META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Project no. 270899

Deliverable D2.2

Report on resources (actually or potentially) available to the consortium

Version No. 1.0

31/07/2011

Document Information

Deliverable number:	D2.2
Deliverable title:	Report on resources (actually or potentially) available to the consortium
Due date of deliverable:	31/07/2011
Actual submission date of deliverable:	31/07/2011
Main Author(s):	Gyri Smørdal Losnegaard, Anje Müller Gjesdal
Participants:	All
Internal reviewer:	UT
Work package:	WP2
Work package title:	Analysis and selection of language resources
Work package leader:	UT
Dissemination Level:	PU
Version:	1.0
Keywords:	Resources, meta data, data, IPR

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
0.1	23.05. 2011	Fishbone	UiB		Approved
0.2	01.07. 2011	First draft	UiB	Hanna Westerlund, Sigrún Helgadóttir, Krister Lindén, all META-NORD project participants	Approved
0.3	25.07. 2011	Final version	UiB	Hanna Westerlund, Sigrún Helgadóttir, Martha Dís Brandt, Dorte Haltrup Hansen, Kaili Müürisep, all META-NORD project participants	Approved
1.0	31.07. 2011	Final version	Tilde	Andrejs Vasiļjevs, Aivars Bērziņš	Submitted to PO

EXECUTIVE SUMMARY

This report describes the LRTs that have been identified and collected by the META-NORD consortium by project month M6. The resources to a large extent correspond to the set of resources described in DoW, and most resources are made available by the members of the consortium. As the project progresses, with the continuing of dissemination and the finalizing of the META-SHARE repository and editing tools, the partners are likely to encounter more potential resources. The report most probably includes only part of the resources that will be made available for META-SHARE.

Table of Contents

Abbreviations

Abbreviation	Term/definition
LRT	Language Resources and Technologies
HLT	Human Language Technologies
DoW	The META-NORD Description of Work-document
TILDE	TILDE SIA (Latvia)
UCPH	Københavns Universitet (Denmark)
UT	TARTU ULIKOOL (Estonia)
UIB	UNIVERSITETET I BERGEN ORGANISASJONSEDD (Norway)
UHEL	HELSINGIN YLIOPISTO (Finland)
HI	HASKOLI ISLANDS (Iceland)
LKI	LIETUVIU KALBOS INSTITUTAS (Lithuania)
UGOT	GOETEBORGS UNIVERSITET (Sweden)
CLARIN	Common Language Resources and Technology Infrastructure

Table 0.0.1. Abbreviations

1. Background

1.1 Project objectives: identifying and collecting resources

As stated in the "Description of Work" (DoW), one of the main objectives of the META-NORD project is to contribute to a pan-European digital resource exchange facility by **identifying** and **collecting resources in the Baltic and Nordic countries** and by documenting, processing, linking and upgrading them to agreed standards and guidelines.

The initiative comes from the acknowledgement that multilingual Europe has a rich and diverse linguistic heritage to preserve, while at the same time there is a widening technology gap between "big" and "small" languages. High fragmentation and a lack of unified access to language resources are key factors that hinder European innovation potential in language technology development and research, and "smaller" languages in particular face the danger of falling behind in the future European multilingual society. Enabling communication and cooperation across languages and securing users of any language equal access to information and knowledge is especially important to future-proof under-resourced European languages.

The META-NORD project aims to establish an open linguistic infrastructure in the Baltic and Nordic countries to serve the needs of the industry and research communities. The project will focus on 8 European languages–Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish–that each have less than 10 million speakers.

A particular focus of the project is targeted to the three horizontal action lines: treebanks, wordnets and terminology resources. The competences and resources of project partners will deliver an important contribution not only for the languages and communities of the Baltic and Nordic countries but also for other projects and institutions involved in the creation of an open linguistic infrastructure.

Language resources for META-NORD will be provided by project partner institutions which have a number of key resources in their possession, as well as by other institutions in partner countries addressed by project activities and willing to make their resources accessible through META-NET.

The current deliverable report on the Language Resources and Technologies (LRTs) that are actually and potentially available to the META-NORD consortium by project month M6. The LRTs are listed separately for each country in section 4, and a full description is provided in Appendix I. The resources are described according to a metadata schema being developed in META-NET, and is based on the T4ME/META-NET deliverable D7.2: "Specification of metadata-based description for language resources and technologies".

1.2 Baseline situation

The baseline situation for countries represented by the META-NORD consortium is presented in the DoW-tables 1.3.1 and 1.3.2. The tables include the most important language resources and tools for each country. The combined set of resources counts more than 100 LRTs, distributed between the countries as follows:

- Denmark: 18 (10 resources, 8 tools)
- Estonia: 18 (13 resources, 5 tools)
- Finland: 37 (18 resources, 9 tools)
- Iceland: 28 (19 resources, 9 tools)
- Latvia: 10 (8 resources, 2 tools)
- Lithuania: 9 (8 resources, 1 tool)
- Norway: 21 (12 resources, 9 tools)
- Sweden: 15 (13 resources, 2 tools)

The DoW additionally describes a set of 43 target outcomes with focus on treebanks, wordnets and terminology resources (the three horizontal action lines). Some of these resources, such as the Norwegian Treebank, are under construction or will be developed during the project period.

1.3 Target users and resources

The project aims to provide a solution that will allow different types of target user communities in the area of HLT to use language resources in their activities and provide the European community with innovative and, at the same time, sustainable applications, systems and tools. The target users are developers and researchers both in industry and academia. This includes private and public institutions, companies and individuals involved in HLT research and development: industrial organizations and SMEs, academic institutions, research organizations, universities, individual researchers and students, national governments, EC institutions, and private investors.

META-NORD commits to encompass a large variety of language resources, including language data, such as written and spoken corpora (annotated or in raw form, monolingual as well as multilingual), lexical and terminological databases, grammars, ontologies, etc.; language processing and annotation tools and technologies; metrics and evaluation protocols.

During identification and selection of resources the multilingual aspect of information access, processing and delivery will be taken account, i.e. multilingual resources which are available for all or most of META-NORD languages will have priority in the selection process. However, fully multilingual resources for META-NORD languages are currently available

only to a very limited extent. The target outcomes of the project will improve this situation by providing parallel, multilingual and cross-language linked resources.

The final list of resources accessible by the end of the project will be selected in Task 2.3. Resources are selected according to the intended focus of the project (target outcomes), their relevance and usability in multilingual services, quality and preferences of partners.

