META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Project no. 270899

Deliverable 2.3

Report on methodology and criteria followed for the selection of resources

Version No. 1.0
30/09/2011

Document Information

Deliverable number:	D2.3
Deliverable title:	Report on methodology and criteria followed for the selection of resources
Due date of deliverable:	30/09/2011
Actual submission date of deliverable:	03/10/11
Main Author(s):	Kaili Müürisep
Participants:	All
Internal reviewer:	Tilde
Workpackage:	WP2
Workpackage title:	Analysis and selection of language resources
Workpackage leader:	UT
Dissemination Level:	PU
Version:	1.0
Keywords:	Resources, criteria, meta-data

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
0.1	08/08/ 2011	Fishbone	UT	Kaili Müürisep, Krista Liin	Draft
0.2	09/08/ 2011	First draft	UT	Kaili Müürisep, Krista Liin, Aivars Bērziņš	Stable draft
0.4	27/09/ 2011	Final-near	UT	All partners	Final changes
1.0	03/10/ 2011	Final	Tilde	Submitted to EC	Submitted to EC

EXECUTIVE SUMMARY

This report describes the methodology and criteria for the selection of resources to be used in WP3. The document evaluates the LRTs that have been identified and evaluated by the META-NORD consortium by project month M6. The evaluation has been carried out using the criteria suggested by META-NET Network of Excellence and META-SHARE project. Altogether, 151 LRTs were evaluated based on these criteria.

Table of Contents

1 Background 5

1.1 Project objectives 5

1.2 Baseline situation 5

2 Selection criteria 5

3 Selection process 7

3.1 Evaluation of known resources 7

3.2 Latvia (TILDE) 8

3.3 Denmark (UCPH) 9

3.4 Estonia (UT) 10

3.5 Norway (UIB) 12

3.6 Finland (UHEL) 14

3.7 Iceland (HI) 17

3.8 Lithuania (LKI) 20

3.9 Sweden (UGOT) 21

3.10 Identification of resources which could be potentially included in the database of LRTs 23

4 Conclusions 23

5 References 24

6 List of tables 24

Abbreviations

Abbreviation	Term/definition
LRT	Language resources and tools
DoW	The META-NORD Description of Work document
CC	Creative Commons
TILDE	TILDE SIA (Latvia)
UCPH	Københavns Universitet (Danmark)
UT	Tartu Ülikool (Estonia)
UIB	Universitetet i Bergen Organisasjonsedd (Norway)
UHEL	Helsingin Yliopisto (Finland)
HI	Haskoli Islands (Iceland)
LKI	Lietuviu Kalbos Institutas (Lithuania)
UGOT	Göteborgs Universitet (Sweden)
LRT	Language Resources and Technologies
IPR	Intellectual Property Rights
CLARIN	Common Language Resources and Technology Infrastructure
BLARK	The Basic Language Resource Kit

Table 1. Abbreviations

1 Background

The purpose of this document is to describe the methodology and criteria to be used for the selection of resources in WP3.

1.1 *Project objectives*

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 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 have less than 10 million speakers each.

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 methodology and criteria to be used for the selection of resources for the project.

1.2 *Baseline situation*

The META-NORD consortium has identified and collected the preliminary list of LRTs 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. 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.

2 Selection criteria

Top-level criteria for selection of resources will include availability, popularity, suitability of resources for technology and product/application development, fitness for multilingual purposes, longevity, quality and extensibility. Based upon the agreed criteria the consortium will select the best possible mix of resources that will make the subject of further work.

These criteria have been suggested by META-NET Network of Excellence (Rehm, 2010; Piperidis, 2010).

1. **Availability**: this criterion includes restrictions of uses, licenses, distribution medium. 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. Resources to be included in META-SHARE should ideally be available in the open domain. The copyright conditions of the initial raw resource should be known and documented; ideally they should be copyright free or accompanied by a permissive license. Likewise, processed and derivative resources should ideally be open at least for research purposes, allowing their re-use, reengineering, repurposing, etc. However, commercial use should also be allowed, unless solid justification of restrictions exists. In such a case, resources should be available under fair conditions to all prospective users.
2. **Suitability**: this criterion defines the aim of the resource or the tool describing its target use (for humans or NLP applications), the application for which it has been developed. The preferred resources and tools serve the language technology development.
3. **Multlilinguality**: the resources and tools may be monolingual, parallel, comparable, language independent, etc.; the preferred resources and tools support multilingualism and the linking between languages.
4. **Longevity**: the development of resources and tools may be in different stages: they can be actually in use, depreciated or under development. The important criterion for selection is that they are being maintained or supported to ensure extensibility, reusability and repurposing.
5. **Quality**: LRTs have different quality levels: they may be manually or automatically annotated, have gone through rigorous testing or still under development. The high quality LRTs are given a preference.
6. **Extensibility**: the preferred LRTs have been (ideally) adequately documented and described with a standardized metadata schema.

