Centre for Danish Language Resources and Technology Infrastructure for the Humanities (DK-CLARIN) - Annex 1 rev.

Revised in February 2008, to accommodate the revised budget.

Preamble

The general text below has been kept from the original proposal, although some of the visions have been moved into a more distant future. The work plan has obviously changed, although the overall structure has been kept. WP5, which is the technical work package, has not been touched as much as the rest, as the technical parts are essential for creating and running an infrastructure at all.
WP5.3 has been almost deleted and the rest has been integrated in 5.2. The resource building work packages have been cut by up to 66%; the effect of this is less data, or data that are analysed to a smaller degree etc. The consortium believes this is the best way to implement the reduction, as data can be added later when more funding becomes available. This also means that most of the activities have been kept, just at a lower level.
The consortium had planned to create user panels from the humanities at large, but this is now omitted, apart from the cases where a user panel is specific to a work package and the WP leader has chosen to keep it.

The Challenge

In their "Roadmap for the Social Sciences and Humanities" (2006), the European Strategy Forum for Research Infrastructures offered this assessment of the challenge involved in building a new information technology (IT) infrastructure for the humanities:
"The practice of the social sciences and humanities has slowly but profoundly been transformed along with the emergence of new information technologies. Digital resources, computer networks, and software tools to a great extent, influence the sense of the human record, the way it is understood and the way those understandings are communicated.... Digitizing the products of human culture and society poses intrinsic problems of complexity and scale. The complexity of the record of human cultures — a record that is multilingual, historically specific, geographically dispersed, and often highly ambiguous in meaning — makes digitization difficult and expensive.
The **present major challenge** is therefore to create pan-European infrastructural systems that are needed by the social sciences and humanities to utilize the vast amount of data and information that already exist or should be generated in Europe. Today the social sciences and humanities are however hampered by the fragmentation of the scientific information space. Data, information and knowledge are scattered in space and divided by language, cultural, economic, legal, and institutional barriers."
The present proposal seeks to address this challenge by constructing a Danish IT infrastructure for the humanities. The work proposed here relates to the CLARIN (Common Language Resources and Technology Infrastructure) proposal submitted to the European Commission in May 2007, which received a positive evaluation and is now under negotiation. The University of Copenhagen is a member of the CLARIN consortium, and will be work group leader and participate in the management of CLARIN. The present proposal is related to CLARIN and will follow CLARIN standards and recommendations. However, it is important to realize that the current proposal involves a Danish investment in the construction of a national infrastructure that will stand alone as a vital contribution to the Danish research enterprise, a task that will never be taken on by the EC.

Mission and Vision

Our goal is to provide a research infrastructure for the humanities integrating written, spoken, and visual records into a coherent and systematic digital repository. This repository will form the core of a "virtual institution" where researchers from a wide array of disciplines can use a wide variety of interoperable tools and data conversion interfaces to document and analyze complex patterns of human linguistic, social, artistic, literary, and cultural dynamics. The vision is to establish 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, a researcher's digital toolbox. The key term in this effort is *interoperability*, creating an IT-architecture that supports the integration of the various resources, tools, and formats into the DK-CLARIN platform by web services. The infrastructure will be extensible, providing easy access for adding new resources and tools, and possibly other research-related resources such as papers and other communication.

The Target Area

Traditionally, the development of language resources has been confined to support the development of language technology and computational linguistics tools and models. However, as these tools are now becoming mature, they can be applied to ever-widening circles within the humanities. The fact that a great diversity of data (text, speech, images, information about artefacts) can all be stored and accessed by researchers from different fields opens up enormous new possibilities for integration across this whole target area. The present consortium includes four universities and four institutions from the Danish Ministry of Culture, thereby covering humanities in a very broad sense. The consortium will seek to respond to the interests of researchers in linguistics, translation and literature, archaeologists, ethnologists, ethnographers, anthropologists, art historians, historians, librarians, psychologists, and others. The consortium and its qualifications are described in more detail in Annex 2.

