META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Project no. 270899

Deliverable D3.2

Second batch of resources complying with the project's technical, linguistic, legal, etc.

Version No. 1.0

31/07/2012

Document Information

Deliverable number:	D3.2
Deliverable title:	Second batch of resources complying with the project's technical, linguistic, legal, etc.
Due date of deliverable:	31/07/2012
Actual submission date of deliverable:	31/07/2012
Main Author(s):	Dorte Haltrup Hansen, Lene Offersgaard, Bolette Pedersen, project partners
Participants:	All
Internal reviewer:	Tilde
Workpackage:	WP3
Workpackage title:	Enhancing language resources
Workpackage leader:	UCPH
Dissemination Level:	PU
Version:	V1.0
Keywords:	Documentation, resources

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
0.1	2012-05-27	Fishbone	Dorte Haltrup Hansen, Lene Offersgaard, Bolette Sandford Pedersen (UCPH)	Initial text of deliverable	Submitted for review
0.8	2012-06-27	Pre-final version	Dorte Haltrup Hansen, Lene Offersgaard, Bolette Sandford Pedersen (UCPH)	Corrections after internal review	Submitted for final review
1.0	2012-07-31	Final	Tilde	Final review	Submitted to PO

EXECUTIVE SUMMARY

This report documents the resources delivered in META-NORD, second batch on M18. Each partner gives an overview of the motivation of the selected resources, and general tables of Batch one and Batch two are provided. Further, the plans for the horizontal tasks are given for the rest of the project including plans for collaboration and integration with the other META-NET projects. Finally, in Appendix A, the full set of resources and tools that have been uploaded in META-SHARE are documented in terms of metadata that have been automatically generated from the META-SHARE node at Tilde.

All in all, 207 resources are uploaded by the META-NORD partners in the second batch, including 55 lexical resources, 111 corpora, 11 treebanks, 12 speech resources, 5 monolingual wordnets, four pilot bilingual wordnets, and 9 tools.

Table of Contents

1. Motivation for selected resources and horizontal actions

1.1. Overview

The purpose of this report is to document the resources delivered in second batch, D3.2 in terms of a more general overview than what can be deduced from the metadata in Appendix A.

D3.2 is the second deliverable of WP3 which has the purpose of upgrading and harmonizing national language resources within and across META-NORD languages, in order to make them interoperable w.r.t. their data formats and content. Continued work of documenting, processing, linking, and upgrading META-NORD resources to agreed standards and guidelines has been performed and can be found in this deliverable.

Apart from the monolingual resources, three multilingual actions are undergoing in WP3, namely:

- Horizontal action on treebanking,
- Horizontal action on terminology, and
- Horizontal action on wordnets

Thus, several of these cross-lingual initiatives provide deliverables in batch 2; for instance extracts from 'Sofies verden' have been syntactically annotated and sentence-aligned across languages. Aligned language pairs are: dan-fin, dan-isl, dan-est, dan-swe, dan-deu, dan-nob, deu-nob, deu-isl, deu-swe, deu-est, deu-fin, est-isl, est-nob, est-swe, fin-swe, fin-isl, fin-nob, isl-nob, isl-swe, nob-swe. Furthermore, four pilot cross-lingual wordnets that have been linked via Princeton WordNet Core. The linked language pairs are dan-swe, fin-dan, fin-swe, and fin-est.

1.1. TILDE

The resources uploaded by Tilde have been selected according to the criteria specified in the Task 2.3 (availability, suitability for product/application development, fitness for multilingual purposes, popularity, quality and extensibility). The main focus is on resources that facilitate aim of the META-NORD project to build pan-European digital resource exchange facility and populate it with language resources usable for practical applications and research.

As a result the Latvian and multilingual resources described and made available by Tilde include (a) resources developed or licensed by Tilde (e.g. multilingual electronic dictionaries); (b) resources from EU projects where Tilde is the coordinator or partner responsible for development of particular resource (e.g. outcomes of ACCURAT and EASTIN-CL projects) and (c) public and copyright free resources processed and standardised by Tilde (e.g. *Legislation corpus of Republic of Latvia*).

Tilde also leads *the horizontal action on terminology*. Multilingual terminology is indispensible resource for development of multilingual language technologies, including adaptation of translation tools for particular domain; for professional translators in their every day work as well as for every member of society (e.g. teachers, researchers) where it concerns cross-lingual communication. To provide widely usable terminology resources in the META-SHARE Tilde works on a concept of integration of terminology specific node with the META-SHARE platform through the mutually implemented access interfaces.

