META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Project no. 270899

Deliverable D3.1
First batch of language resources

Version No. 1.0
30/11/2011

Document Information

Deliverable number:	D3.1
Deliverable title:	First batch of language resources (documentation)
Due date of deliverable:	2011-11-30
Actual submission date of deliverable:	2011-12-01
Main Author(s):	Dorte Haltrup Hansen, Bolette Pedersen
Participants:	All
Internal reviewer:	Tilde
Workpackage:	WP3
Workpackage title:	Enhancing language resources
Workpackage leader:	UCPH
Dissemination Level:	PU
Version:	Final
Keywords:	Documentation, resources

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
0.1	2011-10-27	Fishbone	Dorte Haltrup Hansen, Bolette Sandford Pedersen (UCPH)	The structure of deliverable	Fish bone approved
0.2	2011-11-11	1st draft	Dorte Haltrup Hansen, Bolette Sandford Pedersen (UCPH)	Most partners	First description of resources has been added
0.4	2011-11-14	2nd draft	Dorte Haltrup Hansen, Bolette Sandford Pedersen	All partners	Updated description of resources has been added

Prefinal	2011-11-18	Prefinal draft	Dorte Haltrup Hansen, Bolette Sandford Pedersen	All partners	Final description of resources has been added
0.9	2011-11-25	Internal review, Final modifications	Inguna Skadina, Dorte Haltrup Hansen, Bolette Sandford Pedersen	Tilde, UCPH	Final review and modifications
1.0	30-11-2011	Final	Tilde	Final check	Submitted to PO

EXECUTIVE SUMMARY

This report documents the resources delivered in first batch on M10. It includes description of language resources available for download or for access through provided links. Description of resources includes basic and administrative information, technical description, content information and relevant references.

As the META-SHARE software is not yet mature enough for the data upload and download, resources are published in SVN repository created as a temporary solution. When a stable META-SHARE version will be released resources will be uploaded to the META-SHARE repository.

Resources are accessible from the META-NORD webpage: http://www.meta-nord.eu/index.php?p=first-upload.

All in all, 59 resources are uploaded by the META-NORD partners in the first batch, including 32 lexical resources, 12 corpora, 6 treebanks, 5 speech resources, 3 wordnets and one tool.

Table of Contents

Background

The purpose of this report is to document the resources delivered in first batch, D3.1 so that the user can get a more general overview of the content of each resource further than what can be deduced from the metadata. D3.1 is the first deliverable of WP3 which has the purpose of upgrading and harmonizing national language resources within and across META-NORD languages, in order to make them interoperable w.r.t. their data formats and content. Initial work of documenting, processing, linking, and upgrading META-NORD resources to agreed standards and guidelines has been performed and can be found in this deliverable.

Apart from the monolingual resources, three multilingual actions have been embarked in WP3, namely:

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

Several of these cross-lingual initiatives also provide their first deliverables in batch 1; for instance the wordnets have been linked to Princeton WordNet Core as a first step towards multilingual linking between the META-NORD languages.

As the META-SHARE software is not yet mature enough for the data upload and download (issues are reported in the D4.3), resources are published in SVN repository created as a temporary solution. When a stable META-SHARE version will be released resources will be uploaded to the META-SHARE repository.

Resources are accessible from the META-NORD webpage: http://www.meta-nord.eu/index.php?p=first-upload.

1. Latvia (TILDE)

1.1. EuroTermBank

1. **BASIC INFORMATION**

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

Online terminological database

1.2 Representation of the lexicon (flat files, database, markup)

The online terminology database is prepared in proprietary format and is available for browsing.

Internally EuroTermBank data is stored in a standard TBX[1] records.

1.3 Character encoding

UTF-8

2. **ADMINISTRATIVE INFORMATION**

2.1 Contact person (name, e-mail)

Name: Roberts Rozis, e-mail: roberts.rozis@tilde.lv

2.2 Copyright statement and information on IPR

Copyright holder: EuroTermBank Consortium

[1] http://en.wikipedia.org/wiki/TBX#TBX

http://web.archive.org/web/20110102011556/http://www.lisa.org/Term-Base-eXchange.32.0.html

Copyright and Terms of use: http://www.eurotermbank.com/Disclaimer.aspx

3. **TECHNICAL INFORMATION**

3.1 Data structure of an entry

Dictionary entries has proprietary format.

3.2 Lexicon size (num. of lexical items)

2.3M terms, 625345 term entries, 27 languages

4. **CONTENT INFORMATION**

4.1 The natural language(s) of the lexicon

27 languages represented: EU BG HR CS DA NL EN ET FI FR DE EL HU IT LA LV LT MT NO PL PT RO RU SK SL ES SV.

4.2 Entry Type

4.3 Attributes

4.4 Coverage of the lexicon

Multiple domains, classified by Eurovoc 4.2

4.5 Intended application of the lexicon

General terminology use. Translators, researchers. MT. Integration in CAT tools via API (integration to be negotiated)

4.6 Reliability (automatically/manually constructed)

Data of EuroTermBank federated from 100 various terminology resources.

5.**RELEVANT REFERENCES AND OTHER INFORMATION**

Website of the resource: http://www.eurotermbank.eu/

1.2. Corpus of Latvian literature

1 **BASIC INFORMATION**

1.1 Corpus composition

The Corpus of Latvian literature contains literary works of Latvian authors which are not protected by copyright low. It contains works of 21 authors – poems, stories, novels and other literary works, 69 in total which correspond to 15 000 printed pages .

1.2. Representation of the corpora (flat files, database, markup)

The corpus is stored in XML files.

1.3 Character encoding – UTF-8

2. **ADMINISTRATIVE INFORMATION**

2.1. Contact person (name, e-mail)

For further information, please, contact Roberts Rozis (Roberts.rozis@tilde.lv) or Anita Vasiļjeva (anita.vasiljeva@tilde.lv)

2.2. Copyright statement and information on IPR

The corpus is freely available for browsing from http://www.letonika.lv/literatura/

3. **TECHNICAL INFORMATION**

3.1. Data structure of an entry

Corpus is available in proprietary format.

3.2. Corpora size

15 000 printed pages

4. **CONTENT INFORMATION**

4.1. Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual

4.2. The natural language(s) of the corpus

Latvian

4.3. Domain(s)/register(s) of the corpus

Fiction.

4.4. Annotations in the corpus (if an annotated corpus)

4.5 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

4.5. Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

4.6. Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

4.7. Intended application of the corpus

4.8. Reliability of the annotations (automatically/manually assigned)

Content of corpus is manually checked.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Available from the Web for browsing: http://www.letonika.lv/literatura/.

1.3. The Lithuanian-Latvian dictionary

1. **BASIC INFORMATION**

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

The Lithuanian-Latvian dictionary is updated electronic version of bilingual Lithuanian-Latvian dictionary (Lietuviešu-latviešu vārdnīca) by Jons Balkevičs, Laimute Balode, Apolonija Bojāte, Valters Subatnieks, redaktors Alberts Sarkanis, published in 1995.

1.2 Representation of the lexicon (flat files, database, markup)

The dictionary is prepared in proprietary format and is available for browsing.

1.3 Character encoding

UTF-8

2. **ADMINISTRATIVE INFORMATION**

2.1 Contact person (name, e-mail)

For further information, please, contact Roberts Rozis (Roberts.rozis@tilde.lv) or Anita Vasiļjeva (anita.vasiljeva@tilde.lv).

2.2 Copyright statement and information on IPR

The dictionary is freely available for browsing from http://lietuviu.letonika.lv.

3. **TECHNICAL INFORMATION**

3.2 Data structure of an entry

Dictionary entries has proprietary format.

3.3 Lexicon size (num. of lexical items)

The lexicon contains about 60 00 lexical items.

4. **CONTENT INFORMATION**

4.1 The natural language(s) of the lexicon

The source language is Lithuanian and the target language is Latvian.

4. 2 Entry Type

4.3 Attributes

4.4 Coverage of the lexicon

The lexicon covers Modern Lithuanian and Latvian.

4.5 Intended application of the lexicon

The dictionary is mainly used for education and manual translation purposes. However, it could be used in language technology solutions also.

4.6 Reliability (automatically/manually constructed)

Content of the lexicon is manually checked and updated.

5. **RELEVANT REFERENCES AND OTHER INFORMATION**

More information about the dictionary is available at: http://lietuviu.letonika.lv

1.4. Latvian-Lithuanian dictionary

1. **BASIC INFORMATION**

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

The Latvian-Lithuanian dictionary is bilingual on-line dictionary by Alvydas Butkus (Vytautas Magnus University). It is based on printed Latvian-Lithuanin dictionary (Latviešu-lietuviešu vārdnīca), published in 2003.

1.2 Representation of the lexicon (flat files, database, markup)

The dictionary is prepared in proprietary format and is available for browsing.

1.3 Character encoding

UTF-8

2. **ADMINISTRATIVE INFORMATION**

2.1 Contact person (name, e-mail)

For further information, please, contact Roberts Rozis (Roberts.rozis@tilde.lv) or Anita Vasiļjeva (anita.vasiljeva@tilde.lv).

2.2. Copyright statement and information on IPR

The dictionary is freely available for browsing from http://www.letonika.lv/lvlt.

3. **TECHNICAL INFORMATION**

3.1 Data structure of an entry

Dictionary entries has proprietary format.

3.2. Lexicon size (num. of lexical items)

The lexicon contains about 43 00 lexical items

4. **CONTENT INFORMATION**

4.1 The natural language(s) of the lexicon

The source language is Latvian and the target language is Lithuanian.

4. 2 Entry Type

4.3 Attributes

4.4 Coverage of the lexicon

The lexicon covers Modern Latvian and Lithuanian.

4.5. Intended application of the lexicon

The dictionary is mainly used for education and manual translation purposes. However, it could be included in language technology solutions as well.

4.6 Reliability (automatically/manually constructed)

Content of the lexicon is manually checked and updated.

5. **RELEVANT REFERENCES AND OTHER INFORMATION**

More information about the dictionary is available at: http://www.letonika.lv/lvlt.

1.5. Estonian-Latvian dictionary

1. **BASIC INFORMATION**

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

The bilingual online Estonian-Latvian dictionary is based on Karl's Aben's Estonian-Latvian dictionary published in 1967. The initial dictionary has been revised by Andra Kalnača, Ērika Krautmane, Jana Šteinberga-Ranki and Urve Aivare. During revision obsolote words were removed and about 2000 new lexical entries and 1000 terminology entries were added.

1.2 Representation of the lexicon (flat files, database, markup)

The dictionary is prepared in proprietary format and is available for browsing

1.3 Character encoding

UTF-8

2. **ADMINISTRATIVE INFORMATION**

2.1 Contact person (name, e-mail)

For further information, please, contact Roberts Rozis (Roberts.rozis@tilde.lv) or Anita Vasiļjeva (anita.vasiljeva@tilde.lv).

2.2 Copyright statement and information on IPR

The dictionary is freely available for browsing from http://eesti.letonika.lv.

3. **TECHNICAL INFORMATION**

3.1 Data structure of an entry

Dictionary entries has proprietary format.

3.2 Lexicon size (num. of lexical items)

The lexicon contains about 26 00 lexical items.

4. **CONTENT INFORMATION**

4.1 The natural language(s) of the lexicon

The source language is Estonian and the target language is Latvian.

4. 2 Entry Type

4.3 Attributes

4.4 Coverage of the lexicon

The lexicon covers Modern Estonian and Latvian.

4.5 Intended application of the lexicon

The dictionary is mainly used for education and manual translation purposes. However, it could be included in language technology solutions as well.

4.6 Reliability (automatically/manually constructed)

Content of the lexicon is manually revised and updated.

5. **RELEVANT REFERENCES AND OTHER INFORMATION**

More information about the dictionary is available at: http://eesti.letonika.lv

1.6 Multilingual dictionary of person names

1. **BASIC INFORMATION**

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

The multilingual online dictionary of person names contains person names in Czech, French, Italian and German with their translations into Latvian. For each language most popular first names, surnames as well as names of historically important persons are included. The dictionary includes more than 4000 names in total. The dictionary is created by Michal Škrabal (Czech language), Astra Skrābane (French language), Baiba Bankava (Italian language), Valda Rudziša (German language) and Juris Baldunčiks (editor in chief).

1.2 Representation of the lexicon (flat files, database, markup)

The dictionary is prepared in proprietary format and is available for browsing.