2. Identification and collection of resources

2.1 The collection process

The LRTs that have been collected by the META-NORD consortium and that are described in this report to a large extent correspond to the set of resources described in DoW (see section 1.2), with some additions. Most resources are made available by the members of the consortium. The partners have also identified resources owned by third parties on which they know they will be initializing negotiations during the project. With the META-NORD dissemination efforts to come, the partners are likely find more resources for later uploads, both already existing resources and resources resulting from new LT projects.

At the current stage of the project, legal matters related to IPR and restrictions of use, such as user licenses and agreements, are not yet fully resolved. The IPR specifications (found in Appendix I) should thus be considered preliminary and indicative. As legal issues are resolved, the partners will become better equipped to negotiate with third party resource providers. It should also be taken into account that some negotiations might fail even if it now seems that the resource will be potentially available.

The resource collection can be seen as a two-stage process; *identification* and *registration*, where registration is the process of recording resources in a spread sheet along with a description. In META-NORD, the tasks of resource collection, synchronization of metadata, and IPR issues partly overlap. This overlap mainly concerns the preparation of the resource description, which corresponds to the metadata schema developed by T4ME/META-NET for META-SHARE. META-NORD has been (and is still) involved in the process of developing this schema and the UHEL team as the coordinators of task 4.1 “IPR and other legal issues” has played an important part in this work.

During resource registration a question was raised whether resources that are available for search without any restriction and the text available for download under a certain license should be regarded as two different resources. META-NORD adopts the recommendation given when recording resources in CLARIN – treating such resources as different entities. The rationale behind this treatment is that the search interface and the underlying resource are actually two different resources: one is a tool (or a web interface) accessing the resource; the other is a set of data or a corpus. The search tool or web interface may be publicly accessible and the results of the search may be available with very light copyright restrictions, whereas the underlying resource may be severely restrictively available. In some cases the underlying

resource may not realistically be available at all in a foreseeable future. It will still be good to have the metadata on the resource for those who want to initiate negotiations with the resource owner, since they need to know whom to turn to, and the data to make a fair assessment of the difficulties ahead of them.

2.2 Resources made available via META-NORD

More than 50% of the collected resources are developed by the project partners themselves. These also include resources defined as target outcomes and which are not yet developed. Providing resources owned by the consortium partners gives the advantage of high availability, as these resources are already "at hand". Legal issues are also easier solved, unless legal restrictions apply to the underlying source material or the resource was originally developed for specific purposes or users.

2.3 Resources available through third party networks

A crucial measure of the project success is the number of external providers of language resources participating in the third parties networks and contributing to the META-NORD infrastructure as well as the number of resources made available via META-NORD.

Nearly 50% of the identified resources are owned by third parties; of these, 43 % of these are actually and 54% are potentially available, while for some resources the status is unclear. However, as mentioned in section 2.1, the progressing work on metadata and legal issues as well as project dissemination will prepare the grounds for the discovery of more third party resources and negotiation which may eventually lead to upload to META-SHARE towards the end of the project.

3. A common and shared resource description

META-NORD supports the goal of a common and shared resource description, and has consequently adopted the metadata schema developed in META-NET. The schema is an effort towards a shared metadata format for META-NET and the PSP projects CESAR, METANET4U and META-NORD. The metadata work is led by the META-SHARE group at T4ME/META-NET, and is based on the T4ME deliverable D7.2 "Specification of metadata-based descriptions for language resources and technologies" (Appendix II).

The schema used in this report is the one that was prepared for the META-SHARE demo that was launched at the META-FORUM 2011 in Budapest at the end of June. This schema approximates a minimal description and contains only a subset of a larger set of metadata elements which is described in Appendix III. The larger schema is still under development.

The element names used in the resource descriptions in Appendix I differ a bit from the ones used in the full metadata description (Appendix III). In the latter, the element names correspond to the names used in D7.2, which is not always the case with the schema used in this deliverable. It may thus be necessary to adapt our schema at a later stage in the project.

3.1 A Minimal metadata schema developed by T4ME/META-NET

	Definition	Recommended Values
META-SHARE collaborating project		
Source	The Organization providing the relevant info	
Resource Title	The complete title of the resource without any abbreviations	
Resource Name	A short name (e.g. acronym, abbreviation) to identify the language resource.	
IPRholder.organizationShortName		The Organization who holds the IPR
distributor.organizationShortName	the Organization distributing the resource	
contact.Person.surname	Surname of the contact person (anyone who can give further information on the resource); when more than one contact persons repeat the relevant columns	
contact.Person.givenName	Given name of the contact person (anyone who can give further information on the resource)	
contact.Person.email	Email of the contact person	
Availability	Terms of availability; please choose one of the recommended values; if restricted, please specify in restrictionsOfUse	available-unrestricetd use; available-restricted use; notAvailable
restrictionsOfUse	Restrictions o fuse; see recommended values for examples	academic-non Commercial Use; no Derivatives; share Alike; attribution; commercial Use (specify details); evaluation Use (specify details if needed); other
license	A description of the licensing condition under which the resource can be used; see recommended values for examples	Name of licence, e.g. CC Zero, CC-BY, etc. MSC (IF FOR META-SHARE ONLY). ELRA, LDC, GPL, etc.
distribution Medium	Specifies the format used for the delivery of the resource; if possible, use one of the recommended values; use ";" for multiple values	internet Browsing; download; CD-ROM; DVD-R; Blu-ray; hard Disk; paper Copy; other
licenseSignatory.Person.position	The position (director/head of dept/researcher/etc.) of the person in your	

	Definition	Recommended Values
	organisation authorised to sign the licence by which you make the resource available.	
ForeseenUse.foreseenUse	The use for which the resource has been produced. When more than one values use ";" in between	human use; NLP applications
ForeseenUse.useNLPspecific	The application for which it has been constructed; for indicative values, see recommended values. Use ";" for multiple values	speech analysis; Discourse analysis; Language identification; Speaker identification; Speaker verification; Speech recognition; Spoken dialogue systems; Voice control; Speech synthesis; Used in project; Face verification; Speech verification; User authentication; Face recognition; Automatic speech recognition; Automatic person recognition; Talking head synthesis; Avatar synthesis; Multimedia development; Voice control; Speech assisted video control; Information retrieval; Word sense disambiguation; Machine Translation; Named Entity recognition; Question answering; Automatic text generation and summarization; Document classification; Emotion recognition; Sign language recognition;
ActualUse.actualUse	The actual use of the resource in the framework of a specific project or application. Use ";" for multiple values	human use; NLP applications
ActualUse.useNLPspecific	The application in which it has been used; for indicative values, see recommended values. Use ";" for multiple values	speech analysis; Discourse analysis; Language identification; Speaker identification; Speaker verification; Speech recognition; Spoken dialogue systems; Voice control; Speech synthesis; Used in project; Face verification; Speech verification; User authentication; Face recognition; Automatic speech recognition; Automatic person recognition; Talking head synthesis; Avatar synthesis; Multimedia development; Voice control; Speech assisted video control; Information retrieval; Word sense

