NEALT PROCEEDINGS SERIES
VOL. 13

Proceedings of the NODALIDA 2011 workshop

Visibility and Availability of LT Resources

May 11, 2011
Riga, Latvia

Editors
Sjur Nørstebø Moshagen and Per Langgård

NORTHERN EUROPEAN ASSOCIATION FOR LANGUAGE TECHNOLOGY

Proceedings of the NODALIDA 2011 workshop on
Visibility and Availability of LT Resources

SIG-Infra workshop in conjunction with NODALIDA 2011

NEALT Proceedings Series, Vol. 13

ISSN 1736-6305

Published by
Northern European Association for Language
Technology (NEALT)
http://omilia.uio.no/nealt

Electronically published at
Tartu University Library (Estonia)
http://dspace.utlib.ee/dspace/handle/10062/16975

Volume Editors
Sjur Nørstebø Moshagen and Per Langgård

Contents

Workshop on visibility and availability of LT resources

in conjunction with NODALIDA 2011, Riga, Latvia, May 11th 2011

PREFACE

It is our impression that restrictions on the general accessibility of language resources often are caused by simple unawareness of the needs of the language technology providers or by unfamiliarity with the formal procedures for making data available rather than by deliberate attempts to keep resources closed.

The restricted access to many of the available cross- and multilingual Nordic LT resources is not only harmful for the minority languages of the Nordic countries, it also hinders further development of richer tools and end-user aids for the languages used in Nordic collaboration.

The workshop aims at clarifying existing possibilities for easy dissemination of language resources and encourage the collaboration between Nordic LT communities through stronger focus on open-source resources and clear licensing options.

Sjur Nørstebø Moshagen and Per Langgård

PROGRAMME COMMITTEE

ASTIN (The workgroup for Language Technology in the Nordic Countries) acts as programme committee

- Torbjørg Breivik, The Language Council of Norway
- Rickard Domeij, The Language Council of Sweden
- Jakob Halskov, Danish Language Council
- Per Langgård, The Greenlandic Language Secretariat
- Sjur Nørstebø Moshagen, The Sámi Parliament in Norway

WORKSHOP PROGRAMME

Visibility and Availability of LT Resources

SIG-Infra workshop in conjunction with NODALIDA 2011

May 11, 2011

10:00: Welcome

Key-note presentation - Francis Tyers. The only option is open: Why should language resources and technology be free?

11:00-11:30 Coffee break

11:30-12:30:

Atro Voutilainen and Krister Lindén. Finnish Language Bank: A Framework for Depositing and Disseminating Language Resources for R&D

Markus Forsberg. Green Resources in Plain Sight: Opening up the SweFN++ Project

12:30-13:30 Lunch break

13:30-15:00:

Ville Oksanen/Krister Lindén. Open Content Licenses - how to choose the right one

Andrejs Vasilejevs, Bolette Sandford Pedersen, Koenraad De Smedt, Lars Borin, Inguna Skadiņa. META-NORD: Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Koenraad De Smedt and Eiríkur Rögnvaldsson. The META-NORD language reports

15:00-15:30 Coffee break

15:30-16:30:

Per Langgård and Sjur Nørstebø Moshagen. How Open is Open - visibility and accessibility from a minority perspective

Open discussion

The only option is open: Why should language technology and resources be free*?

Francis M. Tyers
Grup Tranducens
Dept. Lleng. i Sist. Inform.
Universitat d'Alacant

28th April 2011

Abstract

I would like to structure this paper in three parts. The first deals with how we use language technology resources, and impress that especially for marginalised and minority languages, these resources cannot exist in a vacuum. The second describes some of the principle problems faced by language technology and resources. Finally, I argue that the only viable option for the language technology sector in the Nordic countries is one of openness and free distribution.

First some definitions, when referring to *language technology*, it is taken to mean the software on which applications are based, for example a machine translation (MT) or spell-checking engine. When referring to *language resources*, it is taken to mean the data on which these application depend. For example, for a spellchecker, the dictionary, morphological rules, and error models. For a machine translation system, either the parallel corpora (if the engine is corpus based), or the dictionaries and rules (if it is rule based).

Both language technology and the resources on which it depends are interdependent. A spellchecking engine is no use without the data to run on it, likewise, a spelling dictionary is of limited use without the engine to run it.

There are three main problems facing language technology and resources. The first is *visibility*, or 'can the people who are looking for the resource find it?', the second is *availability* 'can it be used for what they want to use it for?' and finally *sustainability* 'will the resource still be available next year ... or in ten years?'

Imagine you have developed a spellchecker for a language, but it is not used because no-one knows about it, or worse still. Perhaps there is an existing spellchecker, which is no longer maintained but is more widely used because it is easier to find, or comes pre-installed. This is the problem of *visibility*.

*Free here refers to *freedom*, not *price*.

On the other hand, perhaps you are planning to work on machine translation systems between Swedish and the immigrant languages of Sweden. You find a source of bilingual lexica between Swedish and Kurdish, Swahili and Pashto, but they cannot be used because of prohibitive licensing terms. This is the problem of *availability*.

Finally, you develop a morphological disambiguator during a government-funded project. The project funding expires and work comes to a halt. There is no one left to make sure that the disambiguator is *visible* and *available* to other researchers and developers. This is the problem of *sustainability*.

For larger languages, these problems can be sidestepped by starting from scratch each time. As a result of the amount of funding available, and the larger number of speakers, the amount of effort expended in making a toolchain from scratch can be fairly minimal. One person year from a speaker population of 400 million is substantially more likely to be fundable than one person year from a speaker population of five hundred. Especially if the cost of specialist training is included – there are much more likely to be ready-trained linguists or programmers in a larger population.

This is still a tremendous duplication of effort. Furthermore, *availability* of resources for larger languages can have a direct effect on language technology for minority and marginalised languages. Consider for example the creation of multilingual applications, machine translation and bilingual dictionaries. If we want to create a dictionary of South Sámi and Finnish, then dictionaries of South Sámi and Norwegian and Norwegian and Finnish are likely to be useful – if they are available.

So, what are the solutions? The primary solution to all of these problems has been outlined very effectively by Scannell et al. (2006), the *pool*.

Bibliography

- Pedersen, T. (2008) 'Empiricism Is Not a Matter of Faith'. *Computational Linguistics* 34(3), 465–470.
- Scannell, K., Streiter, O. and Stuflesser, M. (2006) 'Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies for Developers' *Machine Translation*

FIN-CLARIN: A Framework for Depositing and Disseminating Language Resources for R&D

Atro Voutilainen and Krister Lindén
Department of Modern Languages
University of Helsinki
atro.voutilainen@helsinki.fi, krister.linden@helsinki.fi

Abstract

Researchers and developers in academia and industry would benefit from a facility that enables them to easily locate, licence and use the kind of empirical data they need for testing and refining their hypotheses and to deposit and disseminate their data e.g. to support replication and validation of reported scientific experiments. To answer these needs initially in Finland, there is an ongoing project at University of Helsinki and its collaborators to create a user-friendly web service for researchers and developers in Finland and other countries. In our talk, we describe ongoing work to create a palette of extensive but easily available Finnish language resources and technologies for the research community, including lexical resources, wordnets, morphologically tagged corpora, dependency syntactic treebanks and parsebanks, open-source finite state toolkits and libraries and language models to support text analysis and processing at customer site. Also first publicly available results are presented.

1 Introduction

Sharing of digital resources by and for researchers and other types of users is increasingly common worldwide, for instance there are several ongoing projects to create annotated text corpora and treebanks for various languages (Kromann, 2003; Mikulova et al., 2006; Nivre et al., 2006). In Finland, there are various kinds language resources for a number of languages at different organisations, but they are generally difficult to locate and take into use by researchers. Also their interoperability is generally poor due to lack of standardisation. There is an ongoing need for well organised, systematic and readily available language resources and tools. This paper outlines an ongoing effort to answer this need, in particular regarding the Finnish language.

We start with a description of language resources, users and their needs regarding language resources. Then we present an ongoing effort to answer these needs. Finally we outline some Finnish-language resources available currently or in the near future.

2 Resources, users and needs

2.1 Language resources

We use the term "language resource" to refer to a wide range of digital resources:

- small or large samples of naturally occurring text, speech and multimedia, representing different genres and time periods, and possibly annotated with various levels of linguistic analysis or other metadata;
- descriptions of the language (e.g. lexicons, morphologies, syntactic grammars, wordnets, ontologies) for human users;
- formal (linguistic or statistical) models of the language for automatic language processing tasks;
- tools to facilitate use of language resources;
- software and algorithms to enable automatic language processing tasks.

2.2 Types of users

Users of language resources are mainly researchers (in humanities; potentially also other fields such as computer and information science). Also commercial developers of language and information technological applications and services is a potentially large user segment, as development

of high-quality language technological solutions from scratch is a work and expertise intensive task.

2.3 User needs

Language resource users need means to identify and use interoperable language resources. The less effort the researchers and developers need in determining the existence of the required resource and in negotiating the access and use of the resource, the more time and money can be spent on research and innovation. Here is a partial "wish list" of user needs:

- researchers need empirical data to facilitate formulation, testing and evaluation of scientific generalisations;
- to enable replication of published empirical experiments, researchers need a way of sharing their empirical data, documentation and tools;
- researchers also need a facility for persistent storage and sharing of their (annotated) data (i) to help other researchers build on rather than duplicate existing work and (ii) to facilitate evaluation and recognition of an existing contribution, as discussed in (Pedersen, 2008);
- researchers need access to well-documented and modifiable language technological software to enable them to (i) annotate corpora specific to their research need and (ii) provide a "customised" annotation for a better match e.g. with the corpus linguistic research need;
- language technology companies and system integrators need access to well-documented and modifiable language technological software to help them provide a wider range of solutions and services to answer end-user needs in information discovery, multilingual communication, education, etc.

3 Solution in outline

FIN-CLARIN partners with Finnish service providers, research organisations, publishers and archives to set up the following kind of "ecosystem":

- a web service is set up at a service provider (Centre for Scientific Computing CSC) where language resources can be deposited, annotated and licensed for research and commercial uses;
- to help the user (researcher, developer) determine whether the service contains a relevant kind of language resource needed e.g. for formulation and testing of scientific hypotheses, the web service includes a workflow for metadata creation and use in combination with a search functionality;
- to help start use of the relevant resource, the web service sets up a transparent uniform licensing policy using which researchers can optimally access the resource as employee of web service member organisation on a single-access basis. In case the resource is not open source, licensing conditions can be understood easily on the basis of visual "laundry symbol" type classification (Oksanen et al., 2010);
- the service aims to offer various types of language corpora for researchers and developers: text, speech and video with varying levels of manually or automatically assigned linguistic annotation (e.g. morphological, syntactic, ontological). These corpora will represent both present-day Finnish (e.g. publicly available text collections on the Internet, e.g. Europarliament and Wikipedia texts) as well as diachronic corpora (licenced from domestic research institutions);
- in addition to extensive samples of natural language, the service also aims to provide various types of linguistic descriptions of the language, e.g. morphological lexicons, wordnets, name resources and grammatical descriptions (like valency descriptions). Such resources can be used for a variety of academic and practical purposes, e.g. reference material for linguistic studies, language learning solutions, and creation of language analysis software;
- to help researchers and developers efficiently use language corpora and linguistic descriptions, the service also offers a variety of software tools and technologies. One (frequent) type of researcher - a linguist with limited programming skills - needs user-friendly flexible tools to annotate, visualise

and quantitatively analyse the relevant corpus data available at the service (or even other corpora). – Another type of user is a researcher/developer with more extensive programming skills, who will benefit from a wider range of available open-source tools and technologies, e.g. software libraries and statistical modelling and analysis packages.

- the service aims to operate at a large scale, to offer very large quantities of language data (billions of words) to a growing number of users. FIN-CLARIN will partner with publishers, archives and other data providers to increase language resource coverage. FIN-CLARIN and its research partners conduct research to support annotation of the language data with an increasing level of informativeness and accuracy;

- users of the service sometimes enrich the data licenced from the service with additional annotation, e.g. as part of an empirical experiment reported in a scholarly publication. The service will offer a routine for such users to deposit their added annotations to the service for other users e.g. to enable validation and replication of empirical observations; different versions of the language data can be identified with persistent identifier codes (PIDs) and retrieved even a long time after their deposition (continuity of the service).

