META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Project no. 270899

Deliverable D3.3
Third batch of resources
including resources selected in D2.4

Version No. 1.0

31/01/2013

Document Information

Deliverable number:	D3.3
Deliverable title:	Third batch of resources including resources selected in D2.4
Due date of deliverable:	31/1/2012
Actual submission date of deliverable:	31/1/2012
Main Author(s):	Dorte Haltrup Hansen, Bolette Pedersen
Participants:	All partners
Internal reviewer:	Tilde
Workpackage:	WP3
Workpackage title:	Enhancing language resources
Workpackage leader:	UCPH
Dissemination Level:	PU
Version:	V1.0
Keywords:	Documentation, resources

EXECUTIVE SUMMARY

This report documents the resources delivered in META-NORD's third batch in M24. In Chapter 2, each partner gives an overview of the motivation of the selected resources uploaded in META-SHARE, and general tables of Batch 2 and Batch 3 are provided in Chapter 3 and 4. During the project's lifetime, the Consortium has identified and collected more than 500 resources and tools. Considerable work of documenting, processing, linking, and upgrading these resources to agreed standards and guidelines has been performed and is richly documented in Chapter 4. Furthermore, alignment and linking tasks across languages are documented in Chapter 5. Finally, 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 each META-SHARE node hosted by a META-NORD partner (Appendix A).

To sum up the main results of this work package, much more resources for both academic and commercial R&D are now directly available for the Nordic and Baltic languages, and a considerable number of them are furthermore interoperable in *format* and *content*. A broad range of different resources for different languages and language pairs as well as suitable for different purposes have been selected. However, in line of what was planned according to the DoW and further outlined in D2.4 on selected resources, the major focus in META-NORD has been on written resources. Nevertheless, there has been increased focus on audio/video resources in the last project period, and thus the number of such resources has been more than doubled since Batch 2.

Table of Contents

Abbreviations

Table 1 Abbreviations

Abbreviation	Term/definition
DoW	The META-NORD Description of Work document
TILDE	TILDE SIA (Latvia)
UCPH	Københavns Universitet (Danmark)
UT	Tartu Ülikool (Estonia)
UIB	Universitetet i Bergen (Norway)
UHEL	Helsingin Yliopisto (Finland)
HI	Haskoli Islands (Iceland)
LKI	Lietuviu Kalbos Institutas (Lithuania)
UGOT	Göteborgs Universitet (Sweden)
CST	Centre for Language Technology (at UCPH)
LMF	Lexical Markup Framework
TEI	Text Encoding Initiative
LSP	Language for Special Purpose
TBX	TermBAse eXchange
LT	Language Technology
NLP	Natural Language Processing

1 Background and overview of WP3

The aim of work package 3 on Enhancing Language Resources is to upgrade and harmonize national language resources within and across META-NORD languages, in order to make them interoperable w.r.t. their *data formats* and *content*.

During the project's lifetime, more than 500 resources and tools have been identified and collected in the consortium. Considerable work of documenting, processing, linking, and upgrading these resources to agreed standards and guidelines has been performed and is documented in the following sections.

Apart from the considerable work on numerous individual resources, three integrative actions have been completed in this work package, with the aim of:

- Making treebanks for relevant languages accessible through a uniform web interface and state-of-the-art search tool and linking treebanks across languages using a parallel multilingual treebanking;
- Ensuring that Nordic and Baltic wordnets agree to modern standards regarding formats, develop a multilingual web interface which facilitates browsing across resources, derive/create pilot multi-lingual lexicons for IR purposes using cross-language synset linking and validate these pilot cross-lingual resources;
- Linking monolingual and bilingual terminology collections and integrating them into a multilingual terminology bank with elaborated terminology data access and sharing mechanisms.

These cross-lingual initiatives are documented in separate reports D3.4, D3.5, and D3.6, respectively. Further cross-lingual initiatives include:

- An overview of the existing aligned resources containing language material in at least one of the META-NORD languages;
- A collection of Comparable Lithuanian, Latvian and Estonian Laws and Legislations
- Linked Wikipedia;
- A cross-linked collection of comparable sentences from Wikipedia.

These are all described in Section 5.

META-NORD language resources come in many formats, and partners put considerable efforts into making content models as interoperable as possible. This has been done by adopting more strictly structured formats, e.g., LMF rather than proprietary XML or SQL for lexical resources. Mapping to a set of standardized data categories, e.g., ISOcat, was ensured whenever feasible.

The resources and tools that have been registered and uploaded in META-SHARE are documented in terms of metadata. For this deliverable matadata for all resources have been automatically generated from each META-SHARE node (Appendix A).

2 Motivation for selected resources

2.1 TILDE

The resources selected for inclusion in META-SHARE were identified and assessed according to the criteria specified in the Task 2.3: *availability* (openly available LR versus restricted proprietary), *suitability* for LT development, *multilinguality* (in terms of supporting multilingualism or linking between languages), *longevity* (maintenance of LR), *quality*, *extensibility* (level of available documentation and metadata description). The special focus was on resources that facilitate the aim of the META-NORD project to build a pan-European

digital resource exchange facility, i.e., resources that are important for language technology development, especially resources that can be used to build applications that help to overcome language barriers.

The Latvian and multilingual resources described and made available by Tilde include:

a) resources and tools developed by Tilde (e.g. bi/multilingual electronic dictionaries, corpora);
b) resources and tools that are outcomes of EU projects where Tilde was/is the coordinator or partner responsible for development of particular resource (e.g. outcomes of ACCURAT and EASTIN-CL projects);
c) public and copyright free resources developed and standardised by Tilde (e.g. Legislation corpus of Republic of Latvia);
d) resources negotiated by Tilde from the third party creators (e.g. dictionaries, lexical databases).

Before resources were catalogued, IPR issues were clarified and licensing conditions fixed in the metadata. All resources and tools that were planned to be included in D2.4 are now made available.

In addition, the following resources and tools, which were not initially planned at the time of D2.4, have been identified and included by Tilde:

- IDENTIC Indonesian-English parallel corpus;
- German – Russian Gold Standard for knowledge-rich context extraction;
- English-Latvian Road Terms Technical dictionary;
- PIARC Multilingual Terminology database of Road Terms;
- EOPC – Estonian Open Parallel Corpora.

For more information on each resource see Appendix A.

Tilde also leads *the horizontal action on terminology*. Multilingual terminology is indispensible resource for the development of multilingual language technologies, including adaptation of translation tools for particular domains; 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 META-SHARE Tilde has developed an integration service that enables integration of a terminology specific node with the META-SHARE platform through the mutually implemented access interfaces and harvests meta-data from the terminology specific node (the service and work related to multilingual terminology is described in detail in deliverable D3.6 *Interlinked multilingual terminology bank*).

We expect further uploads from external partners to Tilde's META-SHARE node in near future.

2.2 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 eventually contain interesting perspectives wrt. up-grading to agreed standards, extension and/or multilingual linking and validation within the META-NORD project time span and efforts, and iii) they should preferably 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 include primarily resources provided by UCPH (among others LSP Corpora, the Danish Wordnet, DanNet, a computational lexicon for Danish, linked, STO, linked wordnets together with UHEL, UGOT and UT, analysis tools and an annotated audio/visual corpus, NOMCO), supplemented by a few from the Copenhagen Business School (CBS) (treebanks). A majority of the resources are monolingual and focus on the Danish written language. Some of the resources have links to other languages, such as English; this applies to the Danish wordnet, DanNet and The Copenhagen Danish-English Dependency Treebank.

