META-NORD: Towards Sharing of Language Resources in Nordic and Baltic Countries

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

Andrejs Vasiļjevs
Tilde
Riga, Latvia
andrejs@tilde.lv

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

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

Krister Lindén
University of Helsinki
Helsinki, Finland
krister.linden@helsinki.fi

Eiríkur Rögnvaldsson
University of Iceland
Reykjavik, Iceland
eirikur@hi.is

Abstract

This paper introduces the 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. The goals of the project, overall approach and specific action lines on wordnets, terminology resources and treebanks are described. Moreover, results achieved in first five months of the project, i.e. language whitepapers, metadata specification and IPR management, are presented.

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 been developed using different standards (if any) and in many cases are not well documented.

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

To address these issues the European Commission 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 centred 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 the common META-NET network[2]. One of its main activities is creation of META-SHARE – a sustainable network of online repositories for language data, tools and related web services.

At the core of the META-NET is the T4ME project which is funded under FP7 programme. The Central and Southeast part of 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 linguis-

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

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

Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm, pages 107–114,
Chiang Mai, Thailand, November 12, 2011.

tic infrastructure in the Baltic and Nordic countries.

This paper describes the key objectives and activities of the META-NORD project, presents its first results and discusses cooperation with other similar projects, e.g. CLARIN (Váradi et al., 2008).

It is an integral part of the META-NET and other related initiatives like CLARIN to create a pan-European open linguistic resource exchange platform.

2 The META-NORD Project

The META-NORD project focuses on 8 European languages – Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish, – each with less than 10 million speakers. The 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).

META-NORD contributes to the pan-European digital resource exchange facility by mapping and describing the national language technology landscape, identifying and 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 META-NORD is targeted to three horizontal action lines: treebanks, wordnets and terminology resources.

In addition important collaboration with other EU partners is established within the Initial Training Network in the Marie Curie Actions CLARA[3]. 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 Language Whitepapers

The META-NORD consortium has prepared reports of the language service and language technology industry for all the languages targeted by the project: Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian (Nynorsk and Bokmål) and Swedish.

The reports are written as a series of separate publications for each language, but they are closely coordinated in structure. The reports contain information on general facts of the language (number of speakers, official status, dialects, etc.), particularities of the language, recent developments in the language and language technology support, core application areas of language and speech technology, and the situation in the language with respect to these areas.

For each language, 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 are assessed. Legal provisions related to language resources and tools, which may differ from country to country, are outlined.

The reports also present a detailed table with ratings of language technology tools and resources for each language compiled on the basis of the same framework that is used in the whole META-NET network. Experts were asked to rate the existing tools and resources with respect to seven criteria: quantity, availability, quality, coverage, maturity, sustainability, and adaptability. Results are summarized in Figure 3 and Figure 4 for tools and Figure 2 and Figure 5 for resources.

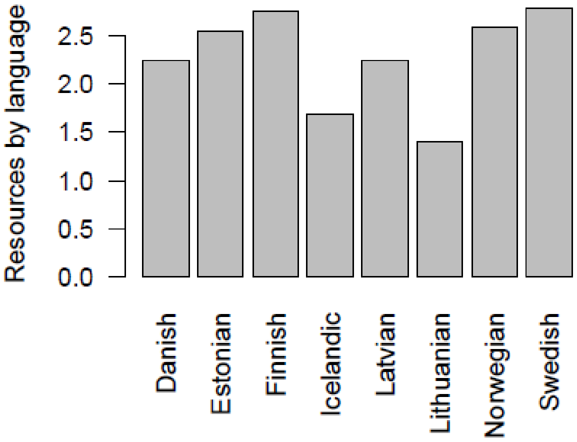

Figure 1. Average scores for resources.

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 status is reasonably positive 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 such as parallel corpora, speech corpora, and grammar, although these tools and resources are rather simple and

[3] http://clara.uib.no/

have limited functionality for some of the languages.

