NEALT Proceedings

Northern European Association for Language Technology

Proceedings of the Workshop on Nordic Language Research Infrastructure

NODALIDA 2013

May 22-24, 2013 • Oslo, Norway

Linköping Electronic Conference Proceedings

Proceedings of the workshop on

Nordic language research infrastructure

at NODALIDA 2013

edited by

Koenraad De Smedt
Lars Borin
Krister Lindén
Bente Maegaard
Eiríkur Rögnvaldsson
Kadri Vider

NEALT Proceedings Series 20 • ISBN 978-91-7519-585-8
Linköping Electronic Conference Proceedings 89
ISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2013

Preface

The workshop is intended to be a forum for presenting and discussing language research infrastructure activities in Northern Europe including the Baltic countries. Recent years have seen a number of large-scale national and international initiatives and projects such as CLARIN, DASISH, EUDAT, META-NORD, INESS, etc. These are aimed at promoting access to big data and enabling eScience in the area of linguistics, language studies, philology and related fields. Some of these activities have already booked preliminary results suitable for dissemination, while others are in the planning or initial implementation stages and want to exchange ideas, learn from each other's experiences and synchronize activities. The Nordic dimension is deemed important in order to overcome local limitations, to stimulate regional cooperation, to secure interoperability, to build on previous Nordic research and to exploit common eInfrastructure solutions. The primary target audience of the workshop consists of everybody who is involved in the planning, implementation, population, operation, support or exploitation of language research infrastructure. Participants in relevant ongoing initiatives and projects as well as their user groups in Northern Europe and the Baltics were especially invited to participate.

The call for papers asked for contributions that address topics including but not limited to the following:

- linguistic web services, tool chaining and workflows
- innovative user interfaces to language resources
- cooperative linguistic annotation and documentation tools
- advances in search, filtering, mapping and mashups of language data
- data categories, pids and metadata for language materials
- 'enhanced' publications, citation and linking of language data to publications
- catalogues, repositories, archiving and curation of language data
- access, licensing, privacy and other legal, ethical and commercial aspects
- national and international networking and eInfrastructure initiatives

Six papers were submitted and all six were accepted for presentation, including one as a poster. There were three papers from Denmark, two from Norway and one from Iceland. They are included in these proceedings.

A website was established at `http://clarin.b.uib.no/nolarein/`

The program committee chairs were the following:

- Koenraad De Smedt (Bergen)
- Lars Borin (Göteborg)
- Bente Maegaard (København)
- Krister Lindén (Helsinki)
- Eiríkur Rögnvaldsson (Reykjavík)
- Kadri Vider (Tartu)

The other members of the program committee were the following:

Kristín Bjarnadóttir
Gerlof Bouma
Daan Broeder
Hanne Fersøe
Sigrun Helgadóttir
Karin Friberg Heppin
Kimmo Koskenniemi
Steven Krauwer
Krista Liin
Kadri Muischnek
Costanza Navarretta
Lene Offersgaard
Bolette S. Pedersen
Inguna Skadiņa
Jürgen Wedekind

Acknowledgements: The workshop has the support of CLARINO, DK-CLARIN and FIN-CLARIN for the participation of delegates. We thank the NoDaLiDa 2013 local organization for their practical support with registration and local arrangements. We have used EasyChair for the process of submission and reviewing.

The workshop organizers:
Koenraad De Smedt
Lars Borin
Bente Maegaard
Krister Lindén
Eiríkur Rögnvaldsson
Kadri Vider

Contents

Author index

Keyword index

Towards Large-Scale Language Analysis in the Cloud

Emanuele Lapponi[1], *Erik Velldal*[1], *Nikolay A. Vazov*[2], *Stephan Oepen*[1]

(1) Language Technology Group, Department of Informatics, University of Oslo
(2) Research Support Services Group, University Center for Information Technology, University of Oslo

{emanuel|erikve|oe}@ifi.uio.no, n.a.vazov@usit.uio.no

ABSTRACT

This paper documents ongoing work within the Norwegian CLARINO project on building a Language Analysis Portal (LAP). The portal will provide an intuitive and easily accessible web interface to a centralized repository of a wide range of language technology tools, all installed on a high-performance computing cluster. Users will be able to compose and run workflows using an easy-to-use graphical interface, with multiple tools and resources chained together in potentially complex pipelines. Although the project aims to reach out to a diverse set of user groups, it particularly will facilitate use of language analysis in the social sciences, humanities, and other fields without strong computational traditions. While the development of the portal is still in its early stages, this paper documents ongoing work towards an already operable pilot in addition to providing an overview of long-term goals and visions. At the core of the current pilot implementation we find Galaxy, a web-based workflow management system initially developed for data-intensive research in genomics and bioinformatics; therefore, an important part of the work on the pilot is to adapt and evaluate Galaxy for the context of a language analysis portal.

KEYWORDS: research infrastructure, High-Performance Computing, web portal, CLARINO.

Emanuele Lapponi, Erik Velldal, Nikolay A. Vazov and Stephan Oepen 2013. Towards large-scale language analysis in the cloud. *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* NEALT Proceedings Series 20 / Linköping Electronic Conference Proceedings 89: 1–10.

1 Introduction

This paper describes ongoing work on building a web portal for natural language analysis, carried out at the University of Oslo (UiO) as a joint effort by the Language Technology Group (LTG) and the Research Computing group at the University Center for Information Technology (USIT). The work forms part of the CLARINO infrastructure initiative,[1] the Norwegian branch of the pan-European CLARIN[2] federation (Common Language Resources and Technology Infrastructure). The aim is to provide an easily accessible web interface that ensures a low bar of entry for users, while at the same time enabling execution of complex workflows and scalability to very large data sets, integrating a wide range of tools to be run on a high-performance computing (HPC) cluster. While the development of the Language Analysis Portal (LAP) is still in its early stages, the current paper documents ongoing work towards an already operable pilot, in addition to providing an overview of long-term goals and visions. A core component of the current pilot implementation is Galaxy (Giardine et al., 2005; Blankenberg et al., 2010; Goecks et al., 2010); a web-based workflow management system initially developed for data-intensive research in genomics and bioinformatics. We here document our efforts on adapting and evaluating Galaxy for the purposes of LAP.

The paper is structured as follows. Section 2 presents a high-level overview of the many aspects related to the overall vision for LAP; ranging from target user groups and interface design to technical specifications and architecture issues. Section 3 surveys other related infrastructure projects, as well as relevant processing frameworks more generally. In Section 4, we present the details of the current LAP pilot and its implementation.

2 LAP: Language Analysis Portal

The efforts described in this paper implement a workpackage of CLARINO, the Norwegian branch of the European CLARIN initiative. CLARINO is dedicated to establishing a shared research infrastructure for language technology (LT) that ensures easy access to persistent and interoperable resources and services. A particularly important part of the mission is to facilitate the use of this infrastructure in the social sciences and humanities. LAP shares this goal in that it aims to boost the availability and usability of large-scale language analysis for researchers both within and outside of the field. In this section we present a high-level view of the kinds of functionality and services that we ultimately aim for in LAP.

Currently, many common LT tools can appear rather daunting to use, requiring a lot of technical knowledge on the side of the user. Apart from the challenge of orienting oneself in the fragmented ecosystem of available tools, many potential users, especially from less technically oriented disciplines, might not be comfortable with command-line interfaces or having to wrestle with difficult and poorly documented installation procedures, or might lack the required knowledge about annotation formats or other dependencies. Many researchers might also not have access to the computing power necessary to process larger data sets. LAP aims to eliminate such obstacles.

The goal is to maintain a large repository of LT tools that are easily accessible through a web portal, offering a uniform graphical interface. Any scholar registered in the system for federated identity management in the Norwegian education sector, Feide,[3] or the CLARIN AAI

[1] The CLARINO website: `http://clarin.b.uib.no/`

[2] The CLARIN website: `http://www.clarin.eu/`

[3] For more information, please see `https://www.feide.no/om-feide`

(Authentication and Authorization Infrastructure) will be able to log into the portal and create a user. Each user will have her own personal *workspace*, allowing data to be stored persistently across sessions. In addition to upload and storage facilities for user-provided data, the portal will also give access to common, pre-existing language resources. It will include tools for content extraction and layout analysis (from common file formats and markup schemes), as well as a comprehensive repository of language analysis tools. The portal will reach out to developers of processing tools, seeking to install the broadest possible range of technologies—ranging from token- to discourse-level analysis and encompassing both rule-based and statistical approaches. In terms of linguistic coverage, LAP will focus on languages actively used in Norway, e.g., Norwegian *Bokmål* and *Nynorsk*, Sámi, other Scandinavian languages, and English—initially at least with a focus on written language.

A central part of the interface will be a *workflow manager*, enabling the user to specify and execute a series of computations. For example, starting with a pdf-document uploaded by the user, she might further want to perform content extraction, sentence segmentation, tokenization, POS tagging, parsing, and finally identification of subjective expressions with positive polarity—all carried out in a consecutive sequence. The output of each component provides the input to the next connected component(s) in the workflow, creating a potentially complex pipeline. Note that the platform we are building on for our pilot implementation, Galaxy, comes equipped with a sophisticated workflow manager, as further described in Section 4.1. Then, after the desired workflow has been specified; at the click of a few buttons, the resources and tools involved will be configured and submitted to the national grid infrastructure, where computational and storage resources are readily available on a scale traditionally inaccessible to academic users.

This latter point, the fact that the portal will be built on top of an HPC cluster, is a crucial feature. Language technology can be computationally quite expensive, often involving sub-problems where known best solutions have exponential worst-case complexity. At the same time, typical language analysis tasks can be trivially parallelized, as processing separate documents (and for many tasks also individual sentences) constitute independent units of computation. The fact that the portal will submit the sub-tasks of a workflow to an underlying HPC cluster—without the need for user knowledge about job scheduling etc.—means that the user will be able to perform analyses that might otherwise not be possible (and faster and on larger data sets). More details about the cluster itself are presented in Section 4.3 below.

Another important requirement of the design is that it should abstract away certain low-level details—the user must be able to design and run workflows without in-depth technical knowledge of the tools or data formats involved. To make this possible the portal interface should itself have built-in awareness of inter-dependencies among component tools, standardized interchange formats and corresponding conversion procedures, etc. For example, in a given step in composing a workflow, the interface should only present the user with options that are compatible with the output of the previous step in the tool chain. Similar context-sensitive menus are also implemented in the WebLicht portal, as further described in Section 3 below.

At the same time, it is important that the abstractions provided by the interface are flexible enough to also allow for detailed control and parametrization for the more advanced users. For a given component in the workflow, the user should be able to "look under the hood" and specify parameters or options in a manner that is closer to the level of command-line interaction. The user should also have access to a recorded *history*, tracking previous actions, making it possible to reuse—or even share—workflows. Similar functionality could be used for supplying built-in or predefined workflows corresponding to typical analysis pipelines.

One of several sources of inspiration for the ideas sketched above is the positive outcome of providing an abstractly similar portal for *computational biology*, BioPortal,[4] also developed and maintained at the University of Oslo. Internationally there are also other related infrastructure projects specifically for LT, such as WebLicht project. In the next section we take a closer look at these and other related efforts and processing frameworks.

3 Related Efforts and Relevant Frameworks

Among the infrastructure projects targeting LT, the WebLicht[5] project—developed within the German branch of CLARIN (CLARIN-D) and its predecessor project D-SPIN—is the one most relevant and similar in spirit to LAP. WebLicht implements an online environment for annotating text and constructing pipelines of various LT tools. A graphical interface allows users to upload text, convert it to a system-internal exchange format, build processing pipelines using the available tools and finally visualize and download the results. The tools offered through WebLicht are distributed across repositories maintained by various CLARIN partners, operating as synchronous REST-style web services. The system-internal representation for exchanging annotations between various components in the tool chain is an XML-based format called TCF (Text Corpus Format, Heid et al., 2010), developed within WebLicht. LAP's HPC-centric design sets it apart from the current implementation of WebLicht, in that all the tools will be hosted locally and adapted to work with the national grid infrastructure. Additionally, given the generous storage and processing facilities available, LAP will allow users to upload and annotate large datasets.

Another CLARIN initiated effort is that of CLARIN-ES-LAB, where a collection of text processing tools have been made available as web services using Soaplab. [6] Finally, the Language Grid[7] initiative should also be mentioned, a Japanese analogue to WebLicht, with a focus on machine translation tools.

Another data intensive field of research that has seen the need for online portal services for computationally demanding applications is that of bioinformatics. Notably, one such portal, BioPortal, is developed and hosted at the University of Oslo. Offering a web-based interface and a high-performance computing (HPC) backbone, the user-friendly web-interface allows the users to easily manipulate their data and run complex and computationally heavy tasks without any HPC knowledge. While having already gained a large and steady user-base, the BioPortal is currently undergoing a complete re-implementation in order to even better meet the requirements of the modern research communities. Among the substantial feature enhancement are better support for reproducibility of research procedures and shareability of the scientific data (both input and output), as well a generally more multi-user centric design. The re-implementation of BioPortal is based on the Galaxy framework (Giardine et al., 2005) which we return to below.

A foundational design decision when building a system like LAP is whether to start from scratch or build on existing frameworks. Several frameworks that offer means to combine analysis services into pipelines have surfaced in the last decade, providing APIs for integrating and

[4] Please see `http://www.bioportal.uio.no` for background.

[5] The WebLicht website: `http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/`

[6] For more information in Spanish, please see `http://clarin-es-lab.org/`

[7] The Language Grid website: `http://langrid.org/en/`

developing tools as well as offering an array of pre-installed services. GATE (Cunningham et al., 2011) and UIMA (Götz and Suhre, 2004) are two noteworthy efforts in the language technology realm, while Taverna (Missier et al., 2010) and Galaxy have mostly been used for research in the life sciences. We found Galaxy to be the most suitable first choice for our pilot experiments for the simple reason that its off-the-shelf feature-set seems, for the most part, compatible with our vision for the portal. While GATE, UIMA and Taverna all offer graphical interfaces that allow end-users to combine tools visually, they are required to install desktop applications which, to different degrees, may be difficult to set up due to, for instance, dependencies to other software packages.

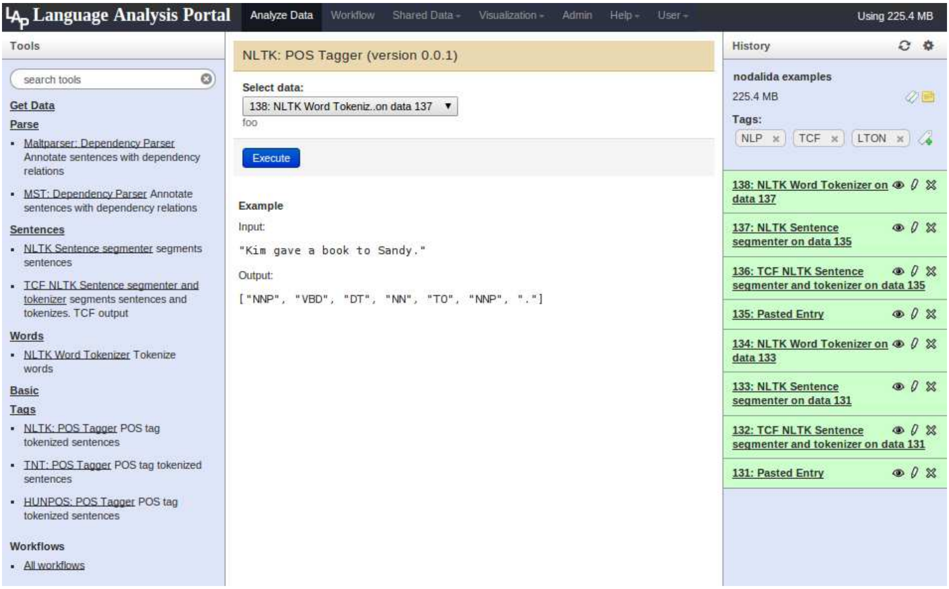
Though systems like UIMA, GATE and Taverna can be ported to work within a browser, the Galaxy framework is natively a web-application, offering a full-blown, intuitive interface, eliminating the need to spend developer time on porting applications to the browser. Taverna and Galaxy take a similar approach to visual pipelining, allowing users to draw directed graphs where each node represents a processing tool with inputs and outputs; Galaxy, however, also implements workspaces that allow users to organize files in different groups, called *histories*, where the outputs generated from each annotator in the workflow are collected.

Another reason for adopting Galaxy for LAP is to cooperate and exchange knowledge with the BioPortal developers at USIT, who are currently re-implementing the system on top of Galaxy. This means we can benefit from existing local expertise when it comes to adapting and integrating tools, connecting it to an HPC backbone, communicating with authentication services, and other technical challenges.

Turning again to BioPortal, it is worth mentioning that the decision to adopt Galaxy as a replacement for the existing implementation was partly due to the large Galaxy developer community and the framework's infrastructure. Galaxy also possesses some characteristics which make it very attractive both for the common user and developers / administrators: Firstly, Galaxy is very "collaborative" in that it allows users to share workflows and intermediate or final results of the data processing at any stage of the computation. This sharing procedure is quick and entirely user-driven. Galaxy supports a complex organization of users into groups (roles) with different levels of cross-group access and permissions over the datasets. Moreover, Galaxy not only facilitates the installation of new resources but also the tuning and maintenance of the existing ones through a web-tool middleware layer. This layer renders the administration of the resources fast and easy. Finally, due to the web interface, Galaxy can easily be adapted as a front-end to large computational resource(s), like storage, grids, and cloud services.

4 Creating a LAP Pilot

In this section we describe the details on implementing a preliminary pilot instance of LAP. While the pilot is already operable, this should still be considered as work in progress. Creating a pilot serves several purposes: Most importantly it will act as a *proof-of-concept*, assessing the viability of involved software as well as ideas before extending the implementation to a larger set of tools and support for a larger set of features. The most important part of this, in turn, is assessing the suitability of the Galaxy platform. Another important use of the pilot will be as a *demonstrator*; for reaching out to tool developers, to illustrate use cases for potential user groups in the humanities and social sciences, and as a foundation for further surveying user-requirements. The pilot is only meant to support a minimal selection of LT tools and will be evaluated in part by a group of test users consisting of master students of the study program *Informatics: Language and Communication* at the University of Oslo.



Get Data

Parse

- Maltparser: Dependency Parser Annotate sentences with dependency relations
- MST: Dependency Parser Annotate sentences with dependency relations

Sentences

- NLTK Sentence segmenter segments sentences
- TCF NLTK Sentence segmenter and tokenizer segments sentences and tokenizes. TCF output

Words

- NLTK Word Tokenizer Tokenize words

Basic

Tags

- NLTK: POS Tagger POS tag tokenized sentences
- TNT: POS Tagger POS tag tokenized sentences
- HUNPOS: POS Tagger POS tag tokenized sentences

Workflows

- All workflows

NLTK: POS Tagger (version 0.0.1)

Select data:
138: NLTK Word Tokeniz..on data 137
foo

Example

Input:

```
"Kim gave a book to Sandy."
```

Output:

```
["NNP", "VBD", "DT", "NN", "TO", "NNP", "."]
```

nodalida examples
225.4 MB
Tags: NLP TCF LTON

138: NLTK Word Tokenizer on data 137
137: NLTK Sentence segmenter on data 135
136: TCF NLTK Sentence segmenter and tokenizer on data 135
135: Pasted Entry
134: NLTK Word Tokenizer on data 133
133: NLTK Sentence segmenter on data 131
132: TCF NLTK Sentence segmenter and tokenizer on data 131
131: Pasted Entry

Figure 1: A screenshot of the current, work-in-progress implementation of LAP within Galaxy.

Currently, our engineering efforts are focused on four areas: integrating and adapting a preliminary selection of LT tools, modifying the Galaxy UI, integrating and evaluating interchange formats, and implementing the connection to the HPC cluster (Abel). Below we discuss these issues in turn.

4.1 Galaxy and Workflows

Figure 1 shows the three main panels of the Galaxy UI as adapted for LAP. The left panel contains a list of the installed tools and available workflows; clicking on an item updates the center panel with its details, allowing users to run (or re-run) the tool using one of the elements present in the history panel on the right. Users can create as many histories as they need, using tags and a short description to keep potentially large collections of files in order. The center panel also houses the workflow manager, where processing pipelines such as the one pictured in Figure 2 are designed. Additionally, Galaxy makes histories and workflows searchable, allowing users to share them and collaborate.

While the larger LAP user-group includes researchers and students from the humanities, social sciencies, linguistics and language technology itself, the pilot release of the portal will only focus on the latter. A typical use case for the working language technologist, and one that the LAP pilot will provide the means to accomplish, could involve annotating a large text corpus, e.g., a snapshot of Wikipedia, with syntactic dependencies. Furthermore, the researcher or student in question could be interested in producing annotations generated using different parsers, that are in turn invoked with part of speech tags that originate from different upstream annotators. Such an endeavour would minimally include the following steps: (a) log into the

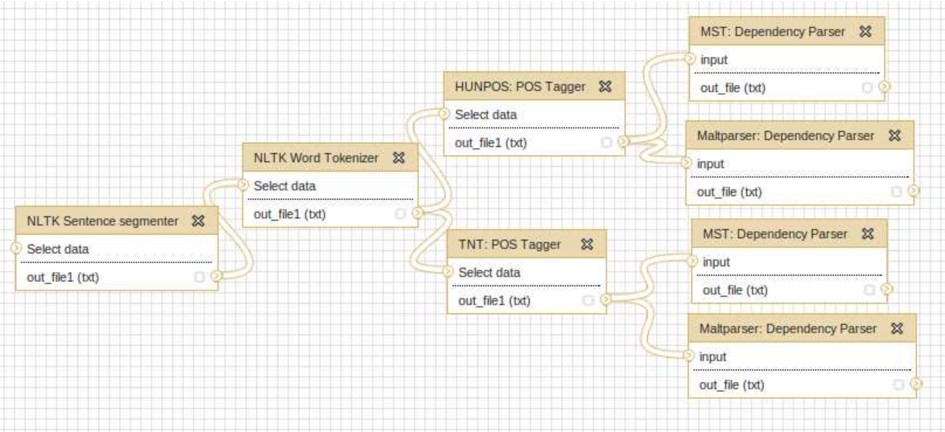


Figure 2: A LAP workflow with four endpoints.

LAP workspace; (b) Create a *history* for the experiment; (c) design and save the workflow, like the example given in Figure 2; (d) run the workflow and (e) download the output files when completion is notified (either on-screen or via email).

4.2 Interchange formats

In order make a heterogeneous set of language processing tools interoperable, datasets have to be converted to and from the required tool-specific representations at each step of the processing. In this context, interchange formats work as a kind of 'trade language' between the tools in the chain. LAP aims to be compatible with other CLARIN-related projects in terms of interchange formats, and provide tools that enable converting to and from widely adopted representations, like the tab-separated CoNLL 2007 format or Penn Tree Bank-style phrase-structure trees.

For LAP's system-internal representation, we are at the moment looking into both TCF (the format used within WebLicht, which comes with a full, albeit closed-source API) and our own in-house JSON-based LTON format (Language Technology Object Notation) which is still currently under development. Figure 3 shows an example of both representations after running the NLTK (Bird et al., 2009) sentence segmenter and tokenizer on the text "*Sandy barks. Kim Snores.*". In TCF, tokens are the smallest unit in the representation and serve as anchor points for downstream processing, while in LTON corpora are annotated according to a notion of *annotation levels* (e.g., sentence, paragraph, document and so on), with lower-level annotations being encapsulated within higher ones. Finding a suitable internal representation is by no means a trivial task, especially given that relevant datasets may potentially be very large, and rapidly increase in size and complexity even further due to annotations accumulating through involved workflows. Evaluation and integration of these two interchange formats is currently ongoing and a full LAP-implementation of both is expected in time for the pilot release.