One of the aspects of the proposed infrastructure is that over time it is meant to contain the necessary resources and tools for language technology. This part of the infrastructure is a prerequisite for the development of language technology for Danish, a necessity not only for humanities research, but also for Danish industry and administration if Denmark is to participate fully in the globalised society. The concept of BLARK (Basic Language Resource Kit, Binnenpoorte et al, 2002) will guide this development [≥]. Apart from Danish monolingual resources, DK-CLARIN encompasses two aspects of multilinguality: parallel corpora and cross-lingual language resources, thereby ensuring that (some of) the resources built are comparable with and - more importantly - can *interact* with parallel resources in other languages. This aspect is relevant not only in a research and standardisation perspective, but also has practical advantages for the development of cross-lingual technologies.

A research infrastructure of this type consists of **1) the data**, including metadata, **2) annotations**, i.e. addition of relevant information. Annotation is normally done automatically, using tools. **3)** The infrastructure needs to be backed by **computer and network systems**. In all cases the proper use of standards is extremely important, and this project will use international standards wherever available, and use the standards chosen by the EU project CLARIN where appropriate. Finally, **4) an organization** to carry out the tasks, and to support the users. Below we describe each of these four components in more detail.

The data

The present proposal for creating an infrastructure aims at 1) taking advantage of already existing digitized material, 2) creating new data for a broad data bank, taking many different types of materi-

[≥] A first attempt of an overview of what exists in terms of language resources and tools for Danish was made in the report *Strategisk satsning på Dansk Sprogteknologi*, 2004, by several of the current applicants.

al into consideration. The use of existing resources has become significantly more important with the budget reduction, as the possibilities for creating new data has diminished.
We can distinguish two types of data resources: *basic resources* and *technological resources*. *Basic resources* are text or speech corpora. Basic resources are the most needed for all types of research: they constitute the 'raw' material. In DK-CLARIN they include modern and old (historical) Danish texts, general language and sublanguages, spoken language and videos as well as text and images. The term *technological resources* is used for those resources that are "man-made", e.g. dictionaries, grammars, ontologies. Existing technological resources will be made available through DK-CLARIN, and an existing wordnet will be extended.

Annotation and other tools

The aim is that a large part of the basic resources will be annotated. Annotation enhances the value of a resource considerably. Annotation will be of several types: linguistic (including transcription), text structural, text historical etc., but general historical, archaeological etc. information is also foreseen. Tools for such annotation exist to a certain extent and will be further developed where possible.
Annotation is done automatically to a very large extent. Existing taggers, lemmatisers etc. will be reviewed and adapted to the different text types. Tools for annotating discourse, and for annotating speech will be reused or developed in accordance with international, in particular European, standards to the extent possible, and where necessary new standards created.

Technology

The infrastructure will be based on a service oriented architecture (SOA) applying a common user web interface to a repository of resources, tools and documentation. The infrastructure will ensure interoperability by using web services for the underlying communication between the tools and resources. The interface will supply the users with a structured overview of the resources and tools available and guide the user to find the relevant resources and annotation, allowing access to tools and resources. To be able to make an integrated search in a number of resources each resource will have to be described using a common metadata infrastructure. The resources have different characteristics and the metadata description will form a basis for a common description of the resources. The metadata will take into consideration the comprehensive work going on in the international research community already and initiatives in EU-CLARIN.
The different resources will have different copyright and access issues; therefore there is a large need for managing user authentication and access rights. This is handled by a central DK-CLARIN authentication server (Broeder 2006). This server will also be hosting the user interface and the integrated search facilities. Tools and resources are located on distributed servers at the institutions already hosting these.
Existing and upcoming resource and tool servers will communicate with the central DK-CLARIN server and register their services with descriptive metadata including access rights at this central server. To ensure all users an overview of the resources the metadata will be searchable for all users, although access rights to the resource may be limited. Expecting that the resources and tools will be improving over time the CLARIN infrastructure has to keep track of different versions of resources and tools. New resources will primarily be pooled by the three service providing partners (KU, DSL and Royal Library) who will constitute a national grid for Danish language resources. All institutions are connected to the Danish Research Network.