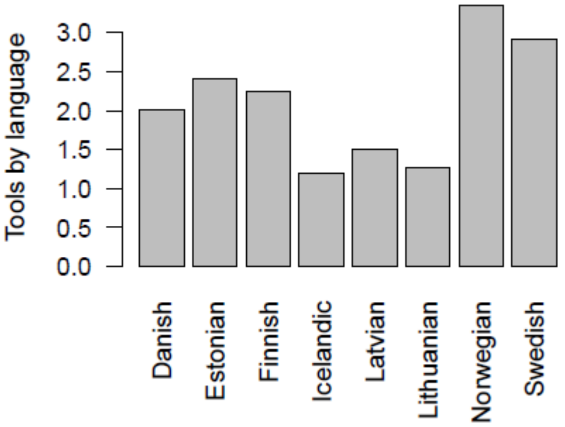

Figure 2. Average scores for tools.

When it comes to more advanced fields such as 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.

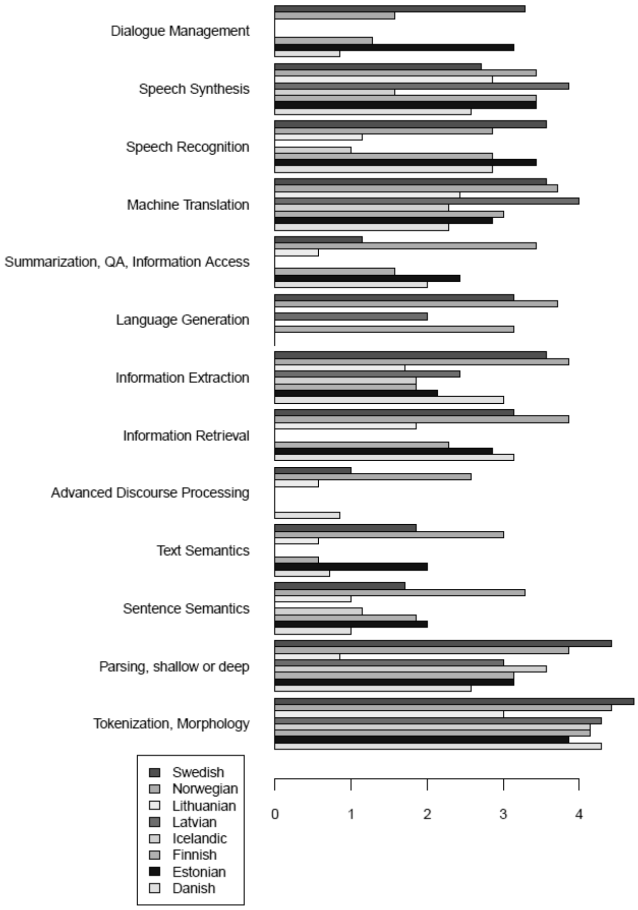

Figure 3. Evaluation results for tools.

For the most advanced tools and resources such as 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.