In order to be integrated in the LAP tool-chain, existing tools are 'wrapped' inside scripts that decode the LAP-internal format, present the tool itself with its expected input and finally re-encode the output so that it is compatible with the next processing step. Additionally, the

TCF

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<?xml-model href="http://de.clarin.eu/images/weblicht-
       tutorials/resources/tcf-04/schemas/latest/d-spin_0_4.
       rnc" type="application/relax-ng-compact-syntax"?>
<ns5:D-Spin xmlns="http://www.dspin.de/data/metadata" xmlns:
       ns2="http://www.dspin.de/data/extdata" xmlns:ns3="http
       ://www.dspin.de/data/textcorpus" xmlns:ns4="http://www.
       dspin.de/data/lexicon" xmlns:ns5="http://www.dspin.de/
       data" version="0.4">
    <MetaData/>
    <ns3:TextCorpus lang="en">
        <ns3:text>Sandy barks. Kim snores.</ns3:text>
        <ns3:tokens>
            <ns3:token ID="t_0">Sandy</ns3:token>
            <ns3:token ID="t_1">barks</ns3:token>
            <ns3:token ID="t_2">.</ns3:token>
            <ns3:token ID="t_3">Kim</ns3:token>
            <ns3:token ID="t_4">snores</ns3:token>
            <ns3:token ID="t_5">.</ns3:token>
        </ns3:tokens>
        <ns3:sentences>
            <ns3:sentence tokenIDs="t_0 t_1 t_2"/>
            <ns3:sentence tokenIDs="t_3 t_4 t_5"/>
        </ns3:sentences>
    </ns3:TextCorpus>
</ns5:D-Spin>
```

LTON