	Definition	Recommended Values
		disambiguation; Machine Translation; Named Entity recognition; Question answering; Automatic text generation and summarization; Document classification; Emotion recognition; Sign language recognition;
description (EN)	Description of the resource in prose	
resourceDocumentationInfo.publications	information on documentation (papers, etc.) of the resource	
resource Type	type of the resource; please use one of the recommended values	corpus; lexicalConceptualResource; languageDescription; technologyToolService
ttsSubtype	subtype of TechnologyToolService resources	tool; service; technology
ldSubtype	subtype of the languageDescription resources	
lcrSubtype	subtype of the lexicalConceptualResources	e.g. computationalLexicon
mediaType	Specification of the media type of the resource; can be multiple if the resource is a multimodal set; please use the recommended values; use ";" for multiple values	text; audio; video; image; tactile
modalityType		written; spoken, ….
noLanguages	An indication of the number of languages that are included in the resource.	if one language, then corpus is monolingual
multilingualityType	Whether the corpus is parallel or comparable	parallel; comparable
languageId	Identifier of the language as defined by ISO 639 that is included in the resource or supported by the tool/service. Use ";" for multiple values	ISO 639-3
Source Lg id	In case of multilingual resources	
Target Lg id	In case of multilingual resources	
languageDependency	for technologyToolService resources; whether they are language dependent or not	languageIndependent; languagedependent
Size	The size of the resource with regard to the SizeUnit measurement in form of a number.	

	Definition	Recommended Values
sizeUnit	Specification of the unit of size that is used when specifying the size; if possible, use one of the recommended values	word; token; byte; sentence; text; …
annotationType	Specification of the types of annotation levels (tiers) provided by the resource; if possible use recommended values; use ";" for multiple values	segmentation; alignment; structural annotation; lemmatization; stemming; PosTagging; bPosTagging…
LexicalConceptualResourceEncoding.encodingLevel	specifically for lexicalConceptualResource	lemma; part of speech; morphology; syntax-subcategorization frames; semantics-definition; semantics-semantic class; semantics-semantic features; semantics-semantic relations; thematic domain
annotationMode		
ResourceCreationInfo.completionDate	the year of the resource completion	
Version		

Table 3.1.1 Minimal metadata schema developed by META-NET for META-SHARE

3.2 Project specific additions to the schema

Table 3.2.1 describes additions to the META-NET schema. The changes are due to project specific documentation needs such as categories that have been used in META-NORD deliverables. The additions also include an element specifying whether the resources are actually or only potentially available to META-NORD. A mapping between the META-NORD type categories (used in the language whitepapers) and the type categories in META-NET has been proposed by the META-SHARE team. However, the mapping is not complete as some relations remain unsolved. It is also not clear whether there is an actual need to keep the whitepaper categories, but we decided to include them in our schema in order to cover all eventualities.

	Definition	Recommended Values
Actually or potentially available to the consortium		P; A
Type	The categories used in the resource evaluation for the language whitepaper	Data; tool
Subtype	The categories used in the resource evaluation for the language whitepaper	Tokenization, Morphology; Parsing; Sentence Semantics; Text Semantics; Advanced Discourse Processing; Information Retrieval; Information Extraction; Language Generation; Summarization, Question Answering, Advanced Information Access Technologies; Machine Translation; Speech Recognition; Speech Synthesis; Dialogue Management; Reference Corpora; Syntax-Corpora; Semantics-Corpora; Discourse-Corpora; Parallel Corpora, Translation Memories; Speech-Corpora; Multimedia and multimodal data; Language Models; Lexicons, Terminologies; Grammars; Thesauri, WordNets; Ontological Resources for World Knowledge; Other
IPR	Informal description of rights and restrictions	ACA; PUB; RES; To be negotiated
IPR-add	additions	NC, Inf, ReD

Table 3.2.1 Project specific additions to the schema

4. Resources actually or potentially available to the consortium

By M6 (July 2011) approximately 155 tools and resources have been identified by the META-NORD project partners. Of these LRTs, 92 are actually and 61 are potentially available to the consortium.

4.1. Latvia (TILDE)

In the table we describe 10 resources of which 10 are potentially available for the project.

5 resources are available for online browsing, 3 resources are available for download, and 1 resource is a tool to be used as a web-service, 1 toolkit will be available upon request.

One resource is a terminology database which can be used online; 4 resources are dictionaries which can be used online. Two resources are monolingual and parallel corpora. One resource - EASTIN-CL multilingual ontology – is a harmonized terminology of Assistive Technology domain in 7 languages.

Seven of the resources are developed or compiled by Tilde. Two resources are developed by the respective project consortium, Tilde being part of it.

The online lookup resources are available "as-is" or they will need to be integrated with META-SHARE. The corpora resources are identified, and need to be collected and converted to the appropriate standards. EASTIN-CL multilingual ontology will be available in spring of 2012.

Licensing of the online lookup resources – terms of use apply. Licensing conditions of the download resources – MSC-BYNCND license applies to resources from Tilde. The licensing conditions of resources developed by partners will need negotiations.