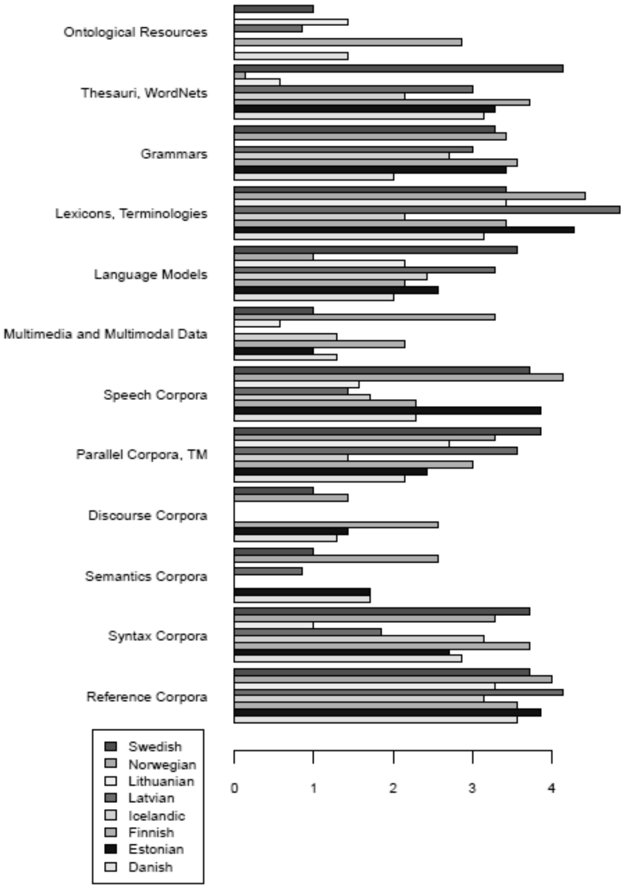

Figure 4. Evaluation results for resources.

The figures for all the languages taken together 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 Horizontal Action on Multilingual 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. In many cases these relations have counterparts in relations among concepts in formal ontologies, so that a straightforward mapping from the one to the other can be established.

According to the BLARK (Basic Language Resource Kit) scheme (Krauwer, 1998), wordnets along with treebanks are central resources when building language enabled applications. The BLARK lists Computer 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. 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.

During the last decades, wordnets have been developed for several languages in the Nordic and Baltic 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 (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 constitutes a prelimary 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.

5 Horizontal Action on Multilingual Terminology

Among specific activities of the 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. The 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.

EuroTermBank platform will be integrated into an open linguistic infrastructure by adapting it to relevant data access and sharing specifications. META-NORD is approaching holders of terminology resources in Nordic countries with the aim of 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 is based on the TBX (TermBase eXchange) standard recently adapted as ISO 30042. It is an open XML-based standard format for terminological data, created by the now dissolved 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.

6 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 information retrieval, parsing and MT systems.

The objective of 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 in Bergen, 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 into 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 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.

A particular goal is to extend the Estonian TreeBank and improve its quality/format/querying interface. The rule based parsing system for Estonian can be used for building Estonian Treebank.

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 consist of approximately one million words (cf. Rögnvaldsson et al., 2011). 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.

7 Management of Intellectual Property Rights

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.

Promoting the use of open data and following the Creative Commons and Open Data Commons principles, 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.

7.1 Open content and open source licenses

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. META-NORD recommends using of CC-licenses for open content resources when the above definition of usage applies.

The **Open source licenses** are specifically designed for software and tools. The only widely translated license is EUPL (European Union Public License) but it is not yet widely used. The most popular license for software programs has lately been the 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 lets the software be combined only with other open source programs, LGPL allows combining the software with proprietary software as well, as long as the open software is distributed with its source. Only an Apache license or similar will also allow distribution of the open software in closed form. Other open source licenses are MsPL and BSD.

7.2 META-SHARE licenses

META-SHARE licenses are 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.

META-SHARE licenses are applicable to 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. The number of potential users is smaller than with CC-licenses. The licenses cover IPR issues in connection with collective works, databases and works of shared authorship.

7.3 CLARIN model agreement templates

CLARIN agreement templates 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 (Oksanen et al., 2010). Without modification, the CLARIN agreement templates do not allow 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.

