

CLARIN in Denmark – European and Nordic perspectives

Hanne Fersøe

University of Copenhagen
Centre for Language Technology
Copenhagen, Denmark
hannef@hum.ku.dk

Bente Maegaard

University of Copenhagen
Centre for Language Technology
Copenhagen, Denmark
bmaegaard@hum.ku.dk

Abstract

This paper gives an overview of the Danish CLARIN project (funding background, national strategic goals, formation of consortium etc.) including the very important priority of aiming the results of the project at researchers from the wide range of all fields of humanities research which is based on language sources, i.e. not exclusively at researchers in the fields of linguistics and language technology, but with a much broader scope. Secondly, it discusses future perspectives of European and Nordic cooperation.

1 The European context

The European Strategy Forum on Research Infrastructures (ESFRI) initiated its Roadmap Process in 2001, and in 2006 it published the first European Roadmap for Research Infrastructures (RI), which was updated in 2008¹. The Roadmap gave its recommendation to 6 SSH-projects (Social Sciences and Humanities), and the European CLARIN project is one of those 6 projects.

At the European Commission level a funding model for European Research Infrastructure (RI) projects was developed in the 7th Framework Programme, a call was opened for those recommended by ESFRI, and 34 projects are currently running, including 5 SSH projects. In parallel, the European Commission has work in progress on a Council Regulation to provide a legal form for the long-term organization to run the pan-European RIs in the construction and deployment phases.

The participation in the construction and deployment of pan-European RIs must be funded nationally, so the 27 EU member states have agreed to develop national Roadmaps. Currently approximately half of these are available.

2 The National Danish Context

2.1 Funding of Danish RI projects or Danish participation in European RI projects

In parallel with the European interest in research infrastructures, the Danish Ministry of Science, Technology and Innovation commissioned the Danish Council for Strategic research to survey the needs and propose a strategy for future research infrastructures. Their report was published in December 2005.

Following these preparatory strategy papers a call for proposals of RIs was published in September 2007 with a pool of 200 million DKK (€27 million) per year for a period of three years.

2.2 Danish Roadmap

Additionally, the Danish Agency for Science, Technology and Innovation is preparing the national Danish roadmap for RIs in agreement with the ESFRI and the Commission process.

3 The Danish CLARIN project

The Danish CLARIN consortium applied for the equivalent of four million euros and was awarded two million for the three year period 2008-2010 for the construction of a national research infrastructure for the humanities, focusing on material expressed in language (written or spoken) and tools to treat this material. This means that Denmark is not in a preparatory phase parallel to the EU-CLARIN project, but that we

¹ <http://cordis.europa.eu/esfri/>

are actually implementing a national research infrastructure.

3.1 The Consortium

The Danish CLARIN consortium has four universities and four cultural institutions as their members with the University of Copenhagen coordinating the consortium. The members are:

- University of Copenhagen
- University of Southern Denmark
- University of Aarhus
- Copenhagen Business School
- Society for Danish Language and Literature
- Danish Language Council
- The Royal Library
- The National Museum of Denmark

A total of 11 research groups are participating with funding, and a 12th group has joined as of January 2009 as an observer.

With these partners the consortium is very strong and to the point, as it has a good combination of the necessary skills and experience: humanities, language technology, language resources, and computer science. The consortium will collaborate with EU-CLARIN where possible, and particularly strive to learn from and adhere to standards as decided at the European level in order to pave the way for Denmark to be an active partner in the construction and exploitation phases of the European project. One of the national tasks for the Danish CLARIN consortium is to propose a strategy for the exploitation at the national level.

3.2 Strategic project goal

The vision is to create a researcher's toolbox by establishing a number of digital Danish text, speech and visual resources and associated tools and to integrate these resources into a web-based environment for research thus creating a much needed support for Danish humanities and enhance its possibilities for European collaboration.