Resource name	Provider	Description	Availability
Eurotermbank	TILDE	101 term collections stored in database + 4 external interlinked termbanks	P
Lithuanian-Latvian dictionary	TILDE	Available as Web application	P
Latvian-Lithuanian dictionary	TILDE	Available as Web application	P
Estonian-Latvian dictionary	TILDE	Available as Web application	P
Latvian-English legislation corpus of Republic of Latvia	TILDE	Latvian-English legislation corpus of Republic of Latvia	P
Multilingual dictionary of person names	TILDE	Available as Web application	P
Tilde's POS-tagger	TILDE	Provides POS-tagging for the Baltic languages. Available as a Web Service	P
Corpus of Latvian literature	TILDE	Collection of novels, short novels, tales etc. by Latvian classics.	P
EASTIN-CL multilingual ontology	TILDE	Harmonized terminology of Assistive Technology domain in 7 languages	P

ACCURAT Toolkit	TILDE	Toolkit for collecting and processing comparable corpora developed in ACCURAT project	P

Table 4.1.1 Latvian resources actually and potentially available to META-NORD

4.2. Denmark (UCPH)

The resources and tools list accounted for by UCPH contains 10 resources and 8 tools. Most of the resources are basically monolingual and focus on Danish, such as the computational lexicon STO and the multimodal corpus NOMCO. 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. It should be observed that resources to be developed during the META-NORD project are also listed even if they have not yet been developed. This is the case for the linked wordnets between Danish, Finnish and Estonian and for some of the treebanks. The resources are available under a series of different licenses, some of which are open source and CC (DanNet), others which are more restricted, such as ACA (STO). Other licenses are still under negotiation such as licenses for the NOMCO Corpus, which is a corpus that is still being developed. With regard to ownership, most resources are owned by UCPH; some exceptions are, however, Reference Corpus for Danish and Copenhagen Dependency Treebanks.

With regard to tools, all listed tools are owned by UCPH and available under restricted use; some however are potentially available as open source (CC BY ...). It should be observed that the tools are generally not versioned since they are continuously optimized.

Resource name	Provider	Description	Availability
Danish wordnet	UCPH and DSL	A Danish lexical semantic wordnet with 65,000 synonym sets (concepts)	A
Cross-lingually linked resource		Cross-lingually linked WordNet	P
Two cross-lingually linked resources		Linked WordNets	P
SprogTeknologisk Ordbase	UCPH	Computational lexicon of Danish	A
Copenhagen Dependency Treebanks	CBS	Treebank for the Danish Parole corpus	A
The Copenhagen Danish-English Dependency Treebank	CBS	Parallel TreeBank for the translated Danish Parole corpus.	A
Danish first encounters NOMCO corpus	UCPH	All interactions are between two persons standing in the same studio. Two different recordings exist for each interaction with cameras placed at different angles. They have all been provided with Praat orthographic transcriptions.	P

Resource name	Provider	Description	Availability
Reference corpus for Danish	Danish Language Council	A contemporary, automatically collected, digitalised and annotated corpus consisting of Danish newspapers, periodicals, literature and specialist texts.	P
Corpus of sublanguage texts (2000 – 2010)	University of Copenhagen - CST and Danish Language Council	Danish corpus of sublanguage texts 2000-2010, 16 mi. words	A
Danish XLE grammar	CBS/UCPH	Advanced grammar for Danish.	A
CstTokeniser	UCPH	Sentence segmenter.	A
CstNER	UCPH	Named entity recognizer for Danish, high performance for Danish named entities.	A
CstTagger	UCPH	POS-tagger building on Brills tagger	A
CstLemma	UCPH	Lemmatiser using affix (= pre- in- and suffix) rules obtained by supervised training.	A
CstKeyExt	UCPH	Keyword extractor for Danish texts, also extracting multiword units.	A
CstNP-Rec	UCPH	NP -recogniser building on the Cass chunker.	A
CstRep	UCPH	Repetitiveness checker. The program finds sequences of tokens that occur more than once and uses a weight function to produce an ordered list.	A
HPSG –grammar	UCPH	Danish grammar based on HPSG using the LKB and cheap parsers	P

Table 4.2.1 Danish resources actually and potentially available to META-NORD

4.3 Estonia (UT)

In the table we describe 20 resources of which 9 are actually available to the project and 11 are potentially available. Four of the resources are tools, 2 of them are reference corpora, 4 are annotated text corpora, and 1 contains transcriptions of spontaneous speech, 1 parallel corpus, 1 speech database, 2 speech corpora, and 5 lexicographic databases. 11 resources or tools have been developed in UT, the other in the Institute of Cybernetics (IOC) at TTU or in the Institute of Estonian Language. One tool has been developed by private company Filosoft OÜ.

Most of the resources developed in UT are available to the consortium, except the Corpus of Spoken Estonian (which contains personal data and requires very secure authentication and licensing conditions) and some sub corpora in Estonian-English parallel corpus whose licensing conditions may be fuzzy. The licensing conditions of resources and tools developed by potential partners need negotiations.

Resource name	Provider	Description	Availability
The Comprehensive Corpus of Estonian	UT	Raw text corpus. The corpus consists of various genres of texts: fiction, newspaper, scientific texts, popular science, stenographic records of parliament speeches, Estonian and European legislative acts, chat room texts.	A
Treebank	UT	Raw /annotated ? Text corpus. A pilot project.	A
Estonian WordNet	UT	Lexicon. Manually compiled thesaurus, uses 45 different semantic relations.	A
BABEL Estonian Database	IOC	Speech corpus. The database consists of three sets: Many Talker Set, Few Talker Set, Very Few Talker Set	P
Corpora of morphologically disambiguated texts	UT	Raw text corpus. Sub corpus of the Comprehensive Corpus, manually annotated.	A
Corpora with shallow syntactic annotation	UT	Raw text corpus. Sub corpus of morphologically annotated corpus, manually annotated	A
Corpus of emotional speech	IEL	Speech corpus. The corpus has two objectives: to form an acoustic basis of corpus-based emotional text-to-speech synthesis; to constitute a reliable database for studying emotions rendered by speech.	P
Corpus of Institute of Estonian Language	IEL	Raw text corpus. Unannotated, mainly newspaper texts	P
Corpus of Spoken Estonian	UT	Speech corpus. Both tapes and transcriptions of spontaneous speech, containing mainly face-to-face conversations and institutional phone calls.	P
Cross-lingually linked resource	UT, UHEL	Cross-lingual WordNet	P
Dictionaries of Estonian-English, Estonian-Russian,	IEL	Dictionary	P
English-Estonian and Estonian-English parallel corpus	UT	Annotated corpus. This corpus contains: 1. Estonian laws and their translations into English. 2. EU legislation translated into Estonian	A
Estonian Foreign Accent Corpus	IOC	Speech corpus. Speech recordings of non-native speakers, 200 speakers, 136 isolated sentences and spontaneous speech.	P
Monolingual dictionaries	IEL	Orthological dictionary, dictionary of explanations, dictionary of idiomatic expressions.	P
Semantically disambiguated corpus	UT	Raw text corpus. Resource for building Estonian Framenet	A
The database of Estonian verbal multi-word expressions	UT	Lexicon. This database contains a subtype of multi-word expressions, namely those consisting of a verb and a particle or a verb and its complements.	A
Estonian text-speech synthesizer	IEL/IOC	Synthesizer. Windows program for synthesizing speech from written texts. Interface for blinds.	P
Morphological analyser	Filosoft	Morph analyzer. The tool finds the lemma and endings and gives all possible morphological readings found in the lexicon, includes a guesser. Correctness more than 99.8%.	P
Morphological analyser	IEL	Morph analyzer. Rule-based system for windows, not supported any more.	P
Morph syntactic	UT	Tagger/parser. Rule-based system uses VISL Constraint	A

Resource name	Provider	Description	Availability
disambiguator and shallow parser		Grammar parser engine developed in the University of Southern Denmark. Also includes a pilot version of dependency annotation.	