UCPH also leads the *horizontal action on wordnets* which is concerned with the validation and pilot linking between Nordic and Baltic wordnets. The aim of the multilingual action is to test the perspective of a *multilingual linking* of the Nordic and Baltic wordnets and via this (pilot) linking to perform a tentative comparison and validation of the wordnets along the measures of taxonomical structure, coverage, and granularity. Four pilot bilingual wordnets have been produced via established links to Princeton Core Wordnet: DA-SE, DA-FI, ET-FI, FI-SE and has been uploaded to META-SHARE. Last but not least, UCPH has developed a common web interface, WordTies, which visualizes the links and enables the user to browse the particular linked wordnets. For documentation, see D3.5 on wordnets.

To sum up, all resources and tools that were planned to be included in D2.4 are now made available. In addition, the following resources and tools, which we had not expected to include or which we were not sure of being able to include at the time of D2.4, have now been fully included (for more info on each resources, see Appendix A):

- Anvil Facetracker;
- HSPG Grammar;
- CST NP grammar;
- NOMCO Corpus (annotated videos);
- Reference Corpus for Danish (CLARIN);
- Corpus of Sublanguages (seven domains);
- CST Keyword Extractor;
- CST Repetitiveness Checker;
- WordTies.

In addition, when the establishment of the Danish META-SHARE node was completed in late autumn 2012, we also invited industrial partners to include links to relevant LT tools and resources. Thus, external partners have uploaded/linked to more than 25 tools and resources to the Danish META-SHARE node. These external resources and tools include:

- Machine translation modules for several languages;
- Danish FrameNet;
- DeepDict – Relational Lexicon for Danish;
- Constraint grammars for several languages;
- Grammatical annotation of the Danish Korpus 2000.

We expect further uploads from external partners to the Danish META-SHARE node in the near future.

2.3 UT

We have aimed to include in META-SHARE all Estonian language resources that are mature enough to be shared, compliant to standards and for which the access terms and user licenses could be negotiated in the timeframe of the project. The resources uploaded are from the three major actors in Estonian LT development: the University of Tartu, the Institute of Estonian Language (IEL) and the Institute of Cybernetics (IoC) who also form the Centre of Estonian Language Resources consortium. Each institution has their own research specialties and together they cover most of Estonian LT landscape. Once the main language resource providers have made their resources available via META-SHARE it should become more attractive to other providers as well.

In the first two batches, the resources for which the preparations were easier, were uploaded – the written corpora and lexical-conceptual resources such as the Estonian WordNet by UT as well as dictionaries, written and speech corpora by IEL.

In Batch 3 we have speech resources from our colleagues as well as from a small Estonian LT company, plus some additional UT resources that were not originally planned to be included in META-SHARE. The metadata records for several resources uploaded in previous batches have been reviewed and filled in with more detailed information, also several of these resources themselves have been upgraded to the level where they could be made downloadable via META-SHARE.

Since most of the resources from the original review list that met our requirements for maturity and availability had been uploaded in previous batches, the resources planned in the third batch were either those by third parties which required more time to get the information about resources or licensing, or those that were only just completed at this time. These resources consist of tools (morphological analyser by Filosoft and text-to-speech synthesis tool by IEL), speech resources (Estonian Foreign Accent Corpus by IoC and Corpus of Spoken Estonian by UT) and two cross-lingually linked resources created from existing monolingual ones in the frame of the META-NORD project (Estonian-Latvian parallel text corpus and Estonian-Icelandic disease names). The cross-lingually linked resources were selected based on a review of available language resources by third parties that could with relatively small effort be made multilingual.

In addition to the initial plans, several resources by UT were added during the course of the project. These include speech corpora (Estonian Dialect Corpus and Phonetic Corpus of Estonian Spontaneous Speech), a text corpus (Corpus of Old Written Estonian) and word lists (Frequency lists, Estonian Frequency Dictionary).

2.4 UIB

The general motivations driving the University of Bergen META-NORD team's selection of resources in Norway have been 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;

- Prioritizing resources that are useful for both industrial and academic R&D, commercially and non-commercially;
- Selecting resources owned and distributed by third party collaborators, thus raising awareness outside of the consortium about resource exchange through META-SHARE.

For the third upload UiB has provided metadata descriptions for downloadable resources. 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 and second batches 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 accessibility and (re-)usability of existing resources by making them available in standard, open, downloadable formats, and some resources have been made downloadable as a direct result of an initiative from UiB.

The *horizontal action on treebanking*, led by the University of Bergen, is motivated by 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.

Resources planned in 2.4 that were not made available in Batch 3

The following resources were planned in 2.4., but could not be made available in time for Batch 3. All resources are provided by third parties.

Leksikografisk bokmålskorpus. The resource is owned by the University of Oslo. META-NORD was unable to negotiate the resource in time for Batch 3, but it will be followed up in CLARINO[1].

Milterm is owned by *Standard Norge*, a private member organisation which is one of three standardisation bodies in Norway. The resource was not made independently available as planned in 2.4, but has been made available as part of the terminological database *Snorre*, which is documented with metadata in META-SHARE under the resource name: Wordlist from SNORRE Terminology Database.

NHH Termbase is owned by the Norwegian School of Economics (NHH). Currently (January 2013) this resource was still not ready from their part and therefore could not be delivered. However, NHH is a member of CLARINO, and the termbase will be made available as part of the CLARINO effort on terminology.

[1] CLARINO is the Norwegian affiliate of CLARIN, a European project which builds an infrastructure for language-related eScience in humanities disciplines. For further information, see http://clarin.b.uib.no/.

Stadsnamnsamlinga is a collection of place names owned by the Department of Linguistic, Literary and Aesthetic Studies at the University of Bergen. The resource is still under development and was not ready in time for Batch 3, but will be followed up by CLARINO.

Translation Corpus Aligner 2 is a tool for alignment of original and translated text in the development of translation corpora. The resource is owned by the University of Bergen and will be made available through CLARINO.

Det nynorske tekstkorpuset is a large Norwegian Nynorsk text corpus owned by the University of Oslo, which forms the basis for a new dictionary on the occasion of the 200 years' anniversary of the Norwegian Constitution. Due to the large numbers of licences that will need to be cleared with a number of publishers, META-NORD was unable to deliver this resource for Batch 3, but the work will be pursued in CLARINO.

International Computer Archive of Modern and Medieval English is a corpus of English owned by the University of Bergen and will be made available through CLARINO.

Resources made available in Batch 3 that were not originally planned in D2.4

The following resources were not originally planned in D2.4, but will be made available in Batch 3 (see table above for further information on each resource and its motivation):

- n-gram for Norwegian Bokmål (based on NNC and NST news text);
- n-gram for Norwegian Bokmål (based on NNC);
- n-gram for Norwegian Bokmål (based on NST news text);
- n-gram for Danish (based on the NST text corpus);
- n-gram for Norwegian Nynorsk (based on NNC and NST);
- n-gram for Swedish (based on the NST Text Corpus);
- NoWaC - Norwegian Web as Corpus;
- Frequency lists (tokens) from NoWaC - Norwegian Web as Corpus;
- Frequency lists (lemmas) from NoWaC - Norwegian Web as Corpus;
- Sofie Parallel Treebank;
- The Wordnet for Norwegian Bokmål;
- The Wordnet for Norwegian Nynorsk;
- The Norwegian Language Council's list over names of historical events and persons;
- The Norwegian Language Council's list over names of inhabitants in Norwegian;
- The Norwegian Language Council's dictionary from Norwegian Bokmål to Nynorsk;
- The Norwegian Language Council's list over language names in Norwegian;
- The Norwegian Language Council's list over state names in Norwegian;
- The Norwegian Language Council's list over geographical names in Norwegian;
- The Morphologically Annotated Part of BulTreeBank;
- The Dependency Part of BulTreeBank;
- Vuosttaš Digisánit - Northern Sámi-Norwegian dictionary;
- Voestes digibaakoeh - Southern Sámi -Norwegian Dictionary;
- Kven-Norwegian-Kven Dictionary;
- Norwegian - Northern Sámi translation memory;
- List of Norwegian - Northern Sámi place names;
- List of mechanical terms (sme, nor, swe, fin, eng);
- The Sámi Parliament's term collection (sme, smj, sma, nor, fin, swe, lat);
- Norwegian Acquis communautaire;
- Snorre termbase (including Milterm).