```
{
    "annotations": [
        {
            "annotators": {
                "token:nltk": [
                    "Sandy",
                    "barks",
                    "."
                ]
            },
            "text": "Sandy barks."
        },
        {
            "annotators": {
                "token:nltk": [
                    "Kim",
                    "snores",
                    "."
                ]
            },
            "text": "Kim snores.\n"
        }
    ],
    "level": "sentence:nltk",
    "name": "/lap-galaxy/database/files/000/dataset_138.dat"
}
```

Figure 3: An example of a toy corpus annotated with sentences and tokens, formatted in the two candidate interchange formats for LAP.

wrapper handles the submission of the job to the Abel cluster.

4.3 The Abel HPC Cluster

Abel[8] is the name of the high performance computing facility at UiO, hosted by the USIT Research Computing group. The powerful Linux cluster is a shared resource for research computing, boasting more than 600 machines, totaling more than 10.000 cores (CPUs). At the time of writing, the cluster ranks at position 134 on the list of top 500 supercomputers worldwide.[9] Among its frequent users, besides the language technology group, we find researchers from the life sciences, astrophysics, geophysics, and chemistry.

When executing a workflow from the LAP instance of Galaxy, each component task will in turn be submitted to the job queue on the Abel cluster. Control is then temporarily delegated to the cluster queue system—using a job scheduler and resource manager called SLURM[10]—before the produced output is finally returned to Galaxy. An important part of the work on adapting Galaxy for LAP (and the BioPortal) is to make this connection as seamless as possible.

For the pilot release, LAP's toolshed will provide enough language processing tools to enable a user test session involving master students of the language technology program at UiO. The inventory will include typical annotators such as sentence segmenters, tokenizers, lemmatizers, chunkers and syntactic parsers; these will, to a varying degree, be configurable in terms of e.g., models, tagsets and syntactic paradigms. In terms of language coverage, the pilot will provide tools for processing Scandinavian languages in addition to English. The test session, which is planned for early Q2 2013, will pave the way for further development.

[8] For more detailed information on the Abel computing cluster, see `http://uio.no/hpc/abel/`.

[9] For more information about the ranking, please see `http://www.top500.org`.

[10] Simple Linux Utility for Resource Management

5 Conclusion and Outlook

This paper has laid out the long-term goals and plans for creating an online portal for language analysis (LAP). The effort forms part of the CLARINO infrastructure project and one of the overall goals is to make language technology readily accessible and usable for researchers from the humanities and social sciences. With easy access through Feide (and CLARIN AAI) authentication, the web portal will provide a uniform graphical interface to a large repository of LT tools installed on a high-performance computing cluster. Users will be able to create complex workflows using an intuitive interface, and each user will have access to a personal workspace for storing data persistently across sessions. In terms of linguistic coverage, LAP will focus on languages most actively used in Norway, e.g., Norwegian *Bokmål* and *Nynorsk*, Sámi, other Scandinavian languages, and English.

The development of the portal is still in its early stages, and in addition to presenting the overall vision of LAP in its final state, this paper has documented the work towards an already functioning pilot. This preliminary version of the portal will act as a proof-of-concept, helping to inform strategic decisions about the remaining work, as well as a demonstrator that can be used when reaching out to tool developers and when surveying user needs.

One of the core components of the current pilot version of the portal is the Galaxy workflow manager—a web-based system initially developed for data intensive research in the biomedical domain. Galaxy is already deployed in a portal for bioinformatics research, BioPortal, also hosted at the University of Oslo. This means we can benefit from existing local expertise when it comes to adapting and integrating tools in Galaxy, connecting it to an HPC backbone, and other technical challenges.

The pilot release of the system will address the requirements of users with a language-technological background, with a closed user-test session planned for early Q2 and an open pilot release in Q3 2013. Further work will address the challenge of shaping LAP into a useful research tool for the humanities and the social sciences, investigating possible use-cases and collecting user-requirements from active researchers from these fields.

References

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, pages 19.10.1–19.10.21.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–5.

Goecks, J., Nekrutenko, A., Taylor, J., and Team, T. G. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86.

Götz, T. and Suhre, O. (2004). Design and implementation of the UIMA common analysis system. *IBM Syst. J.*, 43(3):476–489.

Heid, U., Schmid, H., Eckart, K., and Hinrichs, E. (2010). A corpus representation format for linguistic web services: The D-SPIN Text Corpus Format and its relationship with ISO standards. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 494–499.

Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., Williams, A., Oinn, T., and Goble, C. (2010). Taverna, reloaded. In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management*, pages 471–481.

Workflow Management in CLARIN-DK

Bart Jongejan

Copenhagen University

bartj@hum.ku.dk

ABSTRACT

Clarin.dk, the infrastructure maintained by the CLARIN-DK project, is not only a repository of resources, but also a place where users can analyse, annotate, reformat and potentially even translate resources, using tools that are integrated in the infrastructure as web services. In many cases a single tool does not produce the desired output, given the input resource at hand. Still, in such cases it may be possible to reach the set goal by chaining a number of tools. The approach presented here frees the user of having to meddle with tools and the construction of workflows. Instead, the user only needs to supply the workflow manager with the features that describe her goal, because the workflow manager not only executes chains of tools in a workflow, but also takes care of autonomously devising workflows that serve the user's intention, given the tools that currently are integrated in the infrastructure as web services. To do this, the workflow manager needs stringent and complete information about each integrated tool. We discuss how such information is structured in clarin.dk. Provided that many tools are made available to and through the clarin.dk infrastructure, the automatically created workflows, although simple linear programs without branching or looping constructs, can cover a large swath of users' needs. It is rewarding for both users and tool developers that the infrastructure takes advantage of new tools from the moment they are registered, because there is no need to wait for human expert users to construct and save for later use workflows that incorporate new tools.

KEYWORDS: NoDaLiDa 2013, workflow, tools, automation.

Bart Jongejan 2013. Workflow management in CLARIN-DK. *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* NEALT Proceedings Series 20 / Linköping Electronic Conference Proceedings 89: 11–20.

1 Introduction

Using computer software to analyse resources in the field of humanities can be a difficult to attain goal, because software packages often require a good deal of technical stamina. Even if all software is available to a researcher, the prospect of having to deal with technical details may deter many, for example because output from one piece of software not necessarily can be used as input for another piece of software due to a technical issue such as a format mismatch. Very powerful and versatile workflow managers exist that are used to alleviate the burden of finding viable combinations of tools, but usually this is done by assisting the user in selecting tools for each step in a possibly long chain of tools. Even though such workflow managers take care of the fitting-together, users still have to learn to use a workflow editor to program a tool chain. Examples of such workflow managers are WebLicht (Hinrichs et al., 2010), and Taverna (Kemps-Snijders et al., 2012), Kepler and Triana (Funk et al., 2010).

Reuse of existing technology for workflow management in clarin.dk[1] (Offersgaard et al., 2011, Offersgaard et al., 2013) was seen as problematic. WebLicht was mostly text only, whereas the clarin.dk repository had to be populated with resources of many types, some of which had little to do with text, such as videos. Also, from the outset it was decided to work with stand-off annotations, each annotation being a resource in its own right. This design did not match very well with WebLichts preferred TCF-format that combines all annotations in one file that is passed on from tool to tool, getting enriched on its itinerary until the end of the tool chain is reached. Other workflow managers would require that at least one user knew the tools that were integrated and became expert in using the workflow editor. Not before this expert user had made a set of workflows, other, less technically minded users, could reap the fruits of the integrated tools. In a relatively small NLP community like the Danish, this was not an attractive prospect. The minimal requirement was that new tools for any kind of resource could be integrated into the infrastructure without the need to make changes to the infrastructure's software and with the possibility for all users to easily apply tools to the resources of their choice.

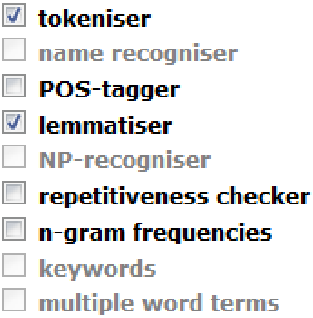
2 Baseline for handling workflows

From a user's perspective, our baseline is the web site[2] where we have made available a number of NLP tools for on-line use. Here, researchers and students, but also consultants and IT-people all over the world, can get acquainted with some of our software. This web site is a small user-friendly laboratory where the user can pick tools and chain them into a workflow. Where needed, tools are automatically, yet visibly, added to the tool chain to fulfil the prerequisites of the selected tools further down the road. Only valid combinations of tools can be made, because tools that are incompatible with the current choices are 'greyed out' and made ineligible. See Fig. 1.

[1] https://www.clarin.dk/

[2] http://cst.dk/tools/index.php?lang=en

The interface is attractive because it is kept very simple and transparent at a level that is interesting for the user, while boring details are hidden away. In this way, users are gently drawn into the world of NLP and can do simple experiments without all the technical fuzz.



☑ tokeniser
□ name recogniser
□ POS-tagger
☑ lemmatiser
□ NP-recogniser
□ repetitiveness checker
□ n-gram frequencies
□ keywords
□ multiple word terms

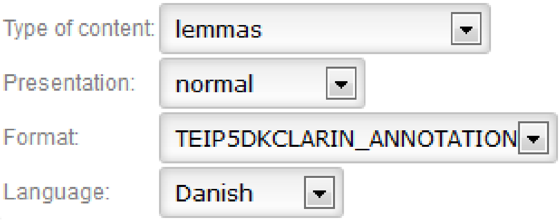
FIGURE 1: The user has chosen the lemmatiser. The system has automatically added the prerequisite tokeniser. The tools in grey text are disabled.

This site has however also its disadvantages. Each new tool has to be integrated by hand, checking out all combinations of tools with the newcomer that are meaningful and above all, disabling all those combinations that aren't. Here, errors are easily made. Another shortcoming is that the web site only takes text, flat or RTF, as input. If we would have to implement support for a wider range of input types, an enormous programming task would lie ahead. And finally, the earlier mentioned focus on tools can also turn into a disadvantage. For many students and researchers the results are what counts and not the tools that are needed to create those results.

3 Tool handling in clarin.dk

Clarin.dk's workflow manager has the same advantages as our base line: tools can be chained into a workflow, tools are automatically added if required by other tools and only valid combinations of tools can be put together in a workflow. There are three big improvements, though. The first is that the system is indifferent to the type of input and output. Not only text, but also pictures, audio and video can be handled. The second improvement is the much easier integration of new tools, which basically takes place by telling the workflow manager how the input must look like, how the output will look like and where the tool, as a web service, is found. This information is stored in a few tables, instead of being expressed in program code as was the case in the baseline. For describing input and output, the tool provider must specify a small number of features by choosing values from predefined lists. If the need arises, an administrator can add values and even features without having to change the workflow algorithms at all. Some feature values can be subspecified, again using any of a number of predefined values. Besides this registration, a tool provider also has to wrap the tool in a web service that is able to communicate with the clarin.dk infrastructure. Because tools are described in a neutral, uniform language, each tool can be seen as isolated from all other tools. Problematic interactions between tools must be solved by correcting the registered information, or by improving the terminology used for the registration and adapting tools so they better match the expectations as registered in the specifications. For each tool, its service URL and its input/output specification is the only contract with the infrastructure that is vital for employing the tool.

The last and, from a user's perspective, most important improvement is that the user only has to specify a goal by choosing values from a few (four, at the moment) drop-down lists (see Fig. 2.) and can leave it to the workflow manager to find a route to that goal from the given input, in the same way that route planning software finds a route between two locations.



Type of content: lemmas

Presentation: normal

Format: TEIP5DKCLARIN_ANNOTATION

Language: Danish

FIGURE 2: Goal specification in clarin.dk

Our approach is reminiscent of the ALPE model (Cristea & Pistol, 2008), which also supports automatic creation of workflows. In contrast to ALPE, however, the I/O specifications in clarin.dk's workflow manager are not hierarchically structured, and the output from a tool is not necessarily an augmented version of the input.

4 Registration of tools

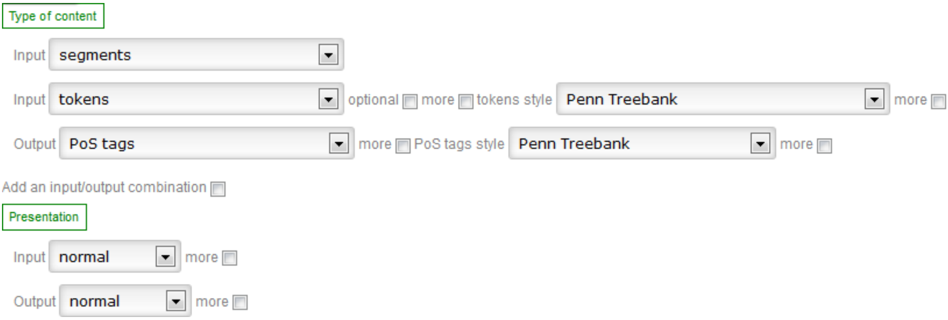
The tool registration facility has two sections, a boilerplate section, and a section dedicated to I/O specification where one or more features are specified. The workflow manager stores all registered tool information in a dedicated database. Most fields in the boilerplate section and a summary of the I/O specifications are replicated as CMDI metadata in the repository. The user can search and view the CMDI metadata, like the metadata of any other resource. Sensitive data are not publicly visible.

A number of boilerplate fields are essential for the operation of the workflow manager: the short *ToolID*, the *PassWord*, which is used to pass maintenance responsibility to another person, the *ServiceURL*, *XMLparms*, to tell the workflow manager that the tool cannot handle simple HTTP parameters but requires parameters in XML-format, *PostData,* telling whether requests should be sent using the GET or POST HTTP method, and *Inactive*, telling whether the tool is off line at the moment.

There can be several I/O specification sections for each boilerplate section, and within an I/O specification section it is also possible to enter several alternative I/O feature specifications. The need for different sections arises in those cases where a choice of one feature has an influence on what can be chosen for another feature. An example is a tool that can take part-of-speech tags as an extra input, but only if the language is Danish or English. In that case there would be two sections, one for Danish and English saying that part-of-speech tags can be taken as input as well, and another section for all other languages, where part-of-speech tags are not mentioned as an option. In other frameworks one may have to register each possible combination of specifications as separate profiles, which can be a lot of work if a tool supports e.g. ten languages, two file formats and three ways of presenting the output, which by multiplication result in tens of profiles.

The workflow algorithms treat all features on a par; there is no ‘most important’ feature. If a feature is not relevant for a tool, it should not be specified. In the current system we have good experiences with the following features: *language*, *file format*, *facet* and *presentation*. The *facet* feature describes the type of content, e.g. whether data is text, a part-of-speech annotation of a text, or gestures occurring in a video. The *presentation* feature is to humans what *file format* is to the computer. With it, we want to express that for example the tokens in a text can be presented in the ‘normal’ way (‘running text’), or as lists sorted alphabetically or according to frequency. The same choice of presentation can be made for several other facets, such as part of speech tags and lemmas, so *presentation* and *facet* are truly orthogonal features.

The last example highlights the working hypothesis that features are and must be orthogonal and vice versa, that characteristics that can be combined rather freely are indicative of the existence of several features. As another example, a resource with *facet* feature *text* can be expressed in several *file format*s, such as an *image* of a page of a book, a *flat* sequence of characters or an *audio* file with spoken words. And from an abstract stance, one could say that some texts exist in several *language*s.



Type of content

Input segments

Input tokens optional □ more □ tokens style Penn Treebank more □

Output PoS tags more □ PoS tags style Penn Treebank more □

Add an input/output combination □

Presentation

Input normal more □

Output normal more □

FIGURE 3: Part of registration form, showing fields for *facet* and *presentation* features. The values *tokens* and *PoS-tags* are subspecified as *Penn Treebank*. Where needed fields can be added by checking the square check boxes.

Besides values for each of these features, a tool provider in some cases also is offered the possibility to further specify a feature, in the same way as MIME types[3] consist of a media type and a media subtype. For example, once the value *image* is chosen for the feature *format*, one can choose image subspecies like *JPEG*, *TIFF*, *GIF*, etc. Here the ruling idea is that subspecifications belong to the realm of technical details that we don’t want to disturb the user with unnecessarily. This is in keeping with everyday software. For example, image viewers handle wide ranges of image formats but do not

[3] http://www.iana.org/assignments/media-types

require that the user knows or even is aware of these formats. Also NLP-software may be able to read and write data in a range of formats. Nonetheless, if it is known that a tool is restricted to or has a preference for a narrow set of formats, it is possible to enumerate these during tool registration. This information helps the workflow manager getting around I/O mismatches between tools. See Fig. 3.

It is clear that devising the lists of features and feature subspecies is more of an art than science, and it is also clear that such lists will evolve in different directions in different communities, if no co-ordination is done. We have chosen to let the lists grow and evolve as new tools pose requirements that are not expressible with the current values. For interoperability between infrastructures, when adding new features we always attempt to reuse existing terminology.

The possible values for each of the four features *format*, *language*, *facet* and *presentation* are as follows.

The *format* feature can take the values *Anvil*, *audio*, *CSV*, *flat*, *HTML*, *image*, *PDF*, *RDF*, *RTF*, *TEIP5*, *TEIP5DKCLARIN*, *TEIP5DKCLARIN_ANNOTATION*, *XML*, or *video*.

The *language* feature can take 50 values. Only one of these, *Xhosa* (*xh*) is not explicitly supported by any currently integrated tool.

The *facet* feature can take the values *anonymized named entities*, *head movements*, *keywords*, *lemmas*, *lexicon*, *multiple word terms*, *N-gram frequencies*, *named entities*, *noun phrases*, *paragraphs*, *PoS tags*, *repeated phrases*, *segments*, *tagged terms*, *text*, *tokens*, and *morphemes*. This list represents the characteristics offered by the currently integrated tools and some tools that are not integrated, but which we are considering.

The *presentation* feature can have the values *alphabetic list*, *frequency list* and the intentionally vague '*normal*'.

Some combinations of feature values may seem far-fetched and can cast doubt on the tenability of the whole idea of orthogonality of features. An example is the combination of the *video* format and the *N-gram frequencies* facet. On the other hand, this may be an acceptable characterization of a video resource that lists short sequences of hand movements, sorted by frequency. Such a resource may be useful in the study of e.g. sign language. That is not to say that *all* combinations must be meaningful, but the great majority of combinations should make sense.

Currently twelve tools are installed as web services on several servers and made integral part of the infrastructure. Some tools have narrow sets of specifications, whereas others would expand into hundreds of profiles, if registering alternative I/O specifications for the same tool hadn't been allowed. The tools cover a range of formats, languages and facets.

1) *Cuneiform*[4], OCR for 23 languages,

2) *RTFreader*, a basically language insensitive program that reads *rtf* or *flat* text and writes *segments*, optionally *tokenised*, in *flat* text format,

[4] http://cognitiveforms.com/ru/products_and_services/Cuneiform.html

3) *Flat2cbf*, a program that converts *flat* text into an implementation of *TEIP5*, the CLARIN-DK basis text format *TEIP5DKCLARIN*.

4) - 7) *tokenisers* and *segmenters* for Danish and for English, taking *TEIP5DKCLARIN*-formatted text as input and producing stand-off annotations in another *TEIP5* implementation, the *TEIP5DKCLARIN_ANNOTATION* format.

8) *Brill's POS-tagger* for Danish and English,

9) *CST's lemmatiser* for 10 languages, also capable of producing a sorted list of lemmas

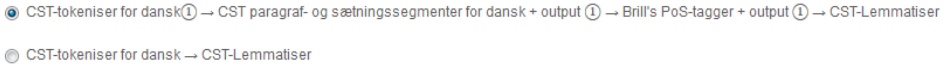
10) *espeak*, a very basic TTS-system for 42 languages,

11) a utility bundling *tokens*, *lemmas* and *part of speech* tags into *CoNLL-X* format, and

12) Bohnets parser[5], a syntactic dependency parser, currently for Danish only.

5 Computation of workflows

The computation of a workflow is a recursive process that starts from the user's goal and that works towards the input specification. Given a goal, the algorithm checks whether it is compatible with what is known about the input. If that is the case, a workflow solution is found. If the goal is not compatible with the know input features, the algorithm finds all tools that produce output that match the goal's features, also taking subspecifications of features into account. The input requirements of these tools are new goals, each of which is analysed in a recursive manner. See Fig. 4 for an example where two workflows were generated that both can satisfy the goal as specified in Fig. 1



● CST-tokeniser for dansk① → CST paragraf- og sætningssegmenter for dansk + output ① → Brill's PoS-tagger + output ① → CST-Lemmatiser

○ CST-tokeniser for dansk → CST-Lemmatiser

FIGURE 4: The workflows that satisfy the goal that is set in Fig. 1. By clicking a radio button, the user can choose between a tool chain of a tokeniser, a segmenter, a PoS-tagger and a lemmatiser, or a tool chain consisting of a tokeniser and a lemmatiser.

We let the recursion halt at an arbitrarily chosen maximum recursion depth of 20, which means that we do not want to ever see workflows with more than 20 tools in a chain. Once an unbroken chain of tools connects the input with the output, any features that are not specified in the goal, but specified in the input or in any of the intermediate stages, are percolated through the chain. For example, if the goal does not state the language, but it is known that the input's language feature has the value *French*, then this value is percolated until it arrives at the goal or until it stumbles into a tool that for example translates from French to Danish, in which case the value *Danish* would percolate further in the direction of the goal. Any parameters (features or subspecifications of features) that are still unresolved hereafter are not decidable for the workflow manager. In such cases the workflow manager constructs *all* fully specified workflows that are compatible with the underspecified workflow and the user is given the choice between these alternatives. If the user did not specify all four goal features,

[5] http://code.google.com/p/mate-tools/

she can normally reduce the number of generated workflow candidates by specifying more features. However, workflows can contain parameters that cannot be resolved by refining the goal specification. Such parameters must either be resolved by the user or by a heuristic implemented in the software. This is not yet done in the current system. Therefore we see that goals, especially those that need many steps, sometimes generate pages full of workflows.

6 Execution of workflows

The workflow manager functions as the hub in the execution of a workflow. The workflow manager starts with sending a request to the web service that takes the initial input. Eventually, the web service returns its result, or, if something went wrong, an HTTP error status code. In the first case, the workflow manager inspects a list of pending workflow steps, picking out and executing those that can be executed given the returned data. When there are no more pending workflow steps, the workflow manager creates a report of the workflow process, step by step naming all tools and the inputs and outputs the tools produced. Also, metadata is constructed for each result, describing how the data was created from earlier results or from the input. These provenance data, together with the intermediary and final results, is compressed in a zip-archive and stored for a limited time, currently a few days. The user receives an email with a link to a web page. When clicking the link, the web page opens in the user's browser. The user will see the report and a download link to the zip file.

For data security reasons, the user is given only one opportunity to download the zip file, which is deleted from the server as soon as it has been fetched by someone. We have chosen this rigid strategy to make it difficult to intercept the data undetected. If there is an eavesdropper on the line who fetches the results before the user tries to do that, the user will notice that something went wrong and likely ring an alarm.

After inspecting the results, the user has the opportunity to deposit the result in the clarin.dk repository, together with the intermediary results, provided that they are of a type that can be deposited. Before depositing results, the user must check and complete the metadata, because the automatically created metadata necessarily lack background knowledge that only the user can provide. The depositing of results is taken care of by a service dedicated to the depositing of resources in general, and not the workflow manager itself.

The workflow manager can asynchronously send requests to several web services at the same time, provided that the web services return immediately with an HTTP status code *202 Accepted* and send results later.

If a web service returns another status code than *200 OK* or *202 Accepted*, the remainder of the workflow is aborted and the user receives an email telling where the workflow got off track. The user still receives the results that were created successfully.

7 Limitations and solutions

The chosen approach has its problems and disadvantages, like any robotic system that tries to replace a human expert.

Whereas user-friendliness and accepted standards are good organizing forces, there is also a bad, but at times inescapable, disorganizing force. Sometimes, when we try to register a new tool, we may discover that a subspecification of a feature value itself needs further specification. But since the architecture only allows two levels of specification, we are forced to heighten the status of the subspecification to that of a feature value, and to introduce the further specifications as subspecifications of the new feature value. If this causes too much terminological pain and exposes too much technical detail to the user, we may have to reconsider the integration of the tool that is in need of the extra distinction. Is the tool really mature, or should it be adapted to accept a wider range of inputs before we attempt to integrate it in the infrastructure? It must be added that changing the existing terminology should not be done light-heartedly, because not only will some tools have to be re-registered, the web services that wrap around these tools will have to be fixed as well, because they will receive parameters with altered names.

On a more theoretical level, we may question whether all goals can be stated in only a few (1-4) words. However, adding even a single extra field to supplement the goal description can easily be more confounding than of help, because after the features *language* and *file format* it is very hard to define generic features that feel natural for users.

Because workflows are created on the fly, they do not have persistent metadata and identity of their own. Yet, together with the output, the user receives a full specification of each step in the workflow, so the same workflow can be chosen the next time the user wants to do a similar task. This workflow documentation only partly compensates for the lack of persistent workflows. Therefore we have plans to make a resource type ‘Workflow’ that can be stored in the repository and reused later. This will make it easier to repeat a workflow at a later time.

Certain tools and kinds of input do not easily fit in the chosen scheme. Tools that require user interaction cannot be integrated, because the execution of workflows is in batch mode. Neither can tools that output metadata rather than data, such as language guessers, be incorporated in our workflows, because they would constitute decision points in workflows that cannot be computed on beforehand. As mentioned before, the automatically generated workflows are linear sequences of instructions, without the possibility to react to conditions that arise at run time, such as taking one or the other workflow branch depending on the outcome of a language guesser.

8 Conclusion and Outlook

The clarin.dk workflow manager is fully and reliably functioning, and response times of the user interface are very fast. We already have a modest number of tools, some of which are very versatile and can be seen as Swiss Army knifes among NLP tools. We have concrete plans to add more tools and also welcome contributions from other tool providers.

