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.