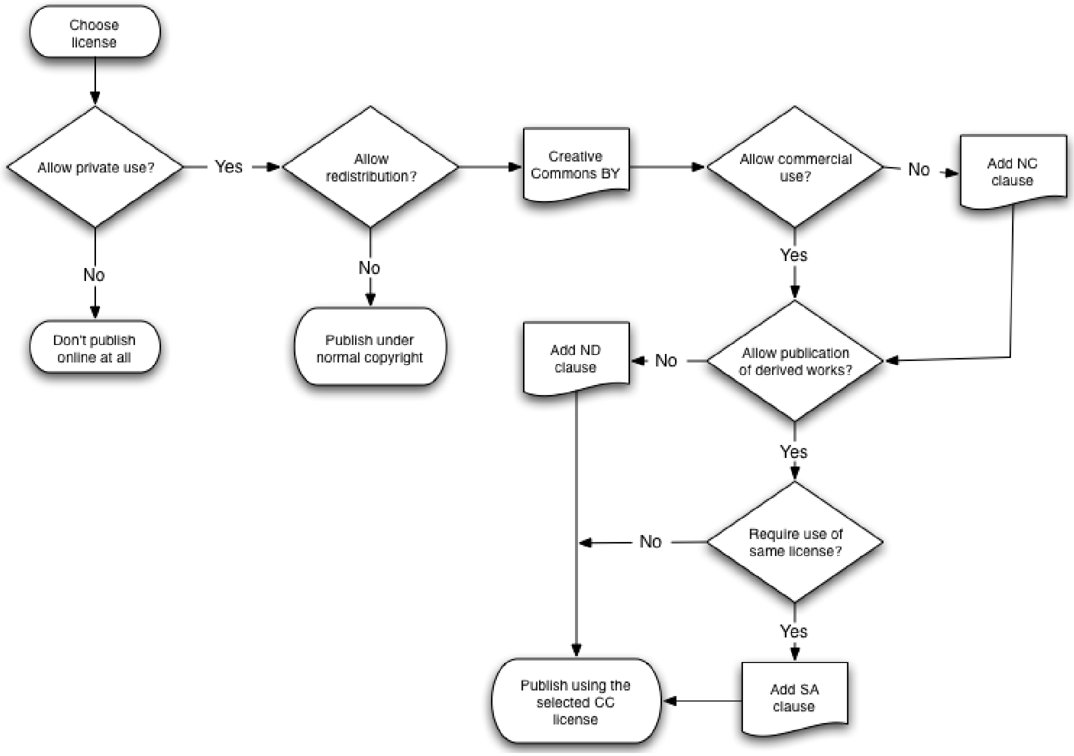
Choose license
Allow private use?
Yes
No
Don't publish online at all
Allow redistribution?
No
Publish under normal copyright
Creative Commons BY
Allow commercial use?
No
Add NC clause
Yes
Allow publication of derived works?
No
Add ND clause
Yes
Require use of same license?
No
Yes
Add SA clause
Publish using the selected CC license

Figure 5. Selection of the appropriate open content license.

The CLARIN model agreements can be modified and are thus applicable to all kinds of purposes. It is, however, advisable to use the existing CC, META-SHARE or CLARIN licenses, if applicable, and modify the CLARIN licenses only for any remaining purpose.

The CLARIN Deliverable D7S-2.1[4] includes two agreements, a deposition agreement and an upgrade agreement. In addition to this, the appendices include other relevant 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

[4] http://www-sk.let.uu.nl/u/D7S-2.1.pdf

agreement (between the content provider and the service provider).

8 Metadata and Content Standards

An important aim of META-NORD is to upgrade and harmonize national language resources and tools in order to make them interoperable, within languages and across languages, with respect to their data formats and as far as possible also as regards their content.

Since resources and to some extent tools normally will remain in one location – one of a number of META-NORD centers – the preferred way of accessing and utilizing resources and tools will be through *metadata* and *APIs*, allowing the assembly of on-the-fly toolchains made up of standardized component language technology tools, processing distributed – and in many cases interlinked – language resources in standardized formats.

8.1 Metadata standards

META-NORD is working on standardized top-level resource descriptions (metadata) for all relevant types of resources, based on a recommended set of metadata descriptors for documenting resources provided by META-NET through META-SHARE. It will produce such descriptions for each and every resource contributed to the shared pool. Metadata sets include mandatory as well as optional elements, together with sets of recommended values whenever possible and appropriate. According to the META-SHARE model[5], metadata must include at least a specified minimum of information in each of the following categories: *identification* (including a persistent identifier); *resource type*; *licensing/distribution*; *validation*; *metadata provenance*; *funding*; *contact information*. The model then allows for extensive further elaboration of each information category, so that metadata records for resources and tools can be arbitrarily informative.

The inspiration for the META-SHARE metadata model comes largely from the CLARIN Metadata Initiative (renamed to *Component Metadata Initiative* (CMDI[6])), which can be seen as building on top of earlier relevant initiatives – e.g., DC and OLAC – and which now aims to become an ISO standard. The data categories, e.g., ISOcat, are the main concern of standardization, not the metadata schema per se.