The chosen approach cannot completely replace workflow strategies as implemented in full scale workflow editors, because workflows created by the clarin.dk workflow manager are chains without branching points that depend on the results of earlier steps, but the workflows that can be realised, are made available in a user friendly manner, not requiring expert knowledge and making optimal use of the available tools.

From a programmer's point of view, it is very rewarding to see that an accurate description of a tool is enough to see it pop up as a step in a workflow, and it is even more rewarding to see that it works. We think that tool providers appreciate this interaction, and that they invest time in improvements to their tools to make them more general, robust and of higher quality.

Acknowledgments

DIGHUMLAB Digital Humanities Lab Denmark is supported by Danish Agency for Science, Technology and Innovation, Ministry of Science, Innovation and Higher Education.

The preparatory project DK-CLARIN was also supported by the Danish Agency for Science, Technology and Innovation, as well as by all eight partner institutions. We want to thank all partners for their contribution.

References

Cristea, D., Pistol, I. (2008): Managing Language Resources and Tools Using a Hierarchy of Annotation Schemas. In *Proceedings of the Workshop on Sustainability of Language Resources*, LREC-2008, Marrakech.

Funk, A., Bel, N., Bel, S., Büchler, M., Cristea, D., Fritzinger, F., Hinrichs, E., Hinrichs, M., Ion, R., Kemps-Snijders, M., Panchenko, Y., Schmid, H., Wittenburg, P., Quasthoff, U. and Zastrow, T. (2010): *Requirements Specification Web Services and Workflow Systems*. Available at: http://www-sk.let.uu.nl/u/D2R-6b.pdf

Hinrichs, E., Hinrichs, M., Zastrow, T. (2010) WebLicht: web-based LRT services for German. In *ACLDemos '10 Proceedings of the ACL 2010 System Demonstrations*, Pages 25-29, Association for Computational Linguistics Stroudsburg, PA, USA.

Kemps-Snijders, M., Brouwer, M., Kunst, J. P. and Visser, T. (2012): Dynamic web service deployment in a cloud environment. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, May 23-25, 2012: 2941-2944, European Language Resources Association (ELRA)

Offersgaard, L. Jongejan, B. and Maegaard, B. (2011). How Danish users tried to answer the unaskable during implementation of clarin.dk. In *SDH 2011 – Supporting Digital Humanities*, Copenhagen.

Offersgaard, L., Jongejan, B., Seaton, M. and Haltrup Hansen, D. (2013). CLARIN DK – status and challenges. In *Proceedings of the Nordic Language Research Infrastructure Workshop at NoDaLiDa, Oslo, May 22, 2013*

CLARIN-DK – status and challenges

Lene Offersgaard, Bart Jongejan, Mitchell Seaton,
Dorte Haltrup Hansen

University of Copenhagen
Njalsgade 140, DK-2300 Copenhagen S
Denmark

leneo@hum.ku.dk, bartj@hum.ku.dk, seaton@hum.ku.dk, dorte@hum.ku.dk

ABSTRACT

The initiative CLARIN-DK (starting as a Danish preparatory DK-CLARIN project) is a part of the Danish research infrastructure initiative, DIGHUMLAB. In this paper the aims, status, and the current challenges for CLARIN-DK are presented. CLARIN-DK focuses on written and spoken language resources, multimodal resources and tools, and involving users is a core issue. Users involved in a preparatory project gave input that led to the current user interface of the resource repository website, clarin.dk. Clarin.dk is now in the transition phase from a repository to a research infrastructure, where researchers and students can be supported in their research, education and studies. Clarin.dk works with a Service-Oriented Architecture (SOA), uses eSciDoc and Fedora Commons, and is primarily based on open source solutions. A key issue in CLARIN-DK is using standards such as TEIP5, IMDI, OLAC, and CMDI for resource metadata. Optional metadata fields suggested by users have been included when it could comply with the standards, allowing for the diversity needed when describing the research material. Current work includes normalising metadata naming in the search pages, and making search more user-friendly by adding selectable pick-lists for query values. Also a consolidation of metadata quality is currently performed by changing some metadata values to a more harmonized set of values. All deposited metadata are maintained. Clarin.dk will apply for assessment as a CLARIN ERIC B centre in 2013 enforcing the sustainability and persistency of the infrastructure. Clarin.dk has already joined the national identity federation WAYF, implemented SSL-certificates, and offers harvesting of metadata via OAI-PMH as part of the CLARIN centre requirements.

KEYWORDS: Infrastructure, Language Resources, Repository, metadata, CLARIN.

Lene Offersgaard, Bart Jongejan, Mitchell Seaton and Dorte Haltrup Hansen 2013. CLARIN DK – status and challenges. *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* NEALT Proceedings Series 20 / Linköping Electronic Conference Proceedings 89: 21–32.

1 CLARIN-DK - a part of the Danish Infrastructure DIGHUMLAB

1.1 DIGHUMLAB

The Danish national, distributed research infrastructure initiative DIGHUMLAB[1] was launched in 2012 and has as goal to integrate and promote digital resources, communities, tools and opportunities to Danish researchers in the humanities and social sciences, and also at European and international levels. The DIGHUMLAB infrastructure initiative funds three focus areas, one being *Language based resources and tools*. CLARIN-DK covers current activities in this area, in close collaboration with the European CLARIN ERIC. One of the other focus areas of DIGHUMLAB is the Danish contribution to the ESFRI project DARIAH.

1.2 From preparatory phase to infrastructure

The initiative CLARIN-DK started as the preparatory Danish DK-CLARIN project (http://dkclarin.ku.dk/english) with the aim of creating a research infrastructure for the humanities, focusing on written and spoken language resources, multimodal resources, and tools. The project was a joint effort of eight leading Danish humanities institutions: four universities and four cultural institutions; at the same time it was a joint effort of researchers and developers. The project specified and implemented the first version of the clarin.dk research repository in the same time frame as applied for the European CLARIN preparatory phase project. This timing issue made it difficult to take full advantage of the findings and solutions of the European CLARIN project. As part of the DIGHUMLAB infrastructure, the opportunity to collaborate with the CLARIN ERIC partners is now available and will be addressed in section 7.

Clarin.dk is now in the transition phase from a repository integrating some tools to a research infrastructure, where researches and students can be supported in their research, education and studies. A part of the work in CLARIN-DK is therefore to carry out a user survey, where researchers on all Danish universities with faculties of humanities are invited to dialogue meetings on their own premises. Involving users is a core issue for CLARIN-DK and the challenge is to rethink the user's plans and wishes into technical feasible solutions given the available resources with respect to data, tools and time.

1.3 Objectives

During the implementation phase until 2017 CLARIN-DK will work on:

- Collect digital material and make it available – specifying metadata and using standards
- Collect and create tools and make them available – focusing on interoperability
- Implement user interface facilities to make it easier for researchers and students to benefit from the current infrastructure

[1] http://dighumlab.dk

- Disseminate knowledge about language resources and tools - facilitating knowledge sharing at both national and European level.
- Provide a CLARIN technical centre – improving the current technical infrastructure with persistency as an important aim.

We anticipate that the research methodology of the researchers using clarin.dk will be influenced by the new opportunities to include data and tools in their research leading to quite new experiments and experiences for the benefit of research.

1.4 Content of repository

The current version of clarin.dk can be found on http://clarin.dk. The available content is a number of resources collected in collaboration with researchers involved in the preparatory DK-CLARIN project. More information can be found in (Fersøe and Maegaard, 2009), (Asmussen, 2011a) and (Asmussen, 2011b).

The diversity of resources can be seen in table 1, also stating the numbers of current resources included. For all resources metadata are publicly available, but the resources are either available for public or academic use. More details about the metadata can be found in section 4.

Resource type	**Description**	**Count**
Text	Contemporary and old, general language and specialised sublanguage texts, as well as parallel corpora with Danish as one of the languages.	45.156
Text Annotation	Annotations of the texts above	140.921
Audio	Audio recordings of spoken language	37
Video	Video recordings of spoken language and gestures	80
Media annotations	Media annotations of audio and video recordings in XML and non-XML-formats	45
Lexicon and knowledge resources	Lexicon resources covering computational dictionaries and dialect dictionary	3
Tools and services	Tools integrated in the infrastructure as services, and tools stored for user download	11
Other Data	A few other resources of various types: zipped corpora and datasets.	9

TABLE 1: Resources available in clarin.dk, Mar2013. Resource types will be added when needed.

2 Researchers’ needs

One concern is to build a repository; to build an infrastructure is in our view quite another assignment. A repository mainly focuses on preservation of data letting users reuse resources. An infrastructure on the other hand, should enable the researchers to work with the resources and to extract new knowledge. We believe that involvement of the potential users when creating a research infrastructure is very important and must play a role during all the phases from specification to test of implementation.

2.1 User dialogue

Already in the interviews with potential users in the preparatory DK-CLARIN project, we realised that it is very difficult for users to imagine their requirements to repository facilities enabling a new digital or data-driven angle on their research. The baseline sketched was to make existing tools and available data integrated in the same platform, thus providing the opportunity to experiment with tools and data. Especially a streamlined common format for as many resources as possible and the possibility to access all available Danish data sources from one single repository was seen as a great benefit. However, the researchers also agreed that for resources that already are available in other databases or through other user-interfaces it should in each case be considered whether it would be beneficial to make the resources and tools available through clarin.dk, freeing the researcher from administering their data.

In collaboration with the researchers a list of issues were prioritized during implementation. The researchers wanted a repository to handle easy storage, sharing and use of resources:

- A repository to deposit data material and tools, to preserve these resources from project to project and to share them with others.
- Standardized ways to specify formats and metadata about resources, without losing diversity needed by research
- Access to the repository without having to use yet another account
- Easy inclusion of new researchers, students and institutions
- Search features for metadata for resources from all institutions even if access rights are restricted
- Combined search in metadata and content for text resources
- Easy access to and use of tools

An on-going series of dialogue meetings has already invited researchers from all the Danish faculties of humanities to meetings on their own campuses. These meetings have been supplemented with a series of meetings with potential data providers, focussing on making new resources available. Currently results from the on-going dialogue meetings, that cover wishes for functionality and the availability of new resources, are being evaluated. New issues may be added to the list above as a result of this process.

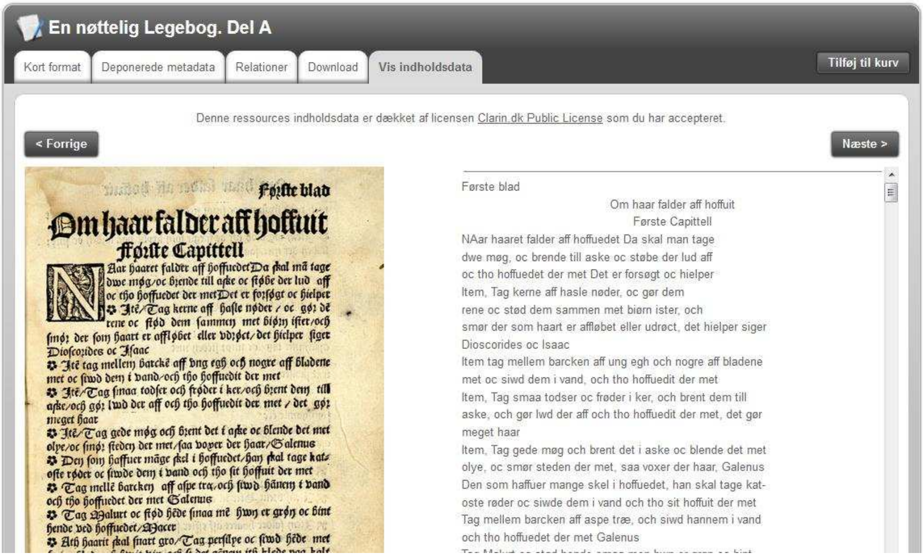
A key desire from the text researchers was an advanced text search facility combining metadata and content search and the possibility for extracting the resulting list of resources, including both texts and annotations. On the basis of this extract the researcher could then create a tailored annotated corpus search application that could be available for research and teaching as long as it was needed. From a user's point of view this seems simple, but for the developers the varieties of text and annotation formats and the possibility of an undefined number of diverse annotations for each text pose a problem. This task will be one of the focus areas in the upcoming work.

3 User functionality of clarin.dk

Currently the resource types mentioned in table 1 can be uploaded in clarin.dk and stored with metadata. An example of the view of a resource, “En nøttelig Legebog”, a book about health from 1533 can be seen in figure 1. As the available version only covers a part of the original book the resource title include “Del A”, indication that the whole book is not available. All resources are before upload described by type specific metadata by the data provider, and the metadata are deposited together with the resource itself. During depositing the platform checks that the resource and its metadata must validate against the CLARIN-DK specific TEIP5, IMDI and RNG schemas that are applicable for the resource type. There is no top down control of the metadata used as long as the content of the metadata comply with the decided standards and schemas.

Clarin.dk offers a toolbox with a number of tools that can be used to process the resources including a workflow planner that allows the user to focus on their goals, letting the workflow planner find the appropriate tools given the requested result, e.g. a frequency list. The current tools mainly offer text annotation and conversion, but also an OCR tool and a speech to text tool are integrated. More details about tools and workflows can be found in (Jongejan, 2013) and (Offersgaard, 2011).

The user can also search for resources based on metadata information, inspect all metadata, and according to permissions: inspect, use or download the resources.



Denne ressources indholdsdata er dækket af licensen Clarin.dk Public License som du har accepteret.

Første blad

Om haar falder aff hoffuit
Første Capittell

NAar haaret falder aff hoffuedet Da skal man tage
dwe møg, oc brende till aske oc støbe der lud aff
oc tho hoffuedet der met Det er forsøgt oc hielper
Item, Tag kerne aff hasle nøder, oc gør dem
rene oc stød dem sammen met biørn ister, och
smør der som haart er affløbet eller udrøct, det hielper siger
Dioscorides oc Isaac
Item tag mellem barcken aff ung egh och nogre aff bladene
met oc siwd dem i vand, och tho hoffuedit der met
Item, Tag smaa todser oc frøder i ker, och brent dem till
aske, och gør lwd der aff och tho hoffuedit der met, det gør
meget haar
Item, Tag gede møg och brent det i aske oc blende det met
olye, oc smør steden der met, saa voxer der haar, Galenus
Den som haffuer mange skel i hoffuedet, han skal tage kat-
oste røder oc siwde dem i vand och tho sit hoffuit der met
Tag mellem barcken aff aspe træ, och siwd hannem i vand
och tho hoffuedet der met Galenus

FIGURE 1: A resource containing part A of the old book “En nøttelig legebog” on the subject health from 1533. Url: https://clarin.dk/clarindk/item.jsp?id=dkclarin:597250#show_content-4

We will now go into detail with the work on metadata for resources and certain changes that are currently carried out in CLARIN-DK.

4 Standardising metadata for resources

It is important for infrastructure initiatives to give high priority to the use of current standards and already known and used formats. When collecting materials the existing resources will appear in a number of standards and formats, and it can be a challenging task to agree on these standards and formats. So streamlining of the resource metadata according to the CLARIN recommendations (Hinrichs, 2009 and Broeder, 2012), creates opportunities when sharing. In CLARIN-DK, the CLARIN recommendations were taken into account by letting clarin.dk enable all resources - independently of resource type – to get metadata in both CMDI and OLAC format; CMDI to fulfil the CLARIN recommendation and OLAC to make the metadata harvestable through the OAI-PMH protocol.

4.1 Resource specific metadata

For each resource type the relevant users were involved in selecting both the relevant metadata and relevant formats. As an effect of the user involvement in the metadata specifications, all user wishes for optional metadata were accepted, i.e. the developers accepted the wish for diversity in ways to describe the research material.

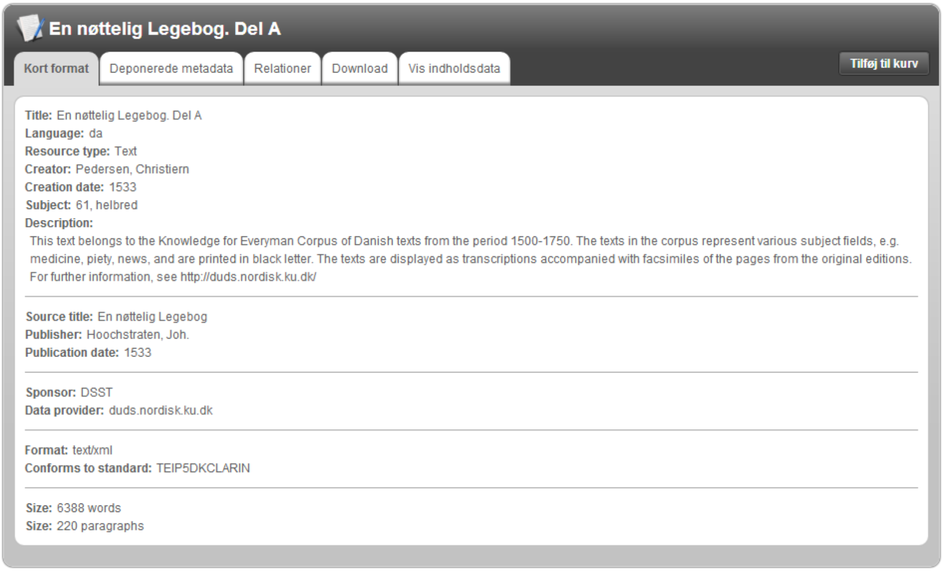
The users have chosen to use different standards for expressing the resource specific metadata. TEI P5[2] is used for simple text, text in a specific TEIP5 DKCLARIN format, text annotations and lexicon metadata. However, TEI P5 is not suitable for all tasks, and IMDI is therefore used for metadata for audio, video and media annotation metadata. The CMDI framework[3] provided from the European CLARIN project was much appreciated for specifying metadata for the resource types “data” and “tools” as no other current standard fulfilled these metadata requirements in a concise manner. The researchers are then able to specify metadata in widely used and well-known metadata formats, and the provided metadata are processed after upload and converted to metadata records in OLAC and CMDI to fulfil the CLARIN recommendations.

Resources in clarin.dk share a set of 15 common metadata elements, a subset of which is obligatory for all resources. The benefit of using a common core set of metadata elements is that it forms a common basis for the metadata search in the user interface. These metadata elements are supplemented with 72 metadata elements that are resource specific, e.g. *Translator* for texts, *DataFormatIn* for tools, *ActorRole* for multimedia and *Dialect* for lexica.

[2] http://www.tei-c.org/Guidelines/P5/

[3] http://www.clarin.eu/cmdi

The obligatory metadata set is basically the core OLAC metadata set. The elements *Coverage* and *Relation* have been omitted and instead the elements *Conforms to Standard* and *Size* have been added. Obligatory metadata are: *Title*, *Language* (not obligatory for data and tools), *Resource type*, *Creator*, *Creation date*, *Subject* (not obligatory for data and tools), *Description*, *Source title* (only for text), *Publisher*, *Publication date*, *Sponsor*, *Data provider*, *Format*, *Conforms to Standard*, and *Size*. In addition all resources get *Id* and *Rights* metadata when they are uploaded to the repository.



Title: En nøttelig Legebog. Del A
Language: da
Resource type: Text
Creator: Pedersen, Christiern
Creation date: 1533
Subject: 61, helbred
Description:
This text belongs to the Knowledge for Everyman Corpus of Danish texts from the period 1500-1750. The texts in the corpus represent various subject fields, e.g. medicine, piety, news, and are printed in black letter. The texts are displayed as transcriptions accompanied with facsimiles of the pages from the original editions. For further information, see http://duds.nordisk.ku.dk/

Source title: En nøttelig Legebog
Publisher: Hoochstraten, Joh.
Publication date: 1533

Sponsor: DSST
Data provider: duds.nordisk.ku.dk

Format: text/xml
Conforms to standard: TEIP5DKCLARIN

Size: 6388 words
Size: 220 paragraphs

FIGURE 2: Screen dump of general metadata for the text document in figure 1. Url: https://clarin.dk/clarindk/item.jsp?id=dkclarin:597250

4.2 Normalising metadata making search more user-friendly

One challenge in clarin.dk is to display the metadata for very different resource types in a uniform way. There is a huge diversity in the contents of metadata fields, because many fields can hold free text and because the validation schemas are non-strict. The benefit is obviously great flexibility when uploading resources, but the cost can be great complexity when searching for resources. Our task is to balance this trade-off which we do by mapping the original metadata to two new set of metadata, currently focussing on OLAC but later also using CMDI. As an example we can look at the rather simple metadata field *Subject*, which can also be named subject domain. In TEIP5 it can be found in both the field *Domain* and to some extent also in the field *CatRef* referring to a *CatRef* Scheme. So first of all we need to decide which metadata field to map to the OLAC metadata *Subject*. Currently we have decided to map *Domain* in TEIP5 to *Subject* in OLAC, but reviewing this with users might lead to a join of both *Domain* and *CatRef* to the same searchable Subject field.

The text providers agreed to a large extent on how to use the special metadata format TEIP5DKCLARIN, see (Asmussen, 2011a) for more details. When users are applying the decisions to real metadata, a diversity not earlier recognized can turn out to be needed. This can be seen for the values for *Subject.* As values we find *9999999999, n/a, general, Denmark in recent times, Health and medicine, Law, Nanotechnology, Agriculture, 20-religion, 61-health, 641-food, 641-food 631-agriculture* and more - a great variety in both naming, structuring, coverage, granularity and point of departure. A fixed list of values would clearly not be satisfactory for neither the data provider nor the future user of the data. Taxonomies are very difficult to use across disciplines.

In clarin.dk we are currently making the metadata search more user-friendly by applying two major changes. In the search interface we are implementing updateable pick lists for the values used in the selected metadata fields *Language, Resource type, Subject, Data provider, Sponsor, Format* and *Conforms to Standard*.

The other change is to rename some values that cover the same to one common value e.g. changing *Health and medicine* and *61-health* to the common term *Health (DK5-61)*, as the numbers used refer to the Danish Library classification system DK5[4]. These changes are not done in the deposited metadata, but in the special common set of metadata that are available for all resources and that are searchable, so the original, deposited metadata will always be available and preserved unchanged. To give an impression on the amount of changes, we will for 21855 resources change the subject from *9999999999* to *Unspecified* and 71 resources subject will be changed from *61-health* to *Health (DK5-61)*. But we will keep the diversity that is specified by data providers by allowing the addition of new subject areas when needed, as we think users from different research areas will need different classification systems.

5 Other challenges

Currently we are also working on making the search interface more intuitively easy to use, a difficult task when the end-users potentially come from all research disciplines in humanities.

When the user searches for relevant resources, it is currently not easy to get a quick overview of the available resources. Therefore re-development of the clarin.dk search pages are carried out during 2013 and more web-pages informing the user what is available and how to use the resources will be implemented.

Implementing a metadata-editor is also planned for this year, as this will make it easier for new users to upload resources. Now the metadata has to be specified in advance. For a text the user will specify metadata in a TEIP5 header which is validating against a special CLARIN-DK TEIP5 schema, before the text is ready to be deposited. In the future users will be able to specify the metadata in a special editor that guides the user to select the right resource type to deposit and the possible metadata fields to fill out.

Currently the resources in clarin.dk can have relations to other resources, but this will be extended by implementing the collection resource type. Collections can be used to

[4] A Danish variant of the Dewey Decimal Classification

bundle homogeneous text resources in a text corpus or to relate all resources from the 1950's to a heterogeneous collection that can hold not only texts but also e.g. images and audio from this period.

Users of texts have strongly recommended that clarin.dk provide a combined view of texts and annotations with search facilities included. As mentioned before this task has some challenges as texts and annotations can be from a number of different data providers and is still under specification.

Important tasks will also be to help researchers deposit their data in clarin.dk, both data that can easily be converted and deposited using the current resource types, but also allowing for new resource types to be added when research work and resource preservation has a need for it.

More tools will be available, but focus will mainly be on tools broadly usable for research tasks. Clarin.dk has planned to apply for assessment as a CLARIN B-centre during 2013, see section 7.

6 Implementation

Clarin.dk works with a Service-Oriented Architecture (SOA), and uses the eSciDoc (The Open Source e-Research Environment[5]) Infrastructure as stable middleware for authentication, access, submission, search and modification of the clarin.dk repository. The clarin.dk architecture is built on top of many open source products and technologies, such as Apache HTTP server, Apache Lucene, JBoss Application Server (AS), PostgreSQL database, and FedoraCommons[6]. MarkLogic[7] is enterprise software, a XML database for performance, reliability and scalability of the large data store required for our XML data storage needs. eSciDoc provides essential middleware services, defines our content model for the repository, and assists with indexing the repository collection so we are able to provide efficient search features and usability to the user functions.

To allow easy login administration the Danish WAYF solution was chosen: a Shibboleth implementation redirecting authentication to the users' home institution. This fits with the recommended solution from CLARIN, and is a flexible and easy solution for user administration.

6.1 Technical details

The backend infrastructure implementation is based on eSciDoc and the FedoraCommons repository system. FedoraCommons is a common component used by two thirds of the current CLARIN centres[8]. All associated XML content files of an

[5] https://www.escidoc.org/

[6] Flexible Extensible Digital Object Repository Architecture http://www.fedora-commons.org/software

[7] http://www.marklogic.com/

[8] Overview CLARIN centres: http://clarin.eu/node/2971 and https://centerregistry-clarin.esc.rzg.mpg.de/

eSciDoc item are referenced and stored in the separate MarkLogic database, which also provides efficient search facilities. More details can be found in (Conrad 2010)

Clarin.dk services are built on top of eSciDoc, to provide four integrated components – *Deposit, Search, Deliver*, and *Tools*. The *Deposit* service enables the submission of packaged resources for processing and then inclusion into the repository. Submitting users are notified via email of the newly published resources once the depositing process is completed. The *Search* service provides an interface to eSciDoc's core SRW/U search service. Clarin.dk utilises SRU/CQL, and Lucene indexes, to query specific, indexed metadata fields. SRU is recognised as an OASIS standard, and CQL is a common formal query language. An offline copy of a resource can be obtained via the *Deliver* service, with the possible inclusion of annotations, and choice of offline format for the selected resources. The *Tools* service enables the ability to run tools services or programs, on a collection of resources to obtain a desired offline output. Backend services are programmed in Java, and deployed in a WAR package through JBoss AS. The PostgreSQL database is used with eSciDoc/FedoraCommons, as well as keeping a record of Clarin.dk deposits and their status.

The front-end environment of clarin.dk is formed in a HTML (HTML5) web application, with cross-browser compatibility. AJAX capabilities of the web browser, event handling and DOM manipulation are used via the jQuery JavaScript library. The JSON data format is used for passing defined static data and configuration to the front-end web-pages. JSP (Java Server Pages) are used primarily to push and display dynamic content to the browser, from server-side HTTP requests, and Session (JavaBean) stored user data. Typically, CSS (Cascading Style Sheets) are used to control and format the web-page aesthetics and layout. Some additional jQuery plugin libraries are used, such as Validation, Cookie, ColorBox (overlay iframe) and a custom UI library. Documentation web-pages are presented on the website, where they provide usable and detailed information to assist the user.

7 Assessment to become a CLARIN ERIC B Center

To ensure interoperability among the CLARIN centres in Europe, national centres can apply for assessment as a CLARIN B centre[9]. The assessment criteria[10] for becoming a B centre includes a number of requirements concerning e.g. having a repository system that can pass a quality assessment procedure as the *Data Seal of Approval*[11], joining a national identity federation such as WAYF, servers having SSL-certificates, offering harvesting of metadata in CMDI format with the OAI-PMH-protocol[12], and associating persistent identifiers to resources and metadata.

Clarin.dk will apply for assessment in 2013 enforcing the sustainability and persistency of the infrastructure. Clarin.dk has already joined WAYF, implemented SSL-certificates, and offers harvesting of metadata via OAI-PMH. Persistent identifiers, PID's, will have

[9] CLARIN center types: http://www.clarin.eu/page/3542

[10] Checklist for CLARIN B Centres: http://www.clarin.eu/page/3577

[11] http://datasealofapproval.org/

[12] Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/pmh

to substitute the current resource id's in clarin.dk as these are not defined as persistent; metadata harvesting will need to be extended from DC and OLAC formats to also include CMDI, and the *Data Seal of Approval* must be achieved.

8 Conclusion and Outlook

Clarin.dk is in the transition from a repository to a research infrastructure, building on the results in the preparatory project DK-CLARIN. Clarin.dk is currently a repository with already over 186000 resources. By a user dialogue processing user requirements from the Danish faculties of humanities are now collected. These requirements will be used as guidelines when prioritising the coming extensions to clarin.dk.

Use of standards are widely implemented in clarin.dk, but as clarin.dk is meant to handle language resources for many disciplines a number of standards have to be brought into use. The data providers and researchers will have different ways to work with the resources and will need different facilities to support their research, so clarin.dk is now going into a development phase where researchers that volunteer to help specifying and using clarin.dk in research and teaching have a chance to influence the future facilities. The focus will still be resources containing language.

Currently some changes to content and display of metadata are carried out, and extensions with a metadata editor and better search facilities are planned for the rest of 2013. Deposited metadata has shown that even given a detailed metadata specification as for the text resources in clarin.dk, data providers fill out the metadata fields with a broad variation. Some variations are simple errors, other are caused by different ways to interpret specifications. In the future the metadata editor should guide the user in specifying metadata but still permitting a large variety of values if needed.