1.3 Character encoding

UTF-8

2. **ADMINISTRATIVE INFORMATION**

2.1 Contact person (name, e-mail)

For further information, please, contact Roberts Rozis (Roberts.rozis@tilde.lv) or Anita Vasiļjeva (anita.vasiljeva@tilde.lv).

2.2 Copyright statement and information on IPR

The dictionary is freely available for browsing from http://www.letonika.lv/personvardi.

3. **TECHNICAL INFORMATION**

3.1 Data structure of an entry

Dictionary entries has proprietary format.

3.2 Lexicon size (num. of lexical items)

The lexicon contains about 4000 lexical items

4. **CONTENT INFORMATION**

4.1 The natural language(s) of the lexicon

The source languages are Czech, French, Italian and German, the target language is Latvian.

4.2 Entry Type

4.3 Attributes

4.4 Coverage of the lexicon

4.5 Intended application of the lexicon

The dictionary is mainly used for education and manual translation purposes. However, it could be included in different applications, e.g., machine translation.

4.6 Reliability (automatically/manually constructed)

Content of the lexicon is manually created.

5. **RELEVANT REFERENCES AND OTHER INFORMATION**

More information about the dictionary is available at: http://www.letonika.lv/personvardi.

1.7. Legislation Corpus of the Republic of Latvia

1. **BASIC INFORMATION**

1.1 Corpus composition

Latvian-English legislation corpus of Republic of Latvia is composed from public legal documents of the Republic of Latvia available in Latvian to English. It contains the Laws of the Republic of Latvia and Cabinet Regulations in the period of 2000-2010. It contains text from the total of 1275 documents: 270 Laws, 2 Cabinet Instructions and 1003 Cabinet Regulations.

1.2 Representation of the corpora (flat files, database, markup)

The corpus is represented in a single TMX[2] standard file and the documents data has been aligned at a sentence level.

1.3 Character encoding – UTF-8

2. **ADMINISTRATIVE INFORMATION**

2.1. Contact person (name, e-mail)

For further information, please, contact Roberts Rozis (Roberts.rozis@tilde.lv)

2.2. Copyright statement and information on IPR

MSC_BYNCND

3. **TECHNICAL INFORMATION**

3.1 Data structure of an entry

The corpus is provided in a singlefile in TMX format, metadata data information is encoded is document header.

3.2 Corpora size (num. of tokens)

The corpus contains about 1 660 000 tokens: 1 million tokens in English, 660 000 tokens in Latvian

4. **CONTENT INFORMATION**

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Parallel

4.2.The natural language(s) of the corpus

Latvian, English

4.3 Domain(s)/register(s) of the corpus

Legal. Legislation.

4.4 Annotations in the corpus (if an annotated corpus)

4.5 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

[2] http://en.wikipedia.org/wiki/Translation_Memory_eXchange

http://web.archive.org/web/20110102010600/http://www.lisa.org/Translation-Memory-e.34.0.html

The corpus is aligned in TMX format, sentence level mark-up.

4.6 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

4.7 Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

The corpus is aligned at sentence level, alignment has been performed by using Hunalign alignment tools

4.8. Intended application of the corpus

NLP application: training of MT systems.

Human use: Analysis of legal language.

4.9. Reliability of the annotations (automatically/manually assigned)

Alignment is done automatically.

5. RELEVANT REFERENCES AND OTHER INFORMATION

All the source documents among many other documents available from the Web for browsing: http://www.likumi.lv/

2. Denmark (UCPH)

2.1 DanNet

1. BASIC INFORMATION

1.1 Lexicon type

Wordnet

1.2 Representation of the lexicon

Two files: csv format and OWL format + README file

1.3 Character encoding

The characters have been encoded in UTF8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Bolette Sandford Pedersen

E-mail: bspedersen@hum.ku.dk

2.2 Copyright statement and information on IPR

Open Source, Princeton license

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

A synset contains a list of lemma(s), a gloss, and a list of relations and features. The information is organized in the following synset files:

synsets.rdf: The synsets are declared in this file.

glossary.rdf: The gloss of each synset.

words.rdf: Words and their lexical form.

wordsenses.rdf: Connects synsets and words.

synset_attributes.rdf: features that describe the synset further (such as connotation ('positive' or 'negative') and 'sex' ('male' or 'female')).

register.rdf: Information about the register of the word senses, e.g. 'slang' or 'sj.' (i.e. 'old-fashioned').

3.2 Lexicon size

58,716 nouns, 8195 verbs and 3680 adjectives. Amounts to 65,000 synsets, 5,000 links to Princeton Wordnet (core wordnet)

4. **CONTENT INFORMATION**

4.1 The natural language(s) of the lexicon

Danish

4. 2 Entry Type

Synonym set (Synset)

4.3 Attributes

The key attributes are the semantic relations and features. The resource contains 32 relation types and seven feature types.

4.4 Coverage of the lexicon

The resource covers 58,716 nouns, 8195 verbs and 3680 adjectives. It represents the basic vocabulary of modern Danish.

4.5 Intended application of the lexicon

Intended applications are advanced applications such as Information Retrieval, Question Answering and Machine Translation.

4.6 Reliability (automatically/manually constructed)

The resource is semi-automatically constructed. 2% has been manually validated.

5. **RELEVANT REFERENCES AND OTHER INFORMATION**

Pedersen, B.S, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. Language Resources and Evaluation, Computational Linguistics Series. 10.1007/s10579-009-9092-1

http:// wordnet.dk

2.2 Danish Treebank - CDT1

1 **BASIC INFORMATION**

1.1 Resource composition

Treebank

1.2 Representation of the resource (flat files, database, markup)

Flat files with markup

1.3 Character encoding

ISO

2 **ADMINISTRATIVE INFORMATION**

2.1 Contact person (name, e-mail)

Name: Matthias Buch-Kromann,

E-mail: Matthias@Buch-Kromann.dk

2.2 Copyright statement and information on IPR

GNU GPL v3.0

3 **TECHNICAL INFORMATION**

3.1 Data structure of an entry

Dependency-annotated sentences

3.2 Resource size (num. of rules)

100.000 word tokens

4 **CONTENT INFORMATION**

4.1 Type of the resource (language (in(dependent)

Treebank, created on the basis of the dependency-based grammar formalism Discontinuous Grammar (Buch-Kromann 2009). Texts are analyzed as a single dependency structure that includes morphology and syntactic dependency.

4.2 The natural language(s) for the resource is applicable (if language dependent)

Danish

4.3 Domain(s)/register(s) of the resource

Danish PAROLE corpus

4.4 Annotations in the resource (if an annotated resource)

4.4.1 Types of annotations

Part-of-speech, syntax (dependency relations)

4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed),

POS-tagged with the PAROLE tag set. The list of dependency relations is contained in annotation manual http://code.google.com/p/copenhagen-dependency-treebank/wiki/CDT

4.5 Intended application of the resource Training of natural language parsers and other statistically based natural language applications

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Manual annotation

5 RELEVANT REFERENCES AND OTHER INFORMATION

Matthias Trautner Kromann, 2003. *The Danish Dependency Treebank and the DTAG treebank tool.* In Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), 14-15 November, Växjö. pp. 217-220.

Matthias Buch-Kromann, 2006. *Discontinuous Grammar. A dependency-based model of human parsing and language learning.* Dr.ling.merc. dissertation, Copenhagen Business School. 432+xvi pp.

Matthias Buch-Kromann, 2009. *Discontinuous Grammar. A dependency-based model of human parsing and language learning.* VDM Verlag. Republication of Buch-Kromann (2006).

2.3 Danish English Treebank – CDT2

1 BASIC INFORMATION

1.1 Resource composition

Parallel treebank

1.2 Representation of the resource (flat files, database, markup)

Flat files with markup

1.3 Character encoding

ISO

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Name: Matthias Buch-Kromann,

E-mail: Matthias@Buch-Kromann.dk

2.2 Copyright statement and information on IPR

GNU GPL v3.0

3 TECHNICAL INFORMATION

3.1 Data structure of an entry

Dependency-annotated aligned sentences

3.2 Resource size (num. of rules)

95.000 word tokens

4 CONTENT INFORMATION

4.1 Type of the resource (language (in(dependent)

Parallel treebank, created on the basis of the dependency-based grammar formalism Discontinuous Grammar (Buch-Kromann 2009). Texts are analyzed as a single dependency structure that includes morphology and syntactic dependency.

4.2 The natural language(s) for the resource is applicable (if language dependent)

Danish – English

4. 3 Domain(s)/register(s) of the resource

Danish – English PAROLE corpus

4.4 Annotations in the resource (if an annotated resource)

4.4.1 Types of annotations

Part-of-speech, syntax (dependency relations)

4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed),

POS-tagged with the PAROLE tag set. The list of dependency relations is contained in annotation manual: http://code.google.com/p/copenhagen-dependency-treebank/wiki/CDT

4.5 Intended application of the resource

Training of natural language parsers, syntax-based machine translation systems, and other statistically based natural language applications

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Semiautomatic annotation and alignment

5 RELEVANT REFERENCES AND OTHER INFORMATION

Matthias Buch-Kromann, Jürgen Wedekind, and Jakob Elming, 2007. *The Copenhagen Danish-English Dependency Treebank v. 2.0.* Parallel dependency treebank for Danish-English with 100,000 words based on the Danish Dependency Treebank.

2.4 STO-LMF

1. BASIC INFORMATION

1.4 Lexicon type

Monolingual lexicon, morphological part, nouns, adjectives and verbs

1.5 Representation of the lexicon

One XML file

1.6 Character encoding

The characters have been encoded in UTF8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Sussi Olsen

E-mail: saolsen@cst.dk

2.2 Copyright statement and information on IPR

CLARIN Academic license, no commercial use

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

The following data structure holds for the data in the first batch. The structure will be expanded with more features for batch 2 and with a syntactic layer for batch 3.

A lexical entry contains part of speech, lemma and word forms.

For the lemma, decomposition specifies if the word is a compound. It is also specified whether the lemma is in accordance with the official Danish orthography.

For each word form, all the relevant grammatical features are specified varying according to the part of speech.

Example of entry

```
<LexicalEntry>
    <feat att="partOfSpeech" val="NOUN_COMMON"/>
    <feat att="mu_id" val="HUKOMMELSE"/>
    <feat att="gmu_id" val="GMU_HUKOMMELSE_1"/>
    <feat att="origin" val="EDB-KORPUS"/>
    <feat att="autonomy" val="YES"/>
    <Lemma>
      <FormRepresentation>
        <feat att="spelling" val="hukommelse"/>
        <feat att="ro_approved" val="YES"/>
        <feat att="fugeResultat" val="hukommelse"/>
        <feat att="decomposition" val=""/>
      </FormRepresentation>
    </Lemma>
    <WordForm>
      <feat att="gender" val="COMMON"/>
      <feat att="grammaticalNumber" val="singular"/>
      <feat att="definiteness" val="indefinite"/>
      <feat att="case" val="unmarked"/>
        <FormRepresentation>
          <feat att="writtenform" val="hukommelse"/>
          <feat att="inflectionalParadigm" val="MFG0076"/>
             </FormRepresentation>
    </WordForm>
```

3.2 Lexicon size

86,935 morphological entries of the STO lexicon: 70305 nouns, 10572 adjectives and 6055 verbs.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Danish

4. 2 Entry Type

Lexical entries in LMF

4.3 Attributes

The key attributes of the lexicon are the lexical entry, the lemma and the word forms.

4.4 Coverage of the lexicon

The entire STO lexicon covers about 88,000 morphological entries, 43,000 syntactic entries and 10,000 semantic entries.

In this batch 86,000 morphological entries (nouns, verbs and adjectives) are included – the entries of the remaining part of speech will be included in the next batch.

4.5 Intended application of the lexicon

Intended applications are all kinds of language technology applications that need morphological information.

4.6 Reliability (automatically/manually constructed)

The original STO lexicon has been 100 % manually validated. The update to LMF has been automatically validated against the LMF DTD (ISO 24613).

5. RELEVANT REFERENCES AND OTHER INFORMATION

ISO 24613, International Standard: *Language resource management - Lexical Markup Framework (LMF)*, ISO 2008

Braasch A. et al.: *STO Sprogteknologisk Ordbase, monolingual lexicon, Documentation, version 2*, 2008. CST, KU, Copenhagen.