Organization

The Organization that will create and run this infrastructure is described in Annex 2.

Work plan overview:

WP1 Coordination and Technical management, incl. copyright and privacy issues
WP2 Basic written language resources
WP3 Spoken language resources and tools
WP4 Technological resources
WP5 Technical Infrastructure

Timetable
The overall time table includes 4 phases: 9 months, 9 months, 10 months and 8 months. As the individual work packages are of quite differing nature, the content of each phase will not be exactly identical for all groups. In general Period 1 will cover: detailed work plan, specification, assessment of existing data and tools, adaptation of tools and methods. Similarly, Period 2 will be devoted to creation of resources, annotation, adaptation of existing resources, first implementation of platform, and Period 3 will be a continuation of period 2. In Period 4, all resources will be finalised, tools and platforms will undergo final user tests and fine tuning, user manuals will be made, and the DK-CLARIN platform and resources will be released to the research community.

Detailed work plan description

WP1 Coordination and Technical management, incl. copyright and privacy issues
This work package consists of overall co-ordination and management. The coordinator will ensure the communication within the centre and between the centre and the Research Agency. A website will be set up for external communication. The Coordinator and the Executive Board will ensure that all WPs have detailed work plans, and follow these up on a regular basis. In collaboration with the WP leaders, the coordinator will set up Quality Assurance measures for all deliverables and make sure that the project runs smoothly.
One of the major problems for a project involving authentic text is the copyright issue which hampers researchers' access to and use of written material (apart from old material), and similarly for spoken language we have the privacy issue. The project will create spoken resources taking into consideration the privacy regulation, so we consider this problem partly solved (existing resources cannot all be made available). For the written language a solution has to be found in order to get full value of the data collected. The project will contact the parties involved as well as Danish authorities, and seek a solution, e.g. inspired by the one The Danish Broadcasting Corporation (DR) recently made with their copyright holders. The project will also take advantage of the fact that the European CLARIN project has a WP devoted to IPR issues and collaborate with this WP. It has to be understood however, that the copyright problem cannot be solved by the present project alone, as it requires regulation at a higher level. Consequently, the project may have to give restricted access to the infrastructure.

WP2 Basic written language resources
This work package deals with written language resources, contemporary and old, general language and specialised sublanguages, literary and professional, as well as parallel corpora with Danish as one of the languages. Collaboration will consist in developing and using the same tools or modified versions of the same tools, as well as exploiting automatic tools for harvesting texts. In particular WP2.1 and WP2.2 will share methods and tools.

WP2.1 Reference corpus of general language
Some corpora for Danish do exist, but there is a need for intensifying the compilation of corpus material both in terms of quantity and with respect to continuity in order to secure the diachronic aspect. The project will collect at least 15 million words of Danish text per year. Material will mainly be taken from newspapers and periodicals. Some basic methods and tools for automatically collecting, structuring and annotating text will be developed, using KB's records of the entire .dk

domain, harvested 4 times per year. Material will be collected taking the copyright into consideration and all the collected text will be made available to the research community without restrictions in so far as copyrights permit, cf. WP1.

Deliverables/Milestones

T9: Ontology of text types, genres. XML-based annotation scheme.
T18: 15 mill. words compiled and annotated according to XML schemas developed in stage 1.
T28: 30 mill words compiled. Corpus hosting, public access and user interfaces implemented.
T36: 45 mill. words compiled. Usability tests. Final evaluation.

WP 2.2 Corpus of sublanguage texts: 11 mill. words from the period 2000-2010

Sublanguage texts will be collected from broadly selected domains (e.g. social sciences, leisure, commerce/finance, medicine) and from different text types. Texts originating mainly from experts and semi-experts with a target readership of semi-experts and laymen will be preferred. The corpus design and selection criteria will be in accordance with best practices. Annotation with metadata and linguistic annotation will be automatic as far as possible. These tasks require some particular processes, e.g. elaboration of appropriate domain ontologies, demanding text selection procedures, and technical solutions tailored to the usually shorter sublanguage texts. A sublanguage corpus supports the development of technological resources, e.g. dictionaries, and will be useful also in linguistic research, language teaching etc. It will also complement the coverage of the general language corpus (cf. WP2.1) by reflecting usages of words within informative, persuasive and instructive texts.