Priority is given to language data and tools, considered the core components of the language technology infrastructure, followed by evaluation packages, services and workflows that integrate them. The above mentioned criteria should not be considered restrictive as they cover a wide range of resource and media types i.e.:

- monolingual text and audio corpora, raw and annotated at any level;
- bi-/multilingual (comparable and parallel) text corpora;
- audio and multimodal corpora;
- mono-/ bi-/ multilingual lexica;
- basic language processing tools (tokenizers, sentence splitters, morphological analyzers, multi-level (sentence-word-phrase) text aligners etc.);
- various text analytics tools (syntactic analyzers, semantic taggers, named entity recognizers etc.);
- audio and multimodal processing tools;
- language models etc.

3 Selection process

3.1 Evaluation of known resources

The partners have evaluated their LRTs using the below mentioned criteria.

Availability is defined as follows:

2 – the LRT is available to the consortium and it is freely, openly available under sensible; Open Source or Creative Commons licenses that allow re-use and re-purposing;

1 – the LRT is potentially available to the consortium and its licenses need negotiations;

0 – the LRT has restricted access.

Suitability of LRT for LT:

1 – LRT serves language technology development;

0 – LRT is theory-oriented or designed for human users.

Multilinguality:

1– LRT supports multilingualism or linking between languages;

0 – LRT does not support it.

Longevity:

1– LRT is actively maintained;

0 – LRT is unmaintained.

Quality of LRT:

2 – high quality, extensively tested;

1 – moderately tested LRT, with some room of improvement;

0 – low quality or untested.

Extensibility of LRT

1 – sufficiently documented and described with standardized metadata schema;

0 – undocumented

3.2 Latvia (TILDE)

Latvian language resources presented in table 3.1.1 contain resources developed or hosted at TILDE as well as created through several EU projects, e.g., FP7 project ACCURAT and CIP-ICT-PSP project EASTIN-CL. Most of these resources are publicly available. However, due to the limitations set by authors on these works (IPR restriction), these resources could be accessed only as web service, but not downloadable. Two corpora – a parallel corpus of legislation of the Republic of Latvia and a corpus of the Latvian literature (containing works which are not IPR protected anymore) – are available under CC licenses.

Latvian language resources and tools listed in the table are well maintained, are of good quality, suitable for LT development and support multilingualism. However, only few of them are well documented and most of them are developed using proprietary metadata schema. Thus, in most cases Latvian LRTs need to be upgraded to the standards and documented.

Table 3.2.1 Latvian resources evaluated by selection criteria

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
Eurotermbank	TILDE	1 (2)*	1	1	1	1	1
Lithuanian-Latvian dictionary	TILDE	1 (2)*	1	1	1	1	0
Latvian-Lithuanian dictionary	TILDE	1 (2)*	1	1	1	1	0
Estonian-Latvian dictionary	TILDE	1 (2)*	1	1	1	1	0
Latvian-English legislation corpus of Republic of Latvia	TILDE	2	1	1	1	1	0
Multilingual dictionary of person names	TILDE	1 (2)*	1	1	1	1	0
Tilde’s POS-tagger	TILDE	1	1	0	1	2	1
Corpus of Latvian literature	TILDE	2	0	0	1	1	0
EASTIN-CL multilingual ontology	TILDE	2(1)**	1	1	0	1	0
ACCURAT Toolkit	TILDE	2	1	1	1	1	1

* Resource is available for online browsing or as web service.

