META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Project no. 270899

http://www.meta-nord.eu/

Annual Public Report

Version No. 1.0
15/11/2011

Contents

META-NORD is a 24 month project with aims to establish an open linguistic infrastructure in the Baltic and Nordic countries to serve the needs of the industry and research communities. The project focuses on 8 European languages - Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish - that each has less than 10 million speakers. The project goal is to assemble, link across languages, and make widely available language resources of different types used by different categories of user communities in academia and industry to create products and applications that facilitate linguistic diversity in the EU.

PROJECT OBJECTIVES

The **main objectives** of the META-NORD project are:

- **to provide a description of the national (resp. language community) landscape** in terms of language use; language-savvy products and services, language technologies and resources; main actors (research, industry, government and society); public policies and programmes; prevailing standards and practices; current level of development, main drivers and roadblocks; in a simple, clear, standardised format;
- **to contribute to a pan-European digital resource exchange facility** by **identifying**, **collecting resources in the Baltic and Nordic countries** and by documenting, processing, linking and upgrading them to agreed standards and guidelines;
- **to collaborate with** other partner projects, in particular **concurrent 6.1 pilot projects and the META-NET Network of Excellence**. Cooperation with other relevant multi-national forums or activities, e.g., FlaReNet, CLARIN, will ensure consistent approach, practices and standards aimed at ensuring a wider accessibility of quality language resources, easier access to them and reuse of the same;
- **to help in building and operating** of broad, non-commercial, community-driven, inter-connected **repositories**, exchanges, and facilities that will be used by different categories of target user communities;
- **to mobilise national and regional actors**, public bodies and funding agencies by raising awareness, organizing meetings as well as other focused events.

Besides the general objectives, META-NORD has set several **specific targets**:

- **to provide expertise** to other 6.1 pilots in fields where the META-NORD partners have outstanding expertise: **treebanks/syntax databases, terminology resources, wordnets**, and **finite-state techniques;**
- **to develop and to document methodologies** for building language resources for the so-called under-resourced languages (i.e., languages with limited language resources) as efficiently as possible, with focus on **semi-automatic/machine-assisted resource generation**;
- **to facilitate the availability of BLARK resources** for the META-NORD languages (Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian, and Swedish);
- to facilitate **knowledge transfer** between CLARIN and META-NORD, especially on standards and intellectual property rights (IPR) issues.

DESCRIPTION OF THE NATIONAL (RESP. LANGUAGE COMMUNITY) LANDSCAPE

The META-NORD partners have compiled 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 Baltic and Nordic geographical area: Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian (Nynorsk and Bokmål), and Swedish. The state of affairs in language technology in the Baltic and Nordic countries is described in the eight reports. Each report describes the situation of a language community and the position of the language service and language technology industry for that particular language. The

reports are written as a series of separate publications for each language, but they are closely coordinated in structure. 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 has been assessed. Legal provisions related to language resources and tools (LRT's), which may differ from country to country, have been outlined. The language reports are available on: http://www.meta-nord.eu/index.php?p=public-deliverables.

COLLECTING RESOURCES IN THE BALTIC AND NORDIC COUNTRIES

The aim of this work is to upgrade and to harmonise national language resources within and across the META-NORD languages in order to make them interoperable with regard to their data formats and content. The evaluation has been carried out using the criteria suggested by META-NET[1] Network of Excellence and META-SHARE[2] project (actual availability, suitability for technology and product development, fitness for multilingual purposes, quality, and potential for re-use, recombination and repurposing).

The analysis of the situation indicates that the criteria of availability and suitability are the most significant, although all the criteria are important.

Issues with extensibility (need for documentation and meta-data description) are generally the easiest to address. In most cases, the LRT's already have documentation and they lack only meta-data descriptions, or, it is easy to add documentation.

The multilinguality criterion is sometimes difficult to meet since some LRT's are of monolingual nature (corpora, specific dictionary/grammar based tools), but these LRT's may be important bases for other tools and resources.

Longevity (active maintenance over longer periods of time) is a preferred feature but in many cases an old LRT, which is freely available, could replace the similar LRT with restricted access.

Altogether, 151 LRT's have been evaluated. 71 of them are available to the consortium, 67 are potentially available and 8 have restricted access. 109 LRT's are well suitable for language technology development, 47 are multilingual, 103 are well maintained, 33 LRT's are of very high quality and 97 of high quality, 93 LRT's have a high-grade documentation and a meta-data schema.