Deliverables/Milestones

T9: Specifications for corpus composition. Technical solutions for corpus collection and annotation.
T18: Compilation and annotation of 2 mill. words corpus.
T28: Compilation and annotation of 5 mill. words corpus.
T36: Compilation and annotation of 4 mill. words corpus. Documentation. Full corpus available.

WP2.3 Knowledge for everyman from the Renaissance to Modern Times

Based on experiences from collecting and investigating texts in the DUDS project (Danish Under Digital Study) a corpus of approximately 0.25 mill. words covering the period 1500 to 1750 will be built and annotated. Existing tools for registering older texts and annotating them will be tested and refined. The foremost challenge when searching both older text and transcribed discourse is variation, orthographical and phonetic, respectively. One general multilevel search and annotation tool can meet the requirements of both written and spoken data types, and we will co-operate with KU-LAN in developing such a tool box. The tool will be XML-based and include conversion between different mark-up principles.

Deliverables/Milestones

T9: Specification of corpus and tools. Pilot texts scanned, OCR-ed, or otherwise digitized.
T18: Specification of content architecture. Digitized text corpus. Annotation and search tool.
T28: Multilevel annotated text corpus. Test user panel.
T36: Refined multilevel annotated text corpus. Corpus and annotation and search tool available.

WP2.4 Enhanced annotation and improved search possibilities to old literary texts

The Archive of Danish Literature (ADL) contains more than 8,000 titles from Saxo to the present day comprising a total of more than 20 mill. words. This WP has three goals: First, the existing text files will be adjusted to the format specified in WP5.1 and partly WP5.2 thereby preparing them for the DK-CLARIN search engine. Secondly, selected works of Johs. V. Jensen (JVJ) will be digitized. Last, the texts will be annotated in order to improve search. Like in WP2.3 words appear in several variants. WP2.3 and this WP will collaborate on sketching a lexicon combining variants with a base form – a minor prototype of such a lexicon will be offered by the project.

Deliverables/Milestones:
T9: Specification and configuration of the search engine requirements.
T18: Digitisation of selected works by JVJ. Annotation of the texts.
T28: Small, prototypical lexicon.
T36: Implementation of search engine and interface.

WP 2.5 Images, artefacts and texts from the National Museum of Denmark.
Images in connection with metadata, archive material such as official documents, material from the museum's tradition archives, descriptions of museum artefacts, etc are an integral part of the cultural heritage, collected and stored at the National Museum for more than 300 years. In WP 2.5 the aim is to uncover the digitized material through language technology and linguistic methods to add new research dimensions to the interpretation of the cultural heritage. Focus will be on **The National Museums Image data base** which will have a size of 150,000 images in its first phase. We will analyse the pictures and connected metadata using different types of interpretation.
The source material consists of images of artefacts, archival images, including images from the Kunstkammer, from the great expeditions in former times, for example to Greenland and Mongolia.
Deliverables/Milestones
T9: Specification of the project
T18: Digitisation of texts of non-digitised material for the project
T28: Annotation (implementation of automatic, semi-automatic and manual annotation)
T36: Testing, access and search for the Image Database

WP2.6 Parallel multi-lingual text resources, with alignment
A parallel text is a text with its translation in another language. Aligned parallel texts will be an important resource for fields like automatic translation, multilingual terminology and multilingual search. The parallel resource will consist of 20 mill. words, arising from other tasks in WP2 and available bilingual texts. Alignment will be at several levels: sentence, word and perhaps part of syntactic structure. The focus will be on Danish and English, but also other languages. Statistical methods will be used for automatic annotation with a special focus on annotating the aligned texts with some entities, e.g. names, titles, dates, objects as these are of general interest to humanities researchers.
Deliverables/Milestones
T9: Specifications for DK-CLARIN annotation of multilingual corpora
T18: 2 mill. words treated. Tools for data alignment, annotation, specification of markup
T28: 12 mill. words treated
T36: 20 mill. words bilingual corpus available in DK-CLARIN