** Resource is under construction, availability is to be clarified.

3.3 Denmark (UCPH)

Danish language LRTs presented in table 3.1.2 contain mainly resources provided by UCPH, but also Copenhagen Business School (CBS) and Danish Language Council. It should be notes that resources to be developed during the META-NORD project are also listed even if they have not been developed yet. The Danish wordnet, Danish Treebank and parallel Treebank are available to the consortium, the licensing conditions of other LRTs need further negotiation. The Danish LRTs serve language technology development, are multilingual, well maintained, are of good quality. Most of LRTs miss meta-data descriptions, although have been well documented.

Table 3.3.1 Danish resources evaluated by selection criteria

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
Danish wordnet	UCPH and DSL	2	1	1	1	1	1
Cross-lingually linked resource	META NORD	n/a	n/a	n/a	n/a	n/a	n/a
Two cross-lingually linked resources	META NORD	n/a	n/a	n/a	n/a	n/a	n/a
SprogTeknologisk Ordbase	UCPH	1	1	0	1	1	1
Copenhagen Dependency Treebanks	CBS	2	1	0	1	1	0*
The Copenhagen Danish-English Dependency Treebank	CBS	2	1	1	1	1	0*
Danish first encounters NOMCO corpus	UCPH	1	1	0	1	1	1
Reference corpus for Danish	Danish Language Council	0	1	0	0	1	1
Corpus of sublanguage texts (2000 – 2010)	University of Copenhagen - CST and Danish Language Council	1	1	0	0	1	1
Danish XLE grammar	CBS/UCPH	1	1	0	0	1	0*
CstTokeniser	UCPH	1	1	0	0	1	0*
CstNER	UCPH	1	1	0	0	1	0*
CstTagger	UCPH	1	1	0	1	1	0*

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
CstLemma	UCPH	1	1	0	1	1	0*
CstKeyExt	UCPH	1	1	0	0	1	0*
CstNP-Rec	UCPH	1	1	0	0	1	0*
CstRep	UCPH	1	1	0	0	1	0*
HPSG –grammar	UCPH	1	1	0	0	1	0*

* Documented, but without standardized meta-data

3.4 Estonia (UT)

The list of resources and tools of Estonia contains 20 items, 9 of them are actually available. Most of LRTs are suitable for language technology development (only 3 of potentially available resources may be described as hard to fit for LT). The monolingual LRTs are dominant (16 monolingual vs. 4 multilingual). Most of the resources are well supported and maintained, except 2. Most LRTs are in active use and well tested, although only 7 are of high quality. The documentation of resources is sufficient and their format is well defined. The documentation of some resources of the third parties network may need further elaboration.

Table 3.4.1 Estonian resources evaluated by selection criteria

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
The Comprehensive Corpus of Estonian	UT	2	1	0	1	2	1
Treebank	UT	2	1	0	1	1	1
Estonian WordNet	UT	2	1	1	1	2	1
BABEL Estonian Database	IOC	1	1	0	1	2	1
Corpora of morphologically disambiguated texts	UT	2	1	0	1	1	1
Corpora with shallow syntactic annotation	UT	2	1	0	1	1	1
Corpus of emotional speech	IEL	1	0	0	1	1	0

Resource name	**Provider**	**Availability**	**Suitability**	**Multilinguality**	**Longevity**	**Quality**	**Extensibility**
Corpus of Institute of Estonian Language	IEL	1	1	0	1	1	0
Corpus of Spoken Estonian	UT	0	1	0	1	2	1
Cross-lingually linked resource	UT, UHEL	1	1	1	1	0	1
Dictionaries of Estonian-English, Estonian-Russian,	IEL	1	0	1	1	2	1
English-Estonian and Estonian-English parallel corpus	UT	2	1	1	1	1	0
Estonian Foreign Accent Corpus	IOC	1	1	0	1	1	0
Monolingual dictionaries	IEL	1	0	0	1	2	0
Semantically disambiguated corpus	UT	2	1	0	0	1	0
The database of Estonian verbal multi-word expressions	UT	2	1	0	1	1	1
Estonian text-speech synthesizer	IEL/IOC	1	1	0	1	1	1
Morphological analyzer	Filosoft	0	1	0	1	2	1
Morphological analyzer	IEL	1	1	0	0	1	0
Morph syntactic disambiguator and shallow parser	UT	2	1s	0	1	1	0

3.5 Norway (UIB)

21 resources are identified for Norwegian, 13 actual and 8 potential. 6 resources have open access, 7 resources are restricted to academic purposes and the rest need to be negotiated. Seven LRTs have been listed twice – as a tool and as data, referring to the underlying material that may be more difficult to access. 13 LRTs fit well for language technology development, 12 LRTs support multilinguality. Most of LRTs are in active use and well tested. The LRTs have been well documented but in some cases they lack meta-data descriptions.