Table 4.3.1 Estonian resources actually and potentially available to META-NORD

4.4 Norway (UIB)

21 resources are identified for Norwegian, 13 actual and 8 potential. 6 resources are open access, 7 resources are restricted to academic purposes and the rest need to be negotiated. The resources cover spoken and written language, as well as both Nynorsk and Bokmål. Seven resources (Leksikografisk bokmålskorpus, Det nynorske tekstkorpuset, NHH Termbase, Norwegian-Vietnamese digital dictionary, Stadsnamnsamlinga, International Computer Archive of Modern and Medieval English, Norwegian Newspaper corpus) have been listed twice; as a tool (i.e. interface for internet browsing) and as data, referring to the underlying material that may be more difficult to access. Apart from these written corpora, there are two terminological databases, one speech corpus and two tools; a tagger and a corpus aligner. The resources are owned by different Norwegian Universities.

The Norwegian META-NORD team collaborates with Norsk språkbank – the Language Technology Resource Collection for Norwegian. Norsk språkbank and META-NORD share some common goals in the assembling and upgrading of HLT resources, and have a mutual benefit from cooperating on issues such as IPR and upgrading.

Resource name	Provider	Description	Availability
Leksikografisk bokmålskorpus	Uni Oslo	Corpus of Bokmål texts from 1985-today. Resource.	P
Leksikografisk bokmålskorpus	Uni Oslo	Corpus of Bokmål texts from 1985-today. Tool.	A
Det nynorske tekstkorpuset	Norsk ordbok 2014	Corpus of Nynorsk texts from 1866-today. Resource.	P
Det nynorske tekstkorpuset	Norsk ordbok 2014	Corpus of Nynorsk texts from 1866-today. Tool.	A
Akustisk database for norsk (NST)	Nasjonalbiblioteket	Acoustic databases. The database consists of five subsets of spoken data available for download via Språkbanken.	A
The Norwegian Spanish Parallel Corpus	Lidun Hareide	Parallell translational corpus (unidirectional).	P
NHH Termbase	NHH	English-Norwegian terminological database for economy and business administration. Resource.	P
NHH Termbase	NHH	English-Norwegian terminological database for economy and business administration. Tool.	A
Norwegian-Vietnamese digital dictionary	Universitetsforlaget, UiB/LLE, Uni Computing?	Bilingual dictionary. Resource.	P
Norwegian-Vietnamese digital	Universitetsforlaget, UiB/LLE, Uni	Bilingual dictionary. Tool.	A

Resource name	Provider	Description	Availability
dictionary	Computing?		
NST lexicon	Joint ownership between University of Oslo, University of Bergen, Norwegian University of Science and Technology, The Norwegian Language Council (Språkrådet) and IBM AS	Meta-lexicon complied of several resources.	A
Stadsnamnsamlinga	Uni Bergen	Place name database with home names from the county of Hordaland, spoken and written. Resource.	P
Stadsnamnsamlinga	Uni Bergen	Place name database with home names from the county of Hordaland, spoken and written. Tool.	A
Oslo-Bergen tagger	Uni Oslo/Uni Research	Grammatical and morphological tagger for Norwegian Nynorsk and Bokmål.	A
Terminology database	Standard Norge	Norwegian-English database of environmental terminology.	A
International Computer Archive of Modern and Medieval English	The collectors of the various corpora	Collection of corpora, written, spoken, historical	P
International Computer Archive of Modern and Medieval English	The collectors of the various corpora	Collection of corpora, written, spoken, historical	P
Norwegian Newspaper corpus	The newspaper publishers	The largest collection of Norwegian texts available for language studies. Dynamic corpus, extraction of new word forms (unregistered earlier). Distribution of hits by newspaper and year. Resource.	P
Norwegian Newspaper corpus	The newspaper publishers	The largest collection of Norwegian texts available for language studies. Dynamic corpus, extraction of new word forms (unregistered earlier). Distribution of hits by newspaper and year. Tool.	A
Translation Corpus Aligner 2	Uni Research	Software to prepare texts for parallel corpora.	A
Sofie Treebank	Uni Bergen	The Sofie Treebank is a parallel treebank that at completion will consist of material from ten North European languages; Danish, Dutch, English, Estonian, Faroese, Finnish, German, Icelandic, Norwegian and Swedish. There are tree-representations of all languages except Dutch, English and Finnish.)	A
Acquis communautaire	Uni Bergen	Aligned multilingual parallel corpus JRC-ACQUIS . The dataset contains resources for the following languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese,	A

Resource name	Provider	Description	Availability
		Romanian, Slovak, Slovene, Swedish.	

Table 4.4.1 Norwegian resources actually and potentially available to META-NORD

4.5 Finland (UHEL)

In the table D2.2 we present a list of 37 resources. 19 of these are available and 18 are potentially available. We have 9 tools, e.g. parsers and speech synthesis. Other resources consist of 4 wordnets, 1 ontological resource, 9 speech corpora and 14 text corpora. The corpora include newspaper texts, geographic names, speech samples of Finnish dialects etc. 3 of the corpora are annotated. 16 of the resources represent also other languages than Finnish (e.g. Russian or Swedish). About half of the resources are either open to public use via internet browsing/downloading or made available for academic purposes. At the moment the number of restricted resources is partly due to the large number of speech corpora and partly to practical issues in access procedure. Most resources are owned by different Finnish universities.