Important tasks are to help researchers deposit their data in clarin.dk and to make available more tools that can extend usability of clarin.dk. Working for assessment as a CLARIN B Centre during 2013 will enforce the sustainability and persistency of the infrastructure.

Acknowledgments

DIGHUMLAB Digital Humanities Lab Denmark is supported by Danish Agency for Science, Technology and Innovation, Ministry of Science, Innovation and Higher Education. The preparatory project DK-CLARIN was also supported by the Danish Agency for Science, Technology and Innovation, as well as by all eight partner institutions. We want to thank all partners for their contribution.

References

Asmussen, J. (2011) Text metadata: What the header of a text item looks like, *DK-CLARIN WP2.1 Technical Report*, http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf

Asmussen, J. (2011) Text formatting: Bringing corpus texts into good shape and enabling flexible annotation of them. *DK-CLARIN WP2.1 Technical Report.*

Asmussen, J. & Halskov, J. (2009) Compiling and annotating corpora in DK-CLARIN. Interpreting and tweaking TEI P5. In *Proceedings of the Corpus Linguistics Conference CL2009*. University of Liverpool, UK 2009. http://ucrel.lancs.ac.uk/publications/cl2009/

Conrad, A. (2010). The use of eSciDoc in Clarin.dk. *eSciDoc Days* Copenhagen, 2010. https://www.escidoc.org/pdf/day1-conrad-clarindk.pdf

Broeder, D. (2012) CMDI: a Component Metadata Infrastructure. CMDI (Component Metadata Infrastructure) workshop, September 13, 2012 MPI for Psycholinguistics, http://www.clarin.eu/sites/default/files/cmdi-daan.pdf

Fersøe, H & Maegaard, B. (2009). CLARIN in Denmark – European and Nordic Perspectives. In: *Nordic Perspectives on the CLARIN Infrastructure on Common Language Resources*, NEALT Proceedings Series, Vol. 5, pp. 6-11. Electronically published at Tartu University Library (Estonia) http://hdl.handle.net/10062/9944.

Halskov, J., Hansen, D. H., Braasch, A., & Olsen, S. (2010). Quality indicators of LSP texts – selection and measurements: Measuring the terminological usefulness of documents for an LSP corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation: LREC 2010* (s. 2614-2620). Valletta, Malta: European language resources distribution agency.

Hinrichs, E. W. (2009). CLARIN Short Guide Standards for Text Encoding. http://www.clarin.eu/files/standards-text-CLARIN-ShortGuide.pdf

Jongejan, B. *Workflow Management in CLARIN-DK.* In *Proceedings of the Nordic Language Research Infrastructure Workshop at NoDaLiDa, Oslo, May 22, 2013*

Offersgaard, L. Jongejan, B. and Maegaard, B. (2011). How Danish users tried to answer the unaskable during implementation of clarin.dk. In *SDH 2011 – Supporting Digital Humanities*, Copenhagen.

Humanities eInfrastructure initiatives in Denmark

Hanne Fersøe, Bente Maegaard, Bolette S. Pedersen

University of Copenhagen, Njalsgade 140, 2300 København S, Denmark

hannef@hum.ku.dk, bmaegaard@hum.ku.dk, bspedersen@hum.ku.dk

ABSTRACT

This paper presents an overview of the policies and decisions which have resulted in the current research infrastructure landscape in Denmark within the humanities, and it describes individual initiatives and how they fit into the landscape.

The landscape was fragmented, and it still is to a degree, but one key factor has played a determining role in shaping it into what it is today: the European strategic processes and policies following from the development of the European Research Area, and the ESFRI (European Strategic Forum for Research Infrastructure) process, which has been adopted by Danish policy and decision makers. The strategies and priorities based on the European ESFRI process has resulted in funding of Danish research infrastructure (RI) initiatives within the humanities in the last 5-6 years, such that we can now define a landscape with national, Nordic and European collaboration and networking, extended into the social sciences community as well.

The list of initiatives is comprehensive: CLARIN, META-NORD, META-NET, DARIAH, DASISH, CLARA, ASTIN, DIGHUMLAB, DIGDAG, LARM.

KEYWORDS: Denmark, eInfrastructure, Humanities, CLARIN, META-NORD, DASISH, DIGHUMLAB.

Hanne Fersøe, Bente Maegaard and Bolette S. Pedersen 2013. Humanities eInfrastructure initiatives in Denmark. *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* NEALT Proceedings Series 20 / Linköping Electronic Conference Proceedings 89: 33–43.

1 Introduction

In the following, we first present a brief overview of priorities and processes leading to funding of Danish research infrastructure (RI) initiatives within the humanities in the last 5-6 years, including the international collaboration and networking leading to or resulting from these initiatives. Following this overview we describe the individual initiatives with regard to funding, national and international collaboration, objectives, results and future plans. The conclusion sums up the status and the future development.

2 European Priorities

This section describes the ESFRI process which is central in establishing a common European approach to RI, and which recommended the national strategic funding schemes for RI currently in process in the member states.

2.1 The ESFRI Process

It is well known, traditionally, that RIs were conceived as single sited physical installations and large scale facilities such as for instance telescopes, electron lasers or particle accelerators, and that these installations were extremely costly to build and maintain. They were therefore often jointly funded by several countries, and scientists from the relevant disciplines would subsequently apply for and be granted access to use the RI facility for a period of time or scientists would be employed as permanent staff at the facilities. In the humanities and social sciences disciplines with different research traditions and methods, national libraries and archives constituted, and still generally constitute, the primary type of research infrastructures. They are, by nature, single sited, but with the emergence of more advanced IT technology they have developed and offer on line access to their databases, thus also defining themselves as virtual facilities (European Commission, 2007).

Steps towards a coordinated European approach on RI were taken in 2001 in connection with the development of the ERA, The European Research Area, leading to the creation in 2002 of ESFRI (European Commission, March 2013).

The first ESFRI Roadmap (European Strategy Forum on Research Infrastructures, 2006) was published in 2006 with updates in 2008 and 2010. The roadmap lists RI initiatives, which are identified as new large scale RIs to strengthen the European Research Area. Five initiatives within the Social Sciences and Humanities are included CESSDA, Council of European Social Science data Archives (CESSDA web site, April 2013), CLARIN, Common Language Resources and Technology Infrastructure (CLARIN web site, April 2013), DARIAH, Digital Research Infrastructure for the Arts and Humanities (DARIAH web site, April 2013), ESS, The European Social Survey (ESS web site, April 2013), SHARE, Survey of Age, Health and Retirement in Europe, (SHARE web site, April 2013). These applied, successfully, for FP7 funding to prepare their implementation phases, and are today ERICs (European Research Infrastructure Consortium) or in the process of becoming ERICs or equivalent structures.

Partner institutions from these five ESFRI RIs, including University of Copenhagen, Department of Language Technology (UCPH-CST), are today collaborating in the so called cluster project DASISH, Data Service Infrastructure for the Social Sciences and Humanities (DASISH web site, March, 2013), where they work on joint activities related to data access, data sharing, data quality, and data archiving to provide common solutions to common problems.

2.2 The META-NET Initiative

There is a long European tradition through the framework programmes for scientific collaboration across Europe in the areas of computational linguistics, language technology, and resource and tool development. This has resulted in the creation of META-NET (META-NET website, March 2013), a research infrastructure initiative for the Humanities, which focusses specifically on language technology. The European Commission's 7th framework programme and the ICT Policy support programme gave support to the 4 projects T4ME, CESAR, METANET4U, and META-NORD, and to the umbrella project META-NET. UCPH-CST was a partner in META-NORD.

Two central awareness activities were accomplished within this initiative: the Language White Paper series on language technology in Europe (META-NET white papers, March 2013) as well as the Strategic Research Agenda (SRA) (META-NET SRA, March 2013). These publications aimed at attracting the attention of and informing politicians and policy makers in their decisions regarding language policy and language technology, especially with regard to the upcoming European funding opportunity Horizon 2020.

The project funding ended 31st January 2013, but the META-SHARE repository (META-SHARE Repository, March 2013), which is an infrastructure facility offering resources, technologies and services is in place and is being maintained. The META projects had partners from a large number of European countries, and many of the partners are also members of the National CLARIN consortium of their country.

2.3 Other supporting initiatives

Among recent European initiatives which have objectives that support the RI agenda, we can mention the Marie Curie programme, which supports the training of 19 research fellows in the area of language resources and their application through the Initial Training Network, CLARA (CLARA web site, March 2013). UCPH-CST employs two of these fellows. Denmark also participated in the FlaReNet (FlaReNet web site, March 2013) project, which developed a common European vision for language resources and technologies with the support from the eContentPlus programme. There were also other projects supported through FP7 which contribute similar types of results, e.g. the MEDAR (MEDAR website, March 2013) project, coordinated by UCPH-CST, which built networking mechanisms within Arabic Speech and Language Technologies, updated the Arabic BLARK, and created Arabic language resources as well as a prototype MT system for English-Arabic.

3 Nordic Collaboration priorities

Formalised collaboration between the Nordic countries takes place through the Nordic Council, one of the oldest and most comprehensive intergovernmental collaborations in the world (Nordic Council website, March, 2013).

Various programmes offer instruments to support collaborative research between the Nordic countries in the humanities area, the overall objective always being to support mutual understanding of Nordic languages and culture. There have been no specific research infrastructure initiatives in the language and culture areas, but the language technology communities in the Nordic and Baltic countries have collaborated in a number of other types of initiatives over the years, so a tradition for knowledge exchange exists, which has been useful in connection with the formation of European RI consortia such as CLARIN, META-NET and DASISH. Examples of such initiatives are e.g. the Nordic Language Councils collaboration in the working group ASTIN (ASTIN website, March 2013), which as one of its goals has "att verka för … tillgång till nödvändig infrastruktur i form av grundläggande språk- och teknikresurser för de nordiska språken", and NEALT the Northern European Association for Language Technology (NEALT web site, March 2013).

4 National Danish Priorities

Already in 2005, a strategy paper, surveying existing Danish infrastructures and needs, and proposing a strategy for future RIs had been published (Forsknings- og Innovationsstyrelsen 2011). Based on this strategy paper calls for RI proposals were published, and two humanities proposals were awarded grants: the Danish CLARIN project proposed by UCPH-CST, which referred itself to the ongoing ESFRI RI initiatives in Europe, and LARM proposed by UCPH-Department of Arts and Cultural Studies.

In parallel, Denmark decided to join the ESFRI roadmap process and became one of the EU member states who agreed to develop a national roadmap. Working groups with members from 6 major research areas were established, and their survey work resulted in publication of the Danish National Roadmap by the Danish Agency for Science, Technology and Innovation (DASTI) in 2011. The roadmap recommended prioritized initiatives, following from the ESFRI process, within 6 scientific fields, humanities and social sciences being one of the fields.

The Digital Humanities Laboratory, DIGHUMLAB, comprising a research infrastructure for the humanities, was awarded a five year grant to the four proposing Danish universities, University of Copenhagen being one of them. Other interested parties such as e.g. the Royal Library and the State and University Library have a standing invitation to join the consortium, at their own cost.

5 Research infrastructure activities in Denmark

This section describes briefly each of the international infrastructure initiatives with Danish involvement within the last six years, and two national projects.

5.1 CLARIN

The CLARIN preparatory project, CLARIN PP, started in January 2008 with the aim of preparing the full scale implementation of the CLARIN RI. The initiative met a huge need in Europe for a unified and concerted approach to a previously fragmented scientific field, and there was and is massive back up from the communities in all the European countries and from CLARIN members worldwide to providing access to resources, tools, and services to all language based sciences, particularly the humanities.

5.1.1 European CLARIN Preparatory Phase

Denmark's role in the CLARIN PP was to lead the work on governance in a future European organization. During the project period the Commission worked out and completed the ERIC Regulation, which sets out the legal framework for a totally new type of European organization with countries and intergovernmental organizations as members. The ERIC Regulation constituted a lever, which enabled the CLARIN consortium to thoroughly prepare the statutes and the application for the future CLARIN ERIC. Other preparatory tasks concerned for instance establishing a technical infrastructure, securing interoperability, and proposing a legal framework for the future use of data resources, tools and services.

5.1.2 Danish CLARIN Preparatory Phase

The first Danish CLARIN project ran in parallel with the European PP project, but not as a preparatory project. The Danish CLARIN consortium had to implement a technical infrastructure and populate it with content. This work is described in detail in (Fersøe et al., 2009).

The project has produced a data repository with a web portal accessible through the user authentication and authorization system WAYF (Where Are You From) (WAYF website, April 2013). The portal contains:

- 46,000 written text files with 143,000 annotation files for them
- 80 videos, 37 audios with 45 annotation files for them
- 3 lexica
- 5 tools

The portal can be accessed at www.clarin.dk.

5.1.3 CLARIN ERIC

In February 2012 CLARIN was granted ERIC status following a successful application process with the Netherlands as the hosting country and 9 founding members. Currently several countries are in the process of applying for membership and one country has obtained observer status.

Denmark's role in the ERIC is twofold: as a member we are obliged to deliver the contribution described in the CLARIN Agreement, which is made between the ERIC and each member (see section 5.2.1), and in addition, following from our governance work in the PP project, the CLARIN ERIC Vice Executive Director is Danish from University of Copenhagen.

The ERIC is established with a very lean administration, and all boards, standing committees and working groups have been established.

5.2 DIGHUMLAB

The DIGHUMLAB initiative is carried out in collaboration between Aarhus University, University of Copenhagen, University of Southern Denmark, and Aalborg University. The DIGHUMLAB project plan is structured into a project management and administration layer and three research themes. The Danish CLARIN and DARIAH fees and in-kind contributions are funded through the research themes 1 and 2.

5.2.1 DIGHUMLAB, Theme 1

Theme 1, Language based Materials and Tools, is anchored at UCPH-CST. The rights obtained by a country through membership, e.g. full access to CLARIN and all its services to the entire Danish research community, are described in the CLARIN ERIC statutes, while the obligations, e.g. in-kind contributions, are described in the CLARIN Agreement between CLARIN ERIC and Denmark.

In addition to paying the annual fee and contributing to the overall goals stated in the statutes, there are country specific in-kind contributions. All members must create a national consortium, provide a national co-ordinator, and provide a data and service center which gives user access in conformance with the access policies defined in the statutes and with an approved user authentication and authorization system; they must also contribute to international knowledge sharing. In addition, they must provide a work programme for the national consortium comprising improvements and enhancements to existing data resources, tools and services, creation of new resources, tools and services for CLARIN, promotion of standards, and a national knowledge sharing infrastructure. All these activities are listed and defined in the agreement, and they are also found in the DIGHUMLAB theme 1 work plan.

The portal developed in the Danish CLARIN preparatory phase constitute the point of departure for the in-kind contributions. Currently, a major activity is preparing for the national centre assessment. Another on-going activity is a revision of the existing metadata scheme. A third activity is to update and enhance the user interface to the web portal, which is the access gate to the Danish CLARIN resources, tools and services.

5.2.2 DIGHUMLAB, Themes 2 and 3

Theme 2 is divided into the sub themes NetLab, and Tools for Sound and Image Media. The sub theme NetLab was a joint initiative of the State University Library and the Royal Library in 2005 aimed at archiving the Danish internet. The NetLab focus in DIGHUMLAB is on the development of adequate tools for research.

The sub theme Tools for Sound and Image Media focusses on extracting metadata, developing tools, and on research maturation in relation to radio, TV, film and related media. Collections of sound and image media represent not only a significant cultural heritage resource for the 20th century in particular, but also a rich research source for the humanities and social sciences. The sub theme is carried out in close collaboration with the LARM research infrastructure project, and serves as a Danish in-kind contribution to DARIAH.

Theme 3 will establish a Danish national research infrastructure with a focus on the experimental environments within the humanities. This will be achieved through i.a. the establishment of contemporary new laboratory facilities with a special focus on interaction between humans and technology. Supporting the development of relevant investigation protocols and development of and compliance with scientific ethical standards will therefore be a major focus area. The material thus produced will additionally be made available through CLARIN.

5.3 DARIAH

The DARIAH PP ran in parallel with the CLARIN PP with 14 partners from 10 countries. The objective of the preparatory phase was to set up physical, strategic and human elements of the RI. The ERIC application process has started with France as the host country and 12 countries signing the Memorandum of Understanding, Denmark is one of them.

UCPH-Department of Scandinavian Research was a partner in the DARIAH PP project leading the work package on dissemination. After the completion of the PP project, there was an unfunded gap where the transition to ERIC status was prepared. DARIAH now operates through its European-wide network of Virtual Competency Centres (VCC). Each VCC is cross-disciplinary, multi-institutional and international and centred on a specific area of expertise. Denmark, University of Aarhus, together with Ireland, coordinates the Virtual Competence Center on Research and Education as a part of DIGHUMLAB, Theme 2, see section 5.2.

5.4 META-NORD

UCPH-CST participated in the Nordic and Baltic branch of the META-NET initiative, META-NORD, during the period 2011-2013. Apart from enhancing and upgrading a considerable number of Danish language technology resources and tools to agreed standards, the main, national focus of this participation consisted in a series of awareness actions concerning Danish language technology. The actions were realized as a white paper on the status of the Danish language in the digital age (Pedersen et al., 2012), a comprehensive press campaign in the Danish media, and as a national language technology workshop for researchers, industry and decision makers. By the end of the project, a Danish META-SHARE node was established, and 56 Danish language technology tools and resources were made accessible via this platform. These include analysis tools and resources provided by UCPH-CST, such as text corpora, a WordNet (DanNet), a computational lexicon for Danish (STO), and an annotated audio-visual corpus (the NOMCO Corpus). Furthermore, tools and resources are provided by University of Southern Denmark/GrammarSoft (constraint grammars, machine translation modules, Danish FrameNet among others) and from Copenhagen Business School (CBS) (treebanks) and The Society for Danish Language and Literature (the CLARIN Corpus). Most of the resources are monolingual and focus on the Danish written language, some, however, include cross-lingual links or alignments.

5.5 DASISH

DASISH is a cluster project between the five ESFRI RIs in the social sciences and humanities area, it has 19 partners, including 6 partners from the Nordic countries, and it is supported by a grant from FP7. The project is managed by the CESSDA president at Swedish National Data Service, University of Gothenburg; the Finnish and one of the Norwegian partners are also from CESSDA, while the second Norwegian partner, the Danish and the Estonian partners are from CLARIN.

The project started in January 2012 and runs for three years. The goal of DASISH is to identify areas of synergy in the infrastructure development of all five RI communities and to work on concrete joint activities in order to propose common solutions to common problems. The idea is to avoid double developments, and that the RIs should mutually benefit from advanced developments done by the others. Examples of such joint activities are: a) understanding the different architectural solutions in RI construction, b) data and metadata quality issues for instance in data collection in European-wide surveys and in the data management and curation methods used in text archives, c) development of a joint shared data access and enrichment framework, comprising for instance AAI, PIDs, joint metadata, workflow implementations, joint annotation framework, d) joint legal and ethical activities, for instance identification of issues and constraints, challenges imposed by new data types, e) development of training and education modules, and f) dissemination of activities and results.

UCPH-CST is leading the work on dissemination and is also involved in the joint activities on PIDs, workflow implementations and joint annotation framework.

5.6 DIGDAG

DIGDAG (DIGDAG web site, March 2013), a digital map, is a cross-institutional research project supported by the Ministry of Science Technology and Innovation under the national RI funding scheme. The name DIGDAG is an abbreviation of *Digitalt atlas over Danmarks historisk-administrative geografi*. The project started in 2009 with the purpose of establishing a historical-geographic database over the administrative subdivision of Denmark and thereby a) creating an infrastructure for research in Denmark's administration history from year 1600 to current time, b) developing a solid search engine to be used by archives, collections and libraries, and c) supporting research in administration history.

The project participants include a large part of the most important Danish cultural and research institutions: The State Archives, University of Copenhagen, University of Southern Denmark, Kort og Matrikelstyrelsen (an agency for maps which is now restructured), the National Museum, The Royal Library, and Kulturarvsstyrelsen (now restructured into the Danish Agency for Culture).

5.7 LARM

LARM[1] Audio Research Archive (LARM website, March 2013) is an interdisciplinary project aiming at producing a digital infrastructure to facilitate researchers' access to the

[1] LARM is the Danish word for noise

Danish radiophonic cultural heritage. The LARM project is a collaboration between a number of research and cultural institutions: The University of Copenhagen, Roskilde University, The University of Southern Denmark, Aalborg University, Aarhus University, The Royal School of Library and Information Science, The Danish Broadcasting Corporation, The State and University Library, Danish e-Infrastructure Cooperation, Kolding School of Design, and The Museum of Media. The project is made possible by a grant from The National RI funding scheme.

The infrastructure is a digital archive with the appropriate research tools which will give access to thousands of hours of national and local radio broadcasts from 1925 and onwards. The platform will enable researchers and university students to stream sound to their own computers directly from the digital archive, which at the conclusion of the project will contain more than one million hours of sound.

User driven innovation is a key element in LARM. The infrastructure and its interface are based on user needs and are developed in close collaboration between technicians, cultural researchers and designers, and the research projects deliver feedback to the development of digital audio search and audio description tools in both effective and innovative ways.

6 Conclusion and Outlook

In conclusion, the status of Danish eInfrastructure within the humanities is that many initiatives have been started in the last 5-6 years, they address various language based data types and media, including some tools and services for the research. The initiatives are anchored in European and national strategies and processes, and they provide excellent opportunities for networking and knowledge sharing across countries and communities and among institutions.

In the future there are a number of important aspects that need to be addressed. Continued funding is an issue; the universities are in the process of preparing their strategies concerning embedding into the university activities. Such an embedding will support the idea that research infrastructure is a general tool for research and consequently be the best possible support for further take up. On the other hand external funding will undoubtedly be necessary as the creation and maintenance of RI is a heavy task, just like the maintenance of a library.

References

ASTIN website (March 2013). http://nordisksprogkoordination.org/links/astin/

CESSDA web site (April 2013). http://www.cessda.org/

CLARA web site (March 2013). http://clara.b.uib.no/

CLARIN web site (April 2013). http://www.clarin.eu/

DARIAH web site (April 2013). http://www.dariah.eu/

DASISH web site (March, 2013). http://dasish.eu/

DIGDAG web site (March 2013). www.digdag.dk

DIGHUMLAB website (March 2013) http://dighumlab.dk/

ESS web site (April 2013). http://www.europeansocialsurvey.org/

European Commission, European Science Foundation (2007). Trends in European Research Infrastructures. Analysis of Data from the 2006/07 survey.

European Commission (March, 2013) http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-background

European Strategy Forum on Research Infrastructures (2006). European Roadmap for Research Infrastructures.

FlaReNet web site (March 2013). http://www.flarenet.eu/

Fersøe, H. and Maegaard, B. (2009). CLARIN in Denmark – European and Nordic Perspectives. In Proceedings of the 17th Nordic Conference of Computational Linguistics - NoDaLiDa 2009, Workshop: Nordic Perspectives on the CLARIN Infrastructure of Language Resources, *NEALT Proceedings series no.5*.

Forsknings- og Innovationsstyrelsen, Ministeriet for Videnskab, Teknologi og Udvikling (2011). Dansk roadmap for forskningsinfrastruktur.

LARM website (March 2013). http://www.larm-archive.org/

MEDAR website (March 2013). http://www.medar.info/

META-NET website (March 2013). http://www.meta-net.eu/

META-NET white papers (March 2013). http://www.meta-net.eu/whitepapers/overview

META-NET SRA (March 2013). http://www.meta-net.eu/sra-en

META-SHARE Repository (March 2013). http://www.meta-net.eu/meta-share/index_html

NEALT web site (March 2013). http://omilia.uio.no/nealt/

Nordic Council website (March, 2013). http://www.norden.org/da/om-samarbejdet

Pedersen, B.S, J. Wedekind, S. Kirchmeier-Andersen, S. Nimb, J.E. Rasmussen, L.B. Larsen, S. Bøhm-Andersen, H. Erdman Thomsen, P. J. Henrichsen, J. O. Kjærum, P. Revsbech, S.Hoffensetz-Andresen, B. Maegaard (2012). The Danish Language in the Digital Age - Det danske sprog i den digitale tidsalder. META-NET White Paper Series, Springer Verlag.

SHARE web site (April 2013). http://www.share-project.org/

WAYF website (April 2013). http://www.wayf.dk/

Linking Northern European Infrastructures for Improving the Accessibility and Documentation of Complex Resources

Gyri Smørdal Losnegaard[1],*Gunn Inger Lyse*[1], *Anje Müller Gjesdal*[1],
Koenraad De Smedt[1], *Paul Meurer*[2], *Victoria Rosén*[1,2]

(1) University of Bergen, Norway
(2) Uni Research, Norway

gyri.losnegaard@lle.uib.no, gunn.lyse@lle.uib.no, anje.gjesdal@uib.no,
desmedt@uib.no, paul.meurer@uni.no, victoria@uib.no

ABSTRACT
This paper describes our integration efforts in two Northern European language infrastructures. Specifically, this work has been a collaboration between the META-NORD team at the University of Bergen and the INESS project, a large treebanking infrastructure project in Norway, in developing and documenting two complex resources, as well as making these accessible to the R&D community.

KEYWORDS: metadata, IPR, Treebanks, research infrastructure, Northern Europe, INESS, META-NORD, CLARIN.

Gyri Losnegaard, Gunn Inger Lyse, Anje Müller Gjesdal, Koenraad De Smedt, Paul Meurer and Victoria Rosén 2013. Linking Northern European infrastructures for improving the accessibility and documentation of complex resources. *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* NEALT Proceedings Series 20 / Linköping Electronic Conference Proceedings 89: 44–59.

1 Introduction

Several large-scale infrastructures are currently under development across Europe for the distribution of research results, data and tools in the Humanities and Social Sciences. The various initiatives differ in the disciplines that they cover and the scope of their goals, but they have the common aim of fostering the reuse and sustainability of resources and tools. Such initiatives require a considerable effort to harmonize metadata schemes, adhere to standards and solve intellectual property rights (IPR) issues (Hinrichs et al., 2010; Duin et al., 2010; Gavrilidou et al., 2011, 2012). Moreover, since different infrastructures co-exist at different levels, infrastructure initiatives will increasingly need to focus on establishing best practice criteria to facilitate the linking of infrastructures and ensure their interoperability.