- the initial user base is expected to be mainly Finnish researchers and developers, but in the longer run the service aims to operate at European level (along with other CLARIN centres);

4 Current Offerings

In this section we outline some ongoing developments and resources available for FIN-CLARIN users.

4.1 FinnWordNet – the Finnish WordNet

FinnWordNet[1] is a lexical database for Finnish. It contains words (nouns, verbs, adjectives and adverbs) grouped by meaning into synonym groups representing concepts. These synonym groups are linked to each other with relations such as hyponymy and antonymy, creating a semantic network. FinnWordNet can be used in language technology research and applications. It can also be used interactively as an electronic thesaurus. The first version of FinnWordNet has been created by having the words of the original English (Princeton) WordNet (version 3.0) translated into Finnish by professional translators.

4.2 FinnTreeBank – a Dependency Syntactic Treebank for Finnish

The FinnTreeBank project[2] is creating a manually annotated dependency syntactic treebank and an automatically created large parsebank for Finnish. This work is licensed under a GNU Lesser General Public License v3.0.

The first version of the treebank (Voutilainen et al., 2011) is annotated by hand and based on 19.000 example senctences in the Large Grammar of Finnish VISK - Iso Suomen Kielioppi (`http://kaino.kotus.fi/visk/etusivu.php`, (Hakulinen et al., 2004)). A parsebank for Finnish based on the Europarl corpus and the JRC-Aquis corpus will be published in late 2011.

4.3 Open Source Morphologies – OMor

The Helsinki Open Source Morphology Project for various languages aims at implementing full-fledged morphological analysers for a number of languages using the Helsinki Finite-State Transducer Technology (HFST).

The first large-scale implemented lexicon is an Open Source Finnish Morphology (OMorFi[3]), but a number of other analyzers and generators based on open source resources for various languages have also been implemented. These works are licensed under the GNU Lesser General Public License v3.0 unless specific restrictions apply to the original lexical resources for a language. The Finnish lexicon has been substantially extended and revised before it was compiled into a finite-state transducer, whereas the other languages are more or less mechanically derived from their repective sources.

[1] `http://www.ling.helsinki.fi/cgi-bin/finclarin/fiwn.cgi`

[2] `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/`

[3] `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/omor/index.shtml`

4.4 Helsinki Finite-State Transducer Technology (HFST)

The Helsinki Finite-State Transducer software[4] is intended for the implementation of morphological analysers and other tools which are based on weighted and unweighted finite-state transducer technology. This work is licensed under a GNU Lesser General Public License v3.0. The feasibility of the HFST toolkit is demonstrated by a full-fledged open source implementation of a Finnish lexicon as well as a number of other languages of varying morphological complexity (OMor) (Lindén et al., 2009).

Acknowledgments

The ongoing project has been funded via CLARIN, FIN-CLARIN, FIN-CLARIN-CONTENT and META-NORD by EU, University of Helsinki and the Academy of Finland.

References

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen and Irja Alho. 2004. *Iso suomen kielioppi* [Large Finnish Grammar]. Helsinki: Suomalaisen Kirjallisuuden Seura. Online version: http://scripta.kotus.fi/visk URN:ISBN:978-952-5446-35-7.

Matthias Kromann. 2003. The Danish Dependency Treebank and the underlying linguistic theory. *Proc. of the TLT 2003*.

Krister Lindén, Miikka Silfverberg and Tommi Pirinen. 2009. HFST Tools for Morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers. *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology 2009*, Zürich, Switzerland.

Marie Mikulova, Alevtina Bemova, Jan Hajic, Eva Hajicova, Jiri Havelka, Veronika Kolarova, Lucie Kucova, Marketa Lopatkova, Petr Pajas, Jarmila Panevova, Magda Razimova, Petr Sgall, Jan Stepanek, Zdenka Uresova, Katerina Vesela, and Zdenek Zabokrtsky. 2006. Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual. Technical Report 30, UFAL MFF UK, Prague, Czech Rep.

Joakim Nivre, Jens Nilsson and Johan Hall. 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*.

[4] `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/index.shtml`

Ville Oksanen, Krister Lindén and Hanna Westerlund. 2010. Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN. *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC2010)*.

Ted Pedersen. 2008. Last Words: Empiricism Is Not a Matter of Faith. *Computational Linguistics, Volume 34, Number 3, September 2008*.

Atro Voutilainen, Krister Lindén and Tanja Purtonen (forthcoming). 2011. Designing a Dependency Representation and Grammar Definition Corpus for Finnish. *Proc. CILC 2011 - III Congreso Internacional de Lingüística de Corpus*.

Green resources in plain sight:
opening up the SweFN++ project

Markus Forsberg
Språkbanken, Department of Swedish
University of Gothenburg, Sweden
markus.forsberg@gu.se

Abstract

SweFN++ is a project focused on the creation and curation of Swedish lexical resources geared towards language technology applications. An important theme of the project is **openness** and its realization as a lexical infrastructure.

We give a short overview of the project, elaborate on what we mean by openness, and present the current state of the lexical infrastructure.

1 The SweFN++ project

SweFN++[1] (Borin et al., 2010a; Borin et al., 2009) is a project conducted at Språkbanken. The objectives of the project are twofold: the creation of a new lexical resource: a Swedish framenet covering at least 50,000 lexical units built on the same principles as the English Berkeley FrameNet; a curation and integration of existing free lexical resources, and thereby reusing the valuable grammatical and semantic information painstakingly collected in these resources.

The core resource to which all other resources are connected is SALDO[2] (Borin and Forsberg, 2009; Borin et al., 2008), a large, freely available lexicon with morphological and semantic information. What makes SALDO suitable as a core resource is partly because of its size, but also because its morphological and sense units have been assigned persistent identifiers (PIDs).

The lexical information of a resource is linked to the sense identifiers of SALDO, which often have the effect that the ambiguity of a resource is explicated: many of the resources associate lexical information to Part-of-Speech tagged headwords, an information that is not always valid for all the senses of the current headword. Another way of expressing this is that the resource contains information requiring human intuition to be understood completely, an undesirable property for a language technology resource.

The linking of all resources to a core resource gives us a "super lexical resource" with a diversity of lexical information. This diversity of information may be used to improve the quality of its parts. For example, the lexicon developed in the EU-project PAROLE (1996-1998) contains syntactic valency information that can be mirrored against the semantic valency information in Swedish framenet, where an inconsistency indicates an error in one of the two resources. We are currently working on a unified test bench for expressing these kinds of dependencies.

SweFN++ also includes historical lexical resources, i.e., it has a diachronic dimension (Borin et al., 2010b). The starting point of the diachronicity is four digitized paper dictionaries: one 19th century dictionary (Dalin, 1853), and three Old Swedish dictionaries (Schlyter, 1887; Söderwall, 1884; Söderwall, 1953).

For computational purposes we need to associate morphological information to the headwords of the dictionaries, a work that has been begun in the CONPLISIT project for 19th century Swedish (Borin et al., to appear) and in a pilot project for Old Swedish (Borin and Forsberg, 2008).

Linking SALDO's identifiers to the entries of Dalin is relatively straightforward because of the closeness of the language varieties. The vocabulary differences are mainly in the compounds, e.g., a word like *bäfverhund* 'dog used for beaver hunt' would not find its way in a modern lexicon since beaver hunt is no longer pursued in Sweden, even though the meaning is still relatively transparent. In cases like this we link to the head of the compound, i.e., for *bäfverhund* it would be *hund* 'dog'.

The work on linking Old Swedish to SALDO is a much more challenging task that we just have

[1] http://spraakbanken.gu.se/swefn
[2] http://spraakbanken.gu.se/saldo

started to think about. An illustrative example is the Old Swedish word *bakvaþi* meaning 'fatal accident resulting from a sword being struck backwards without the striker looking in that direction beforehand'. Naturally, there is no modern variant of this word, and it is an open, empirical question where it is most beneficial to link.

2 Openness

An important theme of the project is **openness**. The theme is a philosophical stance — we believe that research should be carried out in the open to enable scrutinization and increased collaboration. It is, from our point of view, more valuable that anyone is allowed to download and inspect unfinished work today, and, at the same time, run the risk that it is confused with something more mature, rather than taking the safer, but less productive, road of publishing the "finished" product at the end of the project.

The work on openness up until now may be summarized into four goals:

1. *To make resources and related information accessible as soon as possible, preferably at day one.*

A project such as this has it main activity during its project time. This rather obvious observation has the effect that to enable the research community to influence and contribute to the project, access to the resources and tools must be provided as soon as possible, preferably at day one.

2. *To deliver development versions of the resources, tools and related information regularly.*

This goal is related to the first one, since the input of others is only relevant if they have access to up-to-date information. We mentioned the research community, but openness is actually just as important to enable coworkers sitting just a couple of offices away to get involved. Instantaneous updates would be preferred, but for technical reasons we settle for daily updates.

3. *To deliver resources with an open content license, to use open standards for the resources, and to use and produce open source tools*

These are necessary requirements to enable someone to make good use of the resources or to continue the work that the SweFN++ project now started.

4. *To make the resources and tools available through web service APIs*

Web services are convenient ways of making resources and tools available computationally, since they enable instantaneous updates and offers a straight-forward and platform-independent way of including new lexical information into existing systems.

Web services still suffer from network latency; batch processing using web services is only feasible for small materials. On the other hand, the network speed has increased drastically the last few years, so this will probably not be an issue in a not-so-distant future.

3 Openness in practice

We have started the work on a lexical infrastructure to reach the aforementioned goals. The infrastructure has three essential nuts and bolts:

- a versioning system: Subversion[3]
- a content management system: Drupal[4]
- an XML database: eXist-db[5]

The versioning system with anonymous access is our delivery channel for the lexical resources. The use of a versioning system has the advantage that not only the latest version of a resource is available but all of its history. Not to mention the added value of using a versioning system in a collaborative environment such as a research project.

It is not only the resources that are published on a regular basis, but also a set of HTML files that give up-to-date information about such things as change history, test bench output, and statistics. The use of a content management system greatly simplifies the publication of these files.

Many of the resources are developed in CVS format, but are published as XML files[6]. These XML files are every night imported into an XML database. The XML database also has good support for creating web services for the resources, which simplifies the work.

We have developed a simple search interface on top of these web services in the content management system. The interface and the web services is referred to with the collective name *SBLEX*.

[3] `http://subversion.tigris.org/`
[4] `http://drupal.org/`
[5] `http://exist.sourceforge.net/`
[6] We aim for the LMF standard, but have not yet decided on how to best encode all lexical information in LMF.

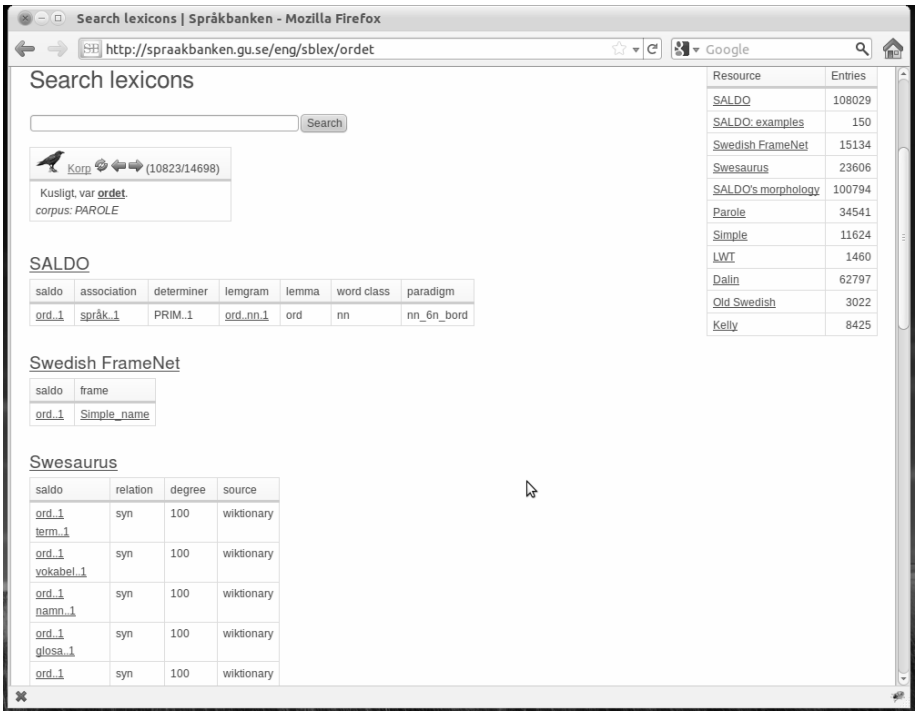


Figure 1: Searching for *ordet* 'the word' in SBLEX

Figure 1 shows a subset of the results when searching for *ordet* 'the word' in SBLEX. On the right hand side there is a table of the lexical resources in the system together with their number of entries. The first table is a random hit in our corpora material that has been annotated with SALDO identifiers, followed by information from the first three resources: SALDO, Swedish Framenet, and Swesaurus, a Swedish wordnet developed in the project.