Braasch, Anna & Olsen, Sussi: STO: A Danish Lexicon Resource - Ready for Applications, 2004 In: *Fourth International Conference on Language Resources and Evaluation, Proceedings*, Vol. IV. Lisbon, pp. 1079-1082.

3. Estonia (UT)

3.1 Estonian Wordnet

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

Estonian WordNet (EstWN) is a lexical ontology following the Princeton WordNet (PWN) organizational principles.

1.2 Representation of the lexicon (flat files, database, markup)

Estonian Wordnet is done up to nowadays with the Polaris tool, we are using the Polaris import-export format text file. We have done conversion to XML, actually into 2 different versions of XML. The KYOTO project format (http://www.kyoto-project.eu/) and VisDic format (http://deb.fi.muni.cz/clients-debvisdic.php). There are no specific reason for these formats, we just were testing the DebVisDic dictionary environment.

1.3 Character encoding

The characters have been encoded in UTF8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Heili Orav
e-mail: heili.orav@ut.ee

2.2 Copyright statement and information on IPR

The resource is free license-based for research purposes and fee license-based for commercial purposes

3 TECHNICAL INFORMATION

3.2 Data structure of an entry

Inside a single import record, one can describe:

- Concepts (word-meanings and word-instances)
- Concept Variants
- Internal Concept Links
- Concept-to-ILI Equivalence Links (links to Princeton Wordnet)
- Properties (in word-meanings) and property values (in word-instances)

The general structure for a word-meaning (noun) record is:

```
0 WORD_MEANING
  1 PART_OF_SPEECH "n"
  1 VARIANTS
    # further detail on the variants go here
  1 INTERNAL_LINKS
    # further detail on the internal links go here
  1 EQ_LINKS
    # further detail on the equivalence links go here
  1 PROPERTIES
    # further detail on the properties go here
```

The general structure for a word-instance record is similar:

```
0 WORD_INSTANCE
  1 PART_OF_SPEECH "pn"
  1 VARIANTS
    # further detail on the variants go here
  1 INTERNAL_LINKS
    # further detail on the internal links go here
```

```
 1 EQ_LINKS
   # further detail on the equivalence links go here
 1 PROPERTY_VALUES
   # further detail on the property values go here
```

The level 0 root field of a concept record identifies its type:
WORD_MEANING or WORD_INSTANCE.

The first child field (level 1) of a concept record should be the PART_OF_SPEECH field, which requires a value. If the record is a WORD_MEANING, then this can be any part-of-speech; if the record is a WORD_INSTANCE, then the part-of-speech must be proper noun ("pn").

The VARIANTS field is the parent of one or more subtrees that are headed by LITERAL fields. The value for a LITERAL field describes the literal string (word or phrase) for the variant. The first required child field for all LITERAL fields is the SENSE field, which provides the sense number that the concept "implements" for that variant.

There are several optional fields:

- a DEFINITION field, providing a single definition field.
- a STATUS field, which can contain a label providing a status indication.
- an EXAMPLES field, heading up a list of one or more EXAMPLE strings.
- a TRANSLATION field, heading up a list of one or more translation strings. A translation string consists of a language code prefix,colon, and the translation itself.

- a USAGE_LABELS field, heading up a list of specific usage labels. Below this you can use the USAGE_LABEL and USAGE_LABEL_VALUE fields. The list of available usage labels is kept in the database itself.
- a FEATURES field, heading up a list of syntactic features. Below this you can use the FEATURE and FEATURE_VALUE fields. The available list of features is stored in the database itself.
- an EXTERNAL_INFO section, providing information on corpus frequency counts, sources and other information.

The INTERNAL_LINKS field heads up the subtree of fields that describes all the links this concept has with other concepts in the Language WordNet. Each individual internal link is headed by a RELATION field. The value for this field is the name of the internal link.

The RELATION field must have a subtree headed by a TARGET_CONCEPT field, which identifies the link's target concept. It can also have an optional subtree headed by a FEATURES field, which add feature information to the link.

A target concept is described by specifying three things: a part-of-speech, a literal, and a sense number. With this combination it is possible to uniquely address any concept.

Below the FEATURES field, a number of optional features may appear as child fields.

The link features available are:

- NEGATIVE: if present, it indicates that the link should be interpreted negatively.

- VARIANT_TO_VARIANT: this is used to specify variants inside both the source and target concepts for links that need it.

The EQ_LINKS field heads up the subtree of fields that describes all the links this concept has with ILI records in the Interlingua. Each equivalence relation must be headed by a EQ_RELATION field, the value for which is the name of the relation. The TARGET_ILI field must be present as a child under each EQ_RELATION field. It identifies which ILI record is the target of the equivalence link.

Identifying a target ILI record can be done in the following ways:

- the combination of a part-of-speech, literal and sense number.
- a combination of part-of-speech and a WordNet 1.5 synset file offset value.
- by specifying the ILI record's database number.

In a WORD_MEANING record, it is possible to specify a list of properties. The properties specified must already have been created as property types.

In a WORD_INSTANCE record, it is possible to assign values to the properties specified in one of the WORD_MEANING records that appears in the hyperonym hierarchy of the instance.

An example synset:

```
0 @5@ WORD_MEANING
  1 PART_OF_SPEECH "n"
  1 VARIANTS
    2 LITERAL "filmifestival"
      3 SENSE 1
      3 DEFINITION "pidulik filmikunsti saavutuste tutvustamine, kogemuste vahetamine ning parimate
filmide väljaselgitamine"
      3 EXTERNAL_INFO
        4 SOURCE_ID 1
          5 TEXT_KEY "32672"
        4 SOURCE_ID 1
          5 TEXT_KEY "27724"
        4 SOURCE_ID 1
          5 TEXT_KEY "29169"
        4 SOURCE_ID 1003
  1 INTERNAL_LINKS
    2 RELATION "has_hyperonym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "festival"
          5 SENSE 1
      3 SOURCE_ID 1003
  1 EQ_LINKS
    2 EQ_RELATION "eq_has_hyperonym"
      3 TARGET_ILI
        4 PART_OF_SPEECH "n"
        4 WORDNET_OFFSET 295927
      3 SOURCE_ID 1003
```

3.3 Lexicon size (nmb. of lexical items)

The current (validated) version contains 47336 synsets (November 10, 2011), with the following distribution:

Noun synsets	Verb synsets	Adj. synsets	Adv. synsets	Total
37973	5305	2337	1570	47336

4 CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the lexical ontology is Estonian.

4. 2 Entry Type

There are five types of entries, all of them having the same structure: entries for nouns, for verbs, for adjectives, for adverbs and for proper names.

4.3 Attributes

See section 3.2:

Concept number is between '@' characters in level 0 record. Unique for the version of wordnet.

Value of PART_OF_SPEECH field is one of "n" (noun), "v" (verb), "a" (adjective), "b" (adverb). Value of PART_OF_SPEECH field in WORD_INSTANCE record is "pn" (proper noun).

4.4 Coverage of the lexicon

The design procedure of the EstWN during more than 10 years has followed different strategies. Firstly, the literals chosen for implementation were selected based on frequency. Secondly, our chosen approach so far for enlarging thesaurus has been domain-specific, i.e we have added semantic fields like architecture, transportation, personality traits and so on. Thirdly, there are some endeavors for automatic additions. For example, a number of words have been derived via suffixes.

The lexical stock covers the basic general language vocabulary of Estonian.

4.5 Intended application of the lexicon

Word sense disambiguation, Information extraction, semantic research for linguistics

4.6 Reliability (automatically/manually constructed)

The lexical ontology has been based on several reference published dictionaries: Explanatory Dictionary of Estonian, Dictionary of Synonyms, Dictionary of Antonyms. Work with Estonian Wordnet has been mostly manual, only small part of words has been derived via suffixes automatically.

5 RELEVANT REFERENCES AND OTHER INFORMATION

References on the Estonian WordNet

1. Kerner, Kadri; Orav, Heili; Parm, Sirli (2010). Semantic Relations of Adjectives and Adverbs in Estonian WordNet. *In: LREC 2010 Proceedings: LREC 2010, Malta, Valetta, 17.-23. mai 2010.* ELRA, 2010, 33 - 37.
2. Kerner, Kadri; Orav, Heili; Parm, Sirli (2010). Growth and Revision of Estonian WordNet. *In: Principles, Construction and Application of Multilingual Wordnets. Proceeding of the 5th Global Wordnet Conference: 5th Global Wordnet Conference; Mumbai, India; 31.jaanuar-4.veebruar 2010. (Toim.) Bhattacharyya, P.; Fellbaum, Ch.; Vossen, P.*. Mumbai, India: Narosa Publishing House, 2010, 198 - 202.
3. Orav, Heili; Õim, Haldur; Kerner, Kadri; Kahusk, Neeme (2010). Main trends in semantic-research in Estonian language technology. *In: Baltic HLT Proceedings: Human Language Technologies — the Baltic Perspective; Riga, Latvia; October 7–8, 2010.* IOS Press, 2010, (Frontiers in Artificial Intelligence and Applications), 201 - 207.
4. Orav, H.; Kerner, K.; Parm, S. (2011). Eesti Wordneti hetkeseisust. Keel ja Kirjandus, 2, 96 - 106.

A larger version of Estonian WordNet can be browsed at the web address http://www.cl.ut.ee/ressursid/teksaurus/

Others references:

Louw, Michael (1988) Polaris User's Guide. The EuroWordNet Database Editor. Deliverable D024, WP6.5, EuroWordNet, LE-4003

3.2 Estonian treebank, esttre

1 BASIC INFORMATION

1.1 Corpus composition

The Corpus consists of ca 1400 sentences (10600 tokens), the text classes represented in the Corpus are fiction, both translated and original, newspaper texts and 20 sentences of transcribed spoken language.

1.2 Representation of the corpora (flat files, database, markup)

TIGER-XML-files

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Kaili Müürisep
e-mail: kaili.muurisep@ut.ee

2.2 Copyright statement and information on IPR

The resource is free for research purposes, local license

3 TECHNICAL INFORMATION

3.1 Data structure of an entry

XML-files

3.2 Corpora size (nmb. of tokens)

The corpus contains about 10600 tokens

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual corpus with TIGER-style syntactic annotations (so-called syntactic trees).

4.2 The natural language(s) of the corpus

The natural language of the corpus is Estonian.

4. 3 Domain(s)/register(s) of the corpus

Corpus represents written Estonian.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations

The corpus is annotated at paragraph, sentence and word level. Word-forms have been lemmatized and tagged for

1) POS and relevant grammatical categories, i.e. case and number for nominals, additionally degree for adjectives and mood, tense, person, voice, positive/negative distinction for verbs;

2) syntactic functions

3) phrase structure

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

POS-tags

```
<value name="n">noun</value>
<value name="prop">proper noun</value>
<value name="art">article</value>
<value name="v">verb</value>
<value name="v-fin">verb</value>
<value name="v-inf">verb</value>
<value name="adj">adjective</value>
<value name="adj-nat">nationality adjective</value>
<value name="adv">adverb</value>
<value name="prp">preposition</value>
<value name="pst">preposition</value>
<value name="conj-s">subordinating conjunction</value>
<value name="conj-c">coordinating conjunction</value>
<value name="conj-p">prepositional conjunction</value>
<value name="pron">pronoun (to be specified)</value>
<value name="pron-pers">personal pronoun</value>
<value name="pron-rel">relative pronoun</value>
<value name="pron-int">interrogative pronoun</value>
<value name="pron-dem">demonstrative pronoun</value>
<value name="pron-indef">indefinite pronoun</value>
<value name="pron-poss">possessive pronoun</value>
<value name="pron-def">possessive pronoun</value>
<value name="pron-refl">reflexive pronoun</value>
<value name="num">numeral</value>
<value name="intj">interjection</value>
```

```
<value name="infm">infinitive marker</value>
<value name="punc">punctuation</value>
<value name="sta">statement</value>
<value name="abbr">abbreviation</value>
<value name="x">undefined word class</value>

phrase type tags
<value name="np">noun phrase</value>
<value name="propp">name phrase</value>
<value name="vp">verb phrase</value>
<value name="ivp">verb phrase</value>
<value name="pp">prepositional phrase</value>
<value name="adjp">adjective phrase</value>
<value name="advp">adverb phrase</value>
<value name="cp">conjunction phrase</value>
<value name="qp">quantifier phrase</value>

clause tags
<value name="fcl">finite clause</value>
<value name="icl">non-finite clause</value>
<value name="acl">averbal clause</value>
<value name="par">paratagma</value>
<value name="cl">clause</value>
<value name="g">group</value>
<value name="x">undefined form</value>
<value name="xx">unspecified form</value>
<value name="VROOT">super node</value>
<value name="partial">partial tree</value>

tags for syntactic functions (edgelabels)
<value name="S">subject</value>
<value name="O">object</value>
<value name="Oaux">argument of auxiliary</value>
<value name="C">(subject) complement</value>
<value name="A">adverbial</value>
<value name="Aneg">negation particle</value>
<value name="P">predicator</value>
<value name="SUB">subordinator</value>
<value name="CO">coordinator</value>
<value name="CJT">conjunct</value>
<value name="H">head</value>
<value name="D">dependent</value>
<value name="DO">dependent</value>
<value name="DA">dependent</value>
<value name="Vmain">main verb</value>
<value name="Vmod">modal verb</value>
<value name="Vpart">particle verb</value>
<value name="Vneg">negation verb</value>
<value name="Vaux">auxiliar verb</value>
<value name="UTT">utterance</value>
<value name="STA">statement</value>
<value name="QUE">question</value>
<value name="COM">command</value>
<value name="EXC">exclamation</value>
<value name="ENUM">exclamation</value>
<value name="X">undefined function</value>
<value name="FST">punctuation</value>
```

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not relevant

4.5 Intended application of the corpus

The corpus can be used for building robust statistical language models and as a source of linguistic information.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Annotations have been assigned manually, every file has been annotated by two persons in parallel and the inconsistencies have been discussed and settled.