This paper describes our integration efforts in two Nordic and Baltic language infrastructures. Specifically, this work has been a collaboration in Norway between the META-NORD and INESS projects in developing and documenting two complex resources, as well as making these accessible to the R&D community.

META-NORD (Vasiļjevs et al., 2012) (2011–2013) has been a CIP ICT-PSP (Information and Communication Technologies Policy Support Programme) project aimed at creating an open infrastructure to promote the accessibility and reuse of language resources and technologies (LRT). Its consortium includes organizations from all the Nordic and Baltic countries. Among its main results has been the documentation, rights clearance, licensing and sharing of many language resources via the META-SHARE[1] catalogue and repository, thereby making LRT more readily available to R&D.

INESS (Rosén et al., 2012) (2010–2016) is an ongoing project at the University of Bergen (Norway) and Uni Computing (a division of Uni Research, also in Bergen), aimed at establishing an Infrastructure for the Exploration of Syntax and Semantics. It is funded by the Research Council of Norway and the University of Bergen. One of its activities is the implementation and operation of a comprehensive open treebanking environment in which a large number of treebanks can be hosted and made accessible through advanced web interfaces for search and visualization[2]. The other is the development of a large parsebank for Norwegian with a wide coverage LFG grammar and lexicon.

INESS has cooperated extensively with META-NORD throughout the project lifetime of the latter. The main field of cooperation has been to collect and develop treebanks, to make these more accessible in standardized ways, to document them through metadata and to link them through alignment in parallel treebanks, as will be explained in more detail below. The results of these activities are also being integrated in CLARINO (the Norwegian part of the CLARIN network) and in *Språkbanken*, a language technology resource collection for Norwegian, hosted at the National Library of Norway.

Among the challenges faced was resolving the sometimes conflicting requirements for creating and integrating treebanks in the INESS treebanking infrastructure, on the one hand, and documenting them with metadata in META-SHARE, on the other. We have made some initial efforts towards consolidating the metadata creation and description between the two infrastructures. While the integration of existing resources is an essential part of building infrastructures, we

[1] `http:/meta-share.tilde.lv`

[2] The different aspects of the INESS treebanking infrastructure, from visualization via interactive annotation of treebanks to treebank search, are described in detail in Rosén et al. (2012).

argue that infrastructure initiatives will increasingly need to focus on establishing best practice criteria to be applied at the data creation stage. This will in turn facilitate the linking of infrastructures and ensure their interoperability.

The rest of this paper is structured as follows: in section 2 we describe the integration and linking of treebanks in the INESS infrastructure. Section 3 addresses the challenges encountered in their documentation: metadata compilation (section 3.1), IPR clearance (section 3.2) and metadata creation in META-SHARE (section 3.4), the latter exposing special challenges in the description of complex resources. We present our work integrating the two infrastructures in section 4, before we finally, in section 5, provide suggestions for best practices in terms of standardization of formats, metadata, IPR and integration between infrastructures.

2 Creating, integrating and linking treebanks in the INESS infrastructure

In the cooperative effort between INESS and META-NORD, two parallel treebanks were constructed: the Sofie Parallel Treebank and the Acquis Parallel Treebank.

The Norwegian novel *Sofies verden* (*Sophie's World*) (Gaarder, 1991) was chosen as a suitable basis for parallel treebanking because it is linguistically rich and professionally translated into many languages, and because some monolingual treebanks already existed for text selections from this material in some languages in the META-NORD area. Existing treebanks for this material had been made in the context of the Nordic Treebank Network (NTN), funded by the Nordic Language Technology Program (2001–2005). Annotation files for Danish, Estonian, German, Icelandic and Swedish were obtained via Tekstlaboratoriet (the Text Laboratory) at the University of Oslo, and a treebank for the English version was obtained from the SMULTRON parallel treebank (Stockholm MULtilingual TReebank)[3] (Adesam, 2012). These treebanks were documented, processed and supplemented with new treebanks for the Norwegian, Georgian and Finnish versions.

The Sofie treebanks made available through INESS and META-NORD show considerable diversity with respect to both the language families that are covered and the linguistic formalisms that are represented. The Sofie Danish Treebank is a dependency treebank, semi-automatically annotated according to the guidelines used to create the Danish Dependency Treebank and automatically converted to TIGER-XML by the DTAG program. The Sofie Estonian Treebank is a constraint grammar (CG) treebank, automatically parsed with a CG parser assigning syntactic function labels and enhanced with manually added constituencies. The Sofie Icelandic Treebank is a constituency treebank which was manually annotated by the late Gunnar Hrafn Hrafnbjargarson. The Sofie Swedish Treebank is a dependency treebank, automatically created with the Maltparser tool. The Sofie German Treebank was annotated with the Annotate tool, followed by an automatic deepening of the flat syntax trees. The Sofie Finnish Treebank is a manually annotated dependency-CG treebank created by the UHEL FinnTreeBank team for FinnTreeBank and META-NORD. The Sofie Norwegian Treebank was automatically parsed with an LFG grammar developed in the NorGram and INESS projects, producing c-structures and f-structures; the analyses were manually (interactively) disambiguated by the use of discriminants (Rosén et al., 2009, 2007). The Sofie Georgian Treebank was similarly processed, but with a Georgian grammar developed by Paul Meurer. The Norwegian and Georgian treebanks are downloadable in Negra/Tiger XML format.

Furthermore, small pilot treebanks were constructed for the JRC Acquis Multilingual Parallel

[3] http://www.cl.uzh.ch/research/paralleltreebanks_en.html

Corpus of EU/EEA law texts,[4] which provides materials from a different genre. The standardized and uniform structure of the corpus and its texts facilitated the selection of a document of appropriate length which was available in all the relevant META-NORD EU-languages, and for which translations existed also for the non-EU languages Icelandic and Norwegian. Dependency annotations were produced for the Danish, Estonian, Finnish, and Swedish Acquis texts, and constituency annotations for Icelandic. INESS provided annotations of the Norwegian and English versions of the selected Acquis document, which were parsed with LFG grammars and manually disambiguated.

These existing and new treebanks were combined and integrated into INESS, providing both long-term physical storage and a platform for research and development of treebanks. INESS supports most of the standard input formats (TigerXML, CoNLL-X, CG3-dependency, Penn Treebank II bracketing or XLE prolog) and with the exception of one treebank that had to be converted from non-standard notation to one of the standard input formats, the integration was seamless. Monolingual annotations for each of the collections were then aligned at sentence level and their alignment was made downloadable in XML stand-off format. The parallel treebanks were made searchable, and individual sentences from the treebanks are visualized side by side, as illustrated in Figure 1.

Besides the Sofie and Acquis monolingual treebanks, which provided the basis for parallel treebanks, the INESS project has also made several other freestanding monolingual treebanks based on different sources available. Some of these were selected for documentation in cooperation with the META-NORD project. These include treebanks for Finnish, Icelandic and Norwegian in the linguistic area of META-NORD as well as for a number of other languages inside and outside the linguistic area of META-NET (including smaller languages in the META-NORD geographical area such as Faroese and Northern Sami).

3 Resource documentation

Adequate documentation is essential both in order to create trustworthy metadata and to resolve IPR issues, and it presupposes correct and reliable information about formats, IPR and resource creation. An important part of resource documentation consists of obtaining this information and clearing the rights for the resource so that it can be used for the intended purposes. In order to exploit the expertise and standards being developed within the META[5] network and to avoid a duplication of efforts, it was decided to delegate metadata and IPR issues in INESS to META-NORD. The parallel treebanks, as well as each monolingual treebank, were documented with structured metadata using META-SHARE.

Although the existing treebanks were originally created in NTN, an advanced research network, their documentation and IPR clearance proved especially challenging, as described in sections 3.1 and 3.2. Moreover, the complexity of the parallel treebanks brought about special documentation requirements which META-SHARE did not allow for in a straightforward way (3.3), forcing us to come up with some expedient solutions (3.4).

3.1 Metadata compilation

For the treebanks developed in NTN, substantial efforts were invested in recovering the information required to make treebanks available for download, directly or indirectly, through

[4] http://langtech.jrc.it/JRC-Acquis.html

[5] http://www.meta-net.eu/

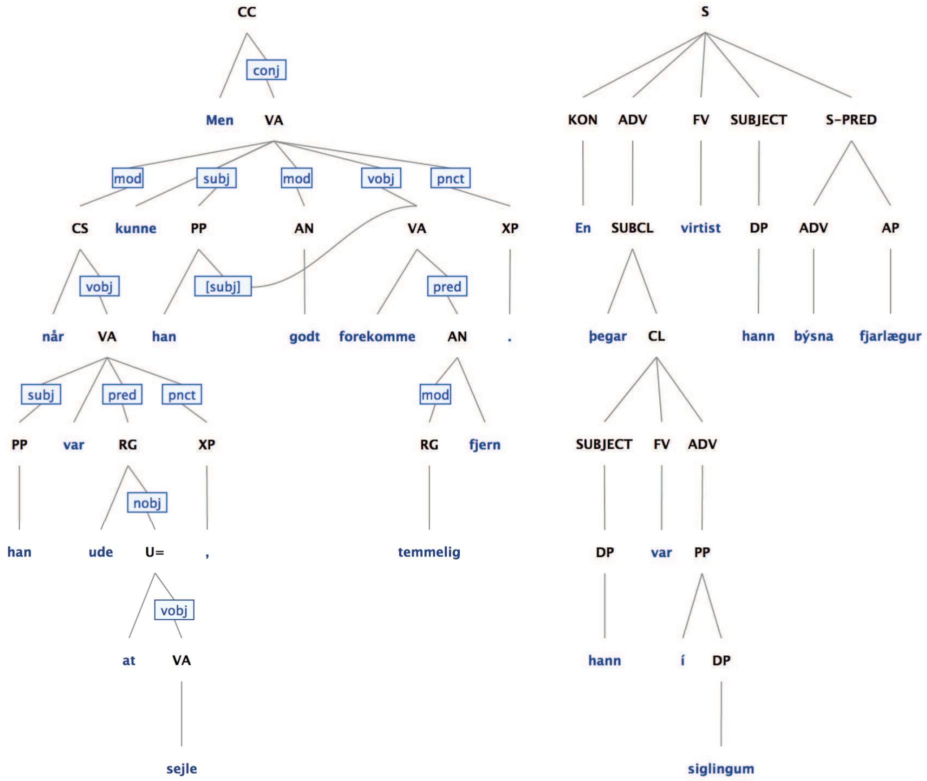


Figure 1: Visualization of sentences from the Parallel Sofie Treebank: Danish and Icelandic

META-SHARE. A major challenge was presented by the fact that NTN project results and documentation were no longer maintained and partly inaccessible. Some information was available on the NTN webpages and as metadata encoded in the XML header of some of the annotation files. However, a large part of the information necessary to create adequate metadata descriptions and to ensure rights clearance had to be recovered otherwise. This was done partly by approaching NTN network participants, and partly by studying the encoding of the annotation files. By searching for tag sets, annotation features, etc. on the web, we identified the treebank types (the formalisms used), input formats and in some cases also the origins of the annotations. Our NTN contacts then verified our educated guesses and supplied them with additional information. Some of the documentation which was missing from the webpages due to inactive URLs was uploaded to the Copenhagen Dependency Treebank's Google Code repository, including a few HTML pages documenting the tools developed in NTN and the common representation formats used in that project. The recovered information was harmonized with META-SHARE and included in the metadata records for the monolingual treebanks, and will be further maintained via the INESS webpages.

For new treebanks, the following metadata were collected from the developers and harmonized with the META-SHARE schema for text corpora description:

- Annotation mode (automatic, semi-automatic, manual)
- Grammar/parser (type and/or name of tool)
- Grammar origin/creator (project, person(s), name(s) of annotator(s))
- Grammar type/formalism (constituency, dependency, LFG, etc.)
- Output format (Tiger XML, CoNLL, etc.)
- Tagset (documentation URL, taglist, etc.)
- Terms of use and license information

3.2 IPR clearance

Rights clearance is an attempt to balance interests. On the one hand copyrighted material must be protected. On the other hand it should be possible to access material for innovative work and to allow uses of copyrighted material that may be beneficial for society. A well-developed and easy-to-grasp legal system that protects tools and resources in a general and transparent way is an incentive for people to create, share and use tools and resources. Infrastructures provide an invaluable framework for establishing best practices for clearing rights using standardized agreements.

Fixed licenses are available for instance through Creative Commons.[6] Moreover, a set of fixed licenses specifically made for the sharing of LRT is available through META-SHARE, including also standardized depositor's agreements that regulate the rights and obligations that hold between the copyright owner and the distributor of a resource. Similarly, the CLARIN infrastructure is developing editable templates for end-user licenses as well as for depositor's agreements. CLARIN licenses, however, were still under development when the work reported here was initiated.

[6] http://creativecommons.org/

In our experience, the copyright holders as well as the researchers negotiating the user terms find standardized legal texts or templates reassuring.

Rights clearance is time-consuming work, and in order to facilitate the long-term reuse of resources, it is essential that rights clearance is done with a long-term perspective in mind. Thus, extensive work has been done on rights clearance both for source texts and for the grammatical annotations of the treebanks. Rights were negotiated separately for source texts and annotations. It was endeavored, to the extent possible, to resolve IPR issues uniformly, using common or similar agreements for resources with a common or similar origin.

For the Sofie treebanks developed under NTN, rights had been cleared for the original source text and its translations in an exemplary way, but only for use in the context of that project. These agreements illustrate a limitation that must be avoided in a long-term infrastructure: in order to secure maximal reuse, permissions must be granted to user groups that are minimally restricted and for as general purposes as possible. For instance, the NTN rights clearance only allowed the acting research group to create one specific *derivative*, namely a corpus to be browsable (but not downloadable) online under certain restrictions.[7] In the context of META-NORD the rights clearance for this material therefore had to be renegotiated to allow the distribution of the treebanks for general use in language technology R&D.

A depositor's agreement was signed with Aschehoug, the publisher of the Norwegian original of *Sophie's World*. Aschehoug also wrote a recommendation letter to the publishers of the translations, which were subsequently contacted. Signed depositor's agreements have so far been obtained for the Swedish, Estonian, Danish, Icelandic, German and Georgian translations, while the English version was already freely available through SMULTRON. For the Finnish translation, unfortunately, the translator who holds the rights to the text did not give permission to distribute the treebank. The depositor's agreements used for the Sofie materials are based on the standard META-SHARE template, and have restrictions on the redistribution of the texts while allowing the use of the texts for R&D purposes in language technology, the most important purpose of META-NORD.

While rights clearance for source texts often requires a certain amount of negotiation since the rights holders did not originally intend the texts to be used for R&D purposes, clearing rights for linguistic annotations is in principle easier, since these annotations are designed specifically for such purposes. The main challenge with linguistic annotations is thus not the rights clearance itself, but the identification of the rights holders in cases where this has not been properly documented. Of the annotations created in NTN, the webpages made no mention of IPR ownership or licenses, and only a few treebanks had creation data encoded in the annotation file itself, so that it remained a challenge to identify the creator(s) and rights holders of several annotations. The solution, reached in agreement with NTN network coordinator Joakim Nivre and project co-workers Mathias Buch-Kromann and Kadri Muischnek (the creators of the Swedish, Danish and Estonian annotations respectively), was for the network coordinator to sign a common depositor's agreement on behalf of the annotation group and for all annotations created within the project. The approach of using one common agreement for all annotations developed under NTN was also adopted for the new annotations developed in META-NORD, with a few exceptions for treebanks made by third-party collaborators. Treebanks borrowed from unrelated projects were released under the conditions specified for that project. New annotations developed for META-NORD by third parties were released under the

[7] A derivative is a product that contains a substantial, or significant, part of an original resource.

same license as treebanks developed by META-NORD partners, but with individual depositor's agreements since the IPR holders were not project members.

In the META-NORD parallel treebanks, treebank alignments constitute pairwise stand-off layers of annotation. These were created in and by INESS and quality assured by META-NORD partners, and the rights to these annotations remain with the project consortium. The alignments are, however, independent annotations, and must also be supplied with a license formalizing the terms of use.

Our experiences with individual rights clearance for each layer of the treebank (source text, linguistic annotation and alignment) clearly demonstrate that the IPR aspect of complex resources is, indeed, very complex. In the process of resolving IPR issues, questions and doubts constantly arose as to whether our approaches were good enough, or whether our solutions were legally sound. Should, for example, the grammatical annotation in a treebank and its source text be considered as separate with respect to licensing? Consider Figure 1, in which the words are 'leaves' in the grammatical sentence analyses. Can the grammatical annotation, being tightly intertwined with the source text, be considered as a resource completely detached from the text that it describes, or is an annotation the *combination* of source text and linguistic marking? The answer to such a question has legal implications, since the linguistic annotations could conceivably be shared for further research under a fairly open license whereas the source text *qua* source text may remain licensed under considerably stricter terms of use. The user perspective is also an important consideration, because it might not be ideal to confront the user of the treebank with several licenses, one for each monolingual treebank and one for each annotation layer.

In project internal discussions it was tentatively concluded that it is possible to provide separate terms of use of the source text and the annotations as long as the user is explicitly told which conditions hold for which part. Annotations with for example a CC-BY license[8] can then be used freely (with attribution) even when the source text is more restricted, as long as the user has been made aware that if the source text is extracted from the annotation, the source text license applies. In other words, it is not *always* the case that the annotation will be restricted by the license of the text.

The discussion originally revolved around the Acquis treebanks, whose source texts are from the Acquis Communautaire[9] and are in the public domain, with no rights holder or restrictions of use. For the META-NORD Acquis treebanks, however, texts for the relevant languages were selected from the JRC-Acquis multilingual corpus[10], where Acquis documents have been aligned at document level. This slightly complicated the picture, since this aligned corpus applies specific terms of use which are stricter than CC-BY. It was made explicit in the META-SHARE description of these treebanks that the Acquis documents are in the Acquis Communautaire, which is available via the EUR-lex webpage, and that the same documents are available in the JRC Acquis corpus under specific conditions.

It is sometimes extremely time-consuming, if not impossible, to establish contact with all copyright owners. Standard corpora collected from many relatively small text excerpts are typical

[8] Public Creative Commons license with attribution, allowing the user to share and to modify the resource, also commercially, provided that the creator and/or licensor of the original resource is attributed; see `http://creativecommons.org/licenses/by/3.0/`.

[9] The total body of European Union (EU) law applicable in the the EU Member States, distributed via EUR-Lex (`http://eur-lex.europa.eu/`).

[10] `http://ipsc.jrc.ec.europa.eu/index.php?id=198`

examples. Fair use of quoted text fragments may sometimes be invoked in such situations. However, if the author of a text does not want the text to be distributed, for whatever reasons, that decision should be respected. In some cases, the original permission from each text contributor is recorded, but only concerned the intended use within the project creating the corpus, not its reuse; such short-sighted arrangements make it necessary to renegotiate the terms of use for new research within a research infrastructure.

The complexity of factors involved in the annotation of (possibly) copyrighted text remains a challenge which calls for juridically skilled scrutiny. We propose as the safest solution to apply one overall license to each treebank as a whole, i.e., to release a treebank under the license with the most restricted conditions of use. For the Sofie treebanks, for instance, even though the linguistic annotations were cleared for a CC-BY license, the user terms of the source texts were applied, restricting the use of these treebanks to language technology R&D purposes. The Acquis treebanks, however, based on source texts that are in the public domain, could be licensed under an open source license (CC-BY) as agreed with the creators of the linguistic annotations.

Despite the ethical and legal considerations, the decision to use only one license per treebank was primarily made out of consideration for the user, who cannot be expected to have expert knowledge about IPR and licenses. A typical treebank user will probably not be interested in the different levels and details of licensing, and will be inconvenienced by having to relate to more than one set of user terms which are often hard to interpret.

An important lesson was learned from a setback experienced when one of the source text rights holders refused to release the annotated source text, even if only for R&D purposes. The problem was identified only *after* a new annotation had been developed, and this demonstrates the importance of establishing work order routines in treebank development. Clearing rights for the source text should ideally be done as a preliminary step, before annotation. This case also suggests the advantage of having a ready-made, consistent and convincing line of arguments for use in the negotiation process. Establishing good routines for treebank development will at worst increase the chances for the resource in question to be released under a restricted license; at best it will allow for unrestricted, attributed distribution.

3.3 The description of complex resources in metadata

The description of complex resources is a general challenge that must be dealt with sooner rather than later in the development of LR infrastructures if we are to avoid a proliferation of ad hoc, nonstandard approaches towards handling them. The concept of a *complex resource* may have different interpretations, but at a very general level we will here define a resource as complex if it has several components, if it is multilingual, or if several tools or methods have been applied in the process of creating it.

Parallel treebanks are complex in at least two different respects. First, they are composed of several monolingual treebanks, which makes them diverse in terms of ‘linguality’ (i.e. a multilingual resource consisting of monolingual ones) and potentially also in terms of provenance, annotation type, IPR and licensing, etc. Second, monolingual treebanks are complex in their own right, having both a text component and one or more layers of annotation. This feature makes them complex in terms of metadata since it should be possible to describe a variable number of components and layers systematically, and to express clearly how the components and layers relate to each other.

The treebanks in question present complexities on all these levels, bringing forward specific requirements for their description in META-SHARE. Within the META-SHARE framework, monolingual treebanks can currently only be described satisfactorily at an appropriate level of detail if described with an individual metadata record for each monolingual treebank. As part of a parallel resource, the individual metadata descriptions must not only account for the range of resource specific features such as type, format, creation details and contact information, they must also represent relations to the other treebanks that constitute a parallel *collection*. META-SHARE allows the definition of relations in metadata, but since there are no standardized relations with fixed meanings, relations in META-SHARE are only meaningful to human users. It is not currently possible to filter or extract resources belonging to a certain collection.

The information common for all components of the complex resource must also be described, ideally without repeating this information for each individual component treebank. In the following section we describe how we ensured that the documentation requirements were met.

3.4 Metadata creation

The implementation of a metadata schema for the description of complex resources in META-SHARE was envisioned, but not accomplished, during the course of the META-NORD project. As suggested in Lyse et al. (2012), for instance, a schema for complex resources should make it possible to search and retrieve all parts of the complex resource, or to retrieve only the subpart that a user is looking for. It was therefore necessary to find an adequate way of representing such resources in a preliminary way until the provision of a more satisfactory solution becomes available.

META-SHARE represents a language resource as a metadata *record*, with mandatory and optional features for different types of resources. The mandatory set of features constitutes the *minimal description* of a resource. A treebank, which is a syntactically annotated corpus, is classified as a “TextCorpus”. A parallel treebank is a set of individual, monolingual treebanks based on texts that stand in a translational relation to each other. Of these, some or all may have been aligned, in our case at sentence level. The representation of parallel treebanks should ideally meet the requirements sketched out in the previous section: each monolingual treebank must be described at an appropriate level of detail, documenting all individual features, while at the same time preserving information about relations holding between the individual treebanks as well as information common for these treebanks, without unnecessary duplication.

Several approaches to the representation of parallel treebanks in META-SHARE were considered. The solution first proposed by the META-SHARE developers was to create one metadata record for the entire parallel treebank, using the feature “sizePerLanguage” to specify the number of sentences for each language. This would have been an acceptable option if all the treebanks had been developed within the same project, if they were of the same type, if they had the same input formats, annotation mode, IPR holder and so on. This was clearly not the case for our treebanks. Another, similar option would be to create one overall record and to add separate “corpusTextInfo” sections for each language module. The “lingualityInfo” feature, which indicates whether the resource is mono- or multilingual, and the “languageInfo” feature, specifying the relevant language, are both described in this section. It would thus be possible to specify language and ‘linguality’ for each component treebank, as well as other annotation features such as creator, type and format. However, the metadata about provenance, IPR holder, distribution, etc., can only be described in the section describing the overall resource. If the component treebanks, as in our case, have been created in different projects, have different source text and annotation rights holders and so on, this information cannot be structured in one metadata record with several "corpusTextInfo" sections. As a consequence, neither of these solutions allows for the level of detail required to describe each treebank properly. Equally important, there is no way of showing that the combination of the component treebanks is multilingual and aligned, and that the treebanks in effect constitute one, multilingual resource.

We thus opted for a resource description with one multilingual parallel 'mother' metadata record, and one record for each of its monolingual components. The metadata records were linked using a "relation" feature: the monolingual treebanks were related to the 'mother' resource with a "partOf" relation, and to their sister resources with an "alignedWith" relation. Our parallel treebanks are now represented as multilingual text corpora which list their language components both in the "sizeInfo" part and in the "relations" part.

4 Metadata links between infrastructures

While META-SHARE collects metadata for a large number of language resources and tools, the INESS system also needs to maintain metadata as documentation of its own resources. These metadata are used for presenting documentation about each treebank to the user, and can also be used for selecting treebanks based on desired features, e.g. language, license, provenance, etc. Importantly, this includes terms of use and licensing information which must be presented to the user and in many cases must be accepted by the user. In terms of usability it should also be made maximally explicit that a parallel treebank is not necessarily a uniform resource, but rather a collection of resources of potentially different provenance, type, and quality, and that aligned treebanks may not be directly comparable.

For reasons of efficiency and consistency, it is therefore important that the metadata in both infrastructures are not created and maintained separately, but are harvested and synchronized. Several solutions were considered. Considering the shortcomings in the META-SHARE schemas described in section 3.4, the ideal solution would be to define parallel treebanks in a proper way in the INESS system and to export relevant metadata to META-SHARE. However, this would necessitate a new metadata editor interface on the INESS side, as well as suitable import mechanisms on the META-SHARE side. An easier solution was adopted, consisting of the export of META-SHARE metadata to INESS. These metadata are further maintained on the INESS server, where they can be edited using any XML editor, and, after validation, uploaded again to a META-SHARE node using a simple http-based communication protocol developed at Tilde.

For licensing purposes, it is also important to implement a trusted authentication and authorization interface. Software was created on the INESS side to allow federated login via Feide, the Norwegian federation of academic ID providers. This authentication solution will be further tested and extended to eduGAIN.[11]

5 Conclusions and suggestions for best practices

5.1 Documentation and metadata