2.5 UHEL

The resources uploaded by UHEL have been selected on the basis of the three main criteria, namely:

i) they should have a certain level of maturity and quality;

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.

Besides these motivations UHEL was also interested in selecting as well as with the intention to select resources owned and distributed by third party collaborators, thus raising awareness outside of the consortium about resource exchange through META-SHARE. Because of this the resources in UHEL's third upload are:

a) UHEL's own resources (Finnish TreeBank 3);

b) Resources stored in the Language Bank of Finland (several written corpora, digital morphology archives, frequency lexicon of the Finnish Newspaper Language);

c) Resources owned by members of FIN-CLARIN (The FinINTAS corpus of spontaneous and read-aloud Finnish speech, ProoF - Pronunciation of Finnish by immigrants in Finland, MultiJur: Multilingual Parallel Corpus of Legal Texts, FiRuLex: Finnish-Russian Comparable Corpus of Legal Texts, The Finnish Broadcasting Company Corpus of Subtitles, Corpus of Spoken Southwestern Finnish, Finnish Telegraphese Corpus, Morfessor, National Semantic Web Ontology Project in Finland, Online Lexicon of Veps Language, Headword List of the Karelian Dictionary, Online Karelian Dictionary, Laws and Directives, New Year's Speeches of the Presidents of the Republic of Finland, Finnish Proverb Collection, Frequencies of Old Literary Finnish Words, Frequencies of Early Modern Finnish Words, Word Collections of Modern Finnish, Names of Countries in Seven Languages, Name Component Lexicon, Topling - Paths in Second Language Acquisition DIALUKI - Diagnosing reading and writing in a second or foreign language);

d) Resources owned by other departments or units of the University of Helsinki (The Bank of Finnish Terminology in Arts and Sciences, ELFA corpus);

e) Resources owned by third party collaborators (The Tampere Bilingual Corpus of Finnish and English, Speech and EGG simultaneous recordings, TEPA – The Finnish Terminology Centre TSK's term bank, Turku Dependency Treebank, variKN Language Modeling toolkit, Suopuhe, Voikko, The EMIME Bilingual Finnish/English German/English Database, FinFlect, eSpeak).

Resources that were negotiated, but are not available on META-SHARE

- UTA Cross-Language Information Retrieval System: UHEL tried to convince the owner to publish the program code of the tool with e.g. GPL licence, but the answer was that the code is obsolete (it was devised to be used for specific resources), i.e. the tool should be recreated from zero;
- ParRus: Russian-Finnish parallel corpus of literary texts: The UHEL team has made a conditional deposition agreement regarding this resource that contains a large number of literary texts and their translations whose IPR status needs to be clarified separately. The work on clarifying the IPR status of these texts and translations is

financed by UHEL through FIN-CLARIN-content funding. This resource, together with ParFin, is a good sample case for UHEL of a carefully documented process involving negotiations with writers, translators and publishers on depositing and further use of language resources. In other words the work done with ParRus and ParFin will facilitate the process of making available for research and perhaps even for wider use of similar resources;

- ParFin: Finnish-Russian parallel corpus of literary texts: same situation as with ParRus;
- Resource Collection of Human-Computer Dialogues (belongs to the category Other Speech Corpora): the owner stated that the resource is outdated. Its new version on the other hand is in a too early stage to publish any metadata on it according to the owner, who promised to get back when publishing its metadata will be timely;
- Collaborative Writing (belongs to the category Other Speech Corpora): the interviewed gave their consent only for the use of the corpora by the researchers participating in the project, thus licencing for (general) research use is impossible according to the owners;
- Northern multilingualism (belongs to the category Other Speech Corpora): same situation as with Collaborative Writing;
- CLIL (content-and-language-integrated learning) corpus (belongs to the category Other Speech Corpora): the contact person, Leila Kääntä from the University of Jyväskylä, promised to discuss with the owner (or owners) in order to find out about the distribution and licencing policy of the resource. In return for this favour UHEL clarified for Leila Kääntä different possibilities to solve the technical issues of one of the University of Jyväskylä's text corpus;
- EFL (English as a Foreign Language) corpus (belongs to the category Other Speech Corpora): see CLIL;
- Talk show corpus (belongs to the category Other Speech Corpora): see CLIL;
- Reality TV corpus (belongs to the category Other Speech Corpora): see CLIL;
- Gaming corpus (belongs to the category Other Speech Corpora): see CLIL;
- The Audio Recordings Archive of Oulu (belongs to the category Other Speech Corpora): according to the owner there are serious personal data protection related issues that make it difficult to publish the resource;
- Emotional speech: same situation as with Collaborative Writing;
- Corpus of translated Finnish: UHEL found out that the original licences made with the right-holders were too tight.

This issue were carefully studied together with the right-holders, the lawyer of the UHEL team and University of Eastern Finland. It turned out that there were a number of ambiguities related to user rights. There were problems like despite the contracts signed the IPR had not been transferred from the authors/translators to the University of Eastern Finland. Trying to make new contracts with all the persons involved would be extremely time-consuming. Because of this the UHEL team decided not to publish metadata on the resource.

2.6 HI

HI wanted to make available as many Icelandic language resources as possible, provided they were considered mature and useful enough to be shared and could be licensed. Most of the resources originate with members of the Icelandic Centre for Language Technology (ICLT);

the University of Iceland (HI), Reykjavik University, and the Árni Magnússon Institute for Icelandic Studies. A few other institutions and independent researchers have also contributed their resources.

None of the Icelandic resources submitted by HI have previously been available at major data repositories like ELRA, LDC and others. In fact, most of them have not been publicly available at all up until now. Most of them are written language resources, many of which are central for their respective fields, such as:

- several corpora, including the *Tagged Icelandic Corpus* (25 million words) and *Íslenskur orðasjóður – Large Corpus* (250 million words);
- lexical resources, including the multilingual dictionary *ISLEX* (50,000 entries), the bilingual *Icelandic Term Bank* (more than 40 terminology lists), the *Database of Modern Icelandic Inflections* (270,000 paradigms, 5.8 million inflectional forms) and a *Pronunciation Dictionary for Icelandic* (65,000 words);
- a treebank, *Icelandic Parsed Historical Corpus* (one million words);
- semantic databases, including *Íslenskt orðanet – Þesárus* (250,000 entries) and *Icelandic Semantic Database* (110,000 words, 2.93 million relations);
- and a number of others.

A few spoken language resources have also been uploaded to META-SHARE, especially the *Hjal Speech Corpus* (around 40,000 sound files) and the *ISLEX Audio Recordings* (50,000 sound files). In addition, the majority of the LT tools that exist for Icelandic have been made available through META-SHARE. This includes especially the *IceNLP* package which contains the PoS tagger *IceTagger*, the parser *IceParser*, and the lemmatizer *Lemmald*.

Most of the LT resources are monolingual and focus on the Icelandic language. However, HI has managed to include two important bi- and multilingual resources that were not included in the list originally prepared for the DoW. One is the *Icelandic Term Bank* which actually is a collection of more than 40 different resources – bilingual (Icelandic–English) terminology lists from various fields. The other is *ISLEX – Icelandic–Scandinavian Web Dictionary* with around 50,000 Icelandic entries and their equivalents in Danish, Swedish, and Norwegian (both Bokmål and Nynorsk).