METADATA, UPGRADE TO AGREED STANDARDS AND HARMONIZATION

During the first months of the project, we have identified and collected an initial pool of language resources including the most important resources for each language, which is up to 100 on the whole. More than a half of these are made available by the consortium, which facilitates negotiations regarding the availability of such resources.

META-NORD language resources come in many formats, and partners put a considerable effort into making content models as interoperable as possible. This can imply adopting more strictly structured formats, e.g., LMF rather than proprietary XML or SQL for lexical resources. In any case, it almost certainly implies mapping to a set of standardized data categories, e.g., ISOcat.

For example, the lexical databases – Danish STO and Swedish SBLEX – are being upgraded to LMF, an XML ISO standard (ISO 24613, 2008). STO consists of morphological, syntactic, and semantic levels. Currently, we are upgrading the morphological level by converting from the original intensional description to the extensional morphology description. It does not explicitly list word forms, but instead, the lexical entry is

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

associated with a morphological pattern. Most of the structural problems have been solved now, and the specification of data categories is under development.

SBLEX is a collection of free Swedish lexical resources integrated within the Swedish Framenet++ project, and is being upgraded to LMF in META-NORD. Currently, it consists of 15 lexical resources describing both modern and historical Swedish, with more resources on the way. The pivot resource of SBLEX, to which all other resources are linked, is SALDO — a large, freely available lexicon with morphological and semantic information. All resources in SBLEX are now in LMF, but there is still some room for improvement in lexical information representation.

Multilingual actions in META-NORD

As previously mentioned, META-NORD is specifically focused on three horizontal action lines: treebanks, wordnets, and terminology resources.

Linking and validating the Baltic and Nordic wordnets

The multilingual task on wordnets is concerned with the validation and pilot linking between the Baltic and Nordic wordnets. The aim is to test the perspective of multilingual linking of the Baltic and Nordic wordnets and, through such pilot linking, tentatively compare and validate the wordnets along the measures of taxonomical structure, coverage, granularity, and completeness. The linking is performed via the so-called Princeton "core wordnet", a subset of the Princeton wordnet containing 5,000 most basic synsets to which all the Baltic and Nordic wordnets have been linked during the first phase of META-NORD. These links will be provided in the first batch opening for a subsequent validation in the next phase.

Horizontal action on treebanks

The horizontal task on parallel treebanks aims to deliver a pilot parallel treebank for a limited number of languages, accessible through a uniform Web interface using state-of-the-art search tools. After clearing rights, the material will be sentence-aligned, syntactically annotated, quality-assured, and made available in a uniform Web interface. In a subsequent phase, phrase alignment will be attempted between at least Danish and Norwegian. In a final phase, linking can be extended to dependency treebanks, e.g., the Finnish treebank, using the technology from FIN-CLARIN. Combining these technologies, a pilot parallel treebank is planned for Norwegian, Danish, Finnish, and English.

Multilingual terminology

META-NORD also addresses a growing demand to consolidate distributed terminology resources across languages and domains by extending the open linguistic infrastructure with multilingual terminology resources. Some META-NORD partners have already established a solid terminology consolidation platform EuroTermBank, which provides a single access point to more than 2 million terms in 27 languages. However, terminology coverage for some languages in EuroTermBank (Latvian, Lithuanian, Polish, and other) surpasses that of other languages, whose terminology resources have not been comprehensively integrated into the EuroTermBank platform. Therefore, META-NORD is approaching the holders of terminology resources in the respective countries that facilitate sharing of the resources through cross-linking and federation approach. The EuroTermBank platform is being integrated into the META-NORD infrastructure by adapting it to the relevant data access and sharing models. META-NORD is also elaborating a mechanism for consolidated multilingual representation of monolingual and bilingual terminology entries. Sharing of terminological data is based on TBX standard (ISO 30042, 2008).

COOPERATION

META-NORD closely cooperates and coordinates its activities with other ICT-PSP Objective 6.1. projects (CESAR and METANET4U) and META-NET Network of Excellence, which is dedicated to fostering the technological foundations of a multilingual European information society to ensure a coherent pan-European open linguistic infrastructure.