WP3 Spoken language resources and tools
This part of the infrastructure involves the construction of a database based on three different spoken language corpora and the development of associated interoperable administrative and analytic tools. Spoken language materials are essential for the development of linguistics, speech technology, and foreign language pedagogy. All of the data in the new database will conform to the TalkBank XML Schema. Conformity to this new international standard will allow each of the projects to make use of a wide range of XML-compliant tools. This format will allow users to access and comment on the database materials through web browsing of streaming hinted media linked to transcripts at the level of the sentence, breath group, and word. In addition, we will where appropriate rely on the XML standard for writing new programs and browser plug-ins that can analyze and search corpora through sockets over the web.
We will configure a common platform for statistical analysis, database querying, and pattern analysis. The partners in WP3 and the partners in WP 2.3, will combine their experience and will confront

the users as far as possible.

3.1 The multimodal interaction database
The first component of the spoken language database is a corpus of audio and videotaped naturally occurring interactions collected from everyday life, from workplaces, institutions, and from the media. The corpus consists of a carefully annotated set of up to 20 hours of conversations in Danish and will set a Gold Standard for the study of interactions in Danish. This corpus will be transcribed using Conversation Analysis (CA) methods for encoding prosody, overlap, pausing, and a wide variety of verbal features. For parts of the corpus, we will use a modified version of the MUMIN system for coding facial gestures, manual gesture, gaze, posture, and proxemics. Once completed, this core reference database will represent the highest quality annotation currently available for any spoken language.
Delivery/Milestones
T 9: Transcription standard fixed, tested, and coordinated. Existing data for Gold Corpus identified.
T 21: Gold corpus part 1 (7 hours video). Activity based search tools tested. File manipulation.
T 30: Gold corpus part 2 (7 hours). Activity based search tools ready. Access to satellite corpora.
T 36: Gold corpus part 3 (6 hours). Corpus and search tools fully integrated into DK-CLARIN.

3.2 Tool box for the treatment of spoken materials for various types of research
The second component of the spoken language data base includes new recordings of 'natural' speech using the best recording equipment and cleared for use as such with the informants. This will be the task of the LANCHART centre which also undertakes to present a tool box for all researchers interested in spoken language data which offers well tested techniques for recording, storing, retrieving and printing both raw and annotated spoken language materials including a generalization of an already existing search engine so that it may become a prototype of such search engines tailored to the needs of specific user types.
Deliverables/Milestones
T9: Discussion with users on the contents of a tool box, conversion programmes operating
T18: Outlines of a tool box presented to user panel, new interviews ready for integration
T28: Generalized search engine prototype presented to user panel, new interviews annotated
T36: Tool box ready for use with new data base

WP3.3 Spoken language resources for speech technology
The Danish PAROLE corpus consists of 300,000 tokens and includes annotations for PoS, syntactic structures, sound files (lab quality), acoustic measurements, phonetic transcription, and more. The sound related data (covering currently 100,000 tokens) are unique in Denmark for phonetic studies and speech technology (e.g. synthesis and recognition). These data are currently in an experimental stage and must be extended, revised and re-organized prior to inclusion in the DK-CLARIN database. Also the tools developed at CBS for PAROLE-analysis must be rewritten to comply with DK-CLARIN standards, such as tools for word-level alignment, verification of phonetic transcription, and acoustically based prosodic analysis.
Deliverables/Milestones
T9: Alignment of tiers for phonetic transcription and acoustic measurements for a substantial subset of PAROLE (at least 25,000 tokens)
T18: Alignment of tiers for phonetic transcription and acoustic measurements for the full 100,000 PAROLE read-aloud corpus
T28: All speech related tiers are completed for PAROLE. Compatibility with CLARIN data formats
T36: Tools for speech analysis are completed. Compatibility with CLARIN standards checked.