Table 3.5.1 Norwegian resources evaluated by selection criteria

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
Leksikografisk bokmålskorpus Downloadable	Uni Oslo	1	1	0	1	1	1
Leksikografisk bokmålskorpus Searchable	Uni Oslo	2	0	0	1	1	1
Det nynorske tekstkorpuset Downloadable	Norsk ordbok 2014	1	0	0	1	1	1
Det nynorske tekstkorpuset Searchable	Norsk ordbok 2014	2	0	0	1	1	1
Akustisk database for norsk (NST)	Nasjonalbiblioteket	2	1	1	1	1	1
The Norwegian Spanish Parallel Corpus	Lidun Hareide	1	1	1	1	1	1
NHH Termbase Downloadable	NHH	1	1	1	0	1	1
NHH Termbase Searchable	NHH	2	1	1	0	1	1
Norwegian-Vietnamese digital dictionary Downloadable	Universitetsforlaget, UiB/LLE, Uni Computing	1	0	0	0	1	1
Norwegian-Vietnamese digital dictionary Searchable	Universitetsforlaget, UiB/LLE, Uni Computing	2	0	0	0	1	1
NST lexicon	Joint ownership between University of Oslo, University of Bergen, Norwegian	2	1	1	1	2	1

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
	University of Science and Technology, The Norwegian Language Council (Språkrådet) and IBM AS						
Stadsnamnsamlinga Downloadable	Uni Bergen	1	0	0	1	1	1
Stadsnamnsamlinga Searchable	Uni Bergen	2	0	0	1	1	1
Oslo-Bergen tagger	Uni Oslo/Uni Research	2	1	1	1	2	1
Terminology database Snorre	Standard Norge	2	0	1	1	2	1
International Computer Archive of Modern and Medieval English Downloadable	The collectors of the various corpora	2	1	0	0	2	1
International Computer Archive of Modern and Medieval English Searchable	The collectors of the various corpora	2	1	0	0	2	1
Norwegian Newspaper corpus Downloadable	The newspaper publishers	1	1	1	1	2	1
Norwegian Newspaper corpus Searchable	The newspaper publishers	2	0	1	1	2	1
Translation Corpus Aligner 2	Uni Research	2	1	1	1	1	0
Sofie Treebank	Uni Berg	2	1	1	1	0	1
Acquis communautaire	Uni Bergen	2	1	1	0	0	0

3.6 Finland (UHEL)

Table 3.6.1 contains 40 LRTs. The evaluation of 6 LRTs will be done later. The 3 cross-lingually linked resources will be developed during the META-NORD project. 11 LRTs serve the language technology well, 13 are multilingual, 18 LRTs are in active use. Most of the LRTs are of high quality but still have room for improvement, also the documentation and meta-data information are often lacking. The LRTs meeting most of the criteria are Finnish Texts Collection (the availability needs negotiation), Finnish Treebank, transducer technology tools, and Finnish Wordnet.

Table 3.6.1 Finnish resources evaluated by selection criteria

Resource name	Provider	Availability	Suitability	Mulitlinguality	Longevity	Quality	Extensibility
Lemmie	CSC	½	0	1	1	1	1
UTA Cross-Language Information Retrieval System	UTA	to be evaluated later					
ParRus: Russian-Finnish parallel corpus of literary texts	UTA	1	0	1	1	1	1/0
MultiJur: Multilingual Parallel Corpus of Legal Texts	UTA	1	0	1	1	1	1/0
FiRuLex: Finnish-Russian Comparable Corpus of Legal Texts	UTA	1	0	1	1	1	1/0
Finnish Text Collection (Kielipankki, Language Bank of Finland)	CSC	1/0	1	0	1	1	1/0
Finland-Swedish Text Collection (Kielipankki, Language Bank of Finland)	CSC	1/0	1	0	1	1	1/0
Other Speech corpora	UHEL; UEF; JYU; OY; UTA	to be evaluated later					
Several written corpora	JYU	to be evaluated later					
Written corpora (old literary Finnish)	KOTUS	2	0	1	1	1	1/0
Finnish TreeBank	UHEL	2	1/0	0	1	1	1
Cross-lingually linked resource	under negotiation	2	n/a	n/a	n/a	n/a	n/a