DISSEMINATION

Websites

META-NORD project website www.meta-nord.eu is the main project communication tool. It contains various materials that reflect the project's aims, progress and impact. This is the place where all information related to META-NORD is stored and made accessible to the Internet sharing community. META-NORD website provides access to localized META-NORD project partners' websites.

Articles

- Andrejs Vasiļjevs, Tatiana Gornostay, Inguna Skadiņa, Daiga Deksne, Raivis Skadiņš, and Mārcis Pinnis. *Recent advances in the development and sharing of language resources and tools for Latvian*. In "Multilingual Processing in Eastern and Southern EU Languages – Less-resourced Technologies and Translation", Cambridge Scholars Publishing, 2011 (in press) (TILDE).
- Anje Müller Gjesdal, Gyri Losnegård. *Språkteknologi i* bakevja. Norway, 2011 (UiB).
- Eiríkur Rögnvaldsson. Article on META-NORD in Hugrás. The online journal of the School of Humanities at the University of Iceland, 2011 (HI).
- Gunn Inger Lyse. *Dårlig språk er dårlig butik* (Poor language means poor business). Feature article in the Paper *Dagens næringsliv*, Norway, 2011 (UIB).
- Krister Lindén. *Combining Statistical Models for POS Tagging using Finite-State Calculus.* Silfverberg, M. & Linden, K. *Combining Statistical Models for POS Tagging using Finite-State Calculus*, p. 183–190. Latvia (UHEL).
- Krister Lindén 2011: *Framework for Compiling and Applying Morphologies.* Linden, K., Silfverberg, M., Axelson, E., Hardwick, S. & Pirinen, T. HFST—Framework for Compiling and Applying Morphologies. In *Systems and Frameworks for Computational Morphology*. Mahlow, C. & Pietrowski, M. (eds.). Vol. 100. Springer p. 67-85. (Communications in Computer and Information Science), Switzerland 2011.08.26 (UHEL).
- Krister Lindén, Atro Voutilainen 2011: *Designing a Dependency Representation and Grammar Definition Corpus for Finnish*. In *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora*. Actas del III Congreso Internacional de Lingüística de Corpus. Valencia, Spain. Spain 2011.05.08 (UHEL).
- Krister Lindén, Ville Oksanen 2011: *Open Content Licenses: How to choose the right oneOpen Content Licenses: How to choose the right one.* Estonia 7 p. 2011.05.11 (UHEL).
- Laura Blėdaitė, Martynas Pumputis 2011: *Project META-NET in Europe and Lithuania.* Journal *Gimtoji kalba* No. 9 (531) (LKI).
- Måns Hulden 2011: *Constraint Grammar parsing with left and right sequential finite transducers - A4.* Article in conference publication (refereed) *Language Finnish Publication year* 2011, Finland (UHEL).
- Miikka Silfverberg, Mirka Hyvärinen, Tommi Pirinen 2011: *Improving Predictive Entry of Finnish Text Messages using IRC Logs.* Logs in Proceedings of the Computational Linguistics-Applications Conference 2011 Publication: *Conference contribution › A4* Article in conference publication (refereed), Poland (UHEL).

Conference papers

- Atro Voutilainen, Tanja Katariina Purtonen. *A Double-Blind Experiment On Interannotator Agreement: The Case Of Dependency Syntax And Finnish Publication*. In NEALT Proceedings Series Vol. 11, Latvia (UHEL).
- Atro Voutilainen. *Creating a research resource and service for language researchers with Constraint Grammar*. In Proc. Workshop on Constraint Grammar at NODALIDA 2011, Riga, Latvia (UHEL).
- Krister Lindén and Atro Voutilainen. *Finnish Language Bank: A Framework for Depositing and Disseminating Language Resources for R&D.* In Proc. Workshop on visibility and availability of LT resources at NODALIDA 2011, Riga, Latvia, 2011 (UHEL).
- Krister Lindén and Atro Voutilainen. *Specifying a linguistic representation with a grammar definition corpus*. In Proc. Corpus Linguistics 2011, Birmingham, United Kingdom, 2011 (UHEL).
- Krister Lindén. Do wordnets also improve human performance on NLP tasks? Muhonen, K. & Linden, K. Estonia 12.05.2011 7 p. (UHEL).
- Andrejs Vasilejevs, Bolette Sandford Pedersen, Koenraad De Smedt, Lars Borin, Inguna Skadiņa. *META-NORD: Baltic and Nordic Branch of the European Open Linguistic Infrastructure*. In Proc. of the NODALIDA 2011 workshop on Visibility and Availability of LT Resources, Riga, Latvia, 2011 (UIB).
- Andrejs Vasiļjevs, Tatiana Gornostay, and Inguna Skadiņa. *From Terminology Database to Platform for Terminology Services*. In Proc. of the NODALIDA 2011 Workshop on Creation, Harmonization and Application of Terminology resources (CHAT 2011), Riga, Latvia, 2011 (TILDE).
- Inguna Skadina, Andrejs Vasiljevs, Lars Borin, Koenraad De Smedt, Krister Linden, and Eirıkur Rognvaldsson. *META-NORD: Towards Sharing of Language Resources in Nordic and Baltic Countries* at IJCNLP 2011, Chiang Mai, Thailand, 2011 (TILDE).