Clicking on any of the resources in the table to the right moves us to the resource page, shown in Figure 2. All resources in SBLEX are downloadable from this page, together with XML schemata and CMDI metadata.

SBLEX is a generic system: adding a new resource requires only that the resource is added to the versioning system in a compatible format together with a few pieces of additional information such as localization.

The fact that SBLEX is generic is both a strength, since a new resource is added with ease, and a weakness, since when assuming little about the resources, it is hard to create a search interface pleasing to the eye. The result of a search is not presented in a unified manner: every resource is listed separately in a tabular format. The weakness can be remedied by creating another interface that sacrifices the function that a new resource becomes visible instantly for the benefit of a more aesthetic and logical presentation of the search results.

4 Final remarks

We have presented SweFN++, a project focused on the creation and curation of Swedish lexical resources, and discussed its theme of **openness** and its realization as a lexical infrastructure.

Openness implies that all members of the SweFN++ project work in plain sight. This can be quite disconcerting at first, but we have experienced nothing but positive effects: we feel that the work has improved in terms of quality and relevance, and that the general interest of the project has increased.

The lexical infrastructure still requires work, especially when it comes to unifying essential functions such as testing and statistics; functionalities that today are supported by a set of ad-hoc scripts for individual resources. In the context of testing we are also adding the functionality of expressing dependencies between different resources to detect inconsistencies and to generate suggestions for new entries.

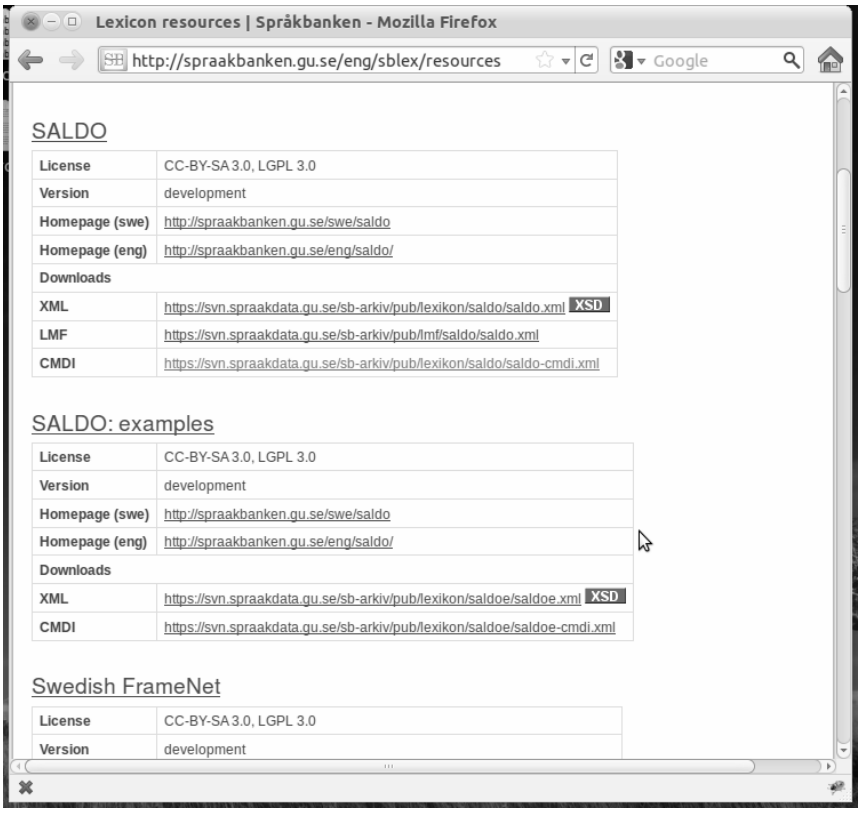


Figure 2: Download page for the resources

References

Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of Old Swedish. In *LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 9–16, Marrakech. ELRA.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In Joakim Nivre, Mats Dahllöf, and Beata Megyesi, editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, number 7 in Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia, pages 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2009. Thinking green: Toward swedish framenet++. In *FrameNet Masterclass and Workshop*.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010a. The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*.

Lars Borin, Markus Forsberg, and Dimitrios Kokkinakis. 2010b. Diabase: Towards a diachronic blark in support of historical studies. In *Proceedings of LREC 2010*.

Lars Borin, Markus Forsberg, and Christer Ahlberger. to appear. Semantic Search in Literature as an e-Humanities Research Tool: CONPLISIT – Consumption Patterns and Life-Style in 19th Century Swedish Literature. In *Proceedings of the Nodalida 2011*, Riga.

Anders Fredrik Dalin. 1853. *Ordbok öfver svenska språket. Vol. I–II.* Stockholm.

C.J. Schlyter. 1887. *Ordbok till Samlingen af Sweriges Gamla Lagar. (Saml. af Sweriges Gamla Lagar 13)*. Lund, Sweden.

Knut Fredrik Söderwall. 1884. *Ordbok Öfver svenska medeltids-språket. Vol I–III*. Lund, Sweden.

Knut Fredrik Söderwall. 1953. *Ordbok Öfver svenska medeltids-språket. Supplement. Vol IV–V*. Lund, Sweden.

Open Content Licenses - How to choose the right one

Ville Oksanen and Krister Lindén
Department of Modern Languages
University of Helsinki
ville.oksanen@tkk.fi, krister.linden@helsinki.fi

Abstract

The EU Directive harmonising copyright, Directive 2001/29/EC, has been implemented in all META-NORD countries[1]. The licensing schemas of open content/open source and META-SHARE as well as CLARIN are discussed shortly. The status of the licensing of tools and resources available at the consortium partners are outlined. The aim of the article is to compare a set of open content and open source license and provide some guidance on the optimal use of licenses provided by META-NET and CLARIN for licensing the tools and resources for the benefit of the language technology community.

1. Background

The aim of the present article is to compare a set of open content and open source licenses as used e.g. in META-NET[2], and some license templates, used e.g. in CLARIN[3], in order to help choosing between them when negotiating the rights for new resources and tools, and also to provide guidance when contacting the right holders of existing resources and tools in case a distributor wishes to take up the task of re-negotiating the rights. The licensed provided by META-NET are ready to use and they cannot be modified whereas the templates from CLARIN can be used after choosing the appropriate conditions or restrictions and they can also be modified to provide the target group with wider or narrower rights than the template does as such, or also to define the group of users entitled to access the resource.

[1] http://www.meta-nord.eu/
[2] http://www.meta-net.eu/
[3] http://www.clarin.eu/

2. Basic concepts of Intellectual Property Rights

This section discusses some of the basic concepts of IPR.

2.1. Copyright

The legislation defines the rights owned by the author of any work. The nature of these rights can be immaterial or material, and the function of copyright is to protect the author, i.e. the copyright holder, so that the rights are realised. The ideas or knowledge in the work is not protected, but the work as such is. Copyright protects the rights of authors, performers, producers and broadcasters. The copyright holder can transfer some of his/her rights to grant a third party certain rights concerning the use of protected material. One option is to issue a license containing information on the conditions under which the use is permitted. The copyright holder can also enter into an agreement stating the conditions of use with a body taking care of the distribution in practice and the agreement then specifies the license under which the administration can give rights to use the work. In the CLARIN and META-NORD context, the work is called resource or sometimes material. There copyright can belong to several authors jointly.

Copyright states that the resource cannot be used, i.e. copied or reproduced, distributed or communicated to the public without the right holder's consent, if no exception in the national legislation applies or there is no license for the resource.

2.2. Related rights for databases

Databases are covered by related rights that have the same function as copyright with the difference in the nature of the protected material (e.g. audiovisual recordings, broadcastings, photographs, databases and lists) and the terms of copyright. Otherwise the rights are similar although some details might differ. The protected issue in these related rights is the work done in compiling these, whereas copyright protects the innovative nature of the work. In the present report, the term copyright is used to cover related rights as well.

2.3. Moral rights and ethical issues

The licenses and agreements do not need to cover such acts that are governed by the legislation. These moral rights include a right to be acknowledged as creator, and a prohibition of distortion of the work. It is therefore not necessary to include a requirement for the user to cite the source in the license or agreement, nor to define that distortion of the work is not allowed. The copyright holder cannot transfer moral rights completely, and naming the author is always a precondition for use of the resource.

2.4. Economical rights

Economical rights include two basic rights: a right to produce copies of the work, and a right to make the work public. There is no requirement for the copy to be identical, and it can also be a translation. Making the work public means distribution, presentation, showing with or without technology. These rights do not mean that there should be payment involved. (Toikkanen & Oksanen, 2011)

2.5. Personal data

The Directive 95/46/EC defines personal data as: *Any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.*

For new personal data, the best approach is to procure sufficient consent for research and secondary use from the research subjects.

If personal data have been collected with insufficient rights for distribution or secondary use, there may still be some options, e.g. anonymisation for distribution or certain exemptions for scientific, historical or statistical research purposes.

In most countries, the data in speech corpora, whether transcript or sound, is regarded sensitive data, and the legislation on private person protection, i.e. the personal data issues, strongly restricts the usage of any resource where the subjects can be identified. Unless the consent from the subjects, i.e. interviewees for example, has been obtained beforehand and explicitly states the right to use it for the specified purposes in a form that the subject/interviewee has understood.

3. Licensing schemes, licenses and agreements

3.1. Open content and open source licenses

The copyright holder typically issues a license for a certain group of people, such as researchers, teachers, individuals, employees of a certain company etc. A license can either give more rights than the user otherwise would have or restrict the rights that the IPR legislation would otherwise provide him/her with. Open content and open source licenses are examples of the former whereas the End User License Agreement usually associated with commercial products such as software is an example of the latter type.

The most widely used **Open content license** system is Creative Commons, CC. The CC licenses do not require that the user be part of any predefined group. The CC-licenses give the user the right to modify, to copy, to present, and to distribute the resource. Recommendation: Use CC-licenses for open content resources when the above definition of usage applies. (Toikkanen & Oksanen, 2011)

The following restrictions can be used to restrict the rights transferred to the user:

BY (Attribution): the creator/ copyright holder must be acknowledged always. Even if the original work constitutes part of the derivative or the work distributed, the original creator needs to be acknowledged. This requirement is always part of all CC-licenses.

SA (ShareAlike): the derivatives based on the resource need to be licensed further with the same license.

NC (NonCommercial): the use towards commercial benefit is prohibited. The resource can still be distributed but no payment can be collected. Defining commercial benefit is very difficult, as the compensation can be indirect e.g. when a resource is part of a website containing commercials providing benefit for the owner. The derivatives cannot be licensed with licenses giving rights to commercial use. (Herkko Hietanen, 2008, pp 75-77).

ND (NoDerivatives): the use of the resource is restricted to the original form. Creating derivatives is prohibited. It is not possible to use parts of a text for example or to join parts of the text with other texts. In practice creating derivatives is realised by distribution.

Recommendation: CC0 offers the widest possible rights for the user

The **Open source licenses** are specifically designed for software and tools. The only widely translated license is EUPL[4] (European Union Public License) but it is not yet widely used. The most popular license for software programs has lately been GNU General Public License (GNU GPL or GPL). It provides anybody a right to use, copy, modify and distribute the software and the source code. If the program is distributed further, or if it is part of a derivative, it has to be licensed with the same license without any additional restrictions. LGPL (Lesser General Public License) differs from the GPL licenses in that where GPL makes the program available for free programs, LGPL allows for proprietary use also. Other open source licenses are MsPL[5] and BSD[6] and the Apache license[7].