Resource name	Provider	Availability	Suitability	Mulitlinguality	Longevity	Quality	Extensibility
Cross-lingually linked resource	under negotiation	2	n/a	n/a	n/a	n/a	n/a
Cross-lingually linked resource	under negotiation	2	n/a	n/a	n/a	n/a	n/a
Helsinki Finite-State Transducer Technology	UHEL	2	1	1	1	1	1/0
Finnish WordNet	UHEL	2	1/0	1	1	1	0*
Samples of Spoken Finnish (Suomen kielen näytteitä)	KOTUS	2	0	1	1	1	1/0
The Finnish Broadcasting Company Corpus of Subtitles (YLE-korpus)	UEF	to be evaluated later					
Geographic Names Register of the National Land Survey	KOTUS	1/0	0	1	1	1	1/0
Corpus of translated Finnish (Käännössuomen korpus)	UEF	1/0	0	1	0	0	1/0
Oulu corpus (Language Bank Of Finland)	CSC	0	0	0	0	1	1/0
International Corpus of Learner Finnish (Kansainvälinen oppijansuomen korpus)	OY	0	0	0	0	0	1/0
ProoF Corpus	UHEL	0	0	0	0	1	1/0
Corpus of Conversational Finnish (Keskusteluntutkimuksen arkisto)	UHEL	1	0	0	0	0	1/0
The Tampere Bilingual Corpus of Finnish and English	UTA	1	0	1	1	1	1/0
INTAS corpus (alias Finnish Dialogue Corpus)	UHEL	0	0	0	0	0	1/0
Corpus of Spoken Southwestern Finnish	UEF	to be evaluated later					
Finnish Telegraphese Corpus	UEF	to be evaluated later					
Emotional speech (Emootiopuheen aineisto)	Aalto	1	1	0	0	1/0	1/0

Resource name	Provider	Availability	Suitability	Mulitlinguality	Longevity	Quality	Extensibility
Speech and EGG (electroglottography) simultaneous recordings (Puheen ja EGG:n samanaikaiset tallenteet)	Aalto	2	0	0	0	1	0
Open Source (Finnish) Morphology	UHEL	2	1	0	1	0	1
Morfessor	Aalto	2	1	0	0	2	1/0
National Semantic Web Ontology Project in Finland	Aalto, UHEL	2	1	0	1	1	1/0
TKK Voice Source Analysis and Parametrisation Toolkit	Aalto	2	0	0	0	2	1
Corpus of early modern Finnish (Varhaisnykysuomen korpus)	Kotus	1	0	0	0	0	1/0
Finnish literature classics (Suomalaisen kirjallisuuden klassikoita)	Kotus	1	0	0	0	0	1/0
Up-to-date word list of modern Finnish (Ajantasainen nykysuomen sanalista)	Kotus	2	1	0	1	1	1/0
Frequency list of words in written Finnish (Kirjoitetun suomen kielen sanojen taajuuslista)	Kotus	2	1/0	0	0	0	1/0
ParFin: Finnish-Russian parallel corpus of literary texts	UTA	1	0	1	1	1	1/0
TamBiC: English-Finnish-English text corpus	UTA	1	0	1	1	1	1/0

* Documented, but without standardized meta-data

3.7 Iceland (HI)

In the table we describe 27 resources of which 16 are actually available to the project and 5 are potentially available. Most of the LRTs are suitable for language technology development purposes, but only some of them are multilingual. 14 LRTs are well maintained. Most of the LRTs are of high or normal quality and have been well documented and formatted. Apertium translation system, Icelandic Wordnet and Termbank meet most of the criteria.