Media

- Eiríkur Rögnvaldsson. *Eiríkur Rögnvaldsson interview on the radio about META-NET and language technology in general Iceland*, 2011 (HI).
- Eiríkur Rögnvaldsson. *Eiríkur Rögnvaldsson interview on the* Icelandic National Broadcasting Service on the Language Whitepapers, 2011 (HI).

Presentations

- Andrejs Vasiļjevs. META-NORD: a brief presentation to the State Language Commission of Latvia and workgroup of the National corpus of Latvian, Latvia, 2010 (TILDE).
- Atro Voutilainen. Building a dependency treebank and other LRs for Finnish Event organized by the Department of Computer Science, University of the Basque Country, San Sebastian, Spain, 2011 (UHEL).
- Atro Voutilainen. EngCG2. Introduction and demo Event organized by the Department of English, Uppsala University, Sweden, 2011 (UHEL).
- Atro Voutilainen. FinnTreeBank. Developing a dependency syntactic treebank and parsebank for Finnish Event organized by the Department of Linguistics and Philology, Uppsala University, Sweden, 2011 (UHEL).
- Atro Voutilainen. Helsinki Corpus Festival 2011. Dept. Modern Languages, University of Helsinki City, Helsinki, Finland, 2011 (UHEL).
- Atro Voutilainen. Treebanking Finnish Event organized by the Institut für Computerlinguistik, Universität Zürich, Switzerland, 2011 (UHEL).
- Aurelija Tamulionienė. Project META-NORD and META-NET. Report about the project META-NORD and META-NET to research community in Lithuania, 2011 (LKI).
- Aurelija Tamulionienė. Project META-NORD and META-NET. Report about the project META-NORD and META-NET to the society and government, Lithuania, 2011 (LKI).
- Aurelija Tamulionienė. Project META-NORD and META-NET. Report about the project META-NORD and META-NET to Media, Lithuania, 2011 (LKI).