Recommendation: The Apache license allows the most unrestricted use of the program.

3.2. META-SHARE licenses

META-SHARE licenses[8] are META-NET licenses based on the CC-licenses discussed above. The only difference is that they are restricted to users within the META-SHARE community. The resource can be distributed via an organisation that is a Member of META-SHARE. All the same restrictions apply.

Recommendation: META-SHARE licenses are applicable for resources where the copyright holder wants the potential users to belong to a predefined group. The distribution is not worldwide but restricted to the META-SHARE community. This can be essential for some copyright holders. Numbers of potential users are smaller than with CC-licenses. The licenses cover issues on collective works, databases and works of shared authorship.

[4] http://www.osor.eu/eupl
[5] http://www.opensource.org/licenses/ms-pl
[6] http://www.opensource.org/licenses/bsd-license.php
[7] http://www.apache.org/licenses/
[8] http://www.meta-net.eu/public_documents/t4me/META-NET-D6.1.1-Final.pdf

If the conditions and requirements of the resource allow, the license can be chosen among the open content licenses as shown in Figure 1 by Tarmo Toikkanen. In practice, the depositor of the resource does not need to create the license but choose from an existing set of licenses. Thus, "Add NC" above effectively means "Choose a license with an NC tag", e.g. META-SHARE BY NC.

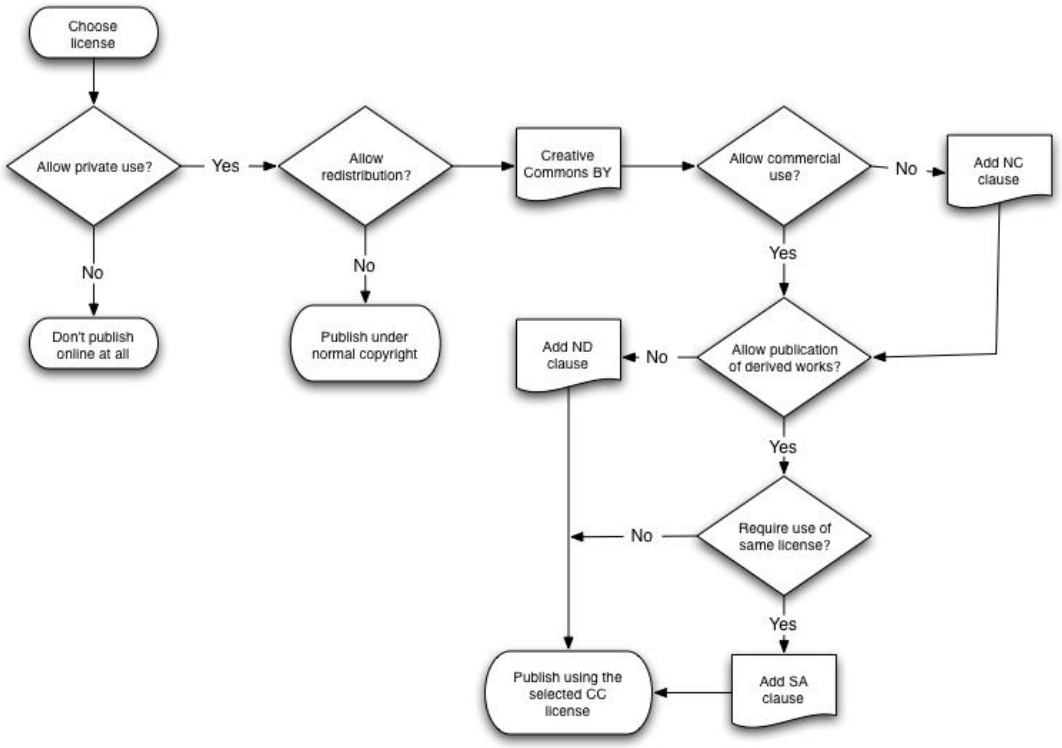


Figure 1 How to choose an open content license.

3.3. CLARIN model agreement templates

CLARIN agreement templates[9] are designed for tools and resources distributed within the research community but the Deposition & License agreement allows commercial use within the scope of the legislation by default when it is not explicitly ruled out. Without modification, the CLARIN agreement templates do not give a right for sub-licensing and they apply within the CLARIN community. The agreements presume that the copyright holder either retains the right to grant usage rights or delegates this task to the repository or some other body but the process can also be more automatic.

The CLARIN agreements are templates. The agreements can be modified to meet the requirements of the copyright holder. This option is not available with the CC-licenses or the META-SHARE licenses as they are fixed licenses.

Recommendation: The CLARIN model agreements can be modified and thus applicable to all kinds of purposes. It is, however, advisable not to make a modified agreement if one of the CC or META-SHARE or standard CLARIN licenses are applicable.

The CLARIN Deliverable D7S-2.1 (Krister Lindén & Ville Oksanen, 2010) includes two model agreements, a deposition agreement and an upgrade agreement. In addition to this, there are

[9] http://www.clarin.eu/deliverables/

other relevant CLARIN agreements, such as terms of service (between the user and the repository), privacy policy issues (for making sure that the details on the user are protected), an application form for use of restricted data from the repository, data user agreement (between the user and the repository) and the data processor agreement (between the content provider and the service provider). The document is available at www.clarin.eu/deliverables.

4. CLARIN classification scheme as a starting point

The resources available or potentially available for the META-NORD consortium have been classified with laundry tags developed for the CLARIN classification scheme. The categories will be discussed here, as well as the potential need of modifying the categories for META-NORD. There is no requirement in the CLARIN agreement templates to allow sub-licensing. Creating derivatives is allowed, but distributing them is not.

The main categories/laundry tags are (Oksanen & al., 2010):

- **Publicly Available** (PUB): No limitations on who can access and use the tools and resources. No limitations on the purpose the tools and resources are used for. No right to distribute the material.
- **Academic Use** (ACA): Available for anyone doing research or studying in an academic institution recognized by an Identity Federation (IdF). Can be used for studying, research and teaching purposes. The user needs to be authenticated.
- **Restricted Use** (RES): Any special conditions included in the deposition agreement and thus contractual in nature, e.g. a requirement to submit detailed information such as an abstract about the planned usage. Specific ethical or data protection - related additional requirements, as content including Personal Data typically falls under the scope of RES. (see section 2.5. above).

Additional restrictions or conditions are labled by NC, Inf, ReD:

- **NC**: A requirement for strictly non-commercial use. A term requiring non-commercial use of the content is commonly found in different licenses. It is problematic because there is no common definition of what non-commercial actually means in different jurisdictions.
- **Inf**: A requirement to inform the Content Owner or the Content Provider regarding the usage of the tools and/or the resources in published articles.
- **ReD**: A requirement to redeposit modified versions of the tools and resources with the Service Provider**.** In certain cases the right holder has an interest to collect the modified versions of the content, e.g. if the user adds annotation to the corpus.

Recommendation: Applying the additional restrictions or conditions should be weighed and the practical implications considered. For example Inf requires that the Content Owner or the Content Provider keep lists of articles and other publications and makes them available for the copyright holder.

The main points to consider when choosing a license or an agreement have been outlined in Figure 2 and, they are:

- Does the copyright holder or the resource itself require **special conditions**? (Use CLARIN RES);
- Is **distribution to third parties** allowed? If yes, how wide is the target group of users? (Use open content/open source or META-SHARE). Is the resource a language resource or a tool (software)? (Use CC and META-SHARE for open content, LGPL etc. for open source tools);
- If distribution to third parties is not allowed, what can the resource or tool be **used for**? (Consider CLARIN ACA

for academic/education, PUB for any kind);

- Are there any **optional requirements**? if yes, select the appropriate paragraphs in the CLARIN agreement template;
- Are there any conditions or requirements that do not have a laundry tag? If yes, modify the CLARIN agreement template accordingly.

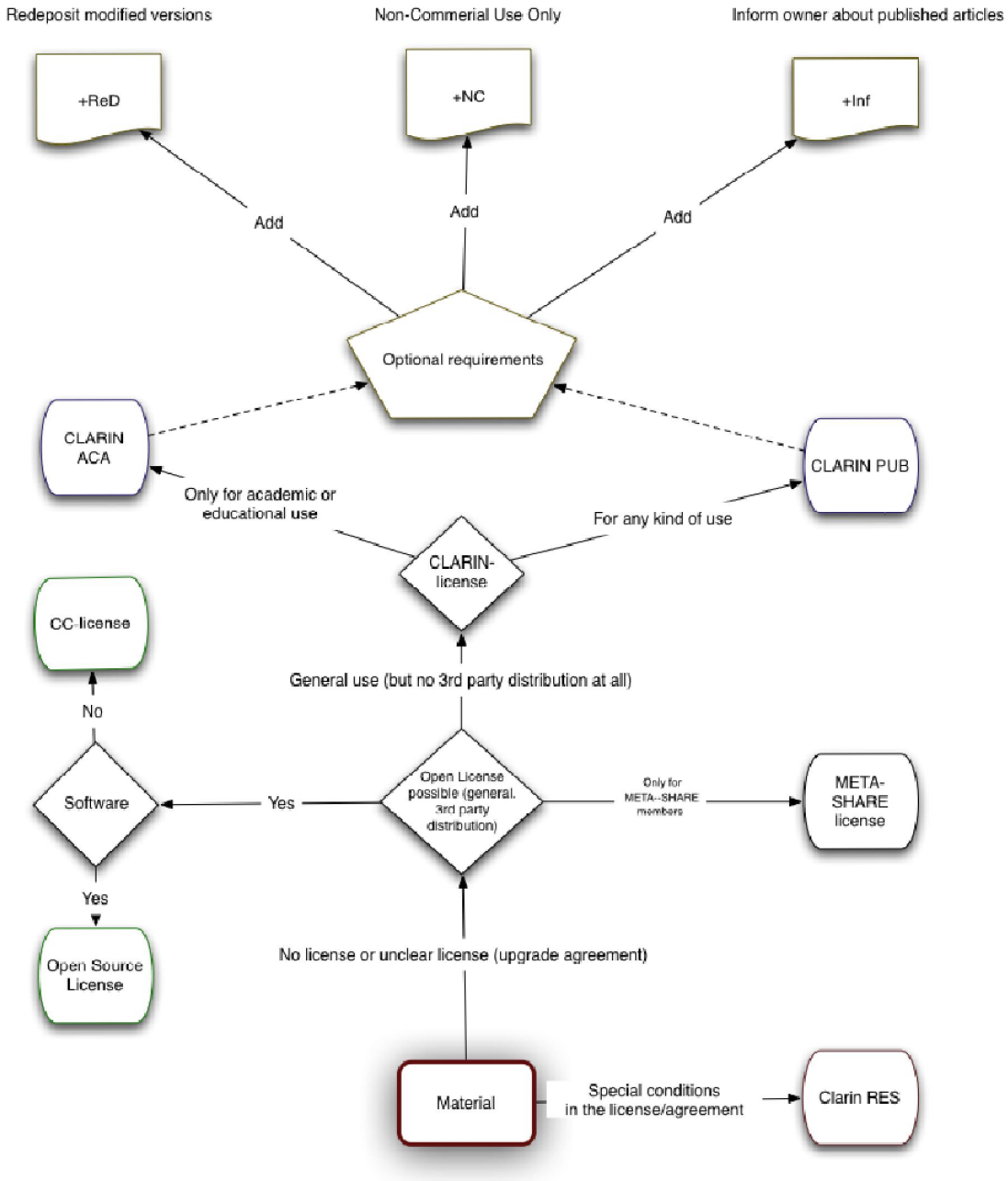


Figure 2 Choosing a license for resources and tools.

5. Conclusion and Future Work

Work with licenses offers two kinds of challenges: one is the terminology that should be common to all parties and as consistent as possible. In practice the terms used in the licenses proposed for META-NORD are not standardised, and the open content and open source licenses, and the CLARIN agreement templates use somewhat differing terms to cover the same concepts. EU wide cooperation would benefit from terminology work on legal terms.