Resource name	Provider	Description	Availability
Lemmie	CSC	Concordance tool for web users.	A
UTA Cross-Language Information Retrieval System	UTA	Utaclir is a Java program, which translates IR queries from the source language to the target language. Utaclir utilizes external resources, for example a translation dictionary and a source language lemmatizer.	A
Russian-Finnish parallel corpus of literary texts	UTA	Russian literary texts (classical literature & 20th century) and their translations into Finnish aligned in paragraph level	P
Comparable Russian-Finnish corpus of juridical texts	UTA	Juridical texts in Russian and Finnish arranged as a comparable text corpus	A
Multilingual parallel corpus of juridical texts	UTA	International conventions and treaties arranged as a parallel corpus aligned on paragraph level	P
Finnish Text Collection (Kielipankki, Language Bank of Finland)	CSC	The Finnish Language Text Collection (Suomen kielen tekstikokoelma) is a selection of electronic research material that contains written Finnish from 1990's	A
Finnish-Swedish Text Collection (Kielipankki, Language Bank of Finland)	CSC	Prose, newspaper texts and administrative texts from 1997-2000	A
Other Speech corpora	UHEL; UEF; JYU; OY; UTA	Various kinds of speech corpora, audio recordings, e.g. argumentation, samples of Finnish dialects, cultural history, modern colloquial language, child language, Finnic minority languages etc.	P
Several written corpora	JYU	Several written corpora, e.g. students' essays, interviews and blogs	P
Written corpora (old literary Finnish)	KOTUS	Various corpora, e.g. early literary Finnish, literary classics, proverbs ...	A
Finnish TreeBank	UHEL	Example sentences covering most grammatical phenomena of Finnish, from VISK (the electronic version of contemporary Finnish grammar)	A
Cross-lingually linked resource	underNegotiation	Cross-lingual WordNet	A
Cross-lingually linked resource	underNegotiation	Cross-lingual WordNet	A
Cross-lingually linked resource	underNegotiation	Cross-lingual WordNet	A
Helsinki Finite-State	UHEL	The Helsinki Finite-State Transducer software is intended for the implementation of morphological analysers and	A

Resource name	Provider	Description	Availability
Transducer Technology		other tools which are based on weighted and unweighted finite-state transducer technology. Open Source Morphology for Finnish.	
Finnish WordNet	UHEL	Finnish WordNet based on Princeton WordNet 3.0	A
Samples of Spoken Finnish (Suomen kielen näytteitä)	KOTUS	Speech samples of several different Finnish dialects	A
The Finnish Broadcasting Company Corpus of Subtitles (YLE-korpus)	UEF	Digital research material of translated subtitles	A
Geographic Names Register of the National Land Survey	KOTUS	Lexicon / Knowledge Source	A
Corpus of translated Finnish (Käännössuomen korpus)	UEF	Monolingual comparable corpus containing both translated Finnish and texts originally written in Finnish.	A
Oulu corpus (Language Bank Of Finland)	CSC	The corpus contains a representative sample of the Finnish language in the 1960's media (excluding TV)	P
International Corpus of Learner Finnish (Kansainvälinen oppijansuomen korpus)	OY	A corpus of written learner language. The texts are written by students of Finnish as a foreign language from various language backgrounds. Consists of several different text types.	P
ProoF Corpus	UHEL	Both read-aloud and spontaneous speech from 100 speakers with Finnish as a foreign language	A
Corpus of Conversational Finnish (Keskusteluntutkimuksen arkisto)	UHEL	A collection of both everyday speech (e.g. phone calls, children playing) and institutional speech (e.g. political debates)	P
The Tampere Bilingual Corpus of Finnish and English	UTA	Corpus consists of fiction and non-fiction, e.g. extracts from both Finnish and English novels together with their translations. Also has an on-line search engine.	P
INTAS corpus (alias Finnish Dialogue Corpus)	UHEL	A corpus of spontaneous discussions and read-aloud performances from native Finnish speakers of different ages	A
Corpus of Spoken Southwestern Finnish	UEF	Audio corpus of spoken Finnish across the traditional Tavastia-Southwest dialect boundary	P
Finnish Telegraphese Corpus	UEF	Finnish telegraphese language (with English interlinears and translation)	P
Emotional speech (Emootiopuheen aineisto)	Aalto	Samples of vocal expression of emotion in continuous speech were gathered using simulated emotion portrayals of 9 professional stage actors.	A
Speech and EGG (electroglottography) simultaneous recordings (Puheen ja EGG:n samanaikaiset tallenteet)	Aalto	Continuous speech during which the functioning of vocal cords was studied	P
Open Source (Finnish) Morphology	UHEL	The Helsinki Open Source Morphology Project for various languages aims at implementing full-fledged morphological analysers for a number of languages using the Helsinki Finite-State Transducer Technology. The first large-scale implemented lexicon is an Open Source Finnish Morphology (OMorFi) but a number of other analyzers and generators based on open source resources for various languages have also been implemented.	A
Morfessor	Aalto	A morphological tool that discovers the regularities behind word forming in natural languages	A
National Semantic Web Ontology Project in Finland	Aalto, UHEL	The goal of this project is to lay a foundation for a national metadata, ontology, ontology service, and linked data framework in Finland	A

Resource name	Provider	Description	Availability
TKK Voice Source Analysis and Parametrisation Toolkit	Aalto	Voice inverse filtering and parameterization software for the MATLAB environment	P
Corpus of early modern Finnish (Varhaisnykysuomen korpus)	Kotus	This corpus includes different kinds of Finnish literature, magazines, newspapers and dictionaries, all published in the 19th century.	P
Finnish literature classics (Suomalaisen kirjallisuuden klassikoita)	Kotus	Includes prose, plays, poetry and aphorisms from important Finnish and Finnish-Swedish writers, dating from 1880's to 1930's.	P
Up-to-date word list of modern Finnish (Ajantasainen nykysuomen sanalista)	Kotus	An annotated word list of modern Finnish. Not supposed to be exhaustive, consists of less than 100 000 words. To be used e.g. as a helpful data for computer programs which can process Finnish.	P
Frequency list of words in written Finnish (Kirjoitetun suomen kielen sanojen taajuuslista)	Kotus	Finnish word frequency list based on the European Parole-corpus	P

Table 4.5.1 Finnish resources actually and potentially available to META-NORD

4.6 Iceland (HI)

In the table we describe 27 resources of which 16 are actually available to the project and 5 are potentially available. Three of the resources are tools, 5 are annotated text corpora, 5 contain synchronized text and speech, 2 contain language description, 3 contain lexicographic data and 3 contain data with semantic relations. Three of the resources will be available with different licenses for internet browsing and download. The resources have mostly been developed by partners of the Icelandic Centre for Language Technology (Reykjavík University, University of Iceland and The Árni Magnússon Institute for Icelandic Studies). Five of the resources come from sources outside these institutes.