5 RELEVANT REFERENCES AND OTHER INFORMATION

E. Bick, H. Uibo, K. Müürisep. Arborest - a VISL-Style Treebank Derived from Estonian Constraint Grammar Corpus. Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004). Tübingen, Germany, Dec 10-11, 2004.

Müürisep, K.; Orav, H.; Õim, H.; Vider, K.; Kahusk, N.; Taremaa, P. (2008). From Syntax Trees in Estonian to Frame Semantics. In: The Third Baltic Conference on Human Language Technologies Procedings: The Third Baltic Conference on Human Language Technologies; Kaunas, Lithuania; 4-5. okt. 2007. (Toim.) Cermak, F.; Marcinkeviciene, R.; Rimkute, E.; Zabarskaite, J.. Vilnius: Vytautas Magnus University, Institute of the Lithuanian Language, 2008, 211 - 218.

3.3 Estinian corpus with morphological annotations, estmorfcorp

1 BASIC INFORMATION

1.1 Corpus composition

The Corpus consists of 300 000 words, the text classes represented in the Corpus are fiction, newspapers and popular science.

1.2 Representation of the corpora (flat files, database, markup)

The corpus is a file with XML markup

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Kadri Muischnek,

e-mail: kadri.muischnek@ut.ee

2.2 Copyright statement and information on IPR

The resource is free for research purposes, local license

3 TECHNICAL INFORMATION

3.1 Data structure of an entry

This is not relevant as the corpus is provided as a text file.

3.2 Corpora size (nmb. of tokens)

The corpus contains about 300 000 tokens

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual corpus with morphological annotations.

4.2 The natural language(s) of the corpus

The natural language of the corpus is Estonian.

4. 3 Domain(s)/register(s) of the corpus

Corpus represents Standard Written Estonian.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations

The corpus is annotated at paragraph, sentence and word level. Word-forms have been lemmatized and tagged for POS and relevant grammatical categories, i.e. case and number for nominals, additionally degree for adjectives and mood, tense, person, voice, positive/negative distinction for verbs.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

POS tags: type="POS" possible values of POS: S (noun), A (adjective), P (pronoun), N (cardinal numeral), O (ordinal numeral), V (verb), D (adverb), X (non-verbal part of the multi-word verb), K (adposition), J (conjunction), I (interjection), T (unknown word), Y (abbreviation), Z (punctuation mark)

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not relevant

4.5 *Intended application of the corpus*

The corpus can be used for building robust statistical language models and as a source of linguistic information.

4.6 *Reliability of the annotations (automatically/manually assigned) – if any*

Annotations have been assigned manually, every file has been annotated by two persons in parallel and the inconsistencies have been discussed and settled.

3.4 Estonian corpus with shallow syntactic annotation, estsyncorp

1 BASIC INFORMATION

1.1Corpus composition

The Corpus consists of ca 300 000 words, the text classes represented in the Corpus are fiction, both translated and original, and newspaper texts.

1.2 *Representation of the corpora (flat files, database, markup)*

flat files

1.3 *Character encoding*

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Kaili Müürisep

e-mail: kaili.muurisep@ut.ee

2.2 *Copyright statement and information on IPR*

The resource is free for research purposes, local license

3 TECHNICAL INFORMATION

3.1 Data structure of an entry

not relevant as the corpus is produces as a flat file

3.2 *Corpora size (nmb. of tokens)*

The corpus contains about 300 000 tokens

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual corpus with Constraint Grammar-style shallow syntactic annotations.

4.2. The natural language(s) of the corpus

The natural language of the corpus is Estonian.

4.3. Domain(s)/register(s) of the corpus

Corpus represents written Estonian.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations

The corpus is annotated at paragraph, sentence and word level. Word-forms have been lemmatized and tagged for

a) POS and relevant grammatical categories, i.e. case and number for nominals, additionally degree for adjectives and mood, tense, person, voice, positive/negative distinction for verbs;

b) syntactic functions

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

a) POS tags: S (noun), A (adjective), P (pronoun), N (cardinal numeral), O (ordinal numeral), V (verb), D (adverb), X (non-verbal part of the multi-word verb), K (adposition), J (conjunction), I (interjection), T (unknown word), Y (abbreviation), Z (punctuation mark)

b) tags for syntactic functions:
tags for verbal chain
@+FMV – main verb, finite form
@-FMV – main verb, infinite form
@+FCV – auxiliary, finite form
@-FCV – auxiliary, infinite form
@NEG – verb negation
tags for phrasal heads
@SUBJ – subject
@OBJ – object
@PRD – predicative
@ADVL – adverbial, also phrasal adverbial
tags for attributes:
@AN> - adjectival and ordinal numeral attribute preceding its head
@<AN - adjectival and ordinal numeral attribute following its head
@AD> - adverbial attribute preceding its head
@<AD – adverbial attribute following its head
@PN> - adpositional attribute preceding its head
@<PN – adpositional attribute following its head
@NN> - nominal, pronominal and cardinal numeral attribute preceding its head
@<NN - nominal, pronominal and cardinal numeral attribute following its head
@VN> - participal attribute preceding its head
@<VN – participal attribute following its head
@INF_N> - infinitival attribute preceding its head
@<INF_N – infinitival attribute following its head
tags for adpositional phrase:
@<P – nominal in a prepositional phrase
@>P – nominal in a postpositional phrase
@<Q – nominal governed by the quantifier and following it
@Q> - nominal governed by the quantifier and preceding it
other tags
@J – conjunct
@I - interjection

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not relevant

4.5 Intended application of the corpus

The corpus can be used for building robust statistical language models and as a source of linguistic information.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Annotations have been assigned manually, every file has been annotated by two persons in parallel and the inconsistencies have been discussed and settled.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Müürisep, K. (2001). Parsing Estonian with Constraint Grammar. *In: Online proceedings of Nordic Conference on Computational Linguistics: 13th Nordic Conference on Computational Linguistics NODALIDA-01; Uppsala, Sweden; May 21-22, 2000.* Uppsala:, 2001, 5 pp.

3.5 Estonian-english parallel corpus, estengparcorp

1. **BASIC INFORMATION**

1.1 Corpus composition

The corpus contains:

a) Estonian laws and their translations into English;

b) EU legislation in English and their translations into Estonian.

1.2 Representation of the corpora (flat files, database, markup)

The corpus is represented as flat files.

1.3 Character encoding

The characters are UTF8 encoded.

2. **ADMINISTRATIVE INFORMATION**

2.2 Contact person

Name: Kadri Muischnek,

e-mail: kadri.muischnek@ut.ee

2.2 Copyright statement and information on IPR

The resource is free for research purposes, local license.

3. **TECHNICAL INFORMATION**

3.1 Data structure of an entry

This is not relevant as the corpus is provided as a text file

3.2 Corpora size (nmb. of tokens)

The corpus contains about 153,500 parallel units (sentences or list items); 1.7 million tokens in Estonian, 2.9 million tokens in English.

4 **CONTENT INFORMATION**

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a semtence-aligned parallel corpus.

4.2 The natural language(s) of the corpus

The natural languages of the corpus are Estonian and English.

4. 3 Domain(s)/register(s) of the corpus

The corpus represents the legislative and bureaucratic language variety (eurospeak).

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The corpus is annotated at sentence level.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

The tags <eesti> and </eesti> delimit the Estonian part; <inglise> and </inglise> delimit the English part.

The subscripts and superscripts are tagged with <hi rend="sub"> and <hi rend="sup">. It often happens that the original or the translated unit contains one of them, but the corresponding parallel unit does not.

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

The texts have been sentence-aligned. The items of lists are treated as equal to sentences. The Estonian and English sentences may be in 1-1, 1-2 or 2-1 alignments. There are no other alignments (like 1-0, 0-1, 2-2 etc) in this corpus.

The aligning was done using the Vanilla aligner (nl.ijs.si/telri/Vanilla). It is a language independent aligner, based on the algorithm from: Gale, W. A. and Church, K. W. (1993) Program for aligning sentences in bilingual corpora. Computational Linguistics 19, 75-102.

4.5 Intended application of the corpus

The corpus can be used for training a machine translation system, bilingual term extraction, bilingual multi-word unit extraction etc

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Annotation and aligning have been done automatically and can contain mistakes.

3.6 Database of estonian multi-word expressions, estmwe

1. **BASIC INFORMATION**

1.1 *Lexicon type*:

lexicon of multi-word units

1.2 Representation of the lexicon:

flat file

1.3 Character encoding

The characters are UTF8 encoded.

2. ADMINISTRATIVE INFORMATION

2.1 *Contact person:*

Name: Kadri Muischnek;
e-mail: Kadri.Muischnek@ut.ee

2.2 Copyright statement and information on IPR

The resource is free for research purposes, local license

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

One entry per line; two fields: 1) the multi-word unit itself and 2) its morphological type, i.e consisting of a case form of a noun and a verb or of an particle and a verb

3.2 Lexicon size

12505 entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Estonian

4.2 Entry Type

All entries have the same entry type

4.3 *Coverage of the lexicon*

unknown

4.4 Intended application of the lexicon

It can be used for linguistic research, a gold standard for multi-word unit extraction task, can be used in lexicon-based tagging of multi-word items in text etc

4.5 Reliability (automatically/manually constructed)

Data has been automatically extracted from 6 dictionaries and from a text corpus

5 RELEVANT REFERENCES AND OTHER INFORMATION

Kaalep, H.-J.; Muischnek, K. Multi-Word Verbs of Estonian: a Database and a Corpus. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions: Marrakech; Morocco; 1. juuni 2008. , 2008, 23 - 26

3.7 Estonian reference corpus, estrefcorp

BASIC INFORMATION

1.1 Corpus composition

The corpus represents the written language and contains 75% newspaper texts, in lesser extent also fiction, science and legislation texts.

1.2 Representation of the corpora (flat files, database, markup)

The corpus is represented in TEI P5 format.

1.3 Character encoding

The characters are UTF8 encoded.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Kadri Muischnek,
e-mail: kadri.muischnek@ut.ee

2.2 Copyright statement and information on IPR

The resource is free for research purposes, local license

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

This is not relevant as the corpus is provided as a text file. It is structured in paragraphs, containing one or more sentences

3.2 Corpora size (nmb. of tokens)

The corpus contains about 245 000 000 tokens

4. CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is an unbalanced monolingual annotated corpus.

4.2 The natural language(s) of the corpus

The natural language of the corpus is standard Estonian.

4. 3 Domain(s)/register(s) of the corpus

Corpus represents Standard Written Estonian and contains 75% newspaper texts, in lesser extent also fiction, science and legislation texts.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The corpus is annotated at text structure (e.g. a novel and its chapters; a newspaper, its articles and sub-parts of these articles), paragraph and sentence level.