Table 3.7.1 Icelandic resources evaluated by selection criteria

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
CombiTagger	Reykjavík University	2	1	0	0	2	1
IceNLP - Tagger, Parser, Lemmatizer	Reykjavík University	2	1	0	1	2	1
Apertium-is-en Translation System	Reykjavík University	2	1	1	1	1	1
Icelandic Frequency Dictionary Corpus (web version)	The Arni Magnusson Institute for Icelandic Studies	2	0	0	0	2	1
Icelandic Frequency Dictionary Corpus (download version)	The Arni Magnusson Institute for Icelandic Studies	1	1	0	0	2	1
Balanced Tagged Icelandic Corpus (web version)	The Arni Magnusson Institute for Icelandic Studies	2	0	0	1	1	1
Balanced Tagged Icelandic Corpus (download version)	The Arni Magnusson Institute for Icelandic Studies	1	1	0	1	1	1
A Gold Standard for PoS Tagging	The Arni Magnusson Institute for Icelandic Studies	1	1	0	1	2	1

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
Icelandic Parsed Historical Corpus	University of Iceland	2	1	0	1	2	1
The Jensson Corpus	Tokyo Institute of Technology	2	1	0	0	1	0
The Thor Corpus	Tokyo Institute of Technology	2	1	0	0	1	0
The Broadcast News RUV-1 Corpus	Tokyo Institute of Technology	2	1	0	0	1	0
Parliament Speech Corpus	The Arni Magnusson Institute for Icelandic Studies	2	1	0	0	1	1
Hjal Speech Corpus	University of Iceland	2	1	0	0	2	1
Pronunciation Dictionary for Icelandic	University of Iceland	2	1	0	0	2	1
Database of Modern Icelandic Inflections (web version)	The Arni Magnusson Institute for Icelandic Studies	2	0	0	1	2	1
Database of Modern Icelandic Inflections (download version)	The Arni Magnusson Institute for Icelandic Studies	1	1	0	1	2	1
Database of Semantic Relations	University of Iceland	2	1	0	0	1	1
Icelandic WordNet - Pilot Project	University of Iceland	2	1	1	1	1	1
Íslenskur orðasjóður - Large Corpus 8web version)	Deutscher Wortschatz, Leipzig University	2	0	0	0	0	1
Íslenskur orðasjóður - Large Corpus (download version)	Deutscher Wortschatz, Leipzig University	1	1	0	0	0	1
Icelandic Term Bank – Terminology (web version)	The Arni Magnusson Institute for Icelandic Studies	2	0	1	1	1	1

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
Icelandic Term Bank – Terminology (download version)	The Arni Magnusson Institute for Icelandic Studies	1	1	1	1	1	1
ISLEX - Icelandic Dictionary Base (werb version)	The Arni Magnusson Institute for Icelandic Studies	2	0	1	0	2	1
ISLEX - Icelandic Dictionary Base (download version)	The Arni Magnusson Institute for Icelandic Studies	1	1	1	0	2	1
Ministry for Foreign Affairs - Translation Centre – Dictionary (web version)	Ministry for Foreign Affairs	2	0	1	1	1	1
Ministry for Foreign Affairs - Translation Centre – Dictionary (download version)	Ministry for Foreign Affairs	1	1	1	1	1	1
Íslenskt orðanet - Thesaurus	The Arni Magnusson Institute for Icelandic Studies	0	0	0	1	2	1

3.8 Lithuania (LKI)

All Lithuanian LRTs are actually or potentially available to the consortium; all of them are suitable for language technology and are monolingual. The listed resources have been maintained by LKI but unfortunately not all of them are in active use. The quality of the LRTs has room for improvement. The documentation and meta-data description is lacking or under development.

Table 3.8.1 Lithuanian resources evaluated by selection criteria

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
Database of the Lexicon of Standard Lithuanian	LKI	1	1	0	1	1	0
The Dictionary of Lithuanian	LKI	1	1	0	1	1	0
Modern Lithuanian Dictionary	LKI	2	1	0	0	1	0
Geoinformational Database of Toponyms	LKI	1	1	0	1	1	0
Database of a historical ethnic place names	LKI, co-authored with the Institute of Mathematics and Informatics	1	1	0	1	1	0
Database of Neologisms	LKI	1	1	0	1	0	0
Database Synonymy of Lithuanian Terms	LKI	1	1	0	1	1	0
Database of proper names	LKI	1	1	0	1	0	0
Morphological analyser, lemmatiser and synthesiser for Lithuanian	LKI	1	1	0	0	1	0

3.9 Sweden (UGOT)

The list of resources and tools for Swedish contains 15 items, of which all but one are actually available and the last one is potentially available in the future. All LRTs are suitable for language technology development, but are also solely monolingual. Furthermore, all the listed Swedish LRTs are actively maintained but most have room for improvement and more testing. Documentation for the LRTs is currently lacking or under development.