License selection tools[10] are available for the open content licenses. The META-SHARE and CLARIN licenses and agreements could be similarly available in a web service application, and such a META-NORD/META-NET/META-SHARE License Machine could be created together with the META-NET project. Especially when one resource can be licensed with several licenses depending on the criteria set by the copyright holder, the applications would help to choose one or more appropriate licenses for both tools and resources.

Acknowledgements

We are grateful to the METANORD and FINCLARIN project for the financial support and to Hanna Westerlund for persistently questioning us to clarify and operationalize the legal concepts.

References

Herkko Hietanen. 2008. The Pursuit of Efficient Copyright Licensing — How Some Rights Reserved Attempts to Solve the Problems of All Rights Reserved, Lappeenranta University of Technology, http://urn.fi/URN:ISBN:978-952-214-721-9)

Krister Lindén and Ville Oksanen. 2011. CLARIN D7S-2.1: A report including Model Licensing Templates and Authorization and Authentication Scheme. www.clarin.eu/deliverables.

Ville Oksanen, Krister Lindén and Hanna Westerlund. 2010. Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN. *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC2010)*.

Tarmo Toikkanen and Ville Oksanen. 2011. *Opettajan tekijänoikeusopas*. FINN LECTURA, Bookwell : Porvoo

[10] For selecting a Creative Commons license, see http://creativecommons.org/choose/?lang=en.

META-NORD: Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Andrejs Vasiļjevs
Tilde
Riga, Latvia
andrejs@tilde.lv

Bolette Sandford Pedersen
University of Copenhagen
Copenhagen, Denmark
bspedersen@hum.ku.dk

Koenraad De Smedt
University of Bergen
Bergen, Norway
desmedt@uib.no

Lars Borin
University of Gothenburg
Gothenburg, Sweden
lars.borin@svenska.gu.se

Inguna Skadiņa
Tilde
Riga, Latvia
inguna.skadina@tilde.lv

Abstract

This position paper presents META-NORD project which develops Nordic and Baltic part of the European open language resource infrastructure. META-NORD works on assembling, linking across languages, and making widely available the basic language resources used by developers, professionals and researchers to build specific products and applications. Goals of the project, overall approach and specific focus lines on wordnets, terminology resources and treebanks are described.

1 Introduction

In the last decade linguistic resources have grown rapidly for all EU languages, including lesser-resourced languages. However they are located in different places, have developed in different standards (if any) and in many cases are not well documented.

High fragmentation and a lack of unified access to language resources are among key factors that hinder European innovation potential in language technology (LT) development and research.

To address these issues European Commission (EC) has dedicated specific activities in its FP7 R&D and ICT-PSP programmes[1]. The overall objective is to ease and speed up the provision of online services centered around computer-based translation and cross-lingual information access and delivery. The focus is on assembling, linking across languages, and making widely available the basic language resources used by developers, professionals and researchers to build specific products and applications.

Several projects have been started to facilitate creation of a comprehensive infrastructure enabling and supporting large-scale multi- and cross-lingual services and applications. These projects closely cooperate and form a common META-NET network.

At the core of the META-NET is TE4ME project which is funded under FP7 programme. The Eastern European part of the META-NET is covered by the CESAR project, United Kingdom and Southern European countries are represented by the METANET4U project, while the META-NORD project aims to establish an open linguistic infrastructure in the Baltic and Nordic countries.

This position paper describes the key objectives and activities of the META-NORD project. Although the project has just started, we believe it is important to introduce it to the Nordic and Baltic research community to encourage cooperation and participation in creation of the European open linguistic infrastructure.

2 META-NORD project

META-NORD project focuses on 8 European languages – Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish, – that each has less than 10 million speakers. It is the integral part of the META-NET and other related initiatives like CLARIN (Váradi et al., 2008) to create a pan-European open linguistic resource exchange platform.

Project partners are University of Copenhagen, University of Tartu, University of Bergen, University of Helsinki, University of Iceland, Institute of Lithuanian Language, University of Gothenburg, and Tilde (coordinator).

[1] http://ec.europa.eu/information_society/activities/ict_psp/documents/ict_psp_wp2010_final.pdf

META-NORD will contribute to a pan-European digital resource exchange facility by describing of the national language technology landscape, identifying, collecting resources in the Baltic and Nordic countries and by documenting, processing, linking and upgrading them to agreed standards and guidelines. A particular focus of the META-NORD is targeted to the three horizontal action lines: treebanks, wordnets and terminology resources.

META-NORD will participate in the building and operating of broad, non-commercial, community-driven, inter-connected repositories, exchanges, and facilities that will be used by language researchers, developers and professionals.

Users will have simple mechanisms for accessing a repository net to search, retrieve and exchange information about language resources as well as to get access to the actual resources. Resource providers will be supported with protocols and mechanisms for making the descriptions of their resources (and the actual resources) harvestable.

The following approaches and technologies will serve as the starting point of the work:

- existing standards (in cooperation with other projects, META-NET and partners, as well as CLARIN); includes Unicode (ISO 10646) for text encoding, ISO 639 for language codes, XML for content and metadata representation;
- digital repositories through the deployment of existing, widely recognised open-source software platforms (such as DSpace, Fedora or Sourceforge);
- metadata descriptors (e.g. Dublin Core metadata, META-SHARE proposal);
- IPR license schemes, e.g. Creative Commons and Open Data Commons principles as well as several legacy or proprietary licensing models. In CLARIN a license classification scheme for language resources has been developed and field tested. The broad categories (PUBlic, ACAdemic or REStricted) of a resource guarantees a minimal but necessary set of rights for the end user (Oksanen et al., 2010), even if a resource on further inspection of its license agreement may come with additional rights;
- open archives initiatives protocol for metadata harvesting (OAI-PHM) used to populate and update the META-SHARE and CLARIN VLO central inventories;
- web service interfaces (REST or SOAP);
- mature, language independent tools developed by the META-NORD partner institutions, e.g. Helsinki Finite-State Transducer software (HFST).

META-NORD will mobilize national and regional actors, public bodies and funding agencies by raising awareness, organizing meetings and other focused events.

In addition important collaboration with other EU partners is foreseen within Initial Training Network in the Marie Curie Actions CLARA. The CLARA project aims to train a new generation of researchers who will be able to cooperate across national boundaries on the establishment of a common language resources infrastructure and its exploitation.

3 Target users

Target users of language resource sharing platform are developers and researchers both in industry and academia. This includes private and public institutions, companies and individuals involved in HLT research and development: industrial organizations and SMEs, academic institutions, research organizations, universities, individual researchers and students, national governments, EC institutions, and private investors.

The size of target user communities is different in the project consortium countries, e.g. Icelandic language community is relatively small and there are 5 commercial companies working in the field of LT. However, META-NORD will try to get more companies interested in the field and will consider alternative possibilities for the LT development (e.g. solutions for handicapped people in collaboration with the Organization of Blind and Partially Sighted, the Icelandic Library for the Blind, and the Communication Centre for the Deaf and Hard of Hearing).

In Norway, for instance, there is as yet no good overview of the number or types of users of currently available language resources. However, based on the user accounts for the resources evaluated by META-NORD, the number of active users in Norway runs in the hundreds rather than thousands, and most users are academic. That is why META-NORD will be mostly aiming to extend the target user community with industrial users.

Similar situation is in Denmark where most users of UCPH's language resources are within academia. To give an example within industry, the Danish official version of OpenOffice now includes the Danish wordnet – DanNet.

A finer-grained analysis of the target user community (with the overview of its size, typology, perceived needs, etc.) in each consortium country will be performed during the project.

4 Open source and data approach

Interoperability between products and services from different sources within the META-NORD will be ensured through the principles and standards proposed and developed by the META-NET and, consequently, exploited by all the projects "under" the META-NET network. This way, interconnection and interoperability of networks and services will be achieved.

META-NORD does not aim at developing approaches, practices and standards within itself. It will, however, contribute to the reliable methodological, organisational and technical solutions of a broadly distributed, community-driven, open source exchange and sharing facility of META-SHARE which is laid by the META-NET. META-NORD will upgrade the chosen resources to standards agreed in cooperation with other projects, META-NET and partners.

The META-NORD linguistic infrastructure will be open and available for European researchers, developers and professionals. An open source approach has been accepted by many (HLT) practitioners, in the area of MT in particular, e. g. since 2005 a number of MT systems have been released as open source solutions and a number of conferences and workshops targeting open source technologies for MT have been held.

Also, there is an OpenNLP organisational centre for open source projects related to the natural language processing. Its primarily role is to encourage and facilitate the collaboration of researchers and developers on such projects. Currently there are more than 25 open source projects in the OpenNLP centre which is meant to provide an "umbrella" for such projects to work with greater awareness and interoperability.

In fact, IPR issues are becoming increasingly important in our field as standardization initiatives advance in the areas of data formats and content structure, making IPR the remaining obstacle to wide-scale reuse of resources. For reproducibility of research results and comparability of research methods, our field requires an open access to resources, in the form of so-called "gold standard" evaluation data. Research is incremental by its nature, and we know that many of our present-day language resources are far from perfect. Thus we rely on being able to incrementally refine language resources and make the modified resources available to the research community. This incrementality of research requires that language resources be made open. Freely available language resources are also good for industry, in particular the SME segment, where freely available resources can allow a relatively low-stakes entry into a market segment.

We would like to underscore at this point that open-source licensing formats do not in any way eliminate the need for language resource service centres, as most users will need assistance in working with resources. Further, resources will need to be periodically migrated to new formats and upgraded in other ways.

Promoting the use of open data and following the Creative Commons and Open Data Commons principles, the META-NORD will apply the most appropriate license schemes out of the set of templates provided by META-NET. Model licenses will be checked by the consortium with respect to regulations and practices at national level, taking account of possibly different regimes due to ownership, type, or pre-existing arrangements with the owners of the original content from which the resource was derived. Resources resulting from the project will be cleared i.e. made compliant with the legal principles and provisions established by META-NET, as completed/amended by the consortium and accepted by the respective right holders.

5 Multilingual action on wordnets

Wordnets organized according to the model of the original Princeton Wordnet for English (Fellbaum 1998) have emerged as one of the basic standard lexical resources in our field. They encode fundamental semantic relations among words, relations that further in many cases have counterparts in relations among concepts in formal ontologies, so that there is in many instances a straightforward mapping from the one to the other.

According to the BLARK (Basic Language Resource Kit) scheme, wordnets along with treebanks, are central resources when building language enabled applications. BLARK lists Com-

puter Assisted Language Learning (CALL), speech input, speech output, dialogue systems, document production, information access and translation applications as dependent of wordnets. The semantic proximity metrics among words and concepts defined by a wordnet are very useful in such applications because in addition to identical words, the occurrence of words with similar (more general or more specific) meanings contribute to measuring of the similarity of content or context or recognizing the meaning. Different translations of the same master wordnet, such as the Princeton WordNet can be linked with each other resulting in a multilingual thesaurus and also a dictionary which is useful e.g. in aligning multilingual parallel documents and other translation oriented tasks.

During the last decades, wordnets have been developed for several languages in the Nordic countries including Finnish, Danish, Estonian, Icelandic and Swedish. Of these wordnets, Estonian WordNet is the oldest one since it was built as part of the EuroWordNet project in the 1990s (see Vossen 1999). In contrast, most of the other wordnets have been recently initiated, e.g. the Danish wordnet has been under development since 2005 (cf. Pedersen et al. 2009).

The builders of these wordnets have applied different compilation strategies: where the Danish, Icelandic and Swedish wordnets are being developed via monolingual dictionaries and corpora and subsequently linked to Princeton WordNet; the Finnish wordnet has applied the translation method by translating Princeton WordNet into Finnish for later adjustment.

From the above mentioned different time perspectives and compilation, there is a need for upgrade of several wordnet resources to agreed standards, which will thus constitute a preliminary task of this META-NORD action.

A prerequisite for multilingual use of the resources is that the monolingually based resources are enhanced with regards to either synsets and/or more links to Princeton WordNet. From these links, which will primarily constitute the so-called "core synsets" extracted at Princeton University, pilot cross-lingual resources will be derived and further adjusted and validated.