1.2. UCPH

The resources uploaded by UCPH have been selected on the basis of three main criteria: i) they should have a certain level of maturity and quality, and ii) they should contain interesting perspectives wrt. up-grade to agreed standards, extension and/or multilingual linking and validation within the META-NORD project time span and efforts, and last but not least iii) they should be clarified wrt IPR related issues to an extent where it was actually feasible to extend and improve them. As a consequence of this mode of priority, Danish language resources contain mainly resources provided by UCPH (DanNet, STO, linked wordnets together with UHEL, UGOT and UT), supplemented by a few from Copenhagen Business School (CBS) (treebanks). Most of the resources are monolingual and focus on Danish written language. Some of the resources have links to other languages, such as English; this counts for the Danish wordnet, DanNet and The Copenhagen Danish-English Dependency Treebank. Since UCPH does not deal with speech processing, and since this field of research is highly commercialized, speech resources have not been included in META-NORD, at least not at the current stage. Finally, it should be noted that since UCPH is the national group with least manpower in META-NORD, we have chosen to select a rather limited set of resources and tools, but on the other hand to focus on the actual improvement of these in order to achieve a level of quality where they are actually useful in practical LT.

UCPH also leads the *horizontal action on wordnets*. The motivation for this choice is given by the fact that wordnets have emerged as one of the basic standard lexical resources in the LT field. They encode fundamental semantic relations among words, relations that further in many cases have counterparts in relations among concepts in formal ontologies. According to the BLARK (Basic Language Resource Kit) scheme, wordnets along with treebanks, are central resources when building language enabled applications. The semantic proximity metrics among words and concepts defined by a wordnet are very useful in applications such as information access systems and authoring tools 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. Another central motivation for selecting wordnets as a horizontal action, is that wordnets have been or are being built for several languages in the Nordic and Baltic countries including Finnish, Danish, Estonian, Icelandic and Swedish. Furthermore, there is the general need for validation schemes for semantic resources, also across languages. Thus, in this action, we validate the wordnets along the lines of *taxonomical structure*, *coverage*, *granularity*, and *completeness* and thereby contribute to the wordnet community with valuable cross-lingual experience.

1.3. UT

The resources selected for upload by University of Tartu consist of those language resources by the main actors in Estonian language technology (University of Tartu, Institute of Estonian Language and the Institute of Cybernetics in Tallinn Technical University), which are considered to have a level of quality and maturity to be shared and for which the IPR related issues can be settled to the level to make them available through the META-SHARE. Once the resources of those main actors are included in META-SHARE, other resource providers will find it easier to share theirs once those resources have reached the level of quality and have been checked against standards.

In the first upload batch we uploaded the corpora and textual resources owned by University of Tartu, to share in first order most of the mature corpora for Estonian and other resources for which the documentation, status and legal aspects were clearer. For the other META-

SHARE uploads we update the metadata on those resources that are still under development as well as correct and enrich the information in metadata.

Two tools (morphological analyser by Institute of Estonian Language (IEL) and syntactic parser by UT) are included in this second batch. The other known mature LT tools for Estonian - speech synthesizers and a morphological analyser by the Filosoft company are to be included in the last upload when the licenses have been cleared.

In the second batch we upload the metadata for corpora and dictionaries by one of the other two largest actors in LR research in Estonia – the Institute of Estonian Language (IEL). This includes a text corpus and a spoken corpus, bilingual and monolingual dictionaries. Since these are written resources and the speech corpus consists of sentences read from paper, the access terms are clearer than for the other speech resources, which are more complex either due to their content or previous legal agreements. Those, as well as some other resources that we have predicted might need longer negotiations about rights clearance and upload terms, have been left for upload in the final batch.

1.4. UIB

General motivations driving the selection of resources in **Norway** are the following:

- Selecting a broad range of different resources for different languages and language pairs and suitable for different purposes, such as speech databases, lexical databases, wordnets, multilingual corpora and treebanks, termbases etc.
- Selecting resources which are not sufficiently visible and not yet available via other main channels such as ELRA and LDC.
- Selecting resources which need work on clearing licenses, clarifying conditions for use, improving metadata and documentation, or conversion to standard formats.
- Prioritizing resources which are maximally open and freely available without cost.
- Selecting resources owned and distributed by third party collaborators, thus raising awareness outside of the consortium about resource exchange through META-SHARE.

For the second upload UiB has provided metadata descriptions for downloadable resources as well as metadata for resources that will be made downloadable during the course of the project. Most of the resources are owned and distributed by third party collaborators. They are both lexical resources such as terminology lists and wordnets, and text corpora including parallel corpora and n-grams. In addition, some resources from the first batch have been extended or further developed.