All the resources and tools that were planned to be included in D2.4 are now made available. In addition, we have managed to get permission to include sound recordings of all the Icelandic entries in the ISLEX dictionary (cf. above).

2.7 LKI

The main motivations for the selection of resources have been their availability through LKI repositories, *suitability* for LT development, *longevity*, *quality*, and *extensibility*.

Initially it was planned to include and use resources mostly which had been created by LKI because they are important for Lithuanian language and have been created on the basis of long time research work containing useful information and to avoid restrictions on their use through META-SHARE. The databases (DB) of those resources have been created at different times, using different programs, which were not suitable for META-SHARE. It took more time than planned to parse data from the old DB, convert it to the new format; for some databases it was necessary to create new web applications for accessing the data. Almost all resources have been extended with additional data and made available through META-SHARE.

During preparation for the second and the third batches of language resources, LKI started negotiations concerning other resources, which were not included in Dow. Possible resources

were selected according to their availability through META-SHARE with required documentation: Lithuanian-Hungarian Dictionary, Term Bank of Lithuanian Republic, Encyclopedic Dictionary of Computing and so on. For more information of these resource see the Appendix A.

For the possibility to continue work with LKI and resources of the third parties the META-SHARE node in the LKI has been established. It was not planned to settle the separate node for META-SHARE resources at the beginning of the project. This plan was worked out during the project in order to sustain the results of the project and opportunity of further work at international level. LKI META-SHARE node will contain resources which have been upgraded and extended to the appropriate level in accordance with META-SHARE.

All resources provided for the META-NORD project can be divided into several groups.

a) lexical conceptual resources developed by LKI (dictionaries, lexical, onomastics, terminology databases);
b) lexical conceptual resources developed by third party creators (dictionaries, lexical, onomastics databases);
c) tool provided by LKI (morphological analyser, lemmatiser and synthesiser).

Most of them are lexical conceptual resources. These resources are major Lithuanian language dictionaries, and databases. A major resource of modern Lithuanian and dialects is **The Dictionary of Lithuanian Language**. This resource is the largest work of twentieth-century Lithuanian linguistics. The dictionary aims to give the words and illustrate their usage by quotations culled from all kinds of writings and dialect records. The twenty volumes of the dictionary make up about 22,000 pages, comprising half a million headwords and over 11,000,000 words of text. This academic edition of the dictionary of the Lithuanian Language (Naktinienė et al., 2005) is significant not only as a major landmark of Lithuanian philology, it is also an authoritative source for comparative Indo-European studies. It presents the origin, history and spread of a word, its grammatical and accentual forms and categories, and its peculiarities with respect to word-formation, semantic structure, stylistic usage, etc. The dictionary contains an abundance of extra-linguistic information: the illustrative material carries much background information about the everyday life of the speakers of the language, their social relations, ethical values, ethnographical details, etc.

There are two resources for modern Lithuanian lexis. One of them is the **Modern Lithuanian Dictionary**. The Modern Lithuanian Dictionary (Keinys et al., 2011) is a universal one-volume explanatory normative work of standard language intended for a wide circle of readers. It contains a huge amount of modern Lithuanian words, some regional dialectal and more widely used spoken language words. Moreover, it also contains words from past and contemporary fiction, especially classical papers, which are necessary for the studying youth to cultivate their language, to reflect on various language styles, often suitable for specific new concept expression.

The database of Standard Lithuanian language lexicon. This database provides a sampling of correct and submitted common words in the standard Lithuanian language. It contains a number of new terms from different areas, word illustrations. The Standard Lithuanian language dictionary will be based on this database.

Database of Lithuanian Neologisms. This database contains Lithuanian neologisms - new words (borrowings and recently-coined words), phrases and abbreviations with the usage examples from media, advertising, fiction, scientific literature and administrative documents, daily spoken language and electronic discourse are stored.

Lithuanian-Hungarian Dictionary (Bojtar et. al., 2007) is a bilingual resource. It is a comprehensive dictionary of modern Lithuanian and Hungarian.

One resource is from old Lithuanian language writings - **Database of Old Lithuanian Writings**. The first writings transferred to electronic files. This database consists of texts, concordances, indices, transcribed facsimiles.

Four resources are databases of proper names. **Geoinformational Database of Lithuanian Toponyms** has been developed mainly for science subjects and applications needs. The database provides information about linguistic units - place names and geographical units. The coordinates of each recorded linguistic-geographical unit have a certain attachment - the attributes of the place: colloquial place name variants, stressing, scientific origin, formation interpretation, the time of recording from the living language, the time of the first mention in historical resources, the sound of the utterance, population density of the residential areas, geographic settings, objects, photographs, and other administrative dependence. In addition, the database includes not only the existing settlements, rivers, lake names, but also extinct objects of those classes, also land names, which were written by the local population during the interwar period, in 1935-1940, and post-war. **Database of proper names** - is a comprehensive database with alphabetically provided current names, followed by the names of its prevalence and incidence of the most Lithuanian population. One of the largest and important resources is the **Lithuanian Names Database**. The author of the resource is The State Commission of the Lithuanian Language. This register provides most of the citizen names of The Republic of Lithuania. There are **Database of Lithuanian Historical Ethnic Place Names**. The core of the database consists of the names of the places near and outside the Lithuanian border gathered from XVI-XIX century historical manuscripts such as administrative-legal documents, church books and the like. Currently, the database consists of 20000 place names, the current Lithuanian genders are set, and notations in the historical documents are given in chronological order.

There are several resources for terminology. The largest in Lithuania is the **Term Bank of Lithuanian Republic** – a very important data base in Lithuania, containing continuously updated information. It is stored in the repository of Seimas (Parliament) of Lithuania and has a restricted access. The meta-data of the resource is available in META-SHARE and it is accessible through an external link for information and learning purposes. For derivate use of the resource the negotiations must be carried out on with the State Commission of the Lithuanian Language. **Database Synonyms of the Lithuanian Terms** includes 34,000 entries. The base contains synonymic lines of terms of different (international, Lithuanian and hybrid) origin that can be found in Lithuanian glossaries and dictionaries of terms and special encyclopaedias from different fields. Articles of terms may include definitions and equivalents of terms in other languages. The structure of term articles depends on the source. Definitions are available for terms found in glossaries and encyclopaedias. Terms provided in dictionaries come with their equivalents in foreign languages. One resource is from computing terminology - (Dagienė et al., 2008). Vilnius University Institute of Mathematics and Informatics is author of this resource. The dictionary describes about 4000 terms and lexical units. It provides definitions, explanations and illustrations. This dictionary is intended for software producers and localizers, it is helpful for computer users who want to better understand or check the software commands and messages.

All the resources are unique in their topics. The resources have been designed for the purposes of inquiry and application purposes of science as well as to satisfy the practical needs of the business and the public. For this purpose LKI presents resources for linguistic usage. **State Language Consultancy Bank** - the author of the resource is The State Commission of the Lithuanian Language. This consultancy bank contains a lot of questions

and answers from a variety of language areas, it also includes the full list of the major language mistakes. LKI prepare **Language Consultancy Bank**. From the collected data was selected and structured relevant matters of the present language usage. **Office Language Recommendations database created** by LKI contains use cases and applications of clerical language.

The Virtual Electronic Heritage System has been created by the National Library of Lithuania, which is an enormous wealth of digital objects, created under the governance of the Lithuanian cultural heritage digitisation, digital preservation and access strategy. The VEPS portal provides quick and convenient access to thousands of works for all who are interested in art, books, newspapers, manuscripts, maps, and sound recordings. All this together creates an unique, rich and vivid panorama of Lithuanian cultural heritage. An integral thesaurus of historical places, personal names and historical chronology is deployed in VEPS. It serves not only as a rich knowledge base, but also as an effective tool for semantic search of Virtual Electronic Heritage System.