The following list includes all the tags and attributes used in the annotation. For more details about the TEI P5 format, see http://www.tei-c.org/Guidelines/P5/ .

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

Not relevant

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not relevant

4.5 Intended application of the corpus

The corpus can be used for building robust statistical language models. It can also be used as a reference corpus for Estonian in various corpus specific types of investigation: quantitative analysis, collocation extraction, grammar induction, etc.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Reliability varies in different subcorpora, but in general the annotation has been carried out automatically and can be erroneous.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Kaalep, H.-J.; Muischnek, K.; Uiboaed, K.; Veskis, K. The Estonian Reference Corpus: its composition and morphology-aware user interface The Fourth International Conference HUMAN LANGUAGE TECHNOLOGIES : THE BALTIC PERSPECTIVE, Riga, Latvia, October 7-8, 2010, 143 - 146

4 Norway (UIB)

4.1 Lexical database for Norwegian

1. **BASIC INFORMATION**
   1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),
   Wordform
   1.2 Representation of the lexicon (flat files, database, markup)
   Database
   1.3 Character encoding

   ISO-8859-1
   Speech Assessment Methods Phonetic Alphabet SAMPA

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)
Arne Martinus Lindstad, ArneMartinus.Lindstad@nb.no
2.2 Copyright statement and information on IPR
Open Source

3. TECHNICAL INFORMATION

3.1 Data structure of an entry
A lexicon entry is associated with information about 51 different features. Each entry covers one line in a text file that comprises the whole database.

3.2 Lexicon size (num. of lexical items)
784 240 items

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Norwegian (nbo)
4. 2 Entry Type
Lexicon item
4.3 Attributes
Pronunciation, POS, morphology, lemma, style, source and other information. 51 features in all.
4.4 Coverage of the lexicon
All words in *Bokmålsordboka* (the reference dictionary for bokmål)
All members of the 100k list (list of the 100 000 most frequent word forms)
All words in the *SpeechDat* database (telephone recordings)
Subsets of place names, person names, etc.
4.5 Intended application of the lexicon
Speech technology
4.5 Reliability (automatically/manually constructed)
Automatically constructed

5. **RELEVANT REFERENCES AND OTHER INFORMATION**
   http://www.nb.no/spraakbanken/tilgjengelege-ressursar/leksikalske-databasar

4.2 Lexical database for Swedish

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),
Wordform
1.2 Representation of the lexicon (flat files, database, markup)
Database
1.3 Character encoding

ISO-8859-1

Speech Assessment Methods Phonetic Alphabet SAMPA

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Arne Martinus Lindstad, ArneMartinus.Lindstad@nb.no

2.2 Copyright statement and information on IPR

Open source

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

A lexicon entry is associated with information about 51 different features. Each entry covers one line in a text file that comprises the whole database.

3.2 Lexicon size (num. of lexical items)

927 167

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Swedish

4. 2 Entry Type

Lexicon items

4.3 Attributes

POS, morphology, lemma, style, source

4.4 Coverage of the lexicon

Telia and SpeechDat material, subsets of person names, place names, etc.

4.5 Intended application of the lexicon

Speech technology

4.6 Reliability (automatically/manually constructed)

Automatically constructed

5. RELEVANT REFERENCES AND OTHER INFORMATION

http://www.nb.no/spraakbanken/tilgjengelege-ressursar/leksikalske-databasar

4.3 Lexical database for Danish - Available, includes documentation and toolset

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

Wordform

1.2 Representation of the lexicon (flat files, database, markup)

Database

1.3 Character encoding

ISO-8859-1
Speech Assessment Methods Phonetic Alphabet SAMPA

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Arne Martinus Lindstad, ArneMartinus.Lindstad@nb.no

2.2 Copyright statement and information on IPR

Open source

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

A lexicon entry is associated with information about 51 different features. Each entry covers one line in a text file that comprises the whole database.

3.2 Lexicon size (num. of lexical items)

237 873

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Danish (dan)

4. 2 Entry Type

Lexicon items

4.3 Attributes

POS, morphology, lemma, style, source

4.4 Coverage of the lexicon

Frequency based 100K list, INSO- og SpeechDat materials

4.5 Intended application of the lexicon

Speech technology

4.6 Reliability (automatically/manually constructed)

There is no automatic inflector for Danish, all lexicon items have been transcribed manually.

5. RELEVANT REFERENCES AND OTHER INFORMATION

http://www.nb.no/spraakbanken/tilgjengelege-ressursar/leksikalske-databasar

4.4 Acoustic database for Norwegian

1. BASIC INFORMATION

1.1 Corpus composition

Databases for speech recognition/ dictation

SpeechDat database (mobile phone and landline data for Norwegian, Swedish and Danish)
Recordings for speech recognition: ADB_OD_Nor.NOR database (recordings in office environment based on phonetically balanced manuscripts produced from sentences from the NST Norwegian corpus)
Recordings for dictation: ADB_D_IBM-N database produced from recordings in office environment based on phonetically balanced manuscripts produced from text material from the newspaper *Aftenposten*. Recordings for telephone services: ADB_T_Nor.NOR landline and mobile telephone recordings Database of non verbal sounds marking hesitation

Databases for speech synthesis

RealSpeak
IBM speech synthesis
IBM Phrase Splicing

1.2 *Representation of the corpora (flat files, database, markup)*

Database

1.3 Character encoding

Not relevant

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Arne Martinus Lindstad, ArneMartinus.Lindstad@nb.no

2.2 Copyright statement and information on IPR

Open Source

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

Soundfiles in PCM/wav-format
Spl-loggfiler in txt

3.2 Corpora size (num. of tokens)

873915 recordings

4. CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual

4.2 The natural language(s) of the corpus
Norwegian
4. 3 Domain(s)/register(s) of the corpus
General speech
Phone conversations
4.4 Annotations in the corpus (if an annotated corpus)
Not relevant
4.5 Intended application of the corpus
Speech recognition, dictation, speech synthesis
4.6 Reliability of the annotations (automatically/manually assigned) – if any
Automatically transcribed, validated by groups led by professionally trained linguists.

5. RELEVANT REFERENCES AND OTHER INFORMATION
http://www.nb.no/spraakbanken/tilgjengelege-ressursar/taledatabasar

4.5 Acoustic database for Swedish

1. BASIC INFORMATION
1.1 Corpus composition

SpeechDat database (mobile phone and landline data for Norwegian, Swedish and Danish)

Telia (recordings for speech recognition made in office environments in the Stockholm area)

Recordings for speech recognition: ADB_OD_Swe.SWE database (recordings in office environment based on phonetically balanced manuscripts produced from sentences from the NST Swedish corpus

Recordings for dictation: ADB_D_IBM-S database produced from recordings in office environment based on phonetically balanced manuscripts produced from news texts in the Swedish NST corpus
Database of non verbal sounds marking hesitation

1.2 Representation of the corpora (flat files, database, markup)

Database

1.3 Character encoding
Not relevant

2. ADMINISTRATIVE INFORMATION
2.1 Contact person (name, e-mail)

Arne Martinus Lindstad, ArneMartinus.Lindstad@nb.no

2.2 Copyright statement and information on IPR

Open Source

3 TECHNICAL INFORMATION
3.1 Data structure of an entry

Soundfiles in PCM/wav-format

Spl-loggfiler in txt

3.2 Corpora size (num. of tokens)

411 920 recordings

4 CONTENT INFORMATION
4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)
Monolingual, speech
4.2 The natural language(s) of the corpus
Swedish
4.3 Domain(s)/register(s) of the corpus
General speech
Phone conversations
4.4 Annotations in the corpus (if an annotated corpus)
Not relevant
4.5 Intended application of the corpus

Speech recognition, dictation, speech synthesis

4.6 Reliability of the annotations (automatically/manually assigned) – if any

No annotation, but the database has been validated by language assistants under the supervision of group leaders.

5 **RELEVANT REFERENCES AND OTHER INFORMATION**

http://www.nb.no/spraakbanken/tilgjengelege-ressursar/taledatabasar

4.6 Acoustic database for Danish

1 **BASIC INFORMATION**

1.1 Corpus composition

SpeechDat database (mobile phone and landline data for Norwegian, Swedish and Danish)

1.2 Representation of the corpora (flat files, database, markup)

Database

1.3 Character encoding

Not relevant

2 **ADMINISTRATIVE INFORMATION**

2.1 Contact person (name, e-mail)

Arne Martinus Lindstad, ArneMartinus.Lindstad@nb.no

2.2 Copyright statement and information on IPR

Open source

3 **TECHNICAL INFORMATION**

3.1 Data structure of an entry

Soundfiles in PCM/wav-format

Spl-loggfiler in txt

3.2 Corpora size (num. of tokens)

320947 recordings

4 **CONTENT INFORMATION**

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual

4.2 The natural language(s) of the corpus

Danish (dan)

4.3 Domain(s)/register(s) of the corpus

General speech
Phone conversations

4.4 Annotations in the corpus (if an annotated corpus)

Not relevant

4.5 Intended application of the corpus

Speech recognition, dictation, speech synthesis

4.6 Reliability of the annotations (automatically/manually assigned) – if any

No annotation, but the database has been validated by language assistants under the supervision of group leaders.

5 **RELEVANT REFERENCES AND OTHER INFORMATION**

http://www.nb.no/spraakbanken/tilgjengelege-ressursar/taledatabasar

4.7 Norsk ordbank

1. **BASIC INFORMATION**

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

Wordform with lemma

1.2 Representation of the lexicon (flat files, database, markup)

Lines with tab-separated fields

1.3 Character encoding

ISO-8859-1, with windows linebreaks (cr/lf)

2. **ADMINISTRATIVE INFORMATION**

2.1 Contact person (name, e-mail)

Arne Martinus Lindstad, ArneMartinus.Lindstad@nb.no

3.3 Copyright statement and information on IPR

Available on the condition of signing a form with conditions of use.

Produced by the University of Oslo

3. **TECHNICAL INFORMATION**

3.1 Data structure of an entry

Seven tab-separated fields, described in documentation.

3.2 Lexicon size (num. of lexical items)

784240 forms

4. **CONTENT INFORMATION**

4.1 The natural language(s) of the lexicon

Bokmål (Nbo) and Norwegian Nynorsk (Nno)

4.2 Entry Type

Wordform

4.3 Attributes

Lemma and morphological information, described in documentation

4.4 Coverage of the lexicon

The lexicon consists of the following resources:

a) Word lists and patterns of inflections produced by IBM Norway
b) Entries and inflection information from *Bokmålsordboka* og *Nynorskordboka* (standard dictionaries for bokmål and nynorsk), produced by ILN, University of Oslo
c) Codes for argument structure produced by NorKompLeks at NTNU (The Norwegian University of Science and Technology)

4.5 Intended application of the lexicon

Language technology

4.6 Reliability (automatically/manually constructed)

Manually and automatically constructed

5. **RELEVANT REFERENCES AND OTHER INFORMATION**

http://www.nb.no/spraakbanken/tilgjengelege-ressursar/leksikalske-databasar

http://www.hf.uio.no/iln/om/organisasjon/edd/forsking/norsk-ordbank/

4.8 Scarrie Lexicon

1. **BASIC INFORMATION**

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),
Wordform

1.2 Representation of the lexicon (flat files, database, markup)

LMF

1.3 Character encoding

Utf-8

2. **ADMINISTRATIVE INFORMATION**

2.1 Contact person (name, e-mail)
Koenraad de Smedt

2.2 Copyright statement and information on IPR
META-SHARE Commons BY SA
The copyright holder is Koenraad De Smedt. Reproduction and use is allowed under share-alike and authorship credit conditions.

3. **TECHNICAL INFORMATION**

3.1 Data structure of an entry
Each LexicalEntry has an unspecified Lemma and a number of WordForm elements.
Normal WordForm elements have elements Feat with att: writtenForm, corrStyle, featureList, replacement and synCat.
Special entries, in particular prefixes and word forms occurring in abbreviations neither have feat with att: replacement nor with att: synCat.
The use of the attributes is explained in the SCARRIE deliverable 3.3.1 Tagset.