The Danish CLARIN project is eager to follow standards and recommendations developed in the preparatory phase of the European CLARIN project, as far as possible, but as the European project is a preparatory project, the recommendations may not all be available when they are needed for implementation in the Danish

project. The European CLARIN project is assessing existing standards and recommendations in order to be able to determine a set of CLARIN specific recommendations and standards in areas such as technical architecture, meta data, interoperability, IPR and copyright issues etc. However, the Danish CLARIN project needs to proceed, in order to make sure to be able to deliver the results foreseen at the end of 2010.

For this reason it was vitally important for the consortium to design the work packages in such a way as to be able to deliver as a result not only the technical infrastructure but also as many types of content as possible. This means that the project plan contains activities both to deliver already existing resources and to produce new resources. The project is organized into thematically defined main work packages, namely written language resources, spoken language resources and collections of constructed data. Each main work package is subdivided into a number of sub work packages, and in each of these the participants are in the process of collecting, annotating and otherwise producing and including different types of resources.

3.3 Written language resources

In the main work package *written language resources* six different written language resources will be created and made available through the Danish CLARIN infrastructure.

The Danish CLARIN partner Society for Danish Language and Literature (DSL)²: is responsible for collecting a contemporary general language corpus of 15 million words of annotated Danish text per year (i.e. a total of 45 million words), mainly from newspapers and periodicals. This new corpus will cover the period around 2010, and as such it will be supplementing the existing KorpusDK³ which contains around 56 million words from the periods around 1990 and around 2000, respectively. The corpus annotations will be expressed according to TEI P5 specifications. Apart from KorpusDK, DSL has many other interesting and relevant digital resources, as can be seen on their web page, and as a part of the project some of these will also be made available through CLARIN.

University of Copenhagen, Centre for Language Technology (CST)⁴, together with the

² <http://dsl.dk/>

³ <http://ordnet.dk/korpusdk>

⁴ <http://english.cst.ku.dk/>

Danish Language Council (DSN)⁵ is responsible for collecting an 11 million words corpus of annotated sublanguage texts from the period 2000-2010 from broadly selected domains such as health care and medicine, IT, agriculture, construction, meteorology. The corpus will be based on texts originating from experts and semi-experts and with a targeted readership of semi-experts and laymen. At present no such corpus exists for Danish so the sublanguage corpus will represent a truly new type of resource for scientists to work with, and as such it will constitute a valuable supplement to the general language corpus. To learn more about the general language corpus and the sublanguage corpus, see Halskov (to appear).

Another corpus of sublanguage texts will be collected by researchers from the DUDS⁶ group at University of Copenhagen. They will create a corpus of 250,000 words composed of extracts from non-literary texts for everyman's use from the period 1500 to 1750. The texts will be extracted from rare books only obtainable from The Royal Library in Copenhagen, and they will cover subjects such as ethics and moral issues, geography and topography, history, housekeeping and cooking, medical science, mathematics and astrology, natural sciences, pedagogics, etc. (Fersøe 2008b). The texts will be scanned and OCR recognized and marked up according to the Multi Level Text (MLT) annotation (Ruus 2002) which handles orthographical variation, and which will be the key to searching the corpus.

The domains covered in the Everyman corpus mentioned above could be richly illustrated by the images found in existing collections belonging to the section Danmarks Nyere Tid (DNT)⁷ of The National Museum of Denmark. A group from this unit is responsible for creating a pilot corpus of 8,000 images with associated textual descriptions and for making them available on the platform. After deciding the best way of capturing and annotating all the available information from the associated texts, including which language technologies to use for this, they will select more images. Currently there are 50,000 digitized images to choose from. It is not the task of this project to link the Everyman corpus and the DNT images, but this is a future research project. Furthermore the linking could also be ex-

tended to the Danish Dictionary of Insular Dialects, DID⁸, see further down.

Older literary texts will be represented through the work of the Danish writer and Nobel Prize winner, Johannes V. Jensen. Of his work 50 books will be digitized, OCR recognized and annotated, the latter a task which implies adapting the tools, e.g. the PoS-tagger, to older Danish. DSL is responsible for this work together with the Johannes V. Jensen Centre of the University of Århus⁹. In addition DSL will also be specifying a prototypical lexicon of orthographical variation.