Language technology tool developed by Vytautas Zinkevičius (LKI). It is a m**orphological analyser, lemmatiser and synthesiser for Lithuanian**. Lemuoklis is a morphological analyser, lemmatiser and tagger for Lithuanian. A word form is characterized grammatically by a combination of properties with respect to 13 categories: part of speech, aspect, reflexiveness, voice, mood, tense, group, degree, definiteness, gender, number, case and person. The database of lexical and grammatical information of the program consists of six lexicons. Using morphological rules together with word-root lexicons enables us to analyse a lots of of theoretically available Lithuanian written forms.

LKI communicated with other research and state institutions concerning planned and not planned resources to submit to META-SHARE. The Lithuanian Standards Board is an author of a Term Base of the Lithuanian Standards Board. Though it was a planned resource, at the moment, there seems to be no possibilities for having it in META-SHARE.

Vilnius University is an author of the Science Language Parallel Corpus. The negotiations and meetings have proceeded, but still there is no final decision concerning availability of this resource in META-SHARE.

Vytautas Magnus University is the author of the Corpus of the Contemporary Lithuanian Language. Negotiations are still underway. It was not planned initially, but during the project life time, possibilities to collaborate have occurred. It is a very big resource with many authors involved and licensing issues of the resource remain still very complicated.

As LKI will maintain the META-SHARE node and sustain the results of the project, there are still possibilities for successful negotiations with other institution for further upload of metadata and resources to the META-SHARE repository at a later stage.

2.8 UGOT

UGOT estimated in D2.4 that 106 resources would be delivered throughout the duration of the META-NORD project, but have already delivered 116 resources and more are still being negotiated or prepared. This includes both resources created in-house (many of the lexical resources) and resources from external sources (many of the corpora) since Språkbanken is considered a national repository. 95 of those resources are downloadable from UGOT (Språkbanken) and 21 are accessible through an external link Negotiations are underway for a number of external resources:

- speech resources from the Center for Speech Technology, Royal Institute of Technology, Stockholm (in progress);

- linguistic experimental data and corpus resources from the Humanities Laboratory, Lund University (in progress);
- the SweDia dialect recording database from a national consortium coordinated by the University of Gothenburg;
- multilingual corpora and corpus tools from Uppsala University(in progress);
- the Gothenburg Dialogue Corpus from the Dialogue Technology Laboratory, University of Gothenburg (in progress);
- the national term bank Rikstermbanken from Terminology Sweden (TNC).

3 Overview of resources in all three batches

Table 2 and 3 sum up the number of resources contained in the two first batches:

Table 2 The total number of resources after first batch

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

Table 3 The total number of resources after second batch

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

Table 4 and 5 sum up the total number of resources provided to META-SHARE by the META-NORD partners after the third batch sorted by different perspectives; including upgrade of those that were provided in the first and second batch.

Table 4 The total number of resources after third

Resources after third batch	
210	Lexical resources for text (excl. wordnets)
182	Text corpora (excl. treebanks)
44	Tools for text
2	Language description (grammars)
31	Treebanks
28	Audio/visual resources and tools
12	WordNets
509	**Total**

Table 5 The total number of resources after third batch, sorted by resource type and media type

Resources after third batch	
222	Lexical resources – text
1	Lexical resources – audio/visual
205	Corpora – text
22	Corpora - audio/visual
8	Corpora – ngram
44	Tool – text
5	Tool - audio/visual
2	Language description – text
509	**Total**

It should be noted that in this report resources are counted by the number of metadata records created in META-SHARE. This number of metadata records accounted in D3.3 differs from the total number of resources provided in D4.5 (509 vs. 493 resources, respectively) because the two parallel treebanks (META-NORD Acquis and META-NORD Sofie) have metadata entries for each language-specific treebank in META-SHARE (16 entries) but only count as two resources in respect to D4.5.

4 Upgrading resources to agreed standards

An important task of all META-NORD partners has been to upgrade selected resources to standards agreed in cooperation with other projects, in particular the META-SHARE initiative of META-NET. The upgrading activities were planned to include at least the following:

- Improvement of resource documentation, both formally structured (as META-SHARE metadata) and narrative documentation;
- Technical format conversion, e.g., a proprietary corpus format into TEI or an internal lexical database format into LMF;
- Content model conversion/mapping/linking, e.g. harmonizing POS tagsets among corpora, or linking word senses among lexical resources with different sense granularities.

All these activities have been carried out by the partners throughout the project period, as reported in the following subsections.

Tilde

A number of resources originating from **Tilde** were already provided in industry standard formats. For other resources necessary conversion and upgrading work was carried out to ensure their usability and interoperability:

- Monolingual *Corpus of Latvian literature* was converted to and provided for download in standardised XML interchange format allowing accessing the structure of the source documents in the corpus. Parallel n-gram corpus was converted to and provided in industry standard Moses format containing the occurrence metric. *Estonian Open Parallel Corpus* was converted to the Moses format.
- Several lexical conceptual resources were converted to and provided for download in generic XML format – *Multilingual Dictionary of Person Names*, *Latvian Russian Personal Names Glossary* and *Latvian Russian Geo Names Glossary* thus allowing reuse both by a human and machine reader.

- To integrate terminology resources from EuroTermBank into META-SHARE we had to ensure that their metadata was compliant with the META-SHARE metadata model. Metadata categories of EuroTermBank were aligned to the META-SHARE model and metadata descriptions of terminology collections were complemented to the META-SHARE schema in order to be interlinked for resource metadata harvesting by META-SHARE.

UHEL reports the following on upgrades:

- Enabling conversation transcripts that have not been aligned with the corresponding audio files be imported to Praat or to a LAT archive in the ELAN Annotation Format, EAF. LAT is a system of tools for the creation, archival, and presentation of complex annotations on video and audio resources, developed by the Max Planck Institute for Psycholinguistics in Nijmegen, and now installed for UHEL at CSC. UHEL's efforts were spent on mapping a TEI form of the highly multi-tiered ELFA corpus onto a heuristic timeline that preserves the synchronisation of the tiers and can be presented in LAT, though without the audio. The EAF form of ELFA may also be later synchronised with the audio by the owners of the corpus.
- Upgrading the Finnish Treebank. Particular effort was spent on making sure that the new FinnTreeBank 3, a new 70-million word dependency-syntactic parse bank of Finnish, is technically close to the CoNLL-X format (now trivially isomorphic at the cost of omitting references to original source locations).
- Upgrading the corpus Samples of Spoken Finnish (a new version of it will be soon available), working on improving its usability in LAT.
- Processing text corpora to be imported to the corpus search interface Korp that is being developed at UGOT. UHEL has Korp installed at CSC.
- Improving and extending the Helsinki Finite State Technology (HFST) software: We offer a new Python interface generated with SWIG through which the HFST library can be used more easily. Compiling HFST with Windows using MinGW tools is now also supported. Two new command line tools are added, a tagger 'hfst-tagger' and a pattern match parser 'hfst-pmatch2fst'. The regular expression parser 'hfst-regexp2fst' supports new operators epenthesis, term negation, containment), identity symbols and many XFST-type rules. The tool 'hfst-fst2txt' also supports new formalisms, now it is possible to write transducers to dot/graphviz and PCKIMMO format. All back-end libraries (SFST, OpenFst and foma) are now bundled with the HFST library, so that incompatible versions will not cause problems. Support for default symbols is added and handling of flag diacritics is improved. We also have a new implementation of the internal transducer format, making conversions between transducer formats more efficient.