UIB's providers include other universities, the most important cooperative body for Norwegian universities and colleges, the Norwegian Language Council and the national Language technology resource collection for Norwegian (Språkbanken). It has also been a goal to increase the (re-)usability of existing resources such by making them available in standard, open, downloadable formats, and some resources are currently being made downloadable as a direct result of an initiative from UiB.

The *horizontal action on treebanking*, led by the University of Bergen, is motivated due to the potential of treebanks as a basis for creating advanced new products and services. Detailed treebanks serve notably as gold standards for inducing grammars or optimizing disambiguating parsers which can serve a range of purposes such as information retrieval, text analysis, document classification and indexing, and high quality machine translation. As

an example, the correct answering of a user question must start with a correct and detailed interpretation of the input. Furthermore, multilingual treebanks serve to identify systematic structural correspondences between languages which is extremely useful for deriving transfer rules in machine translation, as explored by the LOGON project. Some other areas in which treebanking approaches may be useful are the study of ambiguity at all levels and of second language acquisition.

1.5. UHEL

The resources uploaded by UHEL have been selected on the basis of the same three main criteria as described in the chapter on UCPH's motivations. Because of this the resources in the first and secon upload are:

a) UHEL's own resources, like the Open Source (Finnish) Morphology, Helsinki Finite-State Transducer Technology, Finnish TreeBank;

b) Resources stored in the Language Bank of Finland (Oulu corpus, Finland-Swedish Text Collection, Finnish Text Collection, Lemmie);

c) Resources owned by members of FIN-CLARIN (The International Corpus of Learner Finnish, Geographic Names Register of the National Land Survey, Samples of Spoken Finnish);

d) Resources owned by other departments or units of the University of Helsinki (Corpus of Conversational Finnish, Helsinki Corpus).

1.6. HI

We wanted to make as many completed Icelandic LT resources as possible available as well as some of the projects that are in final stages of development. Inclusion in the META-SHARE project could then serve as a milestone on the time line of the project. By including most of the available resources, instead of choosing only a few, the variety of resources is somewhat affected, but that only mirrors the situation of LT in the country where a high percentage of the resources is written corpora and very few are software.

Most of the LT resources are monolingual and focus on the Icelandic language. Some of the resources such as lexical resources and terminological resources contain a link to other languages, e.g. the Nordic languages and English. Members of the Icelandic META-NORD team do not deal with speech processing. However, it was possible to include some spoken data.

1.7. LKI

For the second phase of resource transfer, the Institute of the Lithuanian Language has selected three resources. Two of them pertain to the field of onomastics and one deals with the terminological domain. All the resources are unique in their topics. The resources have been designed for the purposes of the inquiry and application purposes of science as well as to satisfy the practical needs of the business and the people.

The geographic information database of Lithuanian place names has around 40,000 entries. The data base consists of a geographical and a linguistic elements. It features the names of every city, township and village in Lithuania. Every place name is followed by precise geographical and linguistic details.

The database of historical place names consist of roughly 20,000 entries. It is relevant in a way that it features peripheral and out-of-bounds place names, their authentic forms and etymology.

The terminological synonyms database holds some 34,000 entries. It identifies the relationship among terms that denote the same concept in language.

1.8. UGOT

The resources were selected from Språkbanken's repository, which is a national repository for Swedish Language Resources. The resources within Språkbanken originate from other researchers and/or research institutions, e.g. from The Society of Swedish Literature in Finland and Svenska fornskriftsällskapet, as well as new resources which were created by Språkbanken staff. The first upload consisted of fifteen lexical resources and the second upload adds another 2 lexical resources and 77 corpora to the META-SHARE repository. This brings UGOT's total to 79 resources for this task, which is 19 more resources than were previously estimated.

2. Overview of resources in the first batch

The following figure sums up the number of resources contained in first batch:

Table 1 The total number of resource types in the Batch 1

Resources in first batch	
36	Lexical resources
16	Corpora
6	Treebanks
5	Resources for speech
3	WordNet
1	Tool
67	**Total**

3. Overview of resources in the second batch

Table two sums up the total number of resources in the second batch; including upgrade of those that were provided in the first batch. In particular, the number of corpora has been radically increased and more tools have been provided.

Table 2 The total sum of resource types after the Batch 2

Resources in second batch	
55	Lexical resources
111	Corpora
11	Treebanks
12	Resources for speech
9	WordNet
9	Tools
207	**Total**

4. Further work on horizontal actions

4.1. Plans for collaboration with META projects

This section addresses cooperation with other META projects for the three horizontal actions.