3.2 Lexicon size (num. of lexical items)
scarrie-abbrevwords-lmf.xml:23
scarrie-gramwords-lmf.xml:707
scarrie-gramwords2-lmf.xml:48
scarrie-idiomwords-lmf.xml:811
scarrie-main-lmf.xml:359684
scarrie-prefixes-lmf.xml:327
scarrie-suffixes-lmf.xml:562

4. **CONTENT INFORMATION**

4.1 The natural language(s) of the lexicon
Norwegian (nbo)

4. 2 Entry Type
Entries are collections of word forms belonging to the same lemma.

4.3 Attributes
Elements of type feat have attributes att and val.

4.4 Coverage of the lexicon

The lexicon covers the vocabulary in Bokmålsordboka. It was constructed in 1999 and does not cover vocabulary extensions and spelling changes after Jan. 1, 1999.

4.5 Intended application of the lexicon

In its original format, the lexicon was intended for use in the proofreading application developed in the SCARRIE project (LE3-4239), as described in the deliverables from this project. In the current LMF format it is intended to be reused in other applications.

4.6 Reliability (automatically/manually constructed)

The lexicon was mostly automatically constructed. It was partly manually inspected and was used in a series of tests of the proofreading application.

5. **RELEVANT REFERENCES AND OTHER INFORMATION**

http://ling.uib.no/~desmedt/scarrie/

4.9 Sofie Treebank

1. Basic information

1.1 Corpus composition

The novel “Sofies verden” and translations

1.2 Representation of the corpora (flat files, database, markup)

Flat files and markup

2 Administrative information

2.1 Contact person (name, e-mail)

Victoria Rosén, victoria@uib.no

2.2 Copyright statement and information on IPR

Open Source

3 Technical information

3.1 Data structure of an entry

An LFG grammar assigns two representations to each reading of a sentence, a c-structure (a phrase structure tree) and an f-structure (an attribute-value graph).

Prolog file format

3.2 Corpora size (num. of tokens)

200 sentences

4 Content information

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual, annotated

4.2 The natural language(s) of the corpus

Norwegian (nbo)

4.3 Domain(s)/register(s) of the corpus

Fiction

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Syntactic mark-up (LFG)

4.5 Intended application of the corpus

Linguistic research

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Automatically annotated, manually checked

4 RELEVANT REFERENCES AND OTHER INFORMATION

http://iness.uib.no/iness/main-page?session-id=231331326413533

4.10 Oslo-Bergen tagger

1. BASIC INFORMATION

1.1 Resource composition

The Oslo-Bergen tagger is a robust morphological and syntactic tagger developed at the University of Oslo and at Uni Computing in Bergen over several years. The tagger consists of three main modules: a preprocessor with multitagger and compound analyser, a grammar module for morphological and syntactic disambiguation (Constraint Grammar) and a statistical module that removes the last of the remaining morphological ambiguity (only for Bokmål)

1.2 Representation of the resource (flat files, database, markup)

The downloadable resource contains the following elements:

a) Readme file with instructions on how to use the tagger. The readme file also shows where to find and how to install the various parts of the tagger:

b) Multitagger with lexicon for Norwegian Bokmål and Nynorsk. (At the moment, binary versions for Linux Intel 32-bit and 64-bit and og Mac OS X 64-bit are available)

c) CG3-compiler (VISL-CG3) from the University of Southern Denmark in Odense

d) OBT-stat (Statistical module that runs after the CG-rules)

e) HunPos

f) CG-rules for Bokmål and Nynorsk (morphological disambiguation). The Bokmål rules are available in two versions: One version for CG-grammar stand-alone and one version to be used together with OBT-stat.

g) Shell Script for running the tagger

1.3 Character encoding

Not relevant

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Janne Bondi Johannessen, j.b.johannessen@iln.uio.no

2.2 Copyright statement and information on IPR

Downloadable on GPL terms

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

Not relevant

3.2 Resource size (num. of rules)

Not relevant

4. CONTENT INFORMATION

4.1 Type of the resource (language (in(dependant)

Language dependent

4.2 The natural language(s) for the resource is applicable (if language dependent)

Norwegian bokmål and nynorsk (nbo, nno)

4.3 Domain(s)/register(s) of the resource

Not relevant

4.4 Annotations in the resource (if an annotated resource)

Not relevant

4.5 Intended application of the resource

Syntactic and morphological analysis

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Not relevant

5 RELEVANT REFERENCES AND OTHER INFORMATION

http://tekstlab.uio.no/obt-ny/english/read.html

4.11 TRIS corpus

1. **BASIC INFORMATION**

   1.1. *Corpus composition*

   Collection of domain specific national technical regulations of the EU member states

   1.2. *Representation of the corpora (flat files, database, markup)*

   Right now we have paired-MS Word files and pdf scanned documents. For the 30th of November we will upload just around 25 files in tmx format (translation memory exchange).

   1.3. *Character encoding*

   UTF-8

2. ADMINISTRATIVE INFORMATION

   2.1. *Contact person (name, e-mail)*

   Carla Parra Escartín, carla.parra@uib.no

   2.2. *Copyright statement and information on IPR*

   The database will be public and the files will have a special license to avoid commercial usage.

3. **TECHNICAL INFORMATION**

   3.1. *Data structure of an entry*

   TMX standard tags.

   3.2. *Corpora size (num. of tokens)*

   n/a

4. **CONTENT INFORMATION**

   4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

   Bilingual, parallel, aligned at sentence level, raw.

   4.2 The natural language(s) of the corpus

   German (Germany), German (Austria), Spanish (Spain)

   4.3 Domain(s)/register(s) of the corpus

   B00: CONSTRUCTION.

   4.4 Annotations in the corpus (if an annotated corpus)

   4.4.1.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

   none

   4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed)

   none

   4.4.3 Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

   Sentence alignment using the SDL Trados tool “Winalign”. Rresulting files are converted to tmx format automatically with a script.

   4.5 Intended application of the corpus

   Research in word alignment and automatic dictionary extraction. It can be used to other purposes though.

5 **RELEVANT REFERENCES AND OTHER INFORMATION**

For the moment we do not have any. In the future I hope we will have a paper describing the project and additional information.

5. Finland (UHEL)

5.1 Finnish WordNet

1. **BASIC INFORMATION**

   1.1 Lexicon type

   Wordnet

   1.2 Representation of the lexicon

   Online search interface, and downloadable in the Princeton WordNet database format, as lexicographer file (source format), or as various lists.

   1.3 Character encoding

   UTF-8

2. **ADMINISTRATIVE INFORMATION**

   2.1 Contact person

   Name: Krister Lindén
   E-mail: krister.linden@helsinki.fi

   2.2 Copyright statement and information on IPR

   PUB; Finnish translations CC-BY 3.0; Princeton WordNet licence

3. **TECHNICAL INFORMATION**

   3.1 Data structure of an entry

   Same as the Princeton WordNet

   3.2 Lexicon size

   Nouns, verbs, adjectives and adverbs, which make up about 117 000 synsets.

4. **CONTENT INFORMATION**

   4.1 The natural language(s) of the lexicon

   Finnish

   4. 2 Entry Type

   Synset

   4.3 Attributes

   Lexical-semantic relations like synonymy, hyponymy, antonymy.

   4.4 Coverage of the lexicon

   Translation of the English WordNet into Finnish.

   4.5 Intended application of the lexicon

   Can be used in language technology research and applications, and also interactively as an electronic thesaurus.

   4.6 Reliability (automatically/manually constructed)

   Professional translators translated the original Princeton WordNet (version 3.0) into Finnish.

5. **RELEVANT REFERENCES AND OTHER INFORMATION**

   Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. LexicoNordica – Nordic Journal of Lexicography, 17:119–140.

   http://www.ling.helsinki.fi/en/lt/research/finnwordnet/

5.2 Finnish TreeBank: Grammar Definition Corpus

1. **BASIC INFORMATION**

   1.1 Corpus composition

   Dependency-annotated example sentences of the Large Grammar of Finnish

   1.2 Representation of the corpus (flat files, database, markup)

   One tabular CoNLL-X file

   1.3 Character encoding

   UTF-8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Atro Voutilainen

E-mail: atro.voutilainen@helsinki.fi

2.2 Copyright statement and information on IPR

GNU LGPL v3.0

3 TECHNICAL INFORMATION

3.1 Data structure of an entry

CoNLL-X format, token per line with partial morphological annotation and syntactic dependency-links following the running number within the sentence, the surface form and the lemma.

3.2 Corpora size (num. of tokens)

160 000 tokens, 19 000 sentences

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual, annotated with partial morphology and dependency syntax

4.2 The natural language(s) of the corpus

Finnish

4.3 Domain(s)/register(s) of the corpus

Grammatical example sentences

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

sentences separated, lemmas, morphological tags indicate word class and inflection categories for each token, dependency-syntactic functions link dependent words to their heads

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

4.4.3 Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

4.5 Intended application of the corpus

Serves as a model for further grammatical analysis of Finnish

5.1. Reliability of the annotations (automatically/manually assigned) – if any

Manually annotated

5 RELEVANT REFERENCES AND OTHER INFORMATION

http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/index.shtml

http://www.scripta.kotus.fi/visk/etusivu.php (Iso suomen kielioppi on the web)

Atro Voutilainen, Tanja Purtonen, Satu Leisko-Järvinen, Mikaela Klami, 2010, "Suomen kielioppikorpus ja dependenssisyntaktinen kuvausmalli" [A grammar definition corpus and dependency syntactic representation of Finnish], available from the web site.

5.3 Corpus of Old Literary Finnish

1 BASIC INFORMATION

1.1 Corpus composition

Written Finnish texts from the years 1543–1809

1.2 Representation of the corpus (flat files, database, markup)

Browsable and searchable on the web

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Toni Suutari

E-mail: toni.suutari@kotus.fi

2.2 Copyright statement and information on IPR

Freely browsable and searchable

3 TECHNICAL INFORMATION

3.1 Data structure of an entry

Browsable plain text with a search interface on the web

3.2 Corpus size (num. of tokens)

3 428 618 words

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual, raw

4.2 The natural language(s) of the corpus

Finnish

4.3 Domain(s)/register(s) of the corpus

Bible translations and religious texts, legal texts, fairy tales, medical books, various.

4.4 Annotations in the corpus (if an annotated corpus)

No annotations

4.5 Intended application of the corpus

Lexicography

4.6 Reliability of the annotations (automatically/manually assigned) – if any

No annotations

5 **RELEVANT REFERENCES AND OTHER INFORMATION**

http://kaino.kotus.fi/korpus/vks/meta/vks_coll_rdf.xml

5.4 Corpus of Early Modern Finnish

1 BASIC INFORMATION

1.1 Corpus composition

Written Finnish from the 19th century (mostly from the years 1810-1880)

1.2 Representation of the corpora (flat files, database, markup)

Browsable and searchable on the web

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Toni Suutari

E-mail: toni.suutari@kotus.fi

2.2 Copyright statement and information on IPR

Freely browsable on the web

3 TECHNICAL INFORMATION

3.1 Data structure of an entry

Browsable plain text

3.2 Corpora size (num. of tokens)

8 645 700 words

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual, raw

4.2 The natural language(s) of the corpus

Finnish

4.3 Domain(s)/register(s) of the corpus

Published literature, periodicals, newspapers, dictionaries, among others, with a focus on the earliest and most important publications and a wide thematic coverage, and a preference for originally Finnish texts over translations.

4.4 Annotations in the corpus (if an annotated corpus)

No annotations

4.5 Intended application of the corpus

Research

4.6 Reliability of the annotations (automatically/manually assigned) – if any

No annotations

5 **RELEVANT REFERENCES AND OTHER INFORMATION**

http://kaino.kotus.fi/korpus/1800/meta/1800_coll_rdf.xml

5.5 Classics of Finnish Literature

1 **BASIC INFORMATION**

1.1 Corpus composition

Published works by established Finnish authors from the years 1880-1930

1.2 Representation of the corpora (flat files, database, markup)

Browsable on the web site

1.3 Character encoding

UTF-8

2 **ADMINISTRATIVE INFORMATION**

2.1 Contact person

Name: Toni Suutari

E-mail: toni.suutari@kotus.fi

2.2 Copyright statement and information on IPR

Freely browsable and searchable

3 **TECHNICAL INFORMATION**

3.1 Data structure of an entry

Browsable plain text

3.2 Corpus size

1 456 658 words

4 **CONTENT INFORMATION**

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual, raw

4.2 The natural language(s) of the corpus

Finnish

4.3 Domain(s)/register(s) of the corpus

Prose fiction, plays, lyric, aphorisms. Some translated from Swedish.