Finally a parallel multilingual resource of at least 20 million words will be collected from available bilingual texts. The work will build on experience gained from previous work carried out by research groups at the University of Copenhagen (Maegaard, Offersgaard et al. 2006). While this previous work focused on older texts, namely *The Snowman* by the famous Danish fairy tale writer Hans Christian Andersen, the new parallel corpus will focus on contemporary texts. The texts will be collected and subsequently aligned and annotated, and focus will be on Danish-English and Danish-German. CST is responsible for collecting, aligning, and otherwise annotating the multilingual corpus and for making it available.

One of the challenges in connection with collecting and making available current written text resources is the copyright issue. The consortium is asking permission from writers, publishers and other categories of text owners, and only texts for which permission can be obtained will be included.

3.4 Spoken language resources

In the main work package *spoken language resources* three different spoken language corpora, one of them including video recordings, will be collected, annotated and made available with a number of associated tools.

A group of researchers from the University of Southern Denmark. USD¹⁰, in Kolding will collect video and sound recordings of 20 hours of naturally occurring interaction, mostly from face

⁵ <http://www.dsn.dk/>

⁶ <http://duds.nordisk.ku.dk/>

⁷ <http://www.nationalmuseet.dk/sw6796.asp>

⁸

<http://dialektforskning.ku.dk/publikationer/oemaalsordbogen/>

⁹ <http://www.nordisk.au.dk/jensen/index>

¹⁰

http://www.sdu.dk/Om_SDU/Institutter_centre/Isk/Centre/SoPraCon.aspx

to face situations. The corpus will be annotated according to the Conversation Analysis methods (MacWhinney and Wagner, to appear) to encode overlap, pausing, prosody, and a wide variety of non-lexical features. In addition to this, parts of the corpus will also be annotated with multimodality coding according to the MUMIN system (Jokinen, and Navarretta et al., 2008) for facial and manual gestures, gaze, posture, and proximity. The corpus will be accompanied by a search engine which allows the data to be searched for interactional features, mainly combinations of verbal material, timing plus features marked in the transcription.

Another spoken corpus will be collected by the researchers from the Danish National Research Foundation Centre for Language Change in Real Time, LANCHART¹¹, at the University of Copenhagen. This group is working with corpora collected over a long period of time, and they are re-interviewing some of the informants that were interviewed earlier in order to be able to compare their language between then and now and thus study language change (Gregersen, 2007). There are, however, various confidentiality restrictions which are making it very difficult – if not impossible – to offer free availability to these corpora, so in the CLARIN context a new small corpus of spoken young Copenhagen Danish will be collected and annotated according to the LANCHART standards. The group will also deliver a tool that can be used for analysis by all researchers who want to handle and study spoken language materials.

The third spoken corpus to be delivered through the Danish CLARIN infrastructure is created at Copenhagen Business School, CBS¹². The corpus text is the Danish PAROLE corpus¹³ of which currently 100,000 tokens exist as sound files in lab quality (Henrichsen, 2007). This corpus will be made available with the sound files and with annotations for PoS, syntactic structures, acoustic measurements, phonetic transcription, and more. These data are unique in Denmark for phonetic studies and speech technology. The data will be extended, revised and reorganized to be made available through CLARIN, and so will the accompanying tools for word-level alignment, verification of phonetic transcription, and acoustically based prosodic analysis.

¹¹ <http://lanchart.hum.ku.dk/>

¹² <http://isvcbs.dk/~pjuel/index2.html>

¹³ <http://korpus.dsl.dk/e-resurser/parole-korpus.html>

3.5 Collections of constructed data

The term ‘collections of constructed data’, or technological resources as they are also called, is a loose definition we have used in the Danish CLARIN project to cover resources that are not collected and annotated as they are, such as e.g. written or spoken corpora, but which are carefully selected data put together as a collection according to a specific set of requirements, such as e.g. dictionaries. In the main work package *collections of constructed data* three different sets of constructed data will be made available.