Documentation implies the provision of information on representation, provenance and IPR in order to create trustworthy metadata. Different projects produce different data and obviously

[11] http://www.geant.net/service/edugain/pages/home.aspx

have different documentation needs; it does not appear realistic to aim for fixed, predefined metadata solutions that can accommodate any documentation need. Still, there is clearly a need for some level of standardization, and we see a great potential for infrastructure initiatives to actively influence the documentation of future projects. Specifically, initiatives such as CLARIN should provide documentation templates that clearly define the minimal sets of documentation needed. META-SHARE and CMDI(Broeder et al., 2012b,a)offer interesting opportunities here for establishing metadata profiles for different kinds of resources; CLARIN, for example, is currently going through CMDI profiles that have already been created to describe existing resources in order to identify 'families' of metadata profiles.

There should also be clearly defined documentation guidelines regarding *where* to put metadata. It is often the case that a resource is represented by a collection of files. Consider for instance the case of stand-off annotation. Placing information in every document header ensures that a future user looking for information can trust which information applies for the part of a resource represented within a given document. On the other hand this may result in the same information being repeated in several files and thus being redundant. Moreover, in case a resource is upgraded or modified it may be time consuming to update documentation properly in different files or via different channels, unless sychronization is properly automatized. For the time being, we suggest as a general guideline that structured and centralized information must be provided whenever possible, but that annotation files should include, as an absolute minimum, information about *resource creator*, *creation date* and, if relevant, *originating project*. These metadata are invariable and will not become outdated, and they will ensure the future identification of the rights holder in case an annotation should become "orphaned" (i.e., separated from its repository and metadata).

With respect to the metadata schemas, we propose a new schema supporting at least the following requirements for parallel treebanks (possibly also covering other complex resources):

- There must be one metadata record for the resource as a whole, as well as individual, nested descriptions of the monolingual treebank components.
- For each monolingual component (i.e., treebank), individual descriptions of a variable number of layers (i.e., source text and any annotations) must be allowed.
- A description of the validation of each monolingual treebank must be supported, in terms of documenting the number of acceptable analyses, unacceptable analyses, unparsed sentences, etc. (as well as which sentences or parse units this information holds for).

5.2 IPR

Few researchers without legal training are happy to deal with IPR issues without assistance. The development of standardized templates and fixed law texts, such as those developed in CLARIN and META-SHARE, are therefore indispensable. Along with the dissemination of standard depositor's agreements, licenses and the establishment of a basic legal vocabulary, routines and guidelines should be established to enable research seniors and juniors to easily clear the rights for new research material. In the context of META-NORD we tested META-SHARE and Creative Commons licenses as well as the preliminary CLARIN license templates, but our experience is that sufficient guidelines are currently still missing. Among other things, neither META-SHARE nor CLARIN could provide assistance or guidelines to foresee the complex IPR problems encountered in connection with the rights clearance for treebanking. A virtual legal help desk for the CLARIN community similar to the UK JISC Legal Guidance for ICT Use in Education, Research and External Engagement[12] would be a welcome resource for researchers and deposit centers working with language data. A similar virtual competence center and additional training activities are currently planned in the DASISH project.[13]

Among the guidelines we propose is that any efforts toward the creation and distribution of resources should begin with rights clearance of the source texts for the envisaged purpose, audience and distribution scenario before investing any time in annotation and metadata creation. Moreover, license templates as well as fixed licences offer a number of different options, such as prohibiting the distribution of the original resource or allowing derivatives. In hindsight, some of these options turn out to be more decisive than others, if the use and reuse of resources within a long-term infrastructure is truly the aim. Based on our experience it should be prioritized whenever possible to make sure that the licensor accepts derivatives (i.e. allowing modifications of the original resource).

The importance of allowing derivatives may be illustrated by a straightforward scenario for treebank-based research, namely to try out a new parser on material that has previously been analysed with another parser. Unless derivatives are allowed for that source material, the new research product cannot be shared or redeposited for further research unless the new researcher takes the same (typically time-consuming) rights clearance round that was made for clearing the right to distribute the original material. Even though the emerging infrastructure initiatives hopefully will lessen the burden of clearing rights through guidelines, standards and templates, it is a fact that the source texts used for treebanking usually come from some third party that only makes the original text available for research out of goodwill (and not because treebanking research offers the prospect of profit for the source text owner). Under such circumstances repeated rights clearance requests from future researchers may not be welcomed.

5.3 Usability

While META-SHARE is a valuable infrastructure for the availability, description and exchange of resources, it has certain conspicuous shortcomings. First, its level of user-friendliness is still not well tuned towards inexperienced users. For resource owners without previous knowledge about metadata or IPR to be able to register their resource in META-SHARE, guidelines and ideally also tutorials for IPR clearance and licensing as well as metadata description are absolutely necessary. Furthermore, the metadata must be persistent and stable; it must be guaranteed that all metadata is backed up and that the updating of metadata in all META-SHARE nodes is automatic and robust.

On a more general level we insist that resource creators always document their newly created resource, and that they conform to the minimal metadata schemas developed within a collaborative, large-scale infrastructure such as META-SHARE. Integrating metadata creation as part of the resource development routine will, importantly, ensure proper documentation on resource ownership. It will hopefully also force the resource developer to keep best practices with respect to standards and IPR in mind. Drawing on our experience from treebank development in META-NORD and INESS, we claim that resources are not effectively reusable unless they

[12] http://www.jisclegal.ac.uk/

[13] http://dasish.eu

are supplied with an absolute minimum of metadata, as described in section 3.1, and until rights are cleared with an eye towards the long-term perspective, as described in section 5.2. It would, in many cases, require less work to create a new resource than to reuse a poorly documented, existing resource. Adhering to best practices in documentation and IPR clearance is thus a crucial first step towards actual usability, and hence reusability, of language resources.

5.4 Outlook on interoperability

In this paper we have presented cooperative work in two significant infrastructure projects. We have discussed several specific issues including interoperability challenges. In fact, we see interoperability in general as the greatest future challenge for cooperation between infrastructure projects. While META-SHARE has developed a specific metadata editing tool supporting fixed schemas, CLARIN has opted for CMDI as its metadata format. In order to preserve and integrate the metadata created in META-SHARE, it seems that further work on interoperability, specifically between META-SHARE and CLARIN, should have high priority. There are ongoing experiments with mapping META-SHARE schema elements to CMDI and relevant harvesting options. These initiatives hold promise that cooperation between projects, linking of infrastructures and promotion of interoperability will increasingly occupy the agenda of the research community.

References

Adesam, Y. (2012). *The Multilingual Forest: Investigating High-quality Parallel Corpus Development*. PhD thesis, Stockholm University, Stockholm, Sweden.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., and Windhouwer, M. (2012a). Standardizing a component metadata infrastructure. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1387–1390, Istanbul, Turkey. European Language Resources Association (ELRA).

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., and Trippel, T. (2012b). CMDI: a component metadata infrastructure. In Arranz, V., Broeder, D., Gaiffe, B., Gavrilidou, M., Monachini, M., and Trippel, T., editors, *Proceedings of the Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*, pages 1–4, Istanbul, Turkey. European Language Resources Association (ELRA).

Duin, P., Durco, M., Olsson, L.-J., Schonefeld, O., and Windhouwer, M. (2010). Registry Infrastructure – v2. Deliverable d2r-5b, CLARIN.

Gaarder, J. (1991). *Sofies verden: roman om filosofiens historie*. Aschehoug, Oslo, Norway.

Gavrilidou, M., Labropoulou, P., Despiri, E., Giannopoulou, I., Hamon, O., and Arranz, V. (2012). The META-SHARE metadata schema: Principles, features, implementation and conversion from other schemas. In Arranz, V., Broeder, D., Gaiffe, B., Gavrilidou, M., Monachini, M., and Trippel, T., editors, *Proceedings of the Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*, Istanbul, Turkey. European Language Resources Association (ELRA).

Gavrilidou, M., Labropoulou, P., Piperidis, S., Monachini, M., Frontini, F., Francopoulo, G., Arranz, V., and Mapelli, V. (2011). A metadata schema for the description of language resources (LRs). In Calzolari, N., Ishida, T., Piperidis, S., and Sornlertlamvanich, V., editors, *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 84–92, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Hinrichs, E., Vogel, I., Bański, P., Beck, K., Budin, G., Caselli, T., Eckart, K., Elenius, K., Faaß, G., Gavrilidou, M., Henrich, V., Quochi, V., Lemnitzer, L., Maier, W., Monachini, M., Odijk, J., Ogrodniczuk, M., Osenova, P., Pajas, P., Piasecki, M., Przepiórkowski, A., Van Uytvanck, D., Schmidt, T., Schuurman, I., Simov, K., Soria, C., Skadina, I., Stepanek, J., Stranak, P., Trilsbeek, P., and Trippel, T. (2010). Interoperability and Standards. Deliverable d5.c-3, CLARIN.

Lyse, G. I., Escartín, C. P., and De Smedt, K. (2012). Applying Current Metadata Initiatives: The META-NORD Experience. In *Proceedings of the Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*, pages 20–27.

Rosén, V., De Smedt, K., Meurer, P., and Dyvik, H. (2012). An open infrastructure for advanced treebanking. In Hajič, J., De Smedt, K., Tadić, M., and Branco, A., editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

Rosén, V., Meurer, P., and De Smedt, K. (2007). Designing and implementing discriminants for LFG grammars. In King, T. H. and Butt, M., editors, *The Proceedings of the LFG '07 Conference*, pages 397–417. CSLI Publications, Stanford.

Rosén, V., Meurer, P., and De Smedt, K. (2009). LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Van Eynde, F., Frank, A., van Noord, G., and De Smedt, K., editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht. LOT.

Vasiļjevs, A., Forsberg, M., Gornostay, T., Haltrup Hansen, D., Jóhannsdóttir, K., Lyse, G., Lindén, K., Offersgaard, L., Olsen, S., Pedersen, B., Rögnvaldsson, E., Skadiņa, I., De Smedt, K., Oksanen, V., and Rozis, R. (2012). Creation of an open shared language resource repository in the Nordic and Baltic countries. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12)*, pages 1076–1083, Istanbul, Turkey. European Language Resources Association (ELRA).

Language Resources for Icelandic

Sigrún Helgadóttir[1],*Eiríkur Rögnvaldsson*[2]

(1)Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík Iceland
(2)University of Iceland, Reykjavík Iceland

sigruhel@hi.is, eirikur@hi.is

ABSTRACT

We describe the current status of Icelandic language technology with respect to available language resources and tools. The recent META-NET survey of the state of language technology support for 30 languages clearly demonstrated that Icelandic lags behind almost all European languages in this respect. However, it is encouraging that as a result of the META-NORD project, almost all basic language resources for Icelandic are now available through the META-SHARE repository and the local site http://www.málföng.is/, many of them in standard formats and under standard CC or GNU licenses. This is a major achievement since many of these resources have either been unavailable up to now or only available through personal contacts. In this paper, we describe briefly most of the major resources that have been made accessible through META-SHARE; their type, content, size, format, and license scheme. It is emphasized that even though these resources are extremely valuable as a basis for further R&D work, Icelandic language technology is far from having become self-sustaining and the Icelandic language technology community will need support from partners in the Nordic countries and Europe if Icelandic is to survive in the Digital Age.

KEYWORDS: Icelandic, Language Resources, Repositories, Licenses.

Sigrún Helgadóttir and Eiríkur Rögnvaldsson 2013. Language resources for Icelandic. *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* NEALT Proceedings Series 20 / Linköping Electronic Conference Proceedings 89: 60–76.

1 Introduction

According to the survey of language technology support for European languages recently conducted by META-NET (http://meta-net.eu) and published in the series "Europe's Languages in the Digital Age", Icelandic is among the European languages that have the least support (Rögnvaldsson et al., 2012). This is not surprising. The Icelandic language community with its 320,000 speakers is by far the smallest in the survey – only Maltese comes close in the number of speakers, and is also on the same level as regards language technology support. It is well known that the cost of preparing a language for the digital age is independent of the number of speakers. The same basic resources are needed, irrespective of the size and capacity of the language community. Icelandic lacks both financial and human resources to be able to follow the changes that the digital revolution has made – and will make in the near future – to the use of human language within information technology, and on human-computer interaction.

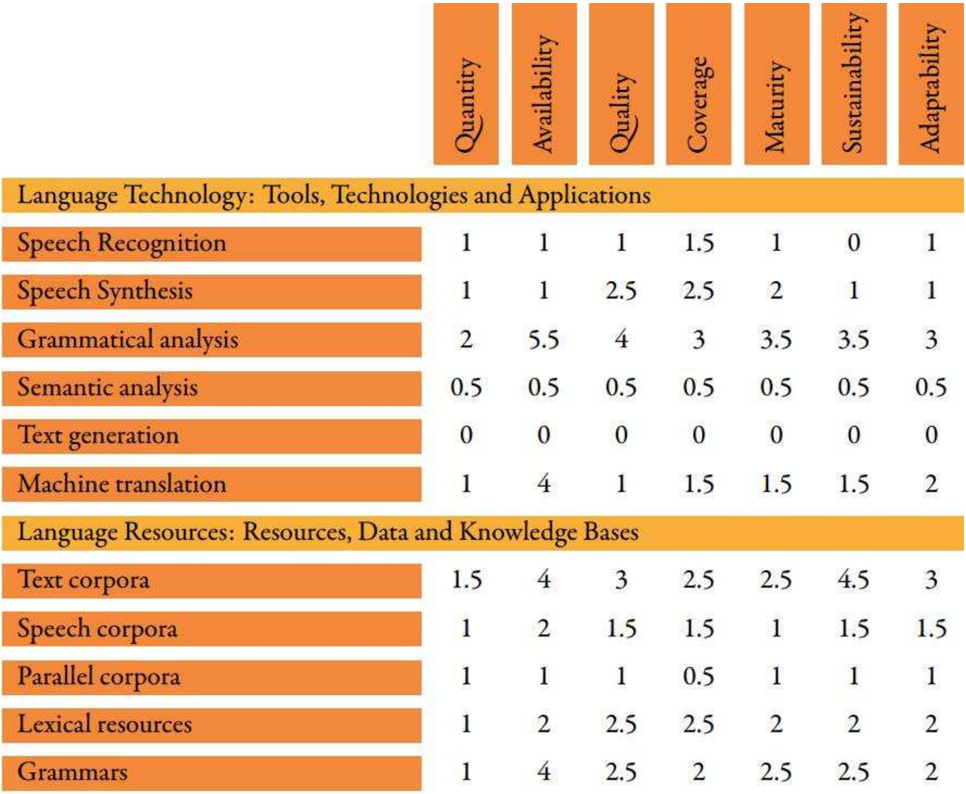
That doesn't mean, however, that the situation of Icelandic is hopeless. Even though most advanced and high-level tools and resources are lacking, a number of basic resources have been built during the last decade. This includes a PoS tagger, a lemmatizer, and a parser; a morphological database containing 270,000 paradigms; a tagged corpus of 25 million words; a multilingual dictionary with 50,000 entries; a large collection of terminologies; a treebank containing one million words; and a number of other smaller yet valuable resources. These resources form a solid ground to build on. Of course, Icelandic will always lag behind languages with millions of speakers in language technology support. It will be necessary to prioritize and select carefully which resources it is absolutely vital to develop.

For the future of Icelandic, it is extremely important that all of the resources mentioned above have now been made accessible and open to the linguistic and language technology communities at large. This should enforce R&D work on Icelandic and contribute to securing the survival of the language in the digital age. In this respect, the situation has changed dramatically during the last two years due to Iceland's participation in the META-NORD project (http://meta-nord.eu). Before the project started, some of the above-mentioned resources were indeed available, but information on their availability was not widely spread. META-NORD has made a large contribution to finalizing and standardizing some of these resources, and the META-NORD team spent a lot of effort convincing the owners of some other of these resources to make them public.

In this paper, we give an overview of the current status of language resources for Icelandic. Section 2 summarizes the main results of the META-NET survey of language technology support. Section 3 deals with language resource repositories, both META-SHARE and the Icelandic repository http://www.málföng.is/. In Section 4, we describe briefly the most important Icelandic language resources; their content, format, availability, etc. Finally, Section 5 is a conclusion.

2 Icelandic language resources in a European perspective

In September 2012, Springer Verlag published a series of 31 white papers entitled "Europe's Languages in the Digital Age" (http://www.meta-net.eu/whitepapers/overview). These volumes present the result of a study conducted by META-NET, a European Network of Excellence dedicated to building the technological foundations of a multilingual European information society. Each white paper describes one European language – its characteristics



	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	1	1	1	1.5	1	0	1
Speech Synthesis	1	1	2.5	2.5	2	1	1
Grammatical analysis	2	5.5	4	3	3.5	3.5	3
Semantic analysis	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Text generation	0	0	0	0	0	0	0
Machine translation	1	4	1	1.5	1.5	1.5	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	1.5	4	3	2.5	2.5	4.5	3
Speech corpora	1	2	1.5	1.5	1	1.5	1.5
Parallel corpora	1	1	1	0.5	1	1	1
Lexical resources	1	2	2.5	2.5	2	2	2
Grammars	1	4	2.5	2	2.5	2.5	2

Figure 1: State of language technology support for Icelandic

and particularities, its status in the society and in an international context. The main purpose of the papers, however, is to describe the status of each language with respect to technological support – language resources and tools. Experts were asked to estimate the status of the language in 11 different subfields on a scale ranging from 0 (very low) to 6 (very high) using seven different criteria. The results for Icelandic are shown in Figure 1 (Rögnvaldsson et al., 2012, page 60).

In these white papers, the state of language technology support is also compared across all 31 languages. This comparison is based on four key areas: speech processing, machine translation, text analysis, and speech and text resources. The languages were placed in one of five possible categories for each area. It turns out that Icelandic is one of only four languages that are placed in the bottom category (categorized as having weak or no support) for all these four areas – the other three being Latvian, Lithuanian and Maltese. If we compare the eight META-NORD languages (Nordic and Baltic) across all four areas, Icelandic ranks lowest of them as shown in Figure 2 (Vasiljevs et al., 2012).

The low ranking of Icelandic in this comparison is hardly surprising. Serious work on Icelandic language technology only started in the beginning of the century (Rögnvaldsson

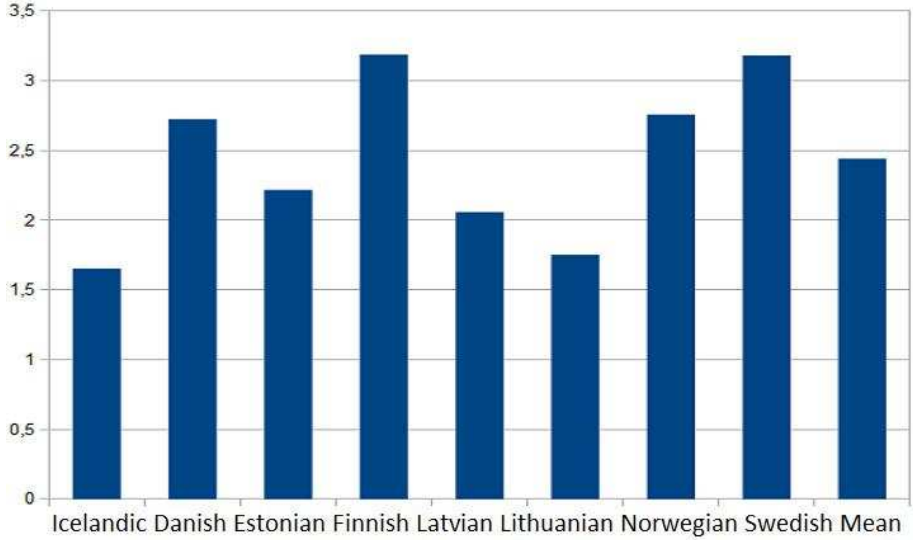


Figure 2: Average scores for each of the META-NORD languages

et al., 2009). At that time, no tools and almost no digitized language resources existed. During the last decade, the tiny Icelandic language technology community has managed to build a number of important tools and resources. Valuable language resources have also been built by Icelandic lexicographers and terminologists. Thus, even though the relative ranking of Icelandic is not encouraging, Icelanders can be proud of what they have achieved.

Two projects have been essential in providing the atmosphere and the financial basis for this work. One is the Language Technology Programme which the Minister of Education, Science and Culture initiated in 2000 (Rögnvaldsson et al., 2009). This program, which lasted for four years, funded the building of several basic resources and tools. The other important project is META-NORD. As a result of Iceland's participation in that project, most of the basic language resources for Icelandic are now easily accessible under open source licenses.

3 Málföng.is

During the last two years, Iceland has participated in META-NET through the META-NORD project which comprises all the Nordic and Baltic countries (Vasiljevs et al., 2012). The main goals of META-NORD, as of the sister projects CESAR in Eastern Europe and METANET4U in Southern Europe, were to strengthen the status of language technology in Europe, to increase awareness about the opportunities and challenges of language technology, and to make language resources and tools for all European languages more open and accessible.

One of the main aims of META-NET and the related projects was to build data repositories where language resources could be stored centrally. This aim was fulfilled by the launching of META-SHARE (`http://www.meta-share.eu/`). META-SHARE is a distributed repository with a number of nodes which are meant to be synchronized. Presently there is a number of managing nodes where all metadata which has been recorded in any META-SHARE node can be accessed. The network will gradually be extended to encompass additional nodes and centers and provide more functionality with the goal of turning into an as largely distributed infrastructure as possible.

With the exception of Iceland, META-SHARE nodes have been established in all the META-NORD countries. Due to lack of human, technical and financial resources, the Icelandic META-NORD partner, the University of Iceland, has not yet been able to establish its own META-SHARE node. Instead, the META-SHARE node established by Tilde serves as a mother node for Icelandic language resources.

However, the University of Iceland is working on the formation of a national consortium, consisting of UI, the Árni Magnússon Institute for Icelandic Studies (AMI), the Reykjavik University, the National and University Library, and perhaps a few others to maintain an Icelandic national repository of language resources. These plans will hopefully be realized in the course of 2013.

Instead of establishing a local META-SHARE node, it was decided to launch a new website called `http://www.málföng.is/` (málföng is a neologism for 'language resources'). The purpose and structure of this website is partly different from that of META-SHARE. First, the user interface and all the documentation is in both Icelandic and English. Second, the scope is meant to be wider than that of META-SHARE. The website `http://www.málföng.is/` will not only accept resources that are "complete" and "standard" in some sense, but also incomplete resources, work in progress, resources that do not follow any established standards, etc. Moreover, `http://www.málföng.is/`will contain resources of different types than those in META-SHARE – all kinds of research results, language technology papers, etc.

4 Main types of Icelandic language resources and tools

4.1 Overview

The Icelandic META-NORD team managed to get hold of almost all of the most important language resources for Icelandic as regards tools and textual resources. As for spoken language resources, a number of valuable resources have been recorded in META-SHARE and are available through `http://www.málföng.is/`, but others are either proprietary or not yet available.

Metadata was entered into the META-SHARE node at Tilde for 23 language resources (11 corpora, 9 lexical conceptual resources and 3 tools services). Of these 2 are tools for processing Icelandic text, one is a language independent tool and 20 are resources containing Icelandic text. None of the resources available for download were uploaded to the META-SHARE node. Instead links are provided as part of the metadata to `http://www.málföng.is/` where some of the downloadable resources are located or to websites such as `http://sourceforge.net/` and `https://github.com/`. At the time of writing 7 additional resources are available through META-SHARE that contain Icelandic text.

In the following, we will review the main types of tools and resources that were harvested, and describe the work carried out by the META-NORD team in order to prepare these resources for inclusion in META-SHARE.

4.2 Tools

The first PoS tagger for Icelandic was developed by Stefán Briem during the preparatory work for the *Icelandic Frequency Dictionary* (IFD, Íslensk orðtíðnibók (Pind et al., 1991)), see Section 4.3. This tagger was never publicly released and information on its structure and performance is scanty. During the years 2001–2003, several taggers were trained on Icelandic texts (the source files from the IFD). Thorsten Brant's TnT gave the best results (Helgadóttir, 2007). The training model developed in this project, together with the source texts from the IFD, has been available for researchers under certain conditions. However, it has not been publicly advertised nor uploaded to any software repository.

Hrafn Loftsson started developing an open source software package for analyzing and processing Icelandic texts during his Ph.D. studies from 2004–2007 (Loftsson, 2007, 2008; Loftsson and Rögnvaldsson, 2007). Since then, students at the University of Reykjavík and the University of Iceland have helped in developing individual components. The software, which goes by the name of IceNLP, is rule-based and uses heuristic methods which guess prepositional phrases and syntactic functions and use the acquired knowledge to force feature agreement where appropriate.

IceNLP is implemented in Java and consists of the following components: tokenizer, unknown word guesser, part-of-speech tagger, lemmatizer, parser and named-entity recognizer. Anton Karl Ingason is the main author of the lemmatizer (*Lemmald*, cf. (Ingason et al., 2008)). Individual components of IceNLP can be run independently or the JAVA clusters in question connected directly to software that is being developed.

IceNLP can be used for various tasks, such as breaking up text into individual tokens, tagging each token with its morphosyntactic tag, finding the lemma of a particular word and returning a shallow phrase structure and labels indicating syntactic functions. The package is downloadable under the LGPL (GNU Lesser General Public License), either directly from sourceforge (`http://icenlp.sourceforge.net/`) or via META-SHARE or `http://www.málföng.is/`.

Two rule-based machine translation systems between Icelandic and other languages have been developed. One is *Tungutorg* (`http://tungutorg.is/`) which translates between Icelandic and English, both ways, and also from Icelandic to Danish and from Esperanto to Icelandic. This system, which was developed by Stefán Briem, is closed and the source has not been released. Hence, it has not been registered in META-SHARE.

The other system is *Apertium-is-en* (`http://nlp.cs.ru.is/ApertiumISENWeb/`), a prototype of a shallow-transfer machine translation system that translates Icelandic text into English. The system is based on the Apertium translation system (`http://www.apertium.org/`). It was developed in the years 2009–2010 at the University of Reykjavík as the MSc project of Martha Dís Brandt as well as in independent projects of two other students, under the guidance of Hrafn Loftsson (Brandt et al., 2011). The Apertium system is downloadable under the GPL (GNU General Public License) from sourceforge (`http://sourceforge.net/projects/apertium/`) or via META-SHARE or `http://www.málföng.is/`.

4.3 Corpora

In the META-NORD project metadata for 6 Icelandic text corpora were entered into META-SHARE and the corpora made available through `http:www.málföng.is/`. One of those, IcePaHC, is a treebank and will be described in Section 4.4. The other five corpora will be described briefly in this section. Five corpora containing both text and sound files were also made available. These will be described in Section 4.8.

The largest of the text corpora, *Íslenskur Orðasjóður*, is a very large corpus of modern Icelandic that was compiled in two research projects: *Leipzig Corpora Collection* and *Frequency Dictionary Icelandic* (Hallsteinsdóttir et al., 2007; Quasthoff et al., 2012). The corpus consists of 5 sub-corpora. The two largest portions are texts from domains ending in .is collected in the autumn of 2005 by the National and University Library of Iceland (ca. 227 million running words) and the same collected in the autumn of 2010 (ca. 336 million running words). The three remaining portions contain text from a newspaper (Morgunblaðið) collected in 2001 (ca. 18.1 million running words), newspaper text from the Internet crawled in 2011 (ca. 22.6 million running words) and texts from the Icelandic edition of Wikipedia (ca. 2.5 million running words).

The corpus comes with an automatically generated monolingual lexicon, comprising frequency statistics, samples of usage, cooccurring words and a graphical representation of the word's semantic neighbourhood (Hallsteinsdóttir et al., 2007). Despite some limitations, this corpus is the only very large corpus of Icelandic in existence and it has proven to be useful in several projects. Of these, it is worth mentioning a project to create a Database of Semantic Relations (Nikulásdóttir and Whelpton, 2010), and projects to develop context sensitive spelling correction for Icelandic and the correction of OCR texts obtained from old print (ongoing unfinished projects).

The oldest Icelandic tagged corpus is the IFD corpus (Pind et al., 1991) which was compiled for the making of the Icelandic Frequency Dictionary, *Íslensk orðtíðnibók*, published in 1991. The IFD corpus consists of just over half a million running words, containing 100 fragments of texts, approximately 5,000 running words each. The corpus has a heavy literary bias as about 80% of the texts are fiction. The tagset of the IFD is more or less based on the traditional Icelandic analysis of word classes and grammatical categories, with some exceptions where that classification has been rationalized. The underlying tagset contains about 700 tags, of which 639 tags actually appear in the corpus. The tags are character strings where each character has a particular function, denoting a (specific value of a) grammatical category. The tagging and lemmatization of the IFD corpus was manually corrected and hence the corpus can be used as a gold standard for training part-of-speech (PoS) taggers and lemmatizers. All data-driven taggers used now for tagging Icelandic text are trained on the IFD corpus and it was used for the development of the rule-based tagger IceNLP (Loftsson, 2008).

The Tagged Icelandic Corpus (MÍM) was finished during the META-NORD project period (Helgadóttir et al., 2012). The corpus was originally financed by the Language Technology Programme initiated in 2000. The MÍM corpus is a synchronic corpus that contains about 25 million running words compiled at the AMI during the years 2004–2012. The texts were taken from different genres of contemporary Icelandic, i.e. texts produced in 2000–2010. The corpus is intended for use in Language Technology projects and for linguistic research. The aim of the project was to produce a balanced collection of contemporary texts, morphosyntactically tagged and lemmatized and supplied with metadata in TEI-conformant XML format (Burnard and Bauman, 2008). The texts are now available for download and search via META-SHARE or `http://www.málföng.is/`.