4.4 Annotations in the corpus (if an annotated corpus)

No annotations

4.5 Intended application of the corpus

Research and teaching

4.6 Reliability of the annotations (automatically/manually assigned) – if any

No annotations

5 **RELEVANT REFERENCES AND OTHER INFORMATION**

http://kaino.kotus.fi/korpus/klassikot/meta/klassikot_coll_rdf.xml (In Finnish)

5.6 Modern Finnish Word List

1 **BASIC INFORMATION**

1.1 Lexicon type

Word list

1.2 Representation of the lexicon

Lexical entries (lemma and inflection)

1.3 Character encoding

UTF-8

1.4 Contact person
Name: Toni Suutari
E-mail: toni.suutari@kotus.fi

1.5 Copyright statement and information on IPR
GNU LGPL

2 TECHNICAL INFORMATION

2.1 Data structure of an entry
Simple XML record of lemma and inflection types

2.2 Lexicon size
94 110 words

3 CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Finnish

4.2 Entry type
Lexical entries indicate a lemma and its inflection type

4.3 Attributes
Entries indicate the various inflection types for basic words but only the lemma for those compounds whose inflected part has its own entry. Rare inflections and other restrictions are indicated with XML attributes. Examples of the 78 inflection types and the 17 consonant gradation types are available on the web site.

4.4 Coverage of the lexicon
Intended to cover a useful amount of modern Finnish, including compounds and derivatives

4.5 Intended application of the lexicon
NLP applications

4.6 Reliability (automatically/manually constructed)
-

5 **RELEVANT REFERENCES AND OTHER INFORMATION**

http://kaino.kotus.fi/sanat/nykysuomi/ (In Finnish)

5.7 Frequency List of Written Finnish Word Forms

1 BASIC INFORMATION

1.1 Lexicon type
Frequency list

1.2 Representation of the corpora (flat files, database, markup)
Flat file of Unix-style records

1.3 Character encoding
ISO-8859-1 (Latin-1)

2 ADMINISTRATIVE INFORMATION

2.1 Contact person
Name: Toni Suutari
E-mail: toni.suutari@kotus.fi

2.2 Copyright statement and information on IPR
Freely downloadable

3 TECHNICAL INFORMATION

3.1 Data structure of an entry
Three text files of different size (and an HTML page of the 5000 most frequent forms)

3.2 Lexicon size
1 339 787 (full list), 542 521 (frequency > 1), 362 514 words (frequency > 2)

4 **CONTENT INFORMATION**

4.1 The natural language(s) of the lexicon

Finnish

4.2 Entry type

Word form per line

4.3 Attributes

Word form with its rank, absolute frequency, and relative frequency in the Finnish Parole corpus

4.4 Coverage of the lexicon

The inflected vocabulary of the Finnish Parole corpus (17 million tokens)

4.5 Intended application of the corpus

NLP applications

4.6 Reliability (automatically/manually constructed)

Automatically constructed

5 **RELEVANT REFERENCES AND OTHER INFORMATION**

http://kaino.kotus.fi/sanat/taajuuslista/parole.php (In Finnish)

6. Iceland (HI)

6.1 Icelandic Parsed Historical Corpus, IcePaHC

1. **BASIC INFORMATION**

1.1. Corpus composition

Treebank

1.2. Representation of the corpora (flat files, database, markup)

61 texts from 1150 to 2008. For each text there are three files: raw text; PoS tagged text (word, tag, lemma) in a flat file, one word per line; the texts in labelled bracketing format; With the corpus comes *Corpald*, a cross-platform graphical user interface to search corpora in labelled bracketing format. *Corpald* calls *CorpusSearch* by Beth Randall on the command line to execute search queries. (http://corpussearch.sourceforge.net/).

1.3. Character encoding

The characters have been encoded in UTF8.

2. **ADMINISTRATIVE INFORMATION**

2.1. Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Eiríkur Rögnvaldsson
Affiliation: Íslensku- og menningardeild Háskóla Íslands.
Address: Árnagarði, 101, Reykjavík, Iceland
E-mail: eirikur@hi.is

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

Available for download from own web page.

2.3. Copyright statement and information on IPR

Open Source, LGPL license

3. **TECHNICAL INFORMATION**

3.1. Directories and files

The package comes with several directories. The corpus itself is in the directory ‘corpora’ which has four subdirectories, one for bibliographic information for each of the 61 files (‘info’), one for the raw text (‘txt’), one for the tagged text (‘tagged’) and one for the parsed text in labelled bracketing format (‘psd’). There are also directories for information needed to install the software for searching the corpus (‘Corpald’).

3.2. Data structure of an entry

Each text sample is one text file. The parsed files are annotated according to the Penn Treebank format with certain modifications. A detailed description of the annotation is found in http://www.linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC)#Annotation_guidelines.

3.3. Corpora size (num. of tokens)
1 million running words

4. CONTENT INFORMATION

4.1. *Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*
Monolingual, annotated.

4.2. *The natural language(s) of the corpus*
Icelandic.

4.3. *Domain(s)/register(s) of the corpus*
Narratives (sagas, fiction), religious texts (bible, sermons),
science (linguistics, natural sciences, history),
formal texts (law, formal letters), biographical material (biographies, travelogues)

4.4. *Annotations in the corpus (if an annotated corpus)*

4.4.1. *Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)*
Sentence mark-up, syntactic mark-up

4.4.2. *Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),*
POS and syntactic tags.

4.4.3. *Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)*
Not applicable.

4.4.4. *Attributes and their values (if annotated)*
Not applicable.

4.5. Intended application of the corpus

The corpus is intended for use both within language technology and in syntactic research, synchronic and diachronic.

4.6. Reliability of the annotations (automatically/manually assigned) – if any
Mark-up automatically assigned, manually checked

***5.* RELEVANT REFERENCES AND OTHER INFORMATION**

Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson og Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC).* Version 0.9. http://www.linguist.is/icelandic_treebank

Eiríkur Rögnvaldsson, Anton Karl Ingason og Einar Freyr Sigurðsson. 2011. Coping with Variation in the Icelandic Parsed Historical Corpus (IcePaHC). Johannessen, Janne Bondi (ritstj.): *Language Variation Infrastructure. Papers on selected projects*, s. 97-111. Oslo Studies in Language 3.2. University of Oslo, Osló.

6.2 Icelandic Frequency Dictionary Corpus, IFD Corpus

1. BASIC INFORMATION

1.1. Corpus composition
Contemporary Icelandic texts from the years 1980-1990. The corpus contains 100 texts of about 5000 running words each. Texts were collected from printed books containing literary works for adults (20 texts written in Icelandic, 20 translated) and children (10 written in Icelandic, 10 translated), biographies (20 texts) and informative writings (10 from the humanities, 10 from natural sciences). The corpus will be made available in three ways: 1) most of the corpus will be available for online search. 2) 61 files (original Icelandic texts) are available for download under a special license. 3) Ten different disjoint pairs of files where in each pair there is a training set containing about 90% of running words from the corpus and a test set containing about 10% of running words from the corpus. The test sets are independent of each other whereas the training sets overlap and share about 80% of the examples. All words in the texts except proper nouns start with a lower case letter. These sets are compiled from all the texts in the corpus (100 texts). The files contain words and tags.

1.2. Representation of the corpora (flat files, database, markup)
Collection of TEI-conformant XML-files.

1.3. Character encoding
The characters are UTF8 encoded.

2. ADMINISTRATIVE INFORMATION

2.1. *Contact person (name, address, affiliation, position, telephone, fax, e-mail)*

Name: Sigrún Helgadóttir
Affiliation: The Arni Magnusson Institute for Icelandic Studies
Adress: Neshagi 16, 107, Reykjavík, Iceland
E-mail: sigruhel@hi.is

2.2. *Delivery medium (if relevant; description of the content of each piece of medium)*

Available for download from own webpage.

2.3. *Copyright statement and information on IPR*

Freely open for search, own license needed for download.

3. TECHNICAL INFORMATION

3.1. *Directories and files*

44 TEI conformant xml-files for download.

3.2. *Data structure of an entry*

The corpus is provided for download as TEI-conformant xml-files. Each file contains a header with bibliographic information. The text is segmented into sentences and each sentence into tokens where each token is equivalent to a word or a named entity. Each token (running word) is accompanied by a morphosyntactic tag and a lemma.

3.3. *Corpora size (nmb. of tokens, MB occupied on disk)*

About 300 thousand tokens for download, about 550 thousand for online search.

4. CONTENT INFORMATION

4.1. *Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*

Monolingual, annotated.

4.2. *The natural language(s) of the corpus*

Icelandic.

4.3. *Domain(s)/register(s) of the corpus*

Fiction both for adults and children, biographies, informative writings. The search interface will also provide access to translated fiction. Only texts written originally in Icelandic will be available for download.

4.4. *Annotations in the corpus (if an annotated corpus)*

4.4.1. *Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)*

Each file (text) has a header containing bibliographic data. Text is segmented into sentences and sentences into words. Each word is assigned a morphosyntactic tag (MSD), and lemmata.

4.4.2. *Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),*

The texts in the corpus are annotated with morphosyntactic tags and lemmata. The tagset was developed for this corpus that was originally used to make a Frequency Dictionary for Icelandic. (Pind et al., 1991). The corpus was part-of-speech tagged by semi-automatic means, all morphosyntactic tags and lemmas were manually corrected.

4.4.3. *Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)*

Not applicable.

4.4.4. *Attributes and their values (if annotated)*

The *s* tag has one attribute: *n* which identifies the sentence in the text.
The *w* tag is present at the token level and can have two attributes: *type* whose value is the morphosyntactic tag and *lemma* whose value is the dictionary form of the word form.

4.5. *Intended application of the corpus*

The corpus has been used to train PoS taggers for Icelandic. The online search will be useful for teaching purposes, both for Icelanders and foreigners. This will give guidance on language use and morphosyntactic analysis. 2) The downloadable corpus will be useful for various LT projects. 3) The ten disjoint pairs will be used for training PoS taggers.

4.6. *Reliability of the annotations (automatically/manually assigned) – if any*

Annotations are automatically checked.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Jörgen Pind (ed.), Friðrik Magnússon and Stefán Briem. 1991. *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík. [Referred to as the Icelandic Frequency Dictionary, IFD.]

2004h. Sigrún Helgadóttir. Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In H. Holmboe (ed.), *Nordisk Sprogteknologi 2004*. Museum Tusculanums Forlag.

Sigrún Helgadóttir. Mörkun íslensks texta *Orð og tunga* 9:75-107. Reykjavík. 2007.

HRAFN LOFTSSON. 2006. Tagging a morphologically complex language using heuristics. In T. Salakoski, F. Ginter, S. Pyysalo and T. Pahikkala (eds.), *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Proceedings*. Turku, Finland.

HRAFN LOFTSSON. 2007. Tagging Icelandic Text using a Linguistic and a Statistical Tagger. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the ACL*. Rochester, NY, USA.

HRAFN LOFTSSON. 2009. Correcting a POS-Tagged Corpus Using Three Complementary Methods. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece.

6.3 Parliament Speech Corpus

1. BASIC INFORMATION

1.1. Corpus composition

About 20 hours of unprepared speeches in parliament 2004-2005, synchronized speech and text files. The corpus consists of sound files, transcribed speech (output of the software Transcriber) and TEI-conformant xml-files with morphosyntactic tags and lemmas.

1.2. Representation of the corpora (flat files, database, markup)

The corpus consists of sound files (mp3), transcribed speech (output of the software Transcriber) and TEI-conformant xml-files with morphosyntactic tags and lemmas. The corpus is also available for online search.

1.3. Character encoding

The characters have been encoded in UTF8.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Ásta Svavarsdóttir
Affiliation: The Arni Magnusson Institute for Icelandic Studies
Adress: Neshagi 16, 107, Reykjavík, Iceland
E-mail: asta@hi.is

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

Available for download from own webpage.

2.3. Copyright statement and information on IPR

CLARIN PUB license.

3. TECHNICAL INFORMATION

3.1. Directories and files

One directory, set of files (mp3, trs, xml) from 12 periods.

3.2. Data structure of an entry

The corpus is provided for download as (1) TEI-conformant xml-files. Each file contains a header with bibliographic information. The text is segmented into sentences and each sentence into tokens where each token is equivalent to a word or a named entity. Each token (running word) is accompanied by a morphosyntactic tag and a lemma. (2) Transcriber files where text and sound is synchronized. (3) Sound files (mp3-files).