Table 3.9.1 Swedish resources evaluated by selection criteria

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
Dalin's morphological dictionary	Språkbanken	2	1	0	1	0/1	0
Old Swedish morphology	Språkbanken	2	1	0	1	0	0
Loan Word Typology list	Språkbanken	2	1	0	1	2	0
Preparatory Action for Linguistic Resources Organization for Language Engineering	Språkbanken	2	1	0	1	1/2	0
Swedish Associative Thesaurus	Språkbanken	2	1	0	1	2	0
Examples from the Swedish Associative Thesaurus	Språkbanken	2	1	0	1	0	0
Swedish Associative Thesaurus' morphology	Språkbanken	2	1	0	1	2	0
Semantic Information for Multifunctional Plurilingual Lexica	Språkbanken	2	1	0	1	1/2	0
Swedish FrameNet++	Språkbanken	2	1	0	1	1	0
Swesaurus	Språkbanken	2	1	0	1	1	0
SB-LEX	Språkbanken	2	1	0	1	1	0
Språkbanken's corpora	Språkbanken	2	0/1	0	1	1	0
Citation corpora	Språkbanken	1	1	0	1	0	0

Resource name	Provider	Availability	Suitability	Multilinguality	Longevity	Quality	Extensibility
CLT Toolkit	Various	2	1	0	1	0	0
CLT Cloud	Various	2	1	0	1	0	0

3.10 Identification of resources which could be potentially included in the database of LRTs

Evaluation had been carried out for resources already known to the consortium. There may be resources belonging to the third parties network, which are still absent in the database of LRTs. The following algorithm should find the gaps in the list, using the suitability criterion extensively.

(1) Find the most important basic software components for written and spoken language and/or resources for their development taking into account the experience of the CLARIN project, the BLARK matrices of different languages, and White Papers of languages composed in the first phase of META-NET project. One should consider that the resources for CLARIN project were dedicated to the needs of eHumanities, but this project focuses on the requirements of development of multilingual web.

(2) Find out how the modules depend on each other.

(3) Find which of these resources are available for each language of the consortium and which are lacking. Also, clarify licensing issues.

(4) Select modules and resources which are available for most of the languages.

(5) Evaluate the quality and availability of each resource. Assemble information on licensing agreements.

(6) Assess the efforts needed to transform each resource to a format of some well-known standard (proposed by META-NET) and compile a work plan for further developments.

(7) Prepare contracts with owners of the resources originated outside the consortium.

The methodology for identification and selection of unknown resources may be used during the rest of the project by contacting the third parties networks.

4 Conclusions

This report describes and evaluates the LRTs that have been identified and collected by the META-NORD consortium by project month M6. The evaluation has been carried out, using the criteria suggested by META-NET Network of Excellence and META-SHARE project (actual availability, suitability for technology and product development, fitness for multilingual purposes, quality, and potential for re-use, recombination and repurposing).

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

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

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

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

The LRTs selected for WP3 should be of high quality, however, there is a chance that active development will increase the quality of LRT significantly during the project.

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

The evaluation supported the assumption that the potential resources for further integration between languages are wordnets (Danish, Estonian, Finnish, and Icelandic), the multilingual database of terminology, treebanks (monolingual treebanks, accessible in the same format) and finite-state techniques.

5 References

Rehm, Georg (2010). META-NET and META-SHARE: An Overview. Presentation at Human Language Technologies – the Baltic Perspective. Riga, 2010.

Piperidis, Stelios (2010). Building META-SHARE - an Open Resource Exchange and Sharing Facility. Presentation at LREC 2010, Malta.

6 List of tables

Table 1 Abbreviations
Table 3.2.1 Latvian resources evaluated by selection criteria
Table 3.3.1 Danish resources evaluated by selection criteria
Table 3.4.1 Estonian resources evaluated by selection criteria
Table 3.5.1 Norwegian resources evaluated by selection criteria
Table 3.6.1 Finnish resources evaluated by selection criteria
Table 3.7.1 Icelandic resources evaluated by selection criteria
Table 3.8.1 Lithuanian resources evaluated by selection criteria
Table 3.9.1 Swedish resources evaluated by selection criteria