In most cases, the resources and tools to be made available in META-NORD do not come equipped with the required metadata information, let alone encoded as formal metadata. The main exceptions are corpora in TEI or XCES format which often have header elements containing at least some of this information, which can be automatically extracted. Some partners are already publishing structured metadata records for at least some of their resources, e.g., the Language Bank of Finland is publishing OLAC – and the obligatory DC – through OAI-PMH for a number of corpora already. In case existing resources are described using popular metadata sets – OLAC being a case in point – the consortium will upgrade them using converters, mappers and other tools provided by the META-NET, or in some cases developed by the META-NORD.

8.2 Content standards

We can foresee that users will want access to the META-NORD language resources in at least the following three ways:

(1) *In toto*, i.e., the resource can be downloaded. This requires that the resource is in a standardized, well-documented format, or it won't be very useful to our target groups. It also requires that all IPR issues have been cleared and licensing terms stated (see section 7 above).

(2) Online browsing either in a standard web browser or through a dedicated tool. Here, standardized metadata must provide sufficient information for a user to find the URL providing the application. However, the base resource may be in a proprietary format (although any export facility should provide a standardized format).

(3) In the form of a web service or other API. Here, standardized metadata are needed. Further, any data returned by a web service should be in a standard format.

Consequently, metadata and resource formats in META-NORD should support at least these three resource usage scenarios.

META-NORD greatly benefits from the work conducted in CLARIN for best practices and guidelines with respect to formats for language resources, language tools and metadata.

From information provided by partners, it is clear that the META-NORD resources and tools come in many formats. Some resources are in

[5] http://www.meta-net.eu/public_documents/t4me/META-NET-D7.2.1-Final.pdf

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

RDB formats (SQL, Access), some in proprietary formats, etc. For interoperability, such resources should probably be converted into other formats. Data format conversion is generally not a problem, and should be implemented in many cases, since partners may have invested heavily in such formats and in such cases we should simply consider a solution whereby conversion is made on demand into an interoperable export format. The only problem with this solution is that it will add complexity, since any change made to the original format must be accompanied by the corresponding change in the conversion utility.

A point of greater concern is that, according to the provided information, many of the resources and tools lack an explicit and formal content model. This issue will need to be addressed in META-NORD.

META-NORD will put considerable effort into making content models of resources and tools as interoperable as possible. This can imply adopting more strictly structured formats, such as LMF rather than proprietary XML or SQL for lexical resources. Regardless of this, it will almost certainly imply a mapping to a set of standardized data categories, such as that of ISOcat. This can mean a considerable amount of work and careful consideration is needed in order not to waste effort. On the other hand, the rewards of the interoperability achieved in this way are potentially great.

For new resources and tools or for those where conversion of the base resource is desirable, the following formats are recommended:

- corpora: TEI or (X)CES format (standoff annotation in ISO formats will be allowed);
- lexical resources: LMF or Princeton WordNet format;
- terminology resources: TBX;
- tools: at least as web services (if possible), described using WSDL.

9 Conclusions

Language whitepapers prepared by the META-NORD project show that the Nordic and Baltic countries still have a long way to go to implement the vision of making the area a leading region in language technology. META-NORD project lays the ground for a fruitful cooperation in identifying, enhancing and sharing of language tools and resources created in the Nordic and Baltic countries, which will considerably strengthen the field in a near future.

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.

Krauwer, S. 1998. *ELSNET and ELRA: A common past and a common future.* The ELRA Newsletter, Vol. 3, n. 2, Paris.

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.

Rögnvaldsson, E., A. K. Ingason and E. F. Sigurðsson. 2011. Coping with Variation in the Icelandic Diachronic Treebank. In Johannessen, J. B. (ed.): *Language Variation Infrastructure. Papers on selected projects*, pp. 97-111. Oslo Studies in Language 3.2. University of Oslo, Oslo.

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.