WP4 Technological resources

Technological resources are data resources that are constructed; here we cover traditional and electronic dictionaries, as well as dictionaries and semantic wordnets meant for computer systems. We cover the linking between different dictionaries as well as between dictionaries and corpora. The use of dictionaries and wordnets will enhance access to and usability of all types of data.

WP4.1 Danish WordNet, extension from 35,000 to 70,000 synsets
A full-fledged lexical semantic wordnet plays an essential role in HLT research and development, but this resource also has an impact for other researchers in the humanities. From a cognitive viewpoint a wordnet resembles elements of the mental lexicon, and thereby it is an interesting platform for research in cognitive fields of e.g. psychology and computer science.
The Danish wordnet, DanNet, will be expanded to 70,000 synsets (sets of words with similar meaning), corresponding to about three quarters of the coverage of the Princeton WordNet of English. The project will also provide translational links of the basic vocabulary to the EuroWordNet family, i.e. links from Danish to Princeton Wordnet and thereby to a large number of languages, including German, Italian, Spanish and French.
Deliverables/Milestones
T6: 5,000 new synsets
T17: 15,000 new synsets
T28: 15,000 new synsets (this brings the total to 70,000 synsets, incl. previous work)
T36: 5,000 links/translations to Princeton Wordnet (this brings the total to 10,000 links)

WP4.2 Dictionaries, standard, computational, old and dialects
1) Dictionaries covering aspects of Danish language (standard, dialect, historical) are essential guides when constructing, searching, and exploiting corpora. Bringing together different dictionaries is scientifically interesting and has obvious benefits for teaching. The project will investigate and implement an optimal DK-CLARIN dictionary structure and representation, taking as its starting point the web-published Jysk Ordbog (www.jyskordbog.dk). The technical platform of this dictionary will be evaluated and, re-designed/re-implemented taking into account other Danish dictionaries, and the possibilities of interaction with corpora.
2) The Danish wordnet, DanNet, is directly linked to Den Danske Ordbog; resembling much of the semantic structure of this resource wrt. sense definitions and distinctions. However, interrelations to other computational dictionary resources of Danish will highly improve its potential as a computerised representation of the Danish vocabulary as whole, providing not only lexical semantic information, but also syntax and morphology. Based on the positive results of a pilot project (done), we will link Sprogteknologisk Ordbase (STO) with DanNet. The vocabulary to be integrated will mainly comprise frequent words that are encoded in both resources (9,000 words).
1) Deliverables/Milestones
T12: Analysis of requirements to new database (DB) structure for Jysk Ordbog anticipating adoption of DK-CLARIN standards
T18: Transfer of data to new DB compatible with DK-CLARIN standards.
T28: Completed development of basic maintenance tools incl. tools facilitating the ongoing addition of further articles to the dictionary. DB adapted to and integrated in DK-CLARIN framework.
T30: Technical documentation completed in anticipation of further development and/or integration of similar dictionaries
2) Deliverables/Milestones
T6: Principles, methods developed and a basic integration tool designed for mapping STO – DanNet, tool tested.
T17: Common set of vocabulary defined and candidates for integration selected; integration of 20% of the vocabulary selected.
T28: Integration of 60% of the vocabulary selected.

T36: Integration of the rest of the vocabulary selected.

WP5 Technical infrastructure
The task of this WP is to provide the technical framework for the infrastructure, as described in the section "Technology" above, including a single web user interface to serve as the DK-CLARIN platform. This platform comprises access to all the tools and text resources of the infrastructure, as well as a personal workspace, communication facilities, user authentication and rights management, and search and retrieval facilities. Coordination with the management will take place on a continuous basis.