Partial validation of the resources will be performed by means of comparison with bilingual dictionaries for the given languages (where they exist). An additional aim of the multilingual task is to investigate the possibility of making the relevant wordnets accessible through a uniform web interface.

Wordnets provide semantically-based concept hierarchies for specific languages and are therefore ideal resources to use as a starting point for cross- and multilingual resources. With such linked resources, cross- and multilingual IR applying semantically-based query expansion becomes feasible. Another possible application for these resources is Machine Translation (MT). The hierarchical structure of wordnets ensures that a translation can be found (going up or down in the hierarchy) even if a precise equivalent is not present between the specific languages.

6 Horizontal Action on multilingual terminology

Among specific activities of META-NORD project will be consolidation of distributed multilingual terminology resources across languages and domains, and upgrading terminology resources to agreed standards and protocols.

META-NORD will extend an open linguistic infrastructure with multilingual terminology resources. META-NORD partners Tilde, Institute of Lithuanian Language, University of Tartu and University of Copenhagen have already established a solid terminology consolidation platform EuroTermBank (Vasiljevs et al., 2008). This platform provides a single access point to more than 2 million terms in 27 languages. Still terminology coverage for some languages (e.g. Latvian, Lithuanian, Polish, Hungarian) is much stronger than for some others which have limited terminology resources integrated.

EuroTermBank platform will be integrated into an open linguistic infrastructure by adapting it to relevant data access and sharing specifications. META-NORD will approach holders of terminology resources in Nordic countries facilitating sharing of their data collections through cross-linking and federation of distributed terminology systems.

Mechanisms for consolidated multilingual representation of monolingual and bilingual terminology entries will be elaborated. Sharing of terminology data will be based on TBX (TermBase eXchange) standard recently adapted as ISO 30042. It is an open XML-based standard format for terminological data, created by Localization Industry Standard Association (LISA) to facilitate interchange among termbases. This standard is very suitable for industry needs as TBX files can be imported into and exported from most software packages that include a terminological database.

7 Horizontal Action on Treebanking

Treebanks are among the most highly valued language resources. Applications include development and evaluation of text classification, word sense disambiguation, multilingual text alignment, indexation and IR, parsing and MT systems.

The objective of the META-NORD is to make treebanks for relevant languages accessible through a uniform web interface and state-of-the-art search tool. In cooperation with the INESS project, an advanced server-based solution will be provided for parsing and disambiguation, for uploading of existing treebanks, indexing, management, and exploration. The treebanking tools will run on dedicated systems and provide fast turnaround. Existing treebanks available in the consortium will be integrated on this platform.

A second objective is to link treebanks across languages using parallel multilingual treebanking based on existing language and corpora.

Parallel treebanks can be used for translation studies, for bilingual dictionary construction, for identifying and characterizing structural correspondences, for multilingual training and evaluation of parsers, and for the development and test of sophisticated MT systems. Especially multilingual parallel treebanks are useful for developing hybrid MT systems.

Linguistically motivated interactive linking with XPAR technology will initially be performed for LFG-based parsebanks which support f-structure linking. Danish, Norwegian and English will be used in the first pilot, based on the multilingual Sofie-corpus. In the second phase, linking will be extended to dependency treebanks, e.g the Finnish treebank, using technology from FIN-CLARIN. Combining these technologies, a pilot parallel treebank is planned for Norwegian, Danish, Finnish and English.

Particular goal is to extend the Estonian TreeBank and improve its quality/format/querying interface. The Estonian Treebank can be used for training parsers and taggers for Estonian. The rule based parsing system for Estonian can be used for building Estonian Treebank. The rule set for deeper dependency parsing will be extended in order to perform better analyses.

The FinnTreeBank can be used for training parsers and taggers for Finnish. In the META-NORD project the goal is to extend the Finnish treebank with a parser and sample quality testing to a Finnish ParseBank for the Europarl corpus in order to create a multilingual treebank so that it will be applicable to training e.g. MT systems. In particular, the efforts will be coordinated with the Norwegian and Danish treebank projects.

The Icelandic treebank will consist of approximately one million words. The main emphasis is on Modern Icelandic but the treebank will also contain texts from earlier stages of the language. Thus, it is meant to be used both for language technology and for syntactic research. This is a Penn-style treebank but it should be possible to convert it to other formats so that it can be linked to other treebanks via the Norwegian treebanking infrastructure.

In cooperation with the INESS a treebanking infrastructure will be put in place that can be used by all languages. A highly detailed Norwegian treebank will be provided.

8 Acknowledgements

The META-NORD project has received funding from the European Commission through the ICT PSP Programme, grant agreement no 270899.

References

Fellbaum, C. (ed). 1998. *WordNet – An Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, London, England.

Oksanen V., Linden K., Westerlund H. 2010. *Laundry Symbols and License Management – Practical Considerations for the Distribution of LRs based on experiences from CLARIN*. In the Proceedings of LREC 2010.

Pedersen, B.S, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen. 2009. *DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary*. Language Resources and Evaluation, Computational Linguistics Series. Volume 43, Issue 3:269-299.

Vossen, P. (ed). 1999. *EuroWordNet, A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, The Netherlands.

Váradi T., Krauwer S., Wittenburg P., Wynne M., Koskenniemi K. 2008. *CLARIN: common language resources and technology infrastructure*. Proceedings of the Sixth International Language Resources and Evaluation Conference.

Vasiljevs, A., Rirdance, S., Liedskalnins, A., 2008. *EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data*. Proceedings of the First International Conference on Global Interoperability for Language Resources ICGL 2008. Hong Kong, 2008, pp.213-220.

The META-NORD language reports

Koenraad De Smedt
University of Bergen
Bergen, Norway
desmedt@uib.no

Eiríkur Rögnvaldsson
Háskóli Íslands
Reykjavik, Iceland
eirikur@hi.is

Abstract

As part of the META-NORD project, the state of affairs in language technology in the Nordic and Baltic countries is being described in a set of eight reports. Each language report describes the situation of a language community and the position of the language service and language technology industry for that language. This position paper presents our methodology and preliminary findings. The final reports will be published in the META-NET series of white papers for all main languages of Europe.

1 Background

The aim of the recently started META-NORD project is to make basic language resources for the Baltic and Nordic countries more accessible to developers, professionals and researchers in order to build language enabled applications.[1] As part of this effort, the project is compiling overviews of the language service and language technology industry for all the languages targeted by the project. These languages include the main official languages spoken in the Nordic and Baltic geographical area: Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish.

For most of these languages, there have been some previous surveying efforts during the past few decades, mostly in preparation of R&D programmes in language technology or for the establishment of language resources infrastructures. These overviews have had different aims and methodologies and their findings are therefore not fully comparable. In some countries, such as Norway, Sweden and Iceland, plan documents and their overviews of the state of the art have often been tied to official language policy and government propositions, whereas in other countries, such as Denmark, government branches dealing with technology and development have also contributed with stimuli towards plans and surveys.

It is not the first time that a surveying effort is launched across the whole of Northern Europe. In the aftermath of the language technology research programme financed by the Nordic Council of Ministers (2000–2005), a comprehensive report was written, known as *Vismansrapporten* (Lindén et al., 2006). This report presents an analysis of needs, opportunities and policies, identifies key areas, estimates magnitudes of R&D funding, indicates obstacles, notably aspects of rights and licensing, and presents a vision for a future embedding of language technology in the Nordic and Baltic society. *Vismansrapporten* is likely the first wide-ranging overview of the situation of language technology in this area. It was compiled by a careful analysis of documents and research budgets, as well as by a questionaire which was sent out to a large number of experts in the area, and includes literal quotes from the expert's answers to open questions.

While the usefulness of *Vismansrapporten* is recognized, the situation of language technology needs and solutions, and the constellation of technology consumers and providers, is rapidly changing, so that a new effort, five years later, is justified. As an indication of the changed situation, consider that fact that access to social media has boomed during the past five years, and in Norway, access to media content from mobile devices tripled from the beginning of 2009 to the end of 2010.[2] Also, new industrial players (especially SMEs) have emerged during the past five years, producing an increased need for contact be-

[1] See elsewhere in this volume for a more extensive overview of general aims and structure of the META-NORD project.

[2] Source: http://medienorge.uib.no/

tween industry and academia. In the same period, the Nordic Language Councils have successfully established a closer cooperation between countries about language technology through seminars and other communication, but they have not published systematic status reports.

The META-NORD reports are written as a series of separate publications for each language, but they are closely coordinated in their structure. Their data includes numerical estimates of a large number of technological aspects, compiled on the basis of the same framework that is used in the whole META-NET network.[3]

2 Aim and audience

The META-NORD reports aim at raising awareness for language technology support and the benefits of sharing and exchanging resources by depicting the importance of language technology for every individual language as part of the European information society. The function of the reports is to serve as the ground for planning cooperation between the participating countries, and for identifying strengths and weaknesses to be addressed. The target audiences are therefore mainly nonexpert readers such as politicians and journalists, national funding bodies, research councils, language councils, private companies in the technology sector, and also universities and research institutions.

Each report, which is about thirty to forty pages long, is brought out in the respective language under discussion as well as in English. Similar reports are prepared by the other partner projects participating in META-NET in order to cover the main languages of Europe. It is expected that the publication of the whole series of papers in the English version will have considerable impact across Europe and may affect the conception of future language technology R&D programmes.

3 Report structure

For each of the languages, an analysis of the language community has been conducted and the role of the language in the respective country/language community is described. The language technology research community and the language service and language technology industry are identified. The importance of language technology products and services in the language community is assessed. Legal provisions related to language resources and tools, which may differ from country to country, are outlined.

The structure of the language reports for all the META-NET languages is the same. They have three main sections. The first section, which is common to all the reports and written by experts from the DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz) is entitled "A Risk for our Languages — A Challenge for Language Technology", and is intended to explain the opportunities and challenges for language technology in the modern information society.

The remainder of each report is different for each language and written by experts on that language. It contains subsections on general facts on the language (number of speakers, official status, dialects, etc.), particularities of the language, recent developments in the language, language cultivation, language in education, international aspects, and the role of the language on the Internet.

The reports further contain an important section on language technology support for the language in question. It contains subsections on the core application areas of language and speech technology, such as language checking, web search, speech interaction, machine translation, etc. and describes the situation in the language with respect to the application areas. Furthermore, there are language particular subsections on language technology in education and language technology programs in the country in question. The language particular parts of this section are written by experts on each language.

The reports present a detailed table with ratings of language technology tools and resources for each language. Experts were asked to rate the existing tools and resources with respect to seven criteria: quantity, availability, quality, coverage, maturity, sustainability, and adaptability. The experts were asked to rate the following 13 types of tools and 12 types of resources according to these criteria for their language:

1. Tokenization, Morphology (tokenization, PoS tagging, morphological analysis/generation)
2. Parsing (shallow or deep syntactic analysis)
3. Sentence Semantics (WSD, argument structure, semantic roles)
4. Text Semantics (coreference resolution, context, pragmatics, inference)
5. Advanced Discourse Processing (rhetorical

[3] META-NET is a Network of Excellence of which META-NORD forms a part; `http://www.meta-net.eu/`

structure, coherence, argumentative zoning, argumentation, text patterns)
6. Information Retrieval (text indexing, multimedia IR, crosslingual IR)
7. Information Extraction (NER, event/relation extraction, opinion/sentiment recognition)
8. Language Generation (sentence generation, report generation, text generation)
9. Summarization, Question Answering, Advanced Information Access Technologies
10. Machine Translation
11. Speech Recognition
12. Speech Synthesis
13. Dialogue Management (dialogue capabilities and user modelling)
14. Reference Corpora
15. Syntax Corpora (treebanks)
16. Semantics Corpora
17. Discourse Corpora
18. Parallel Corpora, Translation Memories
19. Speech Corpora (raw and annotated)
20. Multimedia and Multimodal data (text data combined with audio/video)
21. Language Models
22. Lexicons, Terminologies
23. Grammars
24. Thesauri, WordNets
25. Ontological Resources for World Knowledge (e.g. upper models, linked data)

A preliminary results are summarized as barplots in the Appendix, where the mean value for all criteria (each rated on a scale from 0 to 6) is given for each language and each tool or resource type. The data are not finalized for all languages, as more input from experts for some language is still expected. Also, it must be taken into account that all values are based on estimates.