HI

HI has spent considerable time on writing and improving resource documentation, since most of the resources were poorly documented – many of them lacking documentation in English altogether. Structured metadata has been written and uploaded for all the selected resources, and narrative documentation now exists for all the resources, both in English and Icelandic.

After completion of the 25 million word Icelandic Corpus that had been in preparation for a number of years, the corpus was tagged and lemmatised. Furthermore, the corpus was enhanced with detailed information on its textual sources. This information is presented in TEI conformant XML format.

Apart from this, the upgrading mainly consisted of the conversion of several resources from proprietary to standard formats. The *Tagged Icelandic Corpus*, the *Icelandic Frequency Dictionary* and *MÍM-Gold*, as well as the transcribed text of the *Parliamentary Speech Corpus*, were converted to TEI conformant XML-files. The terminology wordlists in the *Icelandic Term Bank* were converted to TBX format and the *ISLEX* dictionary to LMF format.

UCPH

Apart from extension and documentation of DanNet, the Danish wordnet, (reported on in D3.5) **UCPH** has mainly focused on upgrading the Danish lexical database STO into a Lexical Markup Framework: This update has made the information of the lexicon more readily accessible because the LMF format makes it much simpler for a potential user to understand and use it. STO is a Danish lexical database based on PAROLE with about 80,000 entries with morphological information, 43,000 entries with syntactic information, and about 10,000 entries with rich semantic information. STO was finished in 2005 and is updated regularly according to the official Danish orthography. The data is stored in a relational database and has until now been exported for users in a comma-separated flat format for morphology and a self-defined XML-format for syntax. LMF is the new export format for both morphology and syntax. Semantics will not be converted to LMF during the META-NORD project but UCPH has plans of converting some of the semantic information of STO to LMF format later this year. The Lexical Markup Language is an internationally well-known and accepted XML format and the ISO standard for Natural Language Processing (NLP) lexicons. (www.lexicalmarkupframework.org).

The upgrade task involved some complex decisions. The STO lexical database provides an intentional morphological description which means that each word form is not explicitly listed anywhere but the lexical entry is associated with a morphological pattern. Thus the word forms will be created on demand. Intentional morphological description is possible in LMF but the UCPH team has chosen to apply an extensional description of the morphology where all the word forms are created when the data is dumped from the database to LMF. The extensional approach was taken because it was seen as the most user friendly approach for an export format. In addition the process of converting to LMF includes structuring the linguistic information according to the predefined LMF schemata. At times these do not correspond very well with the structure used in STO and therefore generalisation or nested information needed to be expressed in new ways by means of features. When converting a lexical resource into LMF, the data categories selected were chosen from the Data Category Registry (DCR), ISOCAT (http://www.isocat.org/defined by ISO 12620). UCPH has used the existing DCR data categories as far as possible and in accordance with the guidelines of LMF and DCR they have also defined a new set of data categories in order to cover data category concepts needed for the STO conversion, which were not available in DCR. UCPH aims at a better exploitation of all the different parts of the lexicon including syntax and semantics. This is another good reason for upgrading STO to Lexical Markup Framework (LMF) in the context of META-NORD. The morphological part of the lexicon is currently being used by various companies and institutions, and an on-line user interface that allows the user to search for the different word forms of the lemma has been used a lot for teaching. Now that we can offer STO in LMF format, there is no doubt that the morphological part of the lexicon will be even more attractive for users in the future. The UCPH team is also planning to develop a morphological analyser/generator that makes use of all the morphological information in STO-LMF.

UiB has invested much time in writing and improving resource documentation, since the documentation of most of the resources was fragmented, scattered, outdated and/or incomplete. As an example of incompleteness, IPR issues and licenses needed to be cleared and formalized for most of the resources, even including many of those that had existing narrative documentation pages.

Only one of the resources documented by UiB in META-NORD had an existing structured documentation (the resource *NoWaC - Norwegian Web as Corpus*, which was documented in the OLAC Language Resource Catalog). Therefore, UiB concentrated on providing updated, structured metadata for each resource in English in META-SHARE. For complex resources, such as the multilingual treebanks (Aquis and Sofie), UiB found META-SHARE to be inadequate to describe the resources comprehensively[2]. Therefore, UiB has also spent considerable time in providing narrative documentation for the treebanks on the INESS pages.

Concerning upgrading, UiB has upgraded the *Sofie analyses* and the *Acquis* analyses to agreed standards in the context of task 3.4. In cooperation with the project INESS (www.iness.uib.no) treebanks for Bulgarian, Danish, Icelandic, English, Estonian, Finnish, Georgian, German, Norwegian and Swedish have been made accessible through a uniform web interface and state-of-the-art search tool. Some treebanks have been collected from the Nordic Treebanking Network and other sources, while some treebanks have been developed specifically for the META-NORD project. Some of the treebanks have been linked across languages creating a parallel treebank. In addition, rights for the original text material for the Sofie analyses, based on the novel 'Sofies verden' by Jostein Gaarder have been cleared with the relevant publishing houses, and metadata for all the treebanks has been provided in META-SHARE by META-NORD. UiB/META-NORD have also aligned the Norwegian translations of the Acquis, provided by the Norwegian Ministry of Foreign Affairs, with the Acquis Communautaire on the document level. For further information on the upgrading of the treebanks, see the report D3.4.

UiB ensured a technical format conversion for some resources: UiB converted the resource *UHR' Termbase for Norwegian higher education institutions* from html format into TBX format and the resource *SCARRIE lexical resource* into LMF format. For the seven resources from the Language Council of Norway, UiB encouraged the distributor, Språkbanken at the National Library of Norway, to convert them from html to csv format.

UGOT

UGOT reports the following on upgrade:

- All local corpora are now available in a standardised XML interchange format based on TEI. All corpora have been uniformly automatically segmented into sentence units and automatically annotated with part of speech, lemma plus morphological information (including some information about multi-word entities), and dependency syntax. Corpus import and export pipelines have been built, converting from the interchange format to the internal format used by the corpus search engine and vice versa, and integrating corpus annotation tools in a uniform framework. All software is freely available under an open-source license.

[2] See for instance UiB's post at the META-SHARE forum: http://www.meta-share.org/portal/forum/question/59/0

- The lexical resources have been interlinked using uniform standardised persistent identifiers for lexical forms (lemgrams), lexical senses, parts of speech, and inflectional paradigms. Interlinking is complete on the form level for all modern resources, but the word sense links have been automatically inferred on the basis of the form links, and have been only partly validated and corrected (manually).
- All lexical resources have been (partly manually) converted to LMF (Lexical Markup Framework; ISO 24613:2008).
- Corpus annotations as well as lexical information have been partly linked to ISOCat categories.
- Corpus search and lexical processing are now accessible through web-service (WS) based APIs, allowing rapid prototyping of language-aware web-based user interfaces, as illustrated by the intelligent computer-assisted language learning prototype application Lärka <http://spraakbanken.gu.se/larka/#exe=linguists&type=all&lang=en>, which communicates with the backend WS APIs of the corpus interface Korp <http://spraakbanken.gu.se/korp/#lang=en> and the lexicon search interface Karp <http://spraakbanken.gu.se/karp/#lang=en>. Korp is now being tried out at UHEL and HI. The WS APIs include an authentication mechanism for access to protected datasets by client applications. All software is freely available under an open-source license.
- Documentation of all resources has been improved. The web page structure of Språkbanken has been restructured so that documentation for local resources is partly automatically generated from the META-SHARE metadata describing the resource.