WP5.1 Technical infrastructure, DK-CLARIN infrastructure for search and access
This task will specify and implement the technical infrastructure, the centrally managed services and the common web interface. The task interacts with every other WP that is offering services. It is the responsibility of WP5.1 to support all aspects of the DK-CLARIN interoperability and to provide the necessary infrastructure to enable the central DK-CLARIN web interface and all decentralized DK-CLARIN services. DK-CLARIN will to the furthest possible extent use existing technologies and standards according to best practice. This could include the use of open source software as well as licensed technology, depending on evaluation in each case.
Deliverables/Milestones
T9: General infrastructure specification, including pilot cases of use. Analysis concerning resource metadata and metadata specification proposal.
T18: Specifications completed, and release of prototype
T28: Beta version of infrastructure
T36: Release of DK-CLARIN for operational phase

WP5.2 Access to existing resources and tools produced by the partners or otherwise available
This WP will provide direct access to relevant existing resources of the consortium participants as well as other available resources or tools. First, a comprehensive overview of available resources is created via the common interface. Next, access is provided either through linking (e.g. dictionaries) or through direct access through the DK-CLARIN interface. Some of the following existing resources are integrated: Parole-DK Corpus and parallel corpora, Phonetic mark-up of Parole-DK, Korpus 2000, Corpus of the Danish Dictionary (Korpus 90), DK87-90 Corpus, the MOVIN database, DSL's morphological lexicon, STO, DanNet, Archive of Danish Literature(ADL). The integration of data resources includes implementing a web service interface as defined in WP 5.1 for each resource. A few basic tools may be integrated. WP 2, 3 and 4 suggest which basic tools to be integrated.
Deliverables/Milestones
T9: Analysis of metadata and technical solutions for text and speech resources.
T18: Specifications of resource and tool integration. Agreements with resource and tool owners about rights and use.
T28: Beta version of accessibility modules, some resources integrated in beta version of platform
T36: Operational version of accessibility modules

Quality, topicality and relevance of the research

The DK-CLARIN research infrastructure will provide a much needed support for Danish humanities and enhance its possibilities for European collaboration. In fact, Danish researchers will only through a state-of-the-art infrastructure like the present one, have the necessary support for their work – the days of pen-and-paper research are over and IT is boosting research in all fields. DK-CLARIN will improve the conditions for Danish language technology research and development by starting a structured approach to a Danish BLARK, by making existing resources available

in a standardised way, and by creating new standards and annotations. But the proposed infrastructure aims in particular at boosting the opportunities for corpus based literary and historical as well as archaeological research. In particular old Danish texts will be made available in forms and quantities not previously seen. Historical sources are made available through the Royal Library. The resources from the National Museum will give rise to totally new possibilities for research in the combination of text and images, and will provide historians, ethnologists, archaeologists etc. with fundamentally new tools. All linguistic and contrastive research will benefit from the availability of general and non-fiction corpora and parallel corpora. The multimodal resources will pave the way for the research into e.g. how language, body movements, and gestures are combined by speakers of Danish. Results from other languages cannot immediately be taken over as the use of gestures and their internal structure is at least partly culture specific.
Finally, the creation of tools and resources for the exploitation of the vast materials of spoken language will enable researchers to contrast the two modes of Danish which will lead them to insights of a statistical and corpus linguistic nature never to be had before.

The consortium comprises the strongest partners in Denmark in the fields: The humanities faculties of the three largest universities (KU, AU, SDU) and the institute of International Language Studies and Knowledge Technology at CBS already count most of the university researchers who will be using the infrastructure. KU is internationally known as a very strong partner in the field of language resources. The Society for Danish Language and Literature (DSL) is a central institution in the area of Danish language and literature. A central activity is the development of lexicographic resources and text corpora based on HLT research. KB is the main research library, and has been working with digitisation of the textual heritage for many years, and with user interfaces. DSN as the Danish Language Council has long experience with the extraction of linguistic data from general language corpora and continuously monitors publications from all areas of general public interest. The National Museum provides textual data and images of objects, the National Museum will be able to use the results of the infrastructure, e.g. advanced facilities for searching in combinations of text and picture, - a facility which can be used by others as well when the research is done. Knowledge about interactional practices employed by speakers of a society is needed to develop software which interacts with user through speech. The DK-CLARIN database will allow fully integrated access to spoken databases which will allow new interactive language software to be build and to be trained. DK-CLARIN will create an internationally unique software infrastructure which will have a crucial role for the Danish language not to fall behind other languages with respect to research, tool development and protection of language domains.