The Danish WordNet, DanNet¹⁴ (Pedersen, Nimb et al. 2008), will be extended from 35,000 to 70,000 synsets in close collaboration between CST and DSL and according to a set of specifications for inclusion of new vocabulary. The extension, more precisely, consists of generation of the new synsets, placing them in the ontological structure of DanNet, determining DanNet equivalents for Base Concepts from Princeton WordNet¹⁵, and establishing the links to Princeton WordNet. The existing coding tool will be slightly enhanced, and an xml-format will be developed.

Researchers from the Jens Peter Skautrup Centre¹⁶ at the University of Århus have developed Jysk Ordbog¹⁷, which is a rich resource of dialects of Jutland. In the CLARIN project the research group will evaluate the current data base format of the dictionary and subsequently redesign it to fit more appropriately with CLARIN standards and formats before making it available through the infrastructure.

Bringing together different types of dictionary resources is scientifically interesting and has obvious benefits for teaching. In the CLARIN project researchers from CST will bring together DanNet and the Danish computational dictionary, STO¹⁸, and thus highly improve the potential of both as a computerized representation of Danish vocabulary, providing not only lexical semantic information, but also syntax and morphology. The work will be based on the positive results of a pilot project (Pedersen, Braasch et al. 2008), and will comprise about 9,000 words.

The research group from Danish Dictionary of Insular Dialects (DID) mentioned earlier is not a CLARIN partner with funding from the grant.

¹⁴ <http://www.wordnet.dk/>

¹⁵ <http://wordnet.princeton.edu/>

¹⁶ <http://www.jysk.au.dk/index.jsp>

¹⁷ <http://www.jysk.au.dk/jyskordbog/jyskordbog>

¹⁸ http://english.cst.ku.dk/sto_ordbase/

The group, however, is currently working with some technical issues similar to those of Jysk Ordbog, i.e. formats, meta data, data structure and tools, and therefore the Danish CLARIN consortium has invited the DID group to become observers in the work package regarding the constructed data.

3.6 Technical platform

The technical infrastructure of the Danish CLARIN platform is in the process of being specified, and it is still too early to give a more detailed account of these matters. Currently the infrastructure is seen as a digital repository with a web user interface managing:

- Access rights given to users based on user verification mechanisms
- Access rights for users to specific content based on resource profiling
- Search and retrieval facilities
- A personal work space
- Communication facilities

3.7 The future after 2010

One of the management tasks of the Danish consortium is to propose a plan for future operation and exploitation of the Danish CLARIN infrastructure. Key elements for which future funding must be found are on the one hand the technical inclusion of Danish CLARIN into EU-CLARIN, and on the other hand the continued inclusion of new resources on to the national infrastructure. Another challenge will be the dissemination of the usefulness of the infrastructure for a wide range of humanities research areas.

4 European and Nordic Perspectives

The history of language technology collaboration among the Nordic countries goes back to the early days of computational linguistics. The first Nordic summer school in computational linguistics was held in Marstrand, Sweden, in 1972, followed up by Bergen 1973 and Copenhagen 1974. These summer schools have been instrumental in the creation of a Nordic computational linguistic community. Later on the Nodalida conferences were started by “Den Nordiske Samarbejdsgruppe for datamaskinel sprogbehandling” with the first conference in Gothenburg 1977, and as the latest step in this direction we have the

creation of NEALT (Northern European Association for Language Technology) in 2007.

The Nordic collaboration has been very important for the building up of the Nordic computational linguistics communities, not least for preparing for European collaboration.

4.1 Content of the Nordic collaboration

Some Nordic countries have languages that are similar and in this case it is highly recommendable to reuse and accommodate tools, standards etc., wherever possible. E.g. the CST lemmatizer for Danish has been trained for Icelandic and is now being used in Iceland. This kind of collaboration will take place only if information about the existence of language technology tools and methods is available. There are several instruments for knowledge sharing and dissemination: the NorDokNet centres (Fersøe, Rögnavaldsson et al. 2005) were supported by the Nordic Council of Ministers, and even if funding has stopped, the collaboration among the centres survives, albeit at a lower level. Similarly the Nodalida conferences are a great help to disseminate knowledge and support Nordic collaboration.