UT has focused on the standardisation of our WordNet and text corpora, which are made available in full, as well as on creating formal metadata and English documentation for our resources. Estonian Wordnet has been converted into MySQL database, and special scripts are worked out to convert EstWN into LMF, VISDIC XML and csv files that are used by UGOT to include EstWN into WordTies. The annotation of the Comprehensive Corpus of Estonian (in the META-SHARE repository it has been divided into the following subcorpora parts: Corpus of Estonian Fiction (5,6 million words), Corpus of Estonian Newspapers (182 million words), Corpus of Estonian scientific texts (5 million words), Corpus of Estonian law texts (11 million words) and Corpus of the Proceedings of Estonian Parliament (13 million words)) has been upgraded to conform to the TEI P5 XML standard. The annotation includes the metadata (TEI header), the structure of the document, i.e. its subparts, paragraphs and sentences; its author(s) and heading(s). In addition, the omissions have been replaced by a special tag. Also the annotation of sentence boundaries has been checked and improved during the project; this results in better quality of morphological disambiguation and parsing. Most of the resources had little or no information about them in English, usually the only information available was in Estonian narrative form. UT has created metadata for the language resources that were to be uploaded to META-SHARE and validated the metadata according to the META-SHARE schema.

Also **LKI** has carried out substantial upgrades for many of its resources, including:

- Office Language Recommendations and Language Consultancy Bank have been converted from legacy formats that utilised outdated encoding and packed in a structured downloadable package conforming to LMF standards.

- Database of proper names has been converted from a legacy format and restructured for improved indexing and searching. Export capabilities in LMF as well as generic XML and JSON formats have been added, as well as a new web-based interface.
- The lexical database of Standard Lithuanian language has been upgraded with export capabilities in LMF as well as generic XML and JSON formats, web interface has been improved making retrieval more efficient.
- Database Synonymy of Lithuanian Terms has been converted form legacy formats that utilised outdated encoding and packed in a structured downloadable package conforming to LMF standards. A web-based interface for easy retrieval has been added, as well as a facility enabling direct export to EuroTermBank compliant format.

5 Extending, linking and aligning resources

Extension of language resources has been taking place throughout the project period. As reported by all partners in the First Year Progress Report, a substantial part of the delivered resources has been extended, gaps have been filled out and their coverage has been improved. For example lexical resources such as FinnWordNet (UHEL) were substantially extended, and all Wordnet partners have established 5,000 links to the Princeton WordNet. The Swedish lexical resources available at UGOT have been continuously extended throughout the project period, with, e.g., the Swedish FrameNet growing by an order of magnitude in number of lexical units covered.

Extension work with corpora has also proceeded: the balanced Tagged Corpus of Icelandic has been extended with new genres (HI). The Latvian-English Legislation Corpus prepared by Tilde for the first batch of resources was upgraded at later states to agreed standards and extended with additional data. The corpus collection of Språkbanken (UGOT) has grown substantially and now contains more than one billion words of modern Swedish, including some added genres (Swedish blogs, Wikipedia, tweets, and a dialogue corpus). Most of the Swedish corpora have been catalogued in META-SHARE and are available for search through a web-based user interface and through a web service API, as well as for downloading as sentence-scrambled "sentence sets".

5.1.1 Extending and linking resources

In early spring 2012, UCPH initiated a questionnaire for the task on extending and linking corpora for language for special purposes (LSP, see Appendix B). To increase the suitability of both research and practical use, cooperation on extension of partners' LSP corpora with common domains and enhanced annotations would be an important step towards having comparable LSP corpora for the languages involved and thus a significant progress for terminology work in these domains. The aim was to find and extend corpora within some common domains, and provide them with a set of commonly agreed keywords, eventually keywords that can be found in Eurotermbank, for all the relevant languages, thus producing a set of comparable corpora.

It turned out that the only domain common to more than one partner in META-NORD was the law domain. As a result of this examination, it was decided to continue with the law domain among LKI, UT and TILDE. These three partners have created a linked set of Lithuanian, Latvian and Estonian legislature. Although each country has its own unique legal infrastructure, a significant number of laws and related documents in each country have their counterparts in other countries on similar topics/domains. This comparison and linking is further facilitated by wide availability of English translations. A list of 272 interlinked sets has been compiled, relevant documents have been collected and consistently labelled, and

their formats have been normalised. LKI consolidated the linked sets in a summary table, complete with relevant titles and/or topics, web links and document names for quick reference and packaged the structured set in an easily downloadable archive.

In the beginning we decided to make a digest for the legislation of the three Baltic countries – Estonia, Latvia and Lithuania – only. These three Baltic countries were not chosen randomly. They all are post-Soviet states. After the restoration of the Independence, the three countries created their new legal systems as well, and therefore legal documents of these countries were selected for comparison.

A total of 276 Latvian laws and their equivalents in English have been picked out. Later, we moved on to similar Lithuanian and Estonian laws and their English translations. Obviously, all these three countries for which documents have been chosen have their individual legal systems and unique, albeit similar law-making traditions.

The resource titled *Collection of Comparable Lithuanian, Latvian and Estonian Laws and Legislations* is aimed at disclosing how law terminology is forged within said legal systems, something that may provide some insights and useful information for the creators of legal terminology in all the three states, when it comes to developing new or amending existing terms. Later linguists and language technology researchers will be able to analyse the terminologies of these legal documents, as well as their very language: both the formal structure of terms, and their semantics as well. Comparative studies of legal documents in several languages both reveal the specifics of forging terms in other languages, and allow seeing the terminology of the national language. Therefore, comparative studies can produce a great deal of important information to terminology developers as well as developers of language technology engaged in designing interfaces for terms in different fields and languages.

Of course, different legal systems use different law terms and therefore, for all practical purposes, there can never be an absolute match between legislation items. Notions that are truly identical can only be defined by terminology of international law documents, whilst terms that can be found in the national law of different countries will always come with a lesser or greater degree of semantic differences.

In addition to linguists and language technology experts, this resource will provide a handy tool for specialists from other domains, such as legal advisers. This resource will serve as an indispensible reference in legal matters for those who operate on the Baltic markets.

The *Collection of comparable Lithuanian, Latvian and Estonian laws and legislations* collected with comparable law documents for 3 language pairs: English-Latvian, English-Lithuanian, English-Estonian. The size of each collection is provided in Table 6.

Table 6 Collection of comparable Lithuanian, Latvian and Estonian laws corpus.

Country	Laws	
	LV/LT/EE	EN
Latvia (LV)	276	276
Lithuania (LT)	232	189
Estonia (EE)	133	70

Terminological Synonyms of the Lithuanian Language database is a product of LKI that comprises sets of Lithuanian synonyms which has been expanded by adding English language counterparts to Lithuanian synonym sets. The database has also been updated and

extended by including new material. Presence of bilingual data has enabled sharing the resource within EuroTermBank framework, by uploading its data in bilingual TBX format per ETB specifications. The resource is available for download through the Meta Share system, as well as accessible on the web through a dedicated interface.

5.1.2 Aligning resources across languages

Task 3.3. was aimed at creating and documenting aligned resources across languages as well as gaining systematic knowledge about aligned resources that could be extended or created in the future. Thus the subtasks of Task 3.3. included:

1. Finding the resources which have already been aligned, documenting their annotation and availability;
2. Finding resources which could be easily aligned; estimating the volume of efforts and if possible, aligning resources;
3. Describing the potential resources which could be aligned in further future.

An extensive review of available language resources to find already aligned or potential multilingual resources was conducted in spring of 2012. The outcomes are presented in a table that can be seen in the Appendix C. The table gives an overview of the existing aligned resources containing language material in at least one of the META-NORD languages plus a few resources that could be the potential material for creating a aligned resource.

For creating (better) aligned resources (i.e. subtask 2) some of the resources that had been identified as candidates for cross-lingual alignment with relatively small effort were chosen from the review results. The following actions were undertaken:

1. Adding a new language to an existing parallel resource.

   UiB has added the Norwegian Acquis to the JRC, aligned at document level and added meta-data.