Treebanks (UIB)

In the area of treebanking, collaboration with the other META projects has started with the common organization of the META-RESEARCH Workshop on Advanced Treebanking at LREC 2012. This event was initiated and organized collectively by Koenraad de Smedt (META-NORD), Jan Hajic (T4ME), Marco Tadic (CESAR) and António Branco (META4U). At the workshop, papers and posters from several META projects were presented. This work is documented in the published proceedings (http://www.lrec-conf.org/proceedings/lrec2012/workshops/12.LREC%202012%20Advanced%20Treebanking%20Proceedings.pdf). At this event, the University in Bergen has promoted the treebanking services offered to the other META projects through the INESS portal. Treebanks are currently available for 22 languages (for authenticated users). This activity will be sustained throughout the rest of the project.

Wordnets (UCPH)

Contact has been taken to META-NORD's sister projects META4YOU and CEASAR with the aim of initiating collaboration on wordnets across the three projects. We suggest to collaborate in two ways: i) for META4YOU and CEASAR's wordnets to join the pilot linking performed in META-NORD via their links to Princeton WordNet - to the degree where such links are available, and ii) to compare and validate the wordnets along the lines that we are using in METANORD: *taxonomical structure*, *coverage*, *granularity*, and *completeness*.

Terminology (TILDE)

The concept of sharing of specific language resources by specialized LR repositories of META-SHARE such as Terminology Repository, has been disseminated in the workshop CHAT 2012 collocated with the conference TKE 2012. Tilde has addressed sibling projects METANET4U and CESAR with a call for collaboration in the area of terminology. We invite other projects to search for not-yet discovered terminology resources, preferably free and open ones in their respective countries, e.g., National termbases, Academia terminology collections, industry terminology resources. Tilde also offered EuroTermBank – a dedicated terminology node of META-SHARE - to other projects to be used as a sharing platform for an integrated entry-level access to the terminology resources – a portal where all resources are available for one stop online search.

4.2. Further work plans for the rest of the project

Treebanks (UIB)

The University of Bergen will continue to offer treebanking services to other META projects. This involves providing solutions for parsing and disambiguation, uploading of existing treebanks, indexing, management and exploration of treebanks, and linking of treebanks across languages. These services will be actively promoted for instance through the META-NORD newsletters. Treebank owners and providers will be encouraged to take advantage of the services provided by INESS and to raise awareness of them in the broader international LRT community.

Wordnets (UCPH)

July 2012: Upload of automatically (via Princeton Core) bilingually linked resources (still unvalidated)
Sep. 2012 Common web-interface ready for use
Oct. 2012 Completed validation of bilingually linked resources
Dec. 2012: Completed general validation of the wordnets
January 2013: Upload of Norwegian Wordnet. Upload of validated, linked wordnets. Uniform web interface for browsing monolingual and linked resource.
Eventual integration and/validation of wordnets in META4you and CEASAR.

Potential use and ways of promoting the wordnet resources to industry: Some of the wordnets are already applied in industry, for instance, DanNet is used by two companies in DK (Magenta Aps for synonymy look-up in OpenOffice and LAT Consulting for information retrieval (IR)). Likewise, FinnWordnet has been integrated with some search tools that have been used in industry R&D projects, but the definite proof that there is industry uptake is still lacking. In general terms, IR is probably where the data could be easiest to integrate in addition to author's tools both for monolingual and multilingual IR purposes. Author's tools and IR tools may benefit from regular and high-frequent words if you provide tools for foreign language learners. For native speakers you do need a rather large wordnet before the benefits show up in IR, because people rarely use high-frequent words to find documents or they use a very specialized vocabulary. Promotion to industry of the result of the horizontal action on wordnets will partly be given via the common web interface which will allow interested parties to browse the linked data and get an intuitive impression of the potential of the resources. Secondly, the wordnets will be promoted via the META-NORD newsletter.

Terminology (TILDE)

Tilde will continue taking the lead in collection of terminology resources – new resources will be uploaded by Tilde and other partners. In Y2 of the project the actual integration and import of terminology data for search will take place as the data becomes shared and available. National workshops will be another way to disseminate META-SHARE and encourage key institutions and resource holders to share their LR data. Tilde will be adding any additional free terminology resources to EuroTermBank if shared via META-SHARE. Tilde will continue to offer terminology hosting and sharing services to other META projects. The owners of terminology resources will be encouraged to take advantage of the services provided by EuroTermBank. Integration with META-SHARE will be implemented in a degree in which the platforms can be interlinked.

5. Appendix A: List of metadata

List of metadata for all resources uploaded so far (downloaded metadata from META SHARE v2.1 converted into xml and further to .pdf).