The results indicate that only with respect to the most basic tools and resources such as tokenizers, PoS taggers morphological analyzers/generators, syntactic parsers, reference corpora, and lexicons/terminologies, the situation is reasonably good for all the META-NORD languages. Furthermore, all the languages seem to have some tools for information extraction, machine translation and speech recognition and synthesis, as well as resources like parallel corpora, speech corpora, and grammars, although these tools and resources are rather simple and have a limited functionality for some of the languages.

When it comes to more advanced fields like sentence and text semantics, information retrieval, language generation, and multimodal data, it appears that one or more of the languages lack tools and resources for these fields. For the most advanced tools and resources like discourse processing, dialogue management, semantics and discourse corpora, and ontological resources, most of the languages either have nothing of the kind or their tools and resources have a quite limited scope. The means for all languages together (final tables) indicate that quantity and availability may be a greater concern than quality; this need is the very *raison d´être* of the META-NORD project.

4 Discussion and conclusion

The closely parallel methodology for writing the META-NORD language reports, in coordination with all of META-NET, secures the representation of the Nordic and Baltic languages in a Europe-wide series of white papers on the status of language technology in all main national language communities.

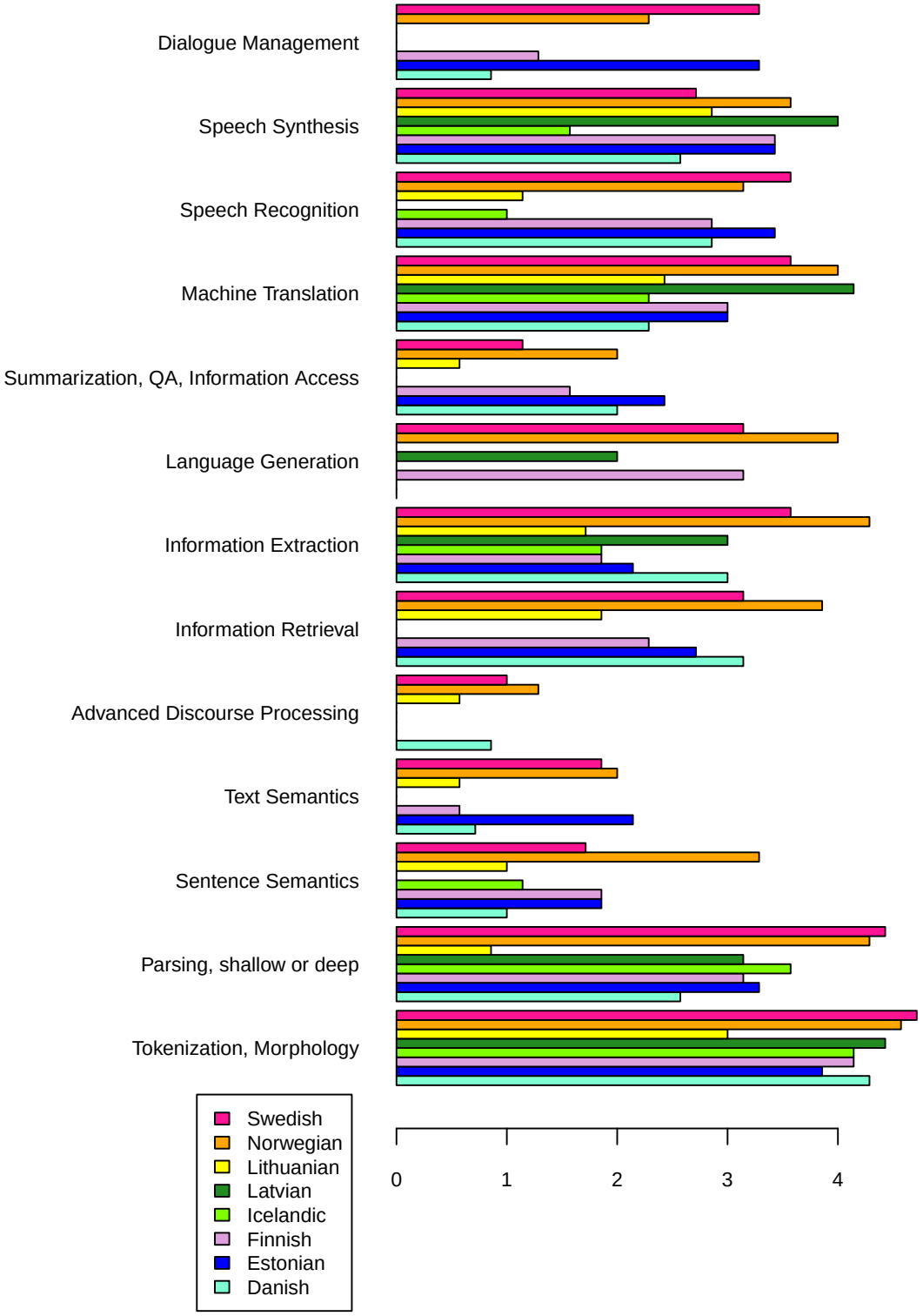
A shortcoming of the current effort is that META-NORD is focusing only on the eight main languages in its geographic area, while minority languages are not explicitly addressed. This means that the smaller Nordic languages Greenlandic, Faroese, Kven and Sami are mentioned only in passing. Also, Russian is not included, even if Northwestern Russia is a part of Northern Europe and Slavic languages are important minority languages in the Baltic countries.

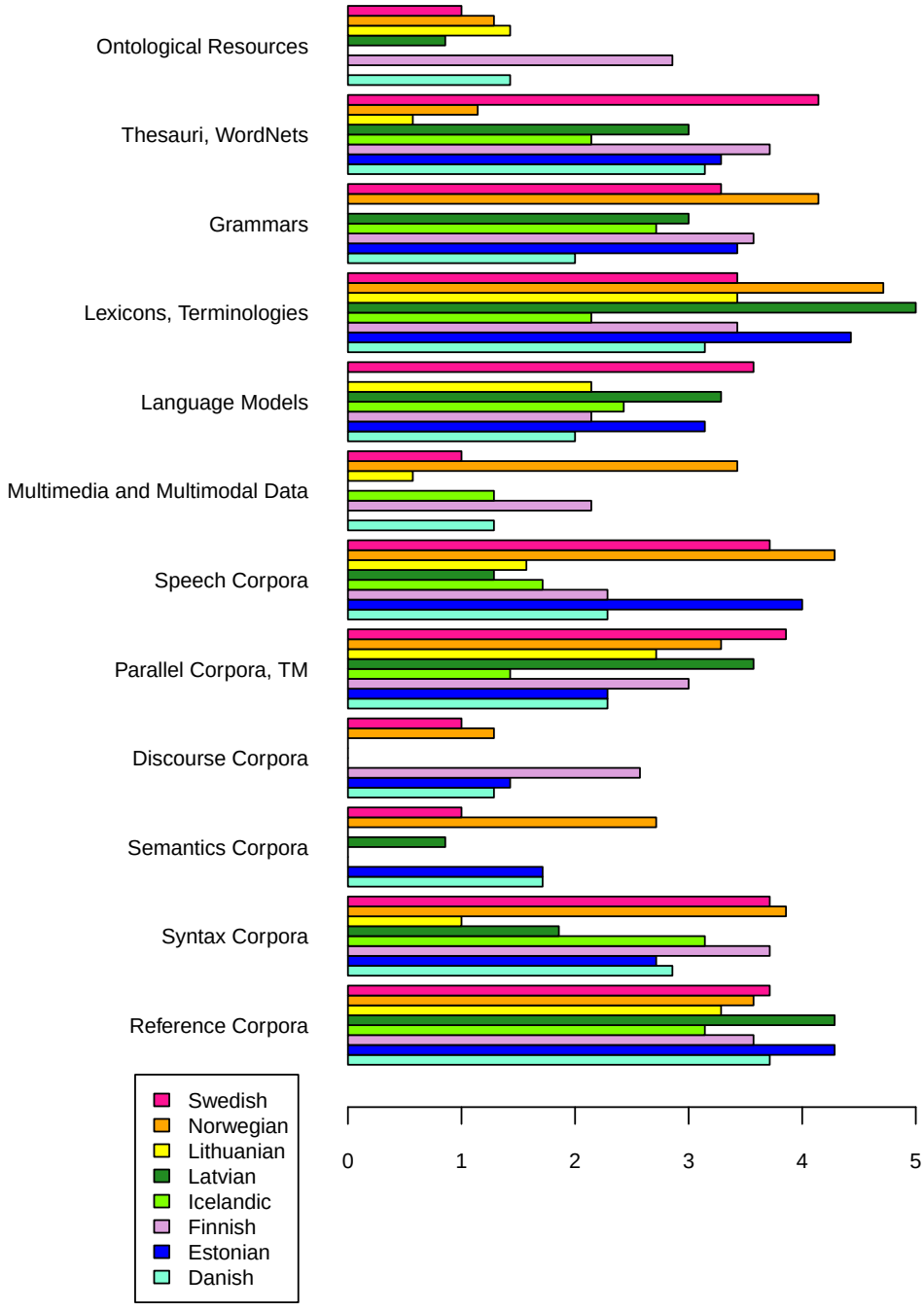
The language reports show that the Nordic and Baltic countries still have a long way to go to realize the vision of making the area a leading region in language technology, which was the aim that *Vismansrapporten* set out for 2016. However, the reports will hopefully enable us to locate our strengths and weaknesses and point to prospective possibilities for fruitful cooperation, in particular sharing of tools and resources, which will considerably strengthen the field in the near future.

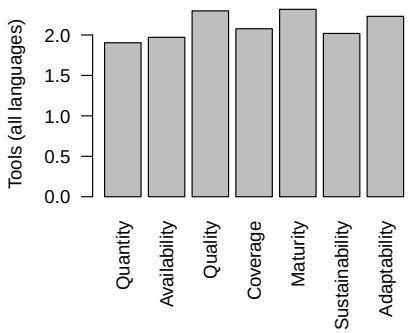
References

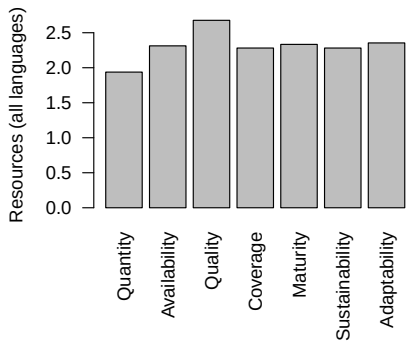
Lindén, Krister, Kimmo Koskenniemi, and Torbjørn Nordgård. 2006. Språkvis — Vismansrapport — Expert Panel Report. The Nordic Countries — A Leading Region in Language Technology. `https://kitwiki.csc.fi/twiki/bin/view/Main/LTExpertPanelBookView`.