Resource name	Provider	Description	Availability
CombiTagger	Reykjavík University	An open source tool, implemented in Java, for developing and evaluating combined taggers according to a given combination method.	A
IceNLP - Tagger, Parser, Lemmatizer	Reykjavík University	An open source Natural Language Processing (NLP) toolkit for analyzing and processing Icelandic text. The toolkit is implemented in Java and includes a tokeniser/sentence segmentiser, an unknown word guesser, a lemmatiser, a named entity recogniser, a linguistic rule-based tagger, a statistical tagger and a shallow parser.	A
Apertium-is-en Translation System	Reykjavík University	A shallow transfer rule-based Icelandic to English machine translation system.	A
Icelandic Frequency Dictionary Corpus	The Arni Magnusson Institute for Icelandic Studies	Tagged Icelandic corpus with about 590 thousand words, tagging hand corrected	A
Icelandic Frequency Dictionary Corpus	The Arni Magnusson Institute for Icelandic Studies	Tagged Icelandic corpus with about 590 thousand words, tagging hand corrected	A

Resource name	Provider	Description	Availability
Balanced Tagged Icelandic Corpus	The Arni Magnusson Institute for Icelandic Studies	25 million word corpus of text and transcribed speech, the corpus is automatically tagged	A
Balanced Tagged Icelandic Corpus	The Arni Magnusson Institute for Icelandic Studies	25 million word corpus of text and transcribed speech, the corpus is automatically tagged	A
A Gold Standard for PoS Tagging	The Arni Magnusson Institute for Icelandic Studies	One million word corpus of text, the corpus is automatically tagged and the tags are hand-corrected	A
Icelandic Parsed Historical Corpus	Reykjavík University	About 440.000 words of Icelandic text, from every century between the 12th and the 19th centuries inclusive annotated for phrase structure, part-of-speech-tagged and lemmatized.	A
The Jensson Corpus	Tokyo Institute of Technology	An Icelandic speech corpus based on a read bi-phonetically balanced text.	A
The Thor Corpus	Tokyo Institute of Technology	An Icelandic speech corpus; Read questions about the weather in Iceland	A
The Broadcast News RUV-1 Corpus	Tokyo Institute of Technology	An Icelandic speech corpus; Read sentences from the news domain	A
Parliament Speech Corpus	The Arni Magnusson Institute for Icelandic Studies	Icelandic speech corpus, 30 minutes of unprepared speeches from the Icelandic Parliament, synchronized text and sound files	A
Hjal Speech Corpus	Reykjavík University	Training material for a speech recognizer, collected and transcribed in 2003.	A
Pronunciation Dictionary for Icelandic	Reykjavík University	Pronunciation dictionary for Icelandic, transcribed in IPA and SAMPA	A
Database of Modern Icelandic Inflections	The Arni Magnusson Institute for Icelandic Studies	Comprehensive full form database of modern Icelandic inflections, containing about 280,000 pardigms with over 5,8 million inflectional forms	A
Database of Modern Icelandic Inflections	The Arni Magnusson Institute for Icelandic Studies	Comprehensive full form database of modern Icelandic inflections, containing about 280,000 pardigms with over 5,8 million inflectional forms	A
Database of Semantic Relations	Reykjavík University	A semantic database of Icelandic words, primarily nouns	A
Icelandic WordNet - Pilot Project	Reykjavík University	Icelandic Core WordNet, based on Princeton WordNet.	A
Íslenskur orðasjóður - Large Corpus	Deutscher Wortschatz, Leipzig University	An Icelandic corpus of more than 250 million running words collected from all domains ending in .is during the autumn of 2005, together with an automatically generated monolingual lexicon, comprising frequency statistics, samples of usage, cooccurring words and a graphical representation of the word's semantic neighbourhood.	A
Íslenskur orðasjóður - Large	Deutscher	An Icelandic corpus of more than 250 million	A

Resource name	Provider	Description	Availability
Corpus	Wortschatz, Leipzig University	running words collected from all domains ending in .is during the autumn of 2005, together with an automatically generated monolingual lexicon, comprising frequency statistics, samples of usage, cooccurring words and a graphical representation of the word's semantic neighbourhood.	
Icelandic Term Bank – Terminology	The Arni Magnusson Institute for Icelandic Studies	A collection of about 53 terminologies from different fields, containing terms in Icelandic, English and for some terminologies also in other languages.	P
Icelandic Term Bank – Terminology	The Arni Magnusson Institute for Icelandic Studies	A collection of about 53 terminologies from different fields, containing terms in Icelandic, English and for some terminologies also in other languages.	P
ISLEX - Icelandic Dictionary Base	The Arni Magnusson Institute for Icelandic Studies	The Icelandic part of a database of modern Icelandic for an online dictionary with translations in Danish, Norwegian, Swedish and Faeroese.	P
ISLEX - Icelandic Dictionary Base	The Arni Magnusson Institute for Icelandic Studies	The Icelandic part of a database of modern Icelandic for an online dictionary with translations in Danish, Norwegian, Swedish and Faeroese.	P
Ministry for Foreign Affairs - Translation Centre – Dictionary	Ministry for Foreign Affairs	Term collection of the Ministry of Foreign Affairs in Iceland, collected in connection with the translation of EU directives and regulations and other documents.	P
Ministry for Foreign Affairs - Translation Centre – Dictionary	Ministry for Foreign Affairs	Term collection of the Ministry of Foreign Affairs in Iceland, collected in connection with the translation of EU directives and regulations and other documents.	P
Íslenskt orðanet - Thesaurus		Database tracing semantic relations based on a large collection of word combinations.	P

Table 4.6.1 Icelandic resources actually and potentially available to META-NORD

4.7 Lithuania (LKI)

The list prepared by LKI consists of 8 resources and 1 tool. The main part of the resources represents the lexical system of Lithuanian: the history of lexica, neologisms, terms, and proper names. Thus they are in textual form and very language dependent. Regarding the legal aspect: IPR of 7 of these resources belong to LKI, IPR of Geoinformational Database of Toponyms to LKI with partner, and IPR of tool to creator. About half of the described resources can be accessed via Internet browsing, the others in hard discs. All of them are appointed for academic (not commercial) purposes. It is important to say that all resources are continually updated in both – technical and matter – sides.