2. Creating new parallel resources.

   Tilde has contributed a parallel resource of comparable sentences for en-lv (116 240 sentences each), en-lt (179 758), lt-lv (29 370) and en-ee (128 939) language pairs extracted from Wikipedia;

   UT has created two aligned resources – an Estonian-Latvian parallel corpus of building product texts (10 276 sentences in each language) and aligned Estonian and Icelandic versions of WHO-s International Classification of Diseases (ICD-10).

3. Improving the quality of some existing resources. UT has done some work to achieve better quality of a parallel resource Open Subtitles corpus compiled by Jörg Tiedemann by finding near-duplicates automatically.

5.2 Linked Wikipedia

Wikipedia, as an online encyclopaedia, is a valuable source of similar or comparable documents as articles in Wikipedia are constantly updated and written in many languages. Documents which contain information on the same topic are also linked to each other. Therefore Wikipedia is widely used resource for different natural language processing tasks, including knowledge driven tasks, named entity recognition, terminology and lexicon extraction, sentence/document alignment and others.

On the other hand, although Wikipedia documents in different languages describe the same topic, articles in different languages differ a lot in size, structure and content, i.e., some

articles could be identical (could be classified as parallel or strongly comparable documents), while others could have very little in common (could be classified as weakly comparable documents). Thus the motivation of this was to provide set of documents which could be classified as strongly comparable and thus could be used for different multilingual NLP tasks.

In the ACCURAT project[3] tool that allows the collection of documents from Wikipedia, which are comparable or contain comparable segments, has been developed (Paramita et al., 2012). A technique to find comparable Wikipedia texts based on the idea that inter-lingually linked Wikipedia text pairs that contain significant numbers of shared anchor texts (i.e., links to other Wikipedia entries where these other documents are also inter-lingually linked by Wikipedia) are likely to be quite similar in content.

The *ACCURAT corpus of Wikipedia texts* collected with this tool contains comparable texts from Wikipedia for 12 language pairs: English-Croatian, English- Greek, English-Estonian, English-Latvian, English-Lithuanian, English-Romanian, English-Slovenian, Greek-Romanian, Latvian-Lithuanian, Romanian-German, Romanian-Lithuanian and German-English. The size of each collection is provided in Table 7.

Table 7 Size of the comparable Wikipedia corpora filtered from the full Wikipedia corpora.

Language pair	Documents	Unique Sentences		Unique Tokens in Unique Sentences	
		Source	Target	Source	Target
EL-EN	4,230	120,257	372,482	1,116,185	4,347,439
EL-RO	841	13,644	27,969	46,503	85,858
ET-EN	20,621	363,649	1,558,751	2,189,911	22,606,320
HR-EN	22,137	507,620	1,242,335	3,845,133	17,345,806
LT-EN	13,906	244,195	839,712	1,420,685	11,944,324
LT-LV	1,541	40,754	31,439	278,037	225,897
LV-EN	6,455	134,270	659,897	860,234	9,137,218
RO-DE	16,246	96,081	942,343	196,275	1,178,060
RO-EN	58,622	566,918	2,200,684	5,305,802	29,678,649
RO-LT	2,209	37,150	55,169	369,039	360,595
SL-EN	28,004	427,524	1,385,581	2,986,077	18,340,872
DE-EN	149,891	4,388,990	4,642,565	52,906,987	66,737,429

In addition to text files an alignment file for each language pair is provided. The alignment file contains linking information between source and target texts: each line contains source and target language filename separated by tabulator.

5.3 Cross-linked collection of comparable sentences from Wikipedia

The *ACCURAT corpus of Wikipedia texts* described in section 5.3.provides links between comparable texts in different languages. However, for multilingual natural language processing (NLP) tasks, especially machine translation, parallel or semi-parallel (parallel and strongly comparable) sentences are much more important. This is especially important for

[3] http://www.accurat-project.eu/

under-resourced languages (including the three Baltic languages of the META-NORD project) which lack sufficient amount of parallel data and thus in many cases data-driven methods are not applicable, e.g., output of statistical machine translation systems trained on these texts have low quality.

Semi-parallel sentences from comparable corpora can be extracted with the freely available *Accurat Toolkit* (Pinnis et al., 2012). The output of this tool is a text file containing semi-parallel and confidence scores for alignment. The confidence scores allows the user to decide which part of the document to use for a particular NLP task.

Four collections of semi-parallel sentences extracted from Wikipedia articles are presented on META-SHARE. These collections cover the following language pairs: Estonian- English (128939 sentence pairs), Latvian-English (116240 sentence pairs), Lithuanian-English (179758 Sentences) and Latvian- Lithuanian (29370 sentence pairs).

The following information is presented for each language pair: sentence in the source language, sentence in the target language and the confidence score (Table 8).

Table 8 Fragment of Latvian-English corpus of comparable sentences with confidence scores

Ir arī mākslīgās salas .	There are also artificial islands .	0.9672564
Maksimālais ātrums : 65 km/h	Maximum speed : 65 km/h	0.9654557
Viņam bija brālis un četras māsas .	He had one brother and four sisters .	0.9639728
Kluba krāsas ir sarkans , zils un balts .	Club colors are red , blue and white .	0.877546

6 Concluding remarks

To sum up the main results of this work package, much more resources for both academic and commercial R&D are now directly available for the Nordic and Baltic languages, and a considerable number of them are furthermore interoperable in *format* and *content.*

All in all, 509 tools and resources have been uploaded within the lifetime of the project. These include a broad range of different resources for different languages and language pairs and suitable for different purposes. However, in line of what was planned according to the DoW and further outlined in D2.4 on selected resources, the major focus in META-NORD has been on written resources. Nevertheless, there has been an increased focus on audio/video resources in the last project period, and thus the number of such resources has more than doubled since Batch 2. An example is the Danish NOMCO video corpus which includes annotations of first encounters. Nevertheless, audio/visual resources are still somewhat underrepresented in the META-NORD META-SHARE nodes, being only five percent of the total number of resources. This could be an indication of the fact that the Nordic and Baltic countries are somewhat under-resourced in this respect, but we also see a tendency of the speech technology field being commercialised to a much larger extent and therefore not as well supported in academia as written resources and tools. Increasing the number of audio/visual resources to provide a more complete picture of language resources in the Nordic and Baltic area could therefore be a fruitful future direction for the project partners.

7 References

- Bojtar, E. *Lithuanian-Hungarian dictionary*. Vilnius: The Institute of the Lithuanian Language, 2007.
- Dagienė, V., Grigas, G. Jevsikova, T. *Encyclopedic Dictionary of Computing*. Vilnius: TEV, 2008.
- Keinys, S. Bilkis, L. Paulauskas, J. Vitkauskas, V. *Modern LIthuanian Language dictionary*: digital versijon. Vilnius: The Institute of the Lithuanian Language, 2006; internet version, 2011. – http://dz.lki.lt
- Naktinienė, G. Paulauskas, J. Petrokienė, R. Vitkauskas, V. Zabarskaitė, J. *The Dictionary of the Lithuanian Language* (t. I–XX, 1941–2002): digital versijon. Vilnius: Institute of the Lithuanian Language, 2005.– www.lkz.lt.
- Paramita, M. L. Clough, P., Aker, A., Gaizauskas, R. Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles // Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, 21-27 May 2012, pp. 790-797.
- Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., Babych, B. Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora // Proceedings of ACL 2012, System Demonstrations Track, Jeju Island, Republic of Korea, 8-14 July 2012.

Appendix A: List of metadata

See document: D3.3-AppendixA.pdf

Appendix B: Questionnaire of LSP Corpora

See document: D3.3-AppendixB.pdf

Appendix C: Existing parallel aligned resources

See document: D3.3-AppendixC.pdf