Appendix: Barplots of the assessment of the status of tools and resources




Ontological Resources
Thesauri, WordNets
Grammars
Lexicons, Terminologies
Language Models
Multimedia and Multimodal Data
Speech Corpora
Parallel Corpora, TM
Discourse Corpora
Semantics Corpora
Syntax Corpora
Reference Corpora
Swedish
Norwegian
Lithuanian
Latvian
Icelandic
Finnish
Estonian
Danish
0
1
2
3
4
5


Tools (all languages)
0.0
0.5
1.0
1.5
2.0
Quantity
Availability
Quality
Coverage
Maturity
Sustainability
Adaptability


Resources (all languages)
0.0
0.5
1.0
1.5
2.0
2.5
Quantity
Availability
Quality
Coverage
Maturity
Sustainability
Adaptability

How open is open - visibility and accessibility from a Greenlandic perspective

Per Langgård
Oqaasileriffik/ The Language Secretariat, Nuuk, Greenland
per@oqaaserpassualeriffik.org

Abstract

Language technology for Inuit languages is vital for language survival. On the surface it should be easily provided since (i) linguistic rights for most of the Inuit dialects are well secured, (ii) Inuit languages maintain a very high status among its speakers, (iii) the need for technological solutions is recognized at the political level and (iv) funding for projects on Inuit culture and language is comparatively easy to obtain. Still, only one working project is found.
A number of reasons for this state of affairs will be identified and a case made to show that an extremely easy access to all kinds of free resources is the only option for Inuit languages to enter into the much needed world of language technology

We are doing well

The Greenlandic language technology project is not very old. Neither is it very big in terms of staff or other resources, and academic achievements are very meagre this far.
Still, the project has attracted vast amounts of attention not only from lay Greenlanders but we have also noted quite a lot of interest among professionals in the field.
We are of course very pleased to see that our efforts pay off and very proud whenever we hear mention of our project in academic circles which we do comparatively often, basically for two reasons:
(i) Language technology programmes for Inuit languages apart from a few attempts conducted by southern scholars are non-existing.
(ii) Greenlandic is notorious for morphological complexity. Until we launched the first finite state automaton in 2006 the standard attitude was a total rejection of language technology for Inuit languages even among the most prominent scholars in the field. A polysynthetic language cannot be computerized. I could add that I even today come across high ranking linguists in the field of Eskimology who maintain that Mother Earth is flat.
We are of course proud and happy to collect the laurels to Greenland but it would be very hypocritical to leave it there, for without computational linguists and computational scientist with a serious wish to share their own achievements with the rest of the world we simply would not be where we are. In our case the guardian angels are situated in Tromsø/ Kautokeino and in Odense, but it could no doubt have been Gothenburg, Oslo, or someplace else had things developed just a bit differently back then in 2005 when the whole project started.
So congratulation Tromsø and Odense, and congratulation to all the rest of you who believe in open resources. We did it together!

Analysis

It is a fact that the need for language technology to support minority languages - especially threatened ones - is generally recognized in the political bodies with direct influence and power like ICC and RAIPON. The Tromsø conference on indigenous languages in the Arctic in October 2008 is an obvious example.
Still, the attempts seldom make it past declarations of intent into concrete projects or good actual projects soon dry out and die. Greenlandic language technology is no doubt one of the few real sunshine stories of its kind, not only among Inuit languages but also among minority languages as such. I would therefore like to take the opportunity to address two of the questions this statement gives rise to at this point:
(i) How come it is so difficult?
(ii) What will it take to pave the way for many more projects like the Greenlandic one?
The analysis to follow is primarily based on my experiences with Greenlandic, Iñupiaq, and Inuktitut but rather many encounters with representatives for American First Nations and RAIPON (Russian Association of Indigenous Peoples of the North) have given me the impression that the observations hereunder are much more widespread than I originally believed them to be.
As a very first answer to the question above most of us will no doubt think in terms of unrecognised linguistic rights, lack of financial and linguistic resources, or the like.
This is no doubt part of the truth and thirty

years ago I would myself gladly have accepted it as the whole truth. But not any more. The situation is simply far more complicated than we believe it to be for an immediate glance.

Linguistic rights have been only partially threatened in Greenland and only for a very limited number of years. Greenlandic as an official/ recognized language has a long history. From the very onset of mission in 1721 Greenlandic speaking Greenlanders have had the saying in language questions apart from a short period in the 50's and 60's when Greenlandic was somewhat stressed by Danish and Danish civil servants. With Home Rule in 1979 language again became an exclusive resort for the local Greenlandic authorities and substantial support was allocated to Greenlandic language and culture in a wider sense.

The situation is not very different in Canada and Alaska, especially in Canadian Inuktitut which in many respects is as well recognized and formally protected as is Greenlandic since the establishment of Nunavut 1999 and full transfer of political power in language questions to the local authorities in Iqaluit.

We are thus dealing with formally recognized languages in high esteem in their own societies and with access to quite substantial funding ear-marked to projects in and for the local languages.

It ought to be comparatively easy to promote language technology in such a setting but the absence of not only language technology but also all kinds of basic language resources is striking.

And still more striking is the irresolution to get going even when offered the means to do so. During the International Polar Year three years ago a joint project between Nuuk, Tromsø and Odense offering language technology to Alaskan Iñupiaq and Canadian Inuktitut based on adaption of the Greenlandic automaton was unsuccessful mainly because local authorities in Alaska and Nunavut did not support the project. To mention one example.

The bottom line is thus a surprising mismatch between high language status and political attitudes in favour of language technology paired with funding possibilities that are not prohibitive on one hand and the fact that nothing happens on the other.

Why is it so difficult to get air-borne?

There are, of course, many factors playing parts in the total explanatory framework but I would like to address one observation that as far as I know never has been treated in the literature before, namely a skewness in Inuit languages' functions.

Until 1950 Greenland was monolingual with all parts of society carried out in Greenlandic. Mother tongue teaching worked according to the same scheme. The development of the subject progressed much like one would expect it to do in a modern society both didactically and technically.

In 1950 this state of affairs was dramatically altered when Greenland was decolonized. The Danish language attracted enormous status and very little attention was devoted to Greenlandic. As a consequence didactic development including production of teaching material for the mother tongue subject almost ceased.

By the mid 70-s it was a general belief that Greenlandic culture was moribund because of the pressure from Denmark and so was the culture's prime manifestation namely the Greenlandic language. The reaction was a culture revolution which ultimately paved the way for Home Rule with Greenlandic as the formal official language in 1979 and its expansion into Self Government in 2009.

Language questions played a significant role in the political movement in those days but the public debate about Greenlandic was thematically very different from the very vivid debate half a century earlier. It evolved around very general issues. Greenlandic was the prime ethnic, national symbol expressing Greenlandic culture and identity as an Inuk but very little energy was devoted to the instrumental and heuristic function of language as opposed to the debate before the war.

The same language view is found in mother tongue teaching in school and at the teacher training college leaving us with a whole generation of students and teachers with literally no descriptive framework for what is repeatedly stressed as one of the constituting parameters for their culture and personality.

To rephrase it we have a language in which the symbolic, artistic etc. half of the language by definition constitute all of the language leaving several functions of language entirely out of account.

Please observe, that I do not postulate such states of affairs to be inherent in Greenlandic and other minority languages for that matter. On the contrary do we know that Greenlandic philology and L1 didactics developed by native Greenlanders met very high standards before 1953 as long as Greenlandic developed at ease at its own pace. The skewness described above is a phenomenon that showed up after a period of pressure on Greenlandic.

The same phenomenon is evidently at play in a number of revitalized languages in Alaska and

Siberia so I believe we are dealing with a general process rather than a language specific one.
It should be obvious that the result inevitably will be a major conflict between language planning at the political level and corpus planning and modernization at the executive level. In a modern society quite a substantial bit of linguistic skills is needed to transform political decisions into everyday life applications. Love for one's language or artistic fluency in the language are important qualities but they do not compensate for lack of description.
Poor description creates poor teaching materials creates poor teaching creates poor motivation .. The descriptive incapability at all levels of society has given rise to a long row of problems and several nasty vicious spirals so that we have ended up in a deadlock situation with lots and lots of work to do but with no one to do it and much too limited tools to do anything really efficiently.
This is the exact reef where almost all projects with the least affinity to language technology are wrecked in spite of all the positive attitudes: Without precise descriptions of a language or staff with the skills needed to provide such descriptions technological solutions cannot be provided.

Problems to address and barriers to overcome

I think we all agree with the director-general of UNESCO when she states that technology is needed for the safeguarding and promotion of minority languages and linguistic diversity. The million dollar question is what we can allow ourselves to expect from inside the minority languages themselves.
Based on my experience expectations to language technology projects in minority languages must have an altered focus. It is next to impossible to find native speakers of for instance Greenlandic with the necessary ability to describe Greenlandic in terms concise enough for use in language technology projects. The approach to language description is simply entirely different - and, unfortunately, unusable with most language technology projects. And what is almost worse is the fact that I see no readiness among the elders who are the decision makers on indigenous culture and language to encourage alternative approaches to languages.
So to sum up: The first step in a language technology project for the next Inuit language in line for language technology is accordingly NOT to identify and define problems and design a project that will deal efficiently with them. There are steps to take before that.
Establishing a language technology project for an Inuit language first of all depends on the elders' acceptance. In Canada and Alaska such acceptance must be formally obtained before establishing the project whereas it in Greenland is not a formal demand but rather an inevitable prerequisite for funding and access to resources and persons needed for the project since all questions concerning Greenlandic language will be passed on to the language board, Oqaasiliortut. Without Oqaasiliortut's approval projects do not have a chance in practice.
Once formalities are cleared and funding secured the question of locating manpower is next. With Canadian Inuktitut this has not at all been possible up till now, with Alaskan Iñupiaq in a joint project between Alaska Native Language Center and Carnegie Mellon a non-Inuk with a certain command of L2 Iñupiaq was hired to develop an automaton after many years of standstill because it wasn't possible to find the know-how needed for the project in Alaska. And in Greenland we have after years of serious problems with locating and retaining staff chosen to design the language technology program to include a formal education in language technology as a hands-on combination with the master of language technology program in Gothenburg.
Third web of problems arise from the lack of basic resources. Almost all linguistic resources at hand for Inuit languages are either rooted in the old missionaries' attempt to propagate Christianity or in attempts to translate foreign words and foreign concepts to the Inuit languages. Resources are accordingly almost exclusively bilingual or focusing on Inuit languages for L2 purposes whereas language internal resources like L1 grammar books or monolingual dictionaries and corpora are rudimentary or non-existing.
Finally, should an Inuit language technology project somehow overcome all the obstacles mentioned above the risk of drying out for lack of funding or drowning in success are both immediate because the public is incredibly attentive in language questions but expectations to technical solutions uttermost unrealistic.
Google Translate can be mentioned as one example. It is an often mentioned in the language debate in Greenland when critical voices rise from the political level as well as from the mediums. Instead of the - as it appears - unnecessary theoretical and tedious work with fst, CG and the like we could simply adapt Google Translate and other "off-the-shell" solutions.
Keep in mind that such opinions are aired by

persons in economic and political power but with basically no understanding of the language in descriptive terms. In that situation it is very hard to sit in the ivory tower and try to explain that data driven technology is not an option for a polysynthetic language and that we need endless years to pave the endless way via tagging and parsing toward rule-driven technologies.

Now, problems as the ones outlined here are of course all too well known to all of you. Still, in a micro first nation state like Greenland processes like these have immediate and direct impacts because we do not have the buffers of academic professionals in bureaucracy and universities to filter public opinions before they are taken to the political level. In Greenland we either have nothing at all between the public and the Parliament or we have institutions manned with lay people without theoretical schooling as is the case with Greenland's powerful Language Board. As a consequence we need to devote very much energy on "staying alive" that is to legitimize our project by answering scores of official memos and public reports, and by feeding the public with information about our doings with very small intervals.

Why free and ready accessibility is crucial to minority languages

Now, after all this lamentation you most likely have started to wonder what it all has to do with visibility and dissemination of language resources.

Very much, actually, so let me once again return to the opening of this presentation.

The Greenlandic project was established with very limited resources in terms of money, manpower, and know-how under the wings of the Sami project. It would obviously not be where it is now without the long-lasting support from Tromsø. Greenland simply cannot itself provide the many tools needed and cannot maintain a forum strong enough to reinforce professionalism, pick up new trends and tools and secure transmission of skills to next generations.

The Sami project's definition of openness to include not only a download button one has to locate oneself but also a deliberate attempt to document and draw attention to resources paired with a willingness to invest time and energy in outsiders like myself paid off. It took quite some effort to launch the Greenlandic project but it functioned. We are still in business.

And it spreads as could be observed last year on Malta where an Iñupiaq project heavily inspired from the Greenlandic project was presented.

So to conclude this talk: Minority languages need language technology badly but very few have the human and linguistic resources needed to get going and the scantiest of resources namely the people in the projects will inevitably find themselves spending most of their time not on language and technology but on human resource development, bureaucracy, and public promotion just to keep a project alive.

The bottom line then is twofold:

(i) We need help, and lots of help at that. Therefore easy access not only to resources but also to actual programs and tools is not only welcomed but rather the very lifeline for a project like the Greenlandic project. We are much too few and we still need all kinds of resources so our only option is to borrow or steal whatever can be borrowed or stolen and limit local resources to deal with language specific and culture specific problems that under no circumstances can be outsourced.

(ii) It works. The Greenlandic language technology program has proven that it can be done in spite of everything when good forces are pooled consciously and deliberately.

Author Index