserad autentiseringsmetod och rättighetsinnehavarna en gemensam auktoriseringsmetod.
3. Med behörigt användartillstånd är alla delar av materialet ändå inte alltid kompatibla med varandra eller med de redskap som finns. Detta kan lösas med standardiserade dataformat, gemensamma applikationsgränssnitt och harmoniserad terminologi.

Med nya redskap kan många problem, som tidigare tog flera veckor att lösa, klaras upp på mindre än en timme eller rentav på några minuter. Många påståenden som tidigare baserade sig på intuition kan baseras

på mer objektiva och välunderbyggda fakta. Nya regelbundenheter och undantag är lättare att upptäcka än med traditionella redskap.

Forskningen inom de humanistiska vetenskaperna kommer också att bli lättare att bekräfta när sådana språkmaterial som argumenten baserar sig på är tillgängliga för andra forskare och påståendena kan verifieras eller falsifieras.

FIN-CLARIN kommer i praktiken genom Språkbanken i Finland och Vetenskapstermbanken i Finland att betjäna hundratals forskare och tusentals studerande i Finland och många fler i resten av EU och världen. Användningen av standardiserad terminologi via Vetenskapstermbanken i Finland kommer att förbättra den vetenskapliga och tvärvetenskapliga diskussionen och rapporteringen om forskningsresultat både på nationell och på internationell nivå. Vetenskapliga upptäckter innebär också terminologiska innovationer som lätt kan förmedlas via termbanken.

Nya typer av forskning kommer att bli möjliga, då många av de nuvarande metoderna för språkbehandling kräver flera miljarder ord för att sammanställa pålitlig statistik för många av sina analysmetoder.

Tillgängligheten på tillräckliga språkresurser är nödvändig för att bygga adekvat språkteknologi för ett språk. Som påvisades av en internationell förfrågan inom META-NET[12], finns det alarmerande små språkresurser[13] för finska jämfört med många grannländer för att inte tala om större länder som Tyskland, Frankrike och Storbritannien. Målet med FIN-CLARIN är att underlätta situationen. (För ytterligare information jämför t.ex. Finnish Language White Paper, Koskenniemi & al., 2012).

Implementering av den internationella forskningsinfrastrukturen

FIN-CLARIN är en del av den europeiska CLARIN-infrastrukturen som bygger ett europeiskt nätverk av center som erbjuder språkmaterial för humanisterna och språkforskarna i Europa. CLARIN var med på den första ESFRI vägkartan med 34 infrastrukturer som skulle byggas. Ett pilotprojekt under 2008–2011 förberedde en mera permanent CLARIN ERIC för att konstruera och driva nätverket av CLARIN center. CLARIN ERIC grundades den 29 februari 2012 med den Europeiska kommissionen beslut 2012/136/EU. För tillfället är Bulgarien, Danmark, Estland, Holland,

12 http://www.meta-net.eu/
13 http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

Polen, Tjeckien, Tyskland och Österrike samt den holländska språkunionen medlemmar och Norge är observatör. Finland avser att gå med under 2014. Detta kräver en initialsatsning på 5 år för att utveckla och upprätthålla CLARIN ERIC och FIN-CLARIN som dess nationella nod.

Dr Krister Lindén är forskningsdirektör vid FIN-CLARIN och nationell koordinator för CLARIN i Finland.

Summary

The virtual distributed research infrastructure FIN-CLARIN is the Finnish national node of CLARIN ERIC. FIN-CLARIN provides two centralized services: the Language Bank of Finland and the Bank of Finnish Terminology in Arts and Sciences. The Language Bank of Finland makes available collections of digital language resources and tools for analyzing them via its on-line service center. It serves a wide research community of humanists as well as social scientists and computer scientists. Its relevance lies in the amount and diversity of materials as well as in the seamless access provided to researchers. The goal is to make available collections from different periods, genres and regions as well as different modalities such as text, audio, pictures and video containing language data. The Bank of Finnish Terminology in Arts and Sciences is a multidisciplinary and multilingual project aiming at developing a permanent and easily updated terminological database for all fields of research in Finland.

Litteratur

Koskenniemi, Kimmo, Lindén, K., Carlson, L., Vainio, M., Arppe, A., Lennes, M., Westerlund, H., Hyvärinen, M., Nuolijärvi, P., Piehl, A., 2012: Suomen kieli digitaalisella aikakaudella [Finnish in the Digital Age], Berlin. 85 p. http://www.meta-net.eu/whitepapers/volumes/finnish

Lindén, Krister, Carlson, L., 2010: FinnWordNet - WordNet på finska via översättning. [FinnWordNet – WordNet in Finnish by Translation] / In: *LexicoNordica, Vol. 17*, 11.

Lindén, Krister and Jyrki Niemi, 2013: Is It Possible to Create a Very Large WordNet in 100 days? – an Evaluation / In: *Language Resources and Evaluation*, Springer Verlag.

Lindén, Krister, Silfverberg, M., Axelson, E., Hardwick, S., Pirinen, T., 2011: HFST—Framework for Compiling and Applying Morphologies. / In

Systems and Frameworks for Computational Morphology. Edited by Cerstin Mahlow and Michael Pietrowski. Vol. 100 Springer. p. 67-85 (Communications in Computer and Information Science).

Lindén, Krister, Silfverberg, M., Pirinen, T., 2009: HFST Tools for Morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers. / In *State of the Art in Computational Morphology* edited by Cerstin Mahlow, Michael Piotrowski. Berlin, Heidelberg, Springer Berlin Heidelberg. p. 28-47 (Communications in computer and information science).

Oksanen, Ville, Lindén, K., Westerlund, H., 2010: Laundry Symbols and License Management – Practical Considerations for the Distribution of LRs based on experiences from CLARIN. In: *Proceedings of LREC 2010 : Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management.*

Voutilainen, Atro, Muhonen, K., Purtonen, T.K., Lindén, K., 2012: Specifying Treebanks, Outsourcing Parsebanks: FinnTreeBank 3. In: *The eighth international conference on Language Resources and Evaluation (LREC).*

Lindén, Krister, Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T., Silfverberg, M., 2013: HFST—a System for Creating NLP Tools. In: *Systems and Frameworks for Computational Morphology Systems and Frameworks for Computational Morphology: Communications in Computer and Information Science,* edited by Cerstin Mahlow and Michael Piotrowski, Springer-Verlag, 2013. (Communications in Computer and Information Science).