Resource name	Provider	Description	Availability
Database of the Lexicon of Standard Lithuanian	LKI	www.lkz.lt Is creating database, with all grammatical markers and component parts of words into the search facility, based on this dictionary.	A
Modern Lithuanian Dictionary	LKI	It is intended for accumulation, administration, and scientific research of lexicographical data as well as other related information. It would be very popular dictionary.	P
Geoinformational Database of Toponyms	LKI	http://www.lki.lt/dlkz/ The most important explanatory universal normative dictionary of the Lithuanian language, accessible for the broadest range of users in Lithuania and foreign countries.	A
Database of a historical ethnic place names	LKI, co-authored with the Institute of Mathematics and Informatics	Information about two major units (toponyms and geographic objects) are provided via internet database. Information provided about a geographic object and information about a toponym.	A
Database of Neologisms	LKI	Collection of historical place names is associated with scientific research and place names selection. It helps to find out peripheral and marginal place names in their original form and origin.	A
Database Synonymy of Lithuanian Terms	LKI	It would be very popular; there would be quickly captured changes in language development.	A
Database of proper names	LKI	It is a lexicographic source of various terms for writing dictionaries. It is a convenient tool to collect the emerging material from the dictionaries. This database can be used for scientific research and can be an aid in the education law matters, editing, consultation and so on.	A
Morphological analyser, lemmatiser and synthesiser for Lithuanian	LKI	Place-names collected from colloquial and historical sources. The surnames collected from colloquial. Creation of the specific proper names database connected with their coordinates. Will be available source cards of dictionaries and proper names.	P

Table 4.7.1 Lithuanian resources actually and potentially available to META-NORD

4.8 Sweden (UGOT)

In the table we describe 13 resources and 2 resource collections. The first resource collection has the Swedish PAROLE corpus, the Talbanken treebank and the SUC corpus currently available. The second resource collection consists of Språkbanken 's other monolingual and parallel corpora which have not yet been made available but will eventually become available. All of these are (or will be) available under share alike licenses from Språkbanken. Two of the thirteen resources are tools, both of which are available to the consortium under the GPL 3.0 share alike license. These tools are owned by various members of CLT Gothenburg. The other eleven resources are also available to the consortium with similar share alike licenses (CC-BY-SA 3.0 and/or LGPL 3.0). One resource can be considered both a tool and a data collection, i.e. SB-LEX, which links the other ten lexical resources together: the Swedish FrameNet++, Swesaurus (i.e. fuzzy synsets from Synlex and SALDO), a Loan Word Typology list, two dictionaries (19th century and Old Swedish), two language engineering resources (one with morphological and syntactic information, the other with semantic information), an extensive lexicon for written modern Swedish (SALDO), a subset of examples from SALDO and a morphological resource of SALDO. These eleven resources and the two resource collections are owned by Språkbanken.

Resource name	Provider	Description	Availability
Dalin's morphological dictionary	Språkbanken	Dictionary of 19th century Swedish. Part of SBLEX.	A
Old Swedish morphology	Språkbanken	Dictionary of Old Swedish (Söderwall & Schlyter). Part of SBLEX.	A
Loan Word Typology list	Språkbanken	Loan Word Typology list. Part of SBLEX.	A
Preparatory Action for Linguistic Resources Organization for Language Engineering	Språkbanken	A language engineering resource with access to morphological and syntactic information in Swedish. Part of SBLEX.	A
Swedish Associative Thesaurus	Språkbanken	SALDO (Swedish Associative Thesaurus version 2) is an extensive lexicon resource for modern Swedish written language. Part of SBLEX.	A
Examples from the Swedish Associative Thesaurus	Språkbanken	SALDO examples. Part of SBLEX.	A
Swedish Associative Thesaurus' morphology	Språkbanken	SALDO's morphology. Part of SBLEX.	A
Semantic Information for Multifunctional Plurilingual Lexica	Språkbanken	A language engineering resource with access to semantic information in Swedish. Part of SBLEX.	A
Swedish FrameNet++	Språkbanken	The Swedish FrameNet++ project is an open-content integrated lexical resource for Swedish. Part of SBLEX.	A
Swesaurus	Språkbanken	Fuzzy synsets for Swedish using the lexical resources Synlex and SALDO. Part of SBLEX.	A
SB-LEX	Språkbanken	Linked lexical resources, including a framenet and a wordnet.	A
Språkbanken's corpora	Språkbanken	The Swedish PAROLE corpus, the Talbanken treebank (from Lund U.), the SUC corpus (from Stockholm U.)	A

Resource name	Provider	Description	Availability
		and the Swedish treebank (from Uppsala U.).	
Citation corpora	Språkbanken	Several of Språkbanken's monolingual and parallel corpora, scrambled on the word or sentence alignment level (so that the original text cannot be reconstructed), and annotated with state-of-the-art tools.	P
CLT Toolkit	Various	A collection of LT tools.	A
CLT Cloud	Various	REST web services.	A

Table 4.8.1 Swedish resources actually and potentially available to META-NORD

5. Conclusions

This report describes the LRTs that have been identified and collected by the META-NORD consortium by project month M6. At his early stage of the project, all resources potentially relevant may not have been covered, but the report should provide a sound basis for further work.

The report covers resources and tools, and indicates that much needs be done to make available the data underlying several of the tools. Most resources will be provided by Consortium partners.

A particular focus of the Meta-Nord project is targeted to the three horizontal action lines: treebanks, wordnets and terminology resources. These target outcomes will be delivered at a later stage so all resources are not listed in WP 2.2. However, the existing data indicate with regards to these targets, the resources reported on indicate a reasonably good situation. There is material for the development of parallel treebanks, wordnet resources are lacking for several languages while terminological resources are concentrated around EuroTermBank.

6. References

META-NORD *Grant agreement for: CIP-Pilot actions. Annex I – "Description of Work"*. December 2010.

META-NET deliverable D7.2. *Specification of metadata-based description for language resources and technologies*. January 2011.

META-NET homepage: <www.meta-net.eu>

Conceptual map from UHEL: <http://cmap.helsinki.fi:8001/rid=1JHDLXN8Z-JS14FY-1B82/META-NET%20metadata.cmap>

7. List of tables

8. Appendices