The close interaction of research institutions will generate new insights, new research questions as well as new perspectives on existing data which will boost basic research in the humanities. Last but not least the data collection will provide a valuable resource for teachers and educators at all levels – as well as for providers of e-learning materials in a large number of areas.

Societal and commercial relevance, benefit and potential

The Danish society and business will profit from the infrastructure through the increased possibilities for producing language technology tools for Danish. The provider industry will be able to use the data provided and produce tools, and all other businesses and administrations will benefit from the existence of tools which will make communication, document production, game production, multilingual communication etc. more efficient and better. As mentioned in Annex 2 the infrastructure will improve education and training at all levels form the schools to the university. DK-CLARIN also wishes to support an unusually broad user-group including pupils and users with dyslectic disabilities. The project therefore includes research and development of highly user-friendly interfaces (standardized access to all databases, simple search procedures, interactive voice-based help

for low-tech users, etc), greatly improving the utilization potential and societal relevance of the associated databases. Existing speech synthesis modules for several languages, as well as conversion to Braille, are made available for research through the infrastructure by a Danish SME.

Data collection of any kind is expensive and so is augmentation of the data with syntactic and semantic annotation. Making the annotated data available to industry, teachers and e-learning providers will inevitably boost the access to information and the production of relevant tools for many citizens. The production costs for new language technology and e-learning products as well as the data collection costs for new research projects can be kept at a minimum, making the products accessible for practically everybody.

References

J. Asmussen, Pedersen, B.S. & Trap-Jensen, L. (2007). DanNet: From Dictionary to WordNet. Kunze, C., Lemnitzer, L. & Osswald, R. (eds.) *GLDV-2007 Workshop on Lexical-Semantic and Ontological Resources* 1-11. Universität Tübingen, Germany.

Binnenpoorte, D., F. De Friend, J. Sturm, W. Daelemans, H. Strik, C. Cucchinari (2002) A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch, In: *Proceedings LREC 2002*, (Third International Conference on Language Resources and Evaluation), Las Palmas de Gran Canaria, Spain.

Broeder, D, Veenendaal, R, Nathan, D, Strömqvist, S (2006) A Grid of Language Resource Repositories In: *e-Humanities workshop in e-Science Conference*, Amsterdam.

Cieri, C., M. Maxwell, S. Strassel (2003): Core Linguistic Resources for the World's Languages. In: *International Roadmap for Language Resources*, Workshop Paris 2003, http://www.enabler-network.org/documents/workshop/Cieri-Maxwell-Strassel.zip

Daelemans, W & H. Strik eds (2002): *Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen*, DLU, Den Haag

Joscelyne, A., R. Lockwood: *Benchmarking HLT Progress in Europe*, The EUROMAP Study, Copenhagen 2003

S. Kirchmeier-Andersen: Dansk Korpusbaseret forskning. Hvordan kommer vi videre? I. *Nydanske Studier 30*, 2002

B. Maegaard, S. Krauwer, K. Choukri, L. Jørgensen: The BLARK concept and BLARK for Arabic. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, Genova, 2006. p. 773-778

B. Maegaard, L. Offersgaard, L. Henriksen, H. Jansen, X. Lepetit, C. Navarretta, C. Povlsen: The MULINCO corpus and corpus platform. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genova, 2006. p 2148-2153.

Roadmap for European Research Infrastructure. Report of the Social Sciences and Humanities Working Group, ESFRI, 2006

Strategisk satsning på dansk sprogteknologi, Statens Humanistiske Forskningsråd, 2004