To make the corpus as useful as possible in LT projects it was considered of utmost importance to secure copyright clearance for the texts to be used. It was anticipated that most of the texts would be protected by copyright (final figure is about 88.5%). Permission was sought from all owners of copyrighted texts included in the MÍM corpus. Official texts (e.g. law, judicial texts, regulations and directives) are not copyrighted (11.5%). All copyright owners signed a special declaration and agreed that their material may be used free of licensing charges. In turn, AMI agrees that only 80% of each published text is included and that copies of the MÍM corpus are only made available under the terms of a standard license agreement. The crucial point in the license agreement is that the licensee can use his results freely, but may not publish in print or electronic form or exploit commercially any extracts from the corpus, other than those permitted under the fair dealings provision of copyright law. Data induced from the corpus, for example by a statistical PoS tagger, is considered results and may be used in commercial products. The license granted to the licensee is non-transferable.

The budget of the project did not allow for extensive collection and transcription of spoken language. Through collaboration with other projects, it was, however, possible to secure some spoken language data. It consists of about 500,000 running words of transcribed text which is about 2.2% of the corpus. The spoken data was obtained through four different projects (Thráinsson et al., 2007) and it includes transcriptions of about 54 hours of natural speech, recorded in different settings in the period 2000–2006. The collection contains monologues, interviews and spontaneous conversations between adults of both sexes and with different backgrounds. All the recordings have been carefully transcribed in a predefined format. The transcribed texts are a part of the downloadable MÍM corpus and are available for search with the rest of the corpus. All names have been substituted with pseudonyms, and other personal data has been removed. The monologues part of the spoken text is debates from unprepared sessions in the Icelandic Parliament, recorded in 2004-2005. The transcribed texts together with the sound files of the spoken text will be made available for search at a later date. Search in all texts except the parliamentary speeches will be password protected. The transcribed text files and sound files with the parliamentary speeches form a separate corpus that will be described in Section 4.8.

Annotation of the corpus was performed in three steps: sentence segmentation and tokenization, morphosyntactic tagging and lemmatization and transcription into TEI-conformant XML format together with relevant metadata. The procedure and software used for sentence segmentation, tokenization, morphosyntactic tagging and lemmatization was explained by (Loftsson et al., 2010) in their work on the *MIM-GOLD* corpus. The tagset used was developed for the *IFD* corpus. The automatic morphosyntactic tagging accuracy has been estimated as 88.1-95.1%, depending on text type (Loftsson et al., 2010). The corpus was transferred into TEI-conformant XML format as a part of the META-NORD project. The Norwegian search interface Glossa (Johannessen et al., 2008), which in turn uses the IMS Corpus Workbench (`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`) as a search engine, was adapted to be used with the MÍM corpus and other tagged corpora that are available through `http://www.málföng.is/`.

The MIM-GOLD corpus (Loftsson et al., 2010) is a corpus of about 1 million running words which has been sampled from MÍM. This corpus is intended as a reliable standard for the development of LT tools. The intention is that tagging and lemmatization of this subcorpus will be manually corrected. At the time of writing the tagging has been corrected by one annotator. The files are made available as version 0.9 through META-SHARE and `www.málföng.is/` for download. In later phases of the MIM-GOLD project tagging will be checked and accuracy estimated. Lemmatization will also be revised.

The last text corpus to be mentioned that is available via META-SHARE and `http://www.málföng.is/` is the *Saga corpus*. The corpus consists of 44 electronic texts of some of the Old Icelandic sagas: Family Sagas (Íslendingasögur), Sturlunga Saga, Sagas of the Kings of Norway (Heimskringla) and the Book of Settlement (Landnámabók). The texts have been normalized to Modern Icelandic spelling (Rögnvaldsson and Helgadóttir, 2011). Several inflectional endings were also changed to Modern Icelandic form.

4.4 Treebanks

The Icelandic Parsed Historical Corpus or IcePaHC (Wallenberg et al., 2011; Rögnvaldsson et al., 2011) is a one million word treebank containing material from both Modern Icelandic and older stages of the language. It is the product of three different projects which originally had different aims. The earliest and largest of these projects was a subpart of a large language technology project which had the aim of developing three different basic language resources for Icelandic. The aim of this subproject was to build a treebank of Modern Icelandic for use in language technology and to develop efficient parsing methods and tools for less resourced languages. Since some of the participants had been involved in historical syntax research, they also wanted to include a few texts from older stages of the language. However, the main emphasis was on language technology use – the corpus was intended to be a suitable training material for a statistical parser for Modern Icelandic.

At the same time, two other projects with the aim of developing resources for studying diachronic Icelandic syntax were in preparation. After some discussion, the participants in these three projects decided to join forces and make a combined effort to build a large parsed corpus covering the history of Icelandic syntax from the earliest sources up to the present. This corpus thus serves the dual purpose of being one of the cornerstones of Icelandic language technology and being an invaluable tool in Icelandic diachronic syntax research.

IcePaHC is a phrase structure treebank which uses the same general type of labeled bracketing as the Penn Treebank (with dash-separated lemmata added). The Penn annotation scheme had already been adapted for Old English (Taylor et al., 2003), which is rather similar to Icelandic in many respects, both as regards the syntax and the morphological system. Thus, the scheme could be applied to Icelandic with only slight modifications.

IcePaHC is designed from the beginning to serve both as a language technology tool and a syntactic research tool, and developed by people with research experience in both diachronic syntax and computational linguistics. Most parsed corpora are developed either for language technology use (such as the Penn Treebank, `http://www.cis.upenn.edu/~treebank/`) or for syntactic research (such as the Penn Parsed Corpora of Historical English, PPCHE, `http://www.ling.upenn.edu/hist-corpora/`.

The usefulness of the corpus as a tool for diachronic syntax research has already been demonstrated in a number of papers. The corpus has not yet been put to use within language technology but there is no reason to doubt that it can serve that purpose too. The corpus contains around 300,000 words which can safely be considered Modern Icelandic – texts from the 19^{th}, 20^{th} and 21^{st} centuries. That is more than enough material to train a statistical parser.

The corpus is completely free and open without any registration or paperwork, and the same goes for all the software that has been used to build it and the software that was developed within the project. Both the software and the corpus itself are distributed under the LGPL license and can be downloaded from the IcePaHC home page (`http://www.linguist.is/icelandic_treebank/Download`) or via META-SHARE. The treebank has also been uploaded to the INESS repository at the University of Bergen (`http://iness.uib.no`) where it may be viewed and searched.

In addition to IcePaHC, two small Icelandic treebanks exist and have been uploaded to META-SHARE. One is a dependency treebank containing Icelandic translation of the first part of the Norwegian novel *Sophie's world* (*Sofies verden*) by Jostein Gaarder. This text was annoted within the Nordic Treebank Network. The other is a small fragment of the JRC-Aquis text which was annotated within META-NORD. Both texts have been aligned with texts from several other languages.

4.5 Dictionaries

In November 2011 *ISLEX – Icelandic Scandinavian Web Dictionary* was opened to the public. The project began in 2005 and is a collaboration of five institutes: AMI in Reykjavík, Iceland, The Danish Society for Language and Literature (DSL) in Copenhagen, Denmark, The Department of Linguistic, Literary and Aesthetic Studies at Bergen University, Norway, and the Department of Swedish at Gothenburg University, Sweden. The project has mainly been financed by the governments of these countries. The administration of ISLEX is in Reykjavík and the Icelandic part of the dictionary is compiled and processed by AMI, and the development of the database and the software is also centred there. The editing of the target languages takes place in the participating countries, each editorial team being responsible for their own target language (Sigurðardóttir et al., 2008).

The dictionary was from the start designed for the web where the possibilities offered by that medium could be used. ISLEX thus contains many images, sounds and hyperlinks. The pronunciation of all Icelandic headwords is given as sounds, and all nouns, adjectives and verbs are linked to the DMII (see Section 4.7). ISLEX is a medium sized dictionary with 50,000 headwords. It describes modern Icelandic with an emphasis on phrases, fixed expressions and examples of use, all of which are translated into the target languages. It is the first comprehensive Scandinavian online dictionary which combines so many languages.

The Icelandic META-NORD team secured consent from the five institutes to make the ISLEX database available through META-SHARE and `http://www.málföng.is/`. A link to the ISLEX search page is provided and the database is downloadable in LMF (Lexical Markup Framework) format (`http://www.lexicalmarkupframework.org/`). The database contains the Icelandic headwords, phrases and fixed expressions and their translations into Danish, Swedish and the two Norwegian language standards and grammatical information such as part-of-speech and gender and number for nouns.

The conversion of the database to the LMF format was done as a part of the META-NORD project. When converting multilingual dictionaries into LMF format, a special record would usually be made for each sense of every word in all the languages of the dictionary. A so-called “Sense Axis” would then be used to link closely related senses in different languages. For ISLEX we had to take a different route. Special records were made for each sense of the words in the source language, Icelandic, which in turn had translations for that sense in each of the target languages.

4.6 Terminologies

A Term Bank was established by the Icelandic Language Institute in November 1997 (Thorbergsdóttir, 2003). The bank contained terminologies from terminological committees and individuals. Some of the terminologies had been published in printed books. In 2006, the Icelandic Language Institute became a part of AMI which now is a curator of the bank.

One of the roles of the Term Bank is to standardize the use of terms within related and unrelated subject fields. The aim is to hinder that many different terms are used for the same concept or phenomenon. The Term Bank provides an overview of Icelandic terminology and topical neologisms and thereby makes it easier to coordinate and standardize term usage. Additionally, the Term Bank provides access to Icelandic translations of foreign terms, and access to definitions of terms in Icelandic and other languages. The Term Bank thus benefits all those who write about specialized topics, such as translators, teachers, students, journalists, government agencies, businesses and any interested people, and last but not least compilers of dictionaries.

As a part of the META-NORD project an agreement was reached with the copyright owners of 41 terminologies in the Term Bank that their terminologies could be made available through META-SHARE and `http://www.málföng.is/`. The terminologies are made available in one package with 41 terminologies with the CC BY-SA license.

Most of the terminologies contain terms in English and Icelandic, just over 103 thousand Icelandic terms in total and just over 104 thousand English terms. Some terminologies also contain terms in other languages e.g. the Nordic languages and German and French. A total of 16 languages are represented in the Term Bank. Some of the terminologies contain definitions or explanations and some even examples of usage and cross-reference between concepts.

The terminologies were transferred into TBX (TermBase eXchange, `http://www.tbxconvert.gevterm.net/`) format which is the standard format used for terminologies. By using the standard the interchange of terminological data including detailed lexical information is made easier. The terminologies from the Icelandic Term Bank in TBX format were easily imported into Eurotermbank.

4.7 Grammars

In 2004 an online version of the *Database of Modern Icelandic Inflection* (DMII; Beygingarlýsing íslensks nútímamáls) was opened to the public on the website of the AMI (`http://bin.arnastofnun.is/`). Earlier the same year the database was made available for use in language technology and lexicography (Bjarnadóttir, 2012). The database was created as a multipurpose resource to serve both the general public, teachers and linguists and the language technology community. The database contains about 270,000 paradigms from Modern Icelandic with over 5.8 million inflectional forms. The DMII was originally financed by the Language Technology Programme initiated in 2000.

As the necessary data for making a productive rule system was not available, the DMII was produced as a database containing the full paradigms for as large a portion of the Icelandic vocabulary as possible. The original source for the DMII was the electronic version of the *Dictionary of Icelandic* (Árnason, 2000) with about 135,000 headwords and the lexicographic archives of the AMI. The author of the DMII, Kristín Bjarnadóttir, has compared the vocabulary of DMII with the vocabulary of the MIM corpus (Section 4.3) and is in the process of adding paradigms to the DMII based on that comparison (Bjarnadóttir, 2012). The vocabulary of the DMII will also be augmented with vocabulary from the Icelandic Term Bank (Section 4.6) (personal communication).

The Icelandic inflectional system is very rich, with up to 16 inflectional forms for nouns, 120 for adjectives and 107 for verbs, not including variants. For each paradigm in the database each lemma is shown in full, including variants. In the version that is used for language technology projects each word form is shown together with the lemma and a morphosyntactic tag.

The database has proven to be extremely useful. The number of visits to the online version of the database has risen every year since its opening. It is used by native speakers of Icelandic, the general public, students and teachers, and students of Icelandic as a foreign language. The DMII has also been used extensively in LT projects such as for search engines, PoS tagging, context sensitive correction, in language teaching and lexicography.

The database has been available for download from the AMI website (`http://ordid.is/forsida/`) free of charge with a proprietary license since 2009. The Icelandic META-NORD team secured permission to include the DMII in META-SHARE and on `http://www.málföng.is/` in such a way that links are provided to the appropriate pages of the website of the AMI.

4.8 Spoken language

Five corpora containing transcribed speech and sound are made available through META-SHARE and `http://www.málföng.is/`. Three of these were developed by the researcher Arnar Jensson (Jensson et al., 2008). All three corpora are based on a read bi-phonetically balanced text. The *Jensson Corpus* is 3.8 hours in length with 5,612 utterances from 20 speakers. The *Thor Corpus* is 2 hours in length with 4000 utterances from 20 speakers, 10 female and 10 male. The *RÚV Corpus* is 46 minutes in length with 400 utterances from 20 speakers and contains read news items that include a large vocabulary. No two speakers read the same text. The files belonging to these three projects can be obtained from the developer under the license CC BY-NC-SA.

As mentioned in Section 4.3 transcribed text and sound files with parliamentary debates form a separate corpus, the *Parliament Speech Corpus* which is available for search and download under CC BY 3.0 license via META-SHARE and `http://www.málföng.is/`. The corpus contains twenty hours of speeches from the Icelandic Parliament during the winter of 2004–2005, in synchronized text and sound files. Information about the recordings and the speakers, such as their age and gender, are provided as well. The data is intended to reflect natural spoken Icelandic under formal conditions. The discussion periods were chosen as they primarily consist of unprepared speeches that are unlikely to have been written in advance and read out loud. In addition, the aim was on diversity of topics and speakers (w.r.t. their origin, age and gender) (Thráinsson et al., 2007).

The *Hjal* corpus is the product of a project to build the first Icelandic speech recognizer during the years 2002–2003 (Rögnvaldsson, 2004). The project was financed by the Language Technology Programme initiated in 2000 and was performed in cooperation with ScanSoft, Inc. Their role was to train the speech recognizer on the basis of the material prepared in the project. The goal of the project was to collect sufficient material to train a speaker independent isolated word recognition system. Since the project was government funded, the data produced are open to all that want to develop a speech recognizer for Icelandic.

Caller sheets were prepared containing words, phrases and sentences for the participants to read. They were to include words and phrases that are likely to be used in ASR (automatic speech recognition) applications; a certain number of person names, place names, company names, numerals, numbers (money amounts etc.), commands, and meaningful fillers (*OK*, *please*, etc.). Furthermore, each sheet should contain five phonetically rich sentences and three strings of isolated letters. The sentences were to be composed in such a way as to get enough samples of all occurring diphones and common triphones in Icelandic.

A word frequency list for Icelandic was also made. The minimum size of the list was to be 30,000 word forms, but due to the inflectional character of Icelandic, it was decided to include about 50,000 word forms in the list. Volunteers were recruited to call in and read the sentences. When valid recordings from 2000 speakers, sufficiently well distributed with respect to gender, age groups, regional dialects, and type of telephone (mobile vs. fixed line), had been obtained, data collection was stopped. The Hjal corpus consists of recordings from 883 speakers of these 2000 speakers together with the transcribed speech. The corpus is available under the license CC BY 3.0.

The word frequency list contains 50–60 thousand word forms. These word forms were transcribed phonetically with both the SAMPA and IPA standards. The list together with the transcriptions is available as an Excel-file under the name *Pronunciation Dictionary for Icelandic* under the license CC BY 3.0.

One of the resources made available through META-SHARE is the Icelandic – Scandinavian web dictionary ISLEX (Section 4.5). The Icelandic headwords have been recorded and the recordings can be accessed through the website (`http://islex.is/`). The resource *ISLEX Recordings* is available through META-SHARE and `http://www.málföng.is/` under the license CC BY-NC-ND.

Finally we will mention two projects that were not a part of META-NORD but are nevertheless very important LT prjojects in the Icelandic context.

The first is *Almannarómur*, a project to develop an Icelandic speech recognizer that was carried out during the winter 2011–2012 (Guðnason et al., 2012). The Almannarómur project is a part of an open source speech project, hosted by Google. The aim of the project was to enable small language communities to generate an open source speech corpus that can be used for research. In the project a database of spoken sentences was created to aid development of automatic speech recognition for Icelandic. However, the database can be used for many other types of spoken language technologies. During the project time 113,547 sentences were recorded by 563 participants through Android phones made available by Google.

The database is not yet available but will be made available to the public via http://www.málföng.is/ in order to develop spoken language technologies for the Icelandic language. For example, the database will be particularly suitable for short utterances in a mobile environment. Google has already used the database to train an acoustic model for Icelandic and made a speech recognizer for Icelandic available through their Android phones.

The second is a project to develop a new speech synthesizer for Icelandic. The Society of the Blind in Iceland instigated in 2010 the development of a new speech synthesizer and made a contract with the Polish company Ivona software (http://www.ivona.com/en/) to carry out the task. The synthesizer was ready in 2012 and contained two voices, a male voice and a female voice. Two actors recorded extensive text material that had been prepared for this purpose. Unfortunately these recordings are not available for other researchers and developers of spoken language technologies for the Icelandic language.

5 Conclusion

Even though considerable achievements have been made in building language resources for Icelandic during the last decade, it is clear that Icelandic language technology is not self-sustaining. As pointed out in the *Strategic Research agenda* (SRA) recently published by META-NET (http://www.meta-net.eu/sra-en; (Rehm and Uzkoreit, 2012)), Iceland needs external support in order to be able to follow the rapid development in language technology.

“Not all countries have the required expertise or human resources to take care of the technology support for their languages. For example, in Iceland there is not a single position in LT at any Icelandic university or college and there is only one company that works in this area. Those colleagues who work on LT at universities and research institutes come from either language or computer science departments; their main duties are not related to LT, still they managed to produce a few basic technologies and resources but advanced types of resources do not exist at all for Icelandic, nor do they for many other under-resourced languages. This is why we need to intensify research and establish techniques, methods and instruments for research and knowledge transfer so that colleagues in countries such as Iceland can benefit as much as possible for their own language from the research carried out in other countries for other languages. Bootstrapping the set of core language technologies and resources for all languages spoken in Europe is not a matter of a few countries joining forces but a challenge on the European scale that must be addressed accordingly to avoid digital exclusion and secure future business development.” (Rehm and Uzkoreit, editors, 2012, p. 66).

It remains to be seen how such bootstrapping and knowledge transfer methods can be implemented and whether they will suffice to secure the establishment of necessary language resources, tools and technologies for survival in the Digital Age.

Acknowledgments

The META-NORD project was supported by the EU ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, grant agreement no 270899 (META-NORD). Th authors would like to thank Kristín Bjarnadóttir, Steinþór Steingrímsson and Halldóra Jónsdóttir for valuable assistance.

References

Árnason, M., editor (2000). *Íslensk orðabók [Dictionary of Icelandic]. 3rd edition, electronic version*. Edda hf., Reykjavík.

Bjarnadóttir, K. (2012). The Database of Modern Icelandic Inflection. In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages – SaLTMiL 8 – AfLaT2012*, pages 13–184, Istanbul.

Brandt, M. D., Loftsson, H., Sigurþórsson, H., and Tyers, F. M. (2011). Apertium-IceNLP: A rule-based Icelandic to English machine translation system. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT-2011)*, pages 217–224. Leuven.

Burnard, L. and Bauman, S. (2008). *Guidelines for Electronic Text Encoding and Interchange P5 edition*. Text Encoding Initiative. `http://www.tei-c.org/Guidelines/P5/`.

Guðnason, J., Kjartansson, O., Jóhannsson, J., Carstensdóttir, E., Vilhjálmsson, H. H., Loftsson, H., Helgadóttir, S., Jóhannsdóttir, K. M., and Rögnvaldsson, E. (2012). Almannarómur: An Open Icelandic Speech Corpus. In *Proceedings of SLTU '12, 3rd Workshop on Spoken Languages Technologies for Under-Resourced Languages*, Cape Town, South Africa.

Hallsteinsdóttir, E., Eckart, T., Biemann, C., Quasthoff, U., and Richter, M. (2007). Íslenskur orðasjóður – Building a Large Icelandic Corpus. In Nivre, J., Kaalep, H.-J., Muischnek, K., and Koit, M., editors, *NODALIDA 2007 Conference Proceedings*, pages 288–291, Tartu. University of Tartu.

Helgadóttir, S. (2007). Mörkun íslensks texta [Tagging Icelandic Text]. *Orð og tunga*, 9:75–107.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (2012). The Tagged Icelandic Corpus (MIM). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages – SaLTMiL 8 – AfLaT2012*, pages 67–72, Istanbul.

Ingason, A. K. Loftsson, H., Helgadóttir, S., and Rögnvaldsson, E. (2008). A Mixed Method Lemmatization Algorithm Using Hierachy of Linguistic Identities (HOLI). In Raante, A. and Nordström, B., editors, *Advances in Natural Language Processing, Lecture Notes in Computer Science*, volume 5221, pages 205–216. Springer, Berlin.

Jensson, A. T., Iwano, K., and Furui, S. (2008). Language model adaptation using machine-translated text for resource-deficient languages. *Eurasip Journal on Audio, Speech, and Music Processing*, 2008. Article ID 573832.

Johannessen, J. B., Nygaard, L., Priestley, J., and Nøklestad, A. (2008). Glossa: a Multilingual, Multimodal, Configurable User Interface. In *Proceedings of LREC 2008*, pages 617–621, Marrakesh, Morocco.

Loftsson, H. (2007). *Tagging and Parsing Icelandic Text*. PhD thesis, Department of Computer Science, University of Sheffield.

Loftsson, H. (2008). Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1):47–72.

Loftsson, H. and Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In Nivre, J., Kaalep, H.-J., Muischnek, K., and Koit, M., editors, *NODALIDA 2007 Conference Proceedings*, pages 128–135, Tartu. University of Tartu.

Loftsson, H., Yngvason, J. H., Helgadóttir, S., and Rögnvaldsson, E. (2010). Developing a PoS-tagged corpus using existing tools. In Sarasola, K., Tyers, F. M., and Forcada, M. L., editors, *7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010*, pages 53–60, Valetta.

Nikulásdóttir, A. B. and Whelpton, M. (2010). Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. In Sarasola, K., Tyers, F. M., and Forcada, M. L., editors, *7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010*, pages 33–39, Valetta.

Pind, J., Magnússon, F., and Briem, S. (1991). *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavik.

Quasthoff, U., Fiedler, S., and Hallsteinsdóttir, E., editors (2012). *Frequency Dictionary Icelandic / Íslensk tíðniorðabók*. Leipziger Universitätsverlag, Leipzig.

Rehm, G. and Uzkoreit, H., editors (2012). *Strategic Research Agenda for Multilingual Europe 2020*. Presented by the META Technology Council. Springer. Berlin.

Rögnvaldsson, E. (2004). The Icelandic Speech Recognition Project Hjal. In Holmboe, H., editor, *Nordisk Sprogteknologi. Nordic Language Technology. Årbog 2003*, pages 239–242. Museum Tusculanums Forlag, Copenhagen.

Rögnvaldsson, E. and Helgadóttir, S. (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In Sporleder, C., van den Bosch, A. P. J., and Zervanou, K. A., editors, *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pages 63–76. Springer, Berlin.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2011). Creating a Dual-Purpose Treebank. In Proceedings of the ACRH Workshop, Heidelberg, 5 Jan. 2012. *Journal for Language Technology and Computational Linguistics*, 26(2):141–152.

Rögnvaldsson, E., Jóhannsdóttir, K. M., Helgadóttir, S., and Steingrímsson, S. (2012). *The Icelandic Language in the Digital Age*. Series editors Uzkoreit, H. and Rehm, G. Springer. Berlin.

Rögnvaldsson, E., Loftsson, H., Bjarnadóttir, K., Helgadóttir, S., Nikulásdóttir, A. B., Whelpton, M., and Ingason, A. K. (2009). Icelandic Language Resources and Technology: Status and Prospects. In Domeij, R., Koskenniemi, K., Krauwer, S., Maegaard, B., Rögnvaldsson, E., and de Smedt, K., editors, *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*, pages 27–32. Northern European Association for Language Technology (NEALT), Tartu University Library, Tartu.

Sigurðardóttir, A., Hannesdóttir, A. H., Jansson, H., Jónsdóttir, H., Trap-Jensen, L., and Úlfarsdóttir, Þ. (2008). ISLEX – an Icelandic-Scandinavian Multilingual Online Dictionary. In Bernal, E. and DeCesaris, J., editors, *Proceedings of the XIII Euralex International Congress (Barcelona, 15-19 July 2008)*, pages 779–790, Barcelona. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Taylor, A., Warner, A., Pintzuk, S., and Beths, F. (2003). The York-Toronto-Helsinki Parsed Corpus of Old English Prose. University of York. http://www.users.york.ac.uk/~lang22/YcoeHome1.htm.

Thorbergsdóttir, Á. (2003). Íslenskt íðorðastarf og orðabanki íslenskrar málstöðvar [Icelandic terminological work and the word bank of the Icelandic Language Institute]. *Málfregnir*, 13:3–12.

Thráinsson, H., Angantýsson, Á., Svavarsdóttir, Á., Eythórsson, T., and Jónsson, J. G. (2007). The Icelandic (Pilot) Project in ScanDiaSyn. *Nordlyd*, 34(1):87–124.

Vasiljevs, A., Forsberg, M., Gornostay, T., Hansen, D. H., Jóhannsdóttir, K. M., Lindén, K., Lyse, G. I., Offersgaard, L., Oksanen, V., Olsen, S., Pedersen, B. S., Rögnvaldsson, E., Rozis, R., Skadina, I., and de Smedt, K. (2012). Creation of an Open Shared Language Resource Repository in the Nordic and Baltic Countries. In *Proceedings of LREC 2012*, pages 1076—-1083, Istanbul.

Wallenberg, J., Ingason, A. K., Sigurðsson, E. F., and Rögnvaldsson, E. (2011). *Icelandic Parsed Historical Corpus (IcePaHC)*. http://www.linguist.is/icelandic_ treebank/.

Sponsors of
NODALIDA 2013 & NEALT

WeSearch

iness

Lingit

max manus

computas

The Center of Estonian
Language Resources

DET HUMANISTISKE FAKULTET
KØBENHAVNS UNIVERSITET

GSLT

Lingsoft®
LANGUAGE
SOLUTIONS

Mikro Værkstedet

National Library of Norway

textUrgy™

www.tungutaekni.is

Design • Joel Priestley
Photo • Arthur Sand

NEALT Proceedings Series 20 • ISBN 978-91-7519-585-8
Linköping Electronic Conference Proceedings 89
ISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2013