4.2 Merging of Nordic and European perspectives

CLARIN is a European initiative, and this means that CLARIN will provide everything which the Nordic collaboration provides, just at the larger, European, scale: standards and tools are shared with many more languages, and it is possible to collaborate with many more research groups and to be inspired by many more researchers around Europe.

In a successful CLARIN we see the Nordic and the European perspective merging.

Acknowledgements

This project is supported by the Danish Agency for Science, Technology and Innovation, as well as by all partner institutions.

We thank all participants in the Danish consortium for their contribution to the project.

We also thank all the work package leaders for their work package descriptions, which have served as input particularly to section 3 of this document.

References

Hanne Fersøe 2008a. *The Danish CLARIN Project*. CLARIN Newsletter, number 2, July 2008.

- Hanne Fersøe 2008b. *Knowledge for Everyman from the Renaissance to Modern Times*. CLARIN Newsletter, number 4, December 2008.
- Hanne Fersøe, Eiríkur Rögnvaldsson and Koenraad de Smedt 2005. *NorDokNet - Network of Nordic Documentation Centres. Contacts to future Baltic Partners*. Nordisk Sprogteknologi. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000 - 2004. København 2005, side 13-23.
- Frans Gregersen 2007. *The LANCHART Corpus of Spoken Danish, Report from a corpus in progress*, in J.Toivanen & P.Juel Henrichsen (eds.): *Current Trends in Research on Spoken Language in the Nordic Countries, Volume II*, Oulu University Press, p.130-143, ISBN 978-951-42-8514-1.
- Jakob Halskov (to appear). *Compiling, annotating and publishing corpora in DK-CLARIN, the Danish incarnation of the pan-European initiative for a common resource infrastructure*. To appear in *Corpus Linguistics 2009*, Liverpool.
- Peter Juel Henrichsen 2007. *The Danish PAROLE corpus - a merge of speech and writing*; in J.Toivanen & al (eds) *Current Trends in Research on Spoken Language in the Nordic Countries, vol II*; Oulu Univ. Press 2007, pp.84-93
- K. Jokinen, C. Navarretta , P. Paggio 2008. *Distinguishing the communicative functions of gestures*. In A. Popescu-Belis and R. Stiefelhagen (eds.) *Proceedings of 5th Joint Workshop on Machine Learning and Multimodal Interaction*, Utrecht, September 2008, Springer, 38-49.
- Brian MacWhinney, Johannes Wagner (to appear): *Transcribing, searching and data sharing: The CLAN software*. To appear in *Gesprächsforschung 2009* (ISSN 1617-1837).
- Bente Maegaard, L. Offersgaard, K.F. Joensen. X. Lepetit, C. Navarretta, J. Pedersen, C. Povlsen. 2006. *MULINCO - Korpusplatform til sprog- og oversættelsesstudier*. Tidsskrift for Universiteternes efter- og videreuddannelse, nr. 7 s. 1-15: E-læring i sprogfag, Danmark.
- Bolette S. Pedersen, S. Nimb, L. Trap-Jensen (2008) *DanNet: udvikling og anvendelse af det danske wordnet*. Nordiske Studier i Leksikografi 9, Rapport fra konference om leksikografi i Norden pp. 353-371, Akureyri, Island.
- Bolette S. Pedersen, A. Braasch, L. Henriksen, S. Olsen, C. Povlsen, 2008. *Merging a Syntactic Resource with a WordNet: A Feasibility Study of a Merge between STO and DanNet*. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*. European Language Resources Association, 2008. 5 s.
- Ruus, Hanne. 2002. *A Corpus-based Electronic Dictionary for (Re)search*, in EURALEX 2002 Proceedings, pages 175-185.