- Daiva Vaišnienė. META-NORD and META-NET. Parliament of the Republic of Lithuania, Lithuania, 2011 (LKI).
- Eiríkur Rögnvaldsson. META-NORD presentation at annual Humanities Conference. the Institute of Humanities at the University of Iceland, 2011 (HI).
- Inguna Skadiņa and Aivars Bērziņš. META-NORD brief presentation. CLARA Career Course on Product Planning for Next Generation Information Access Technology Solutions, Croatia, 2011 (TILDE).
- Inguna Skadiņa and Tatiana Gornostay. META-NORD brief presentation to language workers (translators, editors and terminologist) at Tilde Localization department meeting, Riga, Latvia, 2011 (TILDE).
- Andrejs Vasiļjevs. META-NORD brief presentation to professionals in LT from the Institute of Mathematics and Computer Science, the University of Latvia, internal CLARIN seminar, Riga Latvia, 2011 (TILDE).
- Jolanta Zabarskaitė, Nerijus Butvilas, and Aurelija Tamulionienė. META-NORD and META-NET presentation at Parliament of the Republic of Lithuania at meeting in Committee of the Development of Information Society, Lithuania, 2011 (LKI).
- Jolanta Zabarskaitė. META-NORD and META-NET. The Ministry of Transport and Communications of the Republic of Lithuania, Lithuania 2011.10.28 (LKI)
- Krister Lindén. Forskarnas röst och digitalt material Krister Linden – Speaker: Presenter Arranger National Archive City, Helsinki, Finland, 2011 (UHEL).
- Krister Lindén. Language Bank Organization and Activity Duration of event, National Library of Norway, Mo i Rana, Norway, 2011 (UHEL).
- Kristín M. Jóhannsdóttir. META-NET presentation at a conference arranged by the Vigdís Finnbogadóttir, Institute of Foreign Languages on the European Day of Languages, Iceland, 2011 (HI).
- Raivis Skadiņš. META-NORD: Baltic and Nordic Branch of the European Open Linguistic Infrastructure. Seminar "The Latvian language in the digital environment", University of Latvia, Riga, Latvia, 2011 (TILDE).
- Andrejs Vasiļjevs. META-NORD: Baltic and Nordic Parts of the European Open Linguistic Infrastructure. CHAT 2011 workshop on Creation, Harmonization and Application of Terminology Resources co-located with the NODALIDA 2011 conference, University of Latvia, Riga, Latvia, 2011 (TILDE).
- Mietta Lennes. META-NORD presentation at New Tools and Methods for Very Large-Scale Phonetics: research workshop, University of Pennsylvania, Philadelphia, USA, 2011 (UHEL).
- Mietta Lennes. Praat-training for speech therapists., Åbo Akademi,Turku, Finland, 2011 (UHEL).
- Mietta Lennes. STELARIS workshop – Software to Empower Learning and Research in Speech, University of Pennsylvania, Philadelphia, USA, 2011 (UHEL).
- Inguna Skadiņa. META-NORD: Towards Sharing of Language Resources in Nordic and Baltic Countries: presentation at Workshop on Language Resources, Technology and Services in the Sharing Paradigm co-organised by FLaReNet, Language Grid and META-SHARE in conjunction with IJCNLP 2011, Thailand 2011 (TILDE).

Publications

- Aurelija Tamulionienė and Jolanta Zabarskaitė. European Language Day press release (in Lithuanian), 2011 (LKI).
- Aurelija Tamulionienė. Newsletter on META-NORD in Lithuania, (LKI)
- Hanna Westerlund, Pinja Pennala, Mietta Lennes, and Imre Bartis. European Language Day press release (in Finnish), 2011 (UHEL).
- META-NORD brief description of the project on the partner's website (TILDE), including the link to the project website (TILDE).
- European Language Day press release (in Latvian), 2011 (TILDE).
- Press Release on METAFORUM and white papers, Denmark, 2011 (UCPH).
- Press Release on METAFORUM and white papers, Lithuania, 2011 (LKI).

META-NORD CONSORTIUM AND CONTACT PERSONS

Tilde SIA
Vienibas gatve 75a
Riga, LV1004
Latvia
Project Coordinator:
Andrejs Vasiļjevs
e-mail: andrejs[at]tilde.lv
URL: http://www.tilde.eu

KØBENHAVNS UNIVERSITET
Njalsgade 80, DK-2300
Copenhagen S
Contact person:
Bolette Sandford Pedersen
e-mail: bspedersen[at]hum.ku.dk
URL: http://www.humanities.ku.dk/

University of Tartu
Ülikooli 18, 50090 Tartu
Estonia
Contact person:
Kadri Vider
e-mail: kadri.vider[at]ut.ee
URL: http://www.ut.ee/en

University of Bergen
P.O.Box 7800
5020 Bergen
Norway
Contact person:
Koenraad De Smedt
e-mail: desmedt[at]uib.no
URL: http://www.uib.no/en

University of Iceland
Sæmundargötu 2
101 Reykjavík
Iceland
Contact person:
Eiríkur Rögnvaldsson
e-mail: eirikur[at]hi.is
URL: http://www.english.hi.is/

University of Tartu
P. Vileišio st. 5, LT-10308 Vilnius
Lithuania
Contact person:
Jolanta Zabarskaitė
e-mail: jolanta.zabarskaite[at]lki.lt
URL: http://www.lki.lt/LKI_EN/

University of Gothenburg
PO Box 100, SE-405 30 Gothenburg, Sweden
Contact person:
Lars Borin
e-mail: lars.borin[at]svenska.gu.se
URL: http://www.gu.se/english