3.3. Corpora size (nmb. of tokens, MB occupied on disk)

About 190 thousand running words, 20 hours of speech.

4. CONTENT INFORMATION

4.1. *Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*

Monolingual, annotated.

4.2. *The natural language(s) of the corpus*

Icelandic.

4.3. *Domain(s)/register(s) of the corpus*

Parliamentary talk.

4.4. *Annotations in the corpus (if an annotated corpus)*

4.4.1. *Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)*

Sentence mark-up, morphosyntactic mark-up, synchronization of text and sound.

4.4.2. *Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),*

Morphosyntactic tags, lemmas

4.4.3. *Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)*

Not applicable.

4.4.4. *Attributes and their values (if annotated)*

For the xml-files the following holds: The *s* tag has one attribute: *n* which identifies the sentence in the text. The *w* tag is present at the token level and can have two attributes: *type* whose value is the morphosyntactic tag and *lemma* whose value is the dictionary form of the word form.

4.5. Intended application of the corpus

Web-version will be used for linguistic investigations. Download version will be used for LT projects such as speech analysis, speech recognition, speech synthesis, automatic speech recognition.

4.6. Reliability of the annotations (automatically/manually assigned) – if any

Synchronization performed manually, PoS tagging and lemmatization performed automatically.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Ásta Svavarsdóttir. 2007. Talmál og málheildir — talmál og orðabækur. *Orð og tunga* 9:25-50.

Höskuldur Thráinsson, Ásgrímur Angantýsson, Ásta Svavarsdóttir, Thórhallur Eythórsson and Jóhannes Gísli Jónsson. 2007. The Icelandic (Pilot) Project in ScanDiaSyn. *Nordlyd. Tromsø University Working Papers on Language & Linguistics*, Vol. 34, Nr. 1: 87-124. (Sérhefti um Scandinavian Dialect Syntax 2005). Vefrit á slóðinni http://www.ub.uit.no/baser/nordlyd/viewissue.php?id=11.

6.4 Hjal Speech Corpus, HJAL corpus

1. BASIC INFORMATION

1.1. Corpus composition

Synchronized text and speech, sampled from 2005 individuals, text is transcribed and recorded in SAMPA standard.

1.2. Representation of the corpora (flat files, database, markup)

Sound and text files.

1.3. Character encoding

The characters have been encoded in UTF8.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Eiríkur Rögnvaldsson
Affiliation: Íslensku- og menningardeild Háskóla Íslands.
Address: Árnagarði, 101, Reykjavík, Iceland
E-mail: eirikur@hi.is

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

Available for download from own webpage.

2.3. Copyright statement and information on IPR

CLARIN PUB license.

3. TECHNICAL INFORMATION

3.1. Directories and files

883 subdirectories. Each subdirectory contains data from one speaker, both sound files and a text file.
- usually 47 sound files for each speaker, each file containing one utterance, and one text file containing the transcription of all the sound files.

3.2. Data structure of an entry

The sound files are in .wav format - usually 47 sound files for each speaker, each file containing one utterance. The transcription of all the sound files for each speaker is contained in one text file.

3.3. *Corpora size (nmb. of tokens, MB occupied on disk)*
42,000 sound files, 883 text files, 1,4 GB.

4. CONTENT INFORMATION

4.1. *Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*
Monolingual, annotated?.

4.2. *The natural language(s) of the corpus*
Icelandic.

4.3. *Domain(s)/register(s) of the corpus*
Individual words.

4.4. *Annotations in the corpus (if an annotated corpus)*

4.4.1. *Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)*
Not applicable.

4.4.2. *Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),*
SAMPA phonetic alphabet.

4.5. Intended application of the corpus
For training speech recognizers.

4.6. Reliability of the annotations (automatically/manually assigned) – if any
Transcription and phonetic transcription performed manually.

***5.* RELEVANT REFERENCES AND OTHER INFORMATION**

Rögnvaldsson, Eiríkur (2004). The Icelandic Speech Recognition Project Hjal. In Holmboe, Henrik Ed.), *Nordisk Sprogteknologi. Årbog 2003*. Museum Tusculanums Forlag, University of Copenhagen, pp. 239-242.

6.5 Pronunciation Dictionary for Icelandic

1 BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),
Pronunciation dictionary.

1.2 Representation of the lexicon (flat files, database, markup)
Excel-file.

1.3 Character encoding
Characters are encoded according to ANSI standard

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)
Name: Eiríkur Rögnvaldsson
Affiliation: Íslensku- og menningardeild Háskóla Íslands.
Address: Árnagarði, 101, Reykjavík, Iceland
E-mail: eirikur@hi.is

2.2 Delivery medium (if relevant; description of the content of each piece of medium)
Available for download from own webpage.

2.3 Copyright statement and information on IPR
CLARIN PUB

3 TECHNICAL INFORMATION

3.1 Directories and files
One Excel-file.

3.2 Data structure of an entry
Three columns, word, IPA transcription, SAMPA transcription.

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)
56 thousand lexical items.

4 CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Icelandic

4. 2 Entry Type

Word

4.3 Attributes and their values

Not applicable.

4.4 Coverage of the lexicon

56 thousand words.

4.5 Intended application of the lexicon

Speech recognition, speech synthesis

4.6 POS assignment

Not applicable.

4.7 Reliability (automatically/manually constructed)

The resource is manually created.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Rögnvaldsson, Eiríkur (2004). The Icelandic Speech Recognition Project Hjal. In Holmboe, Henrik (Ed.), *Nordisk Sprogteknologi. Årbog 2003*. Museum Tusculanums Forlag, University of Copenhagen, pp. 239-242.

6.6 The Saga Corpus

1. BASIC INFORMATION

1.1. Corpus composition

Texts from 44 Sagas.

1.2. DanNet

1.3. Representation of the corpora (flat files, database, markup)

Collection of TEI-conformant XML-files.

1.4. Character encoding

The characters are UTF8 encoded.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Eiríkur Rögnvaldsson
Affiliation: Íslensku- og menningardeild Háskóla Íslands.
Address: Árnagarði, 101, Reykjavík, Iceland
E-mail: eirikur@hi.is

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

Available for download from own web page.

2.3. Copyright statement and information on IPR

No copyright, CLARIN PUB license.

3. TECHNICAL INFORMATION

3.1. Directories and files

44 TEI conformant xml-files.

3.2. Data structure of an entry

Each file contains a header with bibliographic information. The text is segmented into sentences and each sentence into tokens where each token is equivalent to a word or a named entity. Each token (running word) is accompanied by a morphosyntactic tag and a lemma.

3.3. Corpora size (nmb. of tokens, MB occupied on disk)

1,659,385 tokens.

4. CONTENT INFORMATION

4.1. *Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*

Monolingual, annotated.

4.2. *The natural language(s) of the corpus*

Icelandic.

4.3. *Domain(s)/register(s) of the corpus*

The language of the Icelandic Sagas, transliterated to modern spelling.

4.4. *Annotations in the corpus (if an annotated corpus)*

4.4.1. *Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)*
Each file (text) has a header containing bibliographic data. Text is segmented into sentences and sentences into words. Each word is assigned a morphosyntactic tag (MSD), and lemmata.

4.4.2. *Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),*
The texts in the corpus are automatically annotated with morphosyntactic tags and lemmata. The tagset was developed for the IFD Corpus (Pind et al., 1991). Morphosyntactic tags and lemmas were not manually corrected.

4.4.3. *Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)*
Not applicable.

4.4.4. *Attributes and their values (if annotated)*
The *s* tag has one attribute: *n* which identifies the sentence in the text.
The *w* tag is present at the token level and can have two attributes: *type* whose value is the morphosyntactic tag and *lemma* whose value is the dictionary form of the word form.

4.5. Intended application of the corpus
For researchers to study Old Icelandic. The corpus will be available for download and web search.

4.6. Reliability of the annotations (automatically/manually assigned) – if any
Annotations are automatically assigned and not manually checked. Tagging accuracy has been estimated as 92.7% (Rögnvaldsson and Helgadóttir, 2011).

5. **RELEVANT REFERENCES AND OTHER INFORMATION**

Eiríkur Rögnvaldsson and Sigrún Helgadóttir. 2011. Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. Sporleder, Caroline, Antal P.J. van den Bosch og Kalliopi A. Zervanou (ritstj:): *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, s. 63-76. Springer, Berlín.

7. Lithuania (LKI)

7.1 Modern Lithuanian dictionary (DŽ)

1 BASIC INFORMATION

1.1 Lexicon type
LexicalConceptualResource

1.2 Representation of the lexicon
The lexicon is stored in a relational database and is accessible via RESTful API (in JSON, LMF and XML with XSD formats) or via user-friendly web interface.

1.3 Character encoding
The characters have been encoded in UTF8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person
Name: Anželika Gaidienė
E-mail: anzelika.gaidiene@gmail.com

2.2 Copyright statement and information on IPR
CLARIN ACA + NC

3 TECHNICAL INFORMATION

3.1 Data structure of an entry
Dictionary entry consists of a headword, word forms, stress, word senses, examples and links to related dictionary entries.

3.2 Lexicon size
About 50000 dictionary entries, covers about 80000 words.

4 CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Lithuanian

4.2 Entry Type
Dictionary entry (headword)
4.3 Attributes
All attributes are listed in XSD.
4.4 Coverage of the lexicon
The lexicon covers Modern Lithuanian.
4.5 Intended application of the lexicon
Human usage and other language technology tools.
4.6 Reliability (automatically/manually constructed)
The lexicon is manually constructed.

7.2 Database of Neologisms (Neologisms)

1. BASIC INFORMATION

1.1 Lexicon type
LexicalConceptualResource
1.2 Representation of the lexicon
The lexicon is stored in a relational database and is accessible via RESTful API (in JSON, LMF and XML with XSD formats) or via user-friendly web interface.
1.3 Character encoding
The characters have been encoded in UTF8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person
Name: Rita Miliūnaitė
E-mail: ritam@lki.lt
2.2 Copyright statement and information on IPR
CLARIN ACA + NC

3. TECHNICAL INFORMATION

3.1Data structure of an entry
Dictionary entry contains a headword, a stressed word, stress form, classification form and type, part of speech, definition, origin, original form, usage domain, examples and sources.
3.2 Lexicon size
4000ry

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Lithuanian
4.2 *Entry Type*
Dictionary entry (headword)
4.3 Attributes
All attributes are listed in XSD
4.3 Coverage of the lexicon
The lexicon covers borrowings and recently-coined words, phrases and abbreviations of Lithuanian.
1.5 Intended application of the lexicon
Human usage and other language technology tools.
4.1 Reliability (automatically/manually constructed)
The lexicon is manually constructed

5.RELEVANT REFERENCES AND OTHER INFORMATION

Monografija (eng. Monograph): Rita Miliūnaitė. Dabartinės lietuvių kalbos vartosenos variantai. Vilnius: Lietuvių kalbos instituto leidykla, 2009. 248 p. ISBN 978-609-411-021-4.

8 Sweden (UGOT)

8.1 SALDO

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon

1.2 Representation of the lexicon
LMF

1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)
Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR
Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.2 Data structure of an entry
*Structured according to the LMF specification (*http://www.lexicalmarkupframework.org/*).*

5.5 Lexicon size
122 213 lexical entries.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Swedish

4. 2 Entry Type
LexicalEntry in LMF.

4.3 Attributes
The resource contains two kinds of lexical-semantic relations, and for every sense, a morphological specification.

4.4 Coverage of the lexicon
A full-scale lexicon covering all parts of speech, including multi-word units.

4.5 Intended application of the lexicon
The intended use is language technology applications.

a. Reliability (automatically/manually constructed)
The resource has been manually constructed with the aid of automatic methods.

5.RELEVANT REFERENCES AND OTHER INFORMATION

Lars Borin, Markus Forsberg 2009. All in the family: A comparison of SALDO and WordNet *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series* [pdf]

8.2 SALDO morphology

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon

1.2 Representation of the lexicon
LMF

1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR

Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

Structured according to the LMF specification (http://www.lexicalmarkupframework.org/*).*

1.2 Lexicon size

114 316 lexical entries., 1 791 527 word forms

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Swedish

4. 2 Entry Type

LexicalEntry in LMF.

4.3 Attributes

The resource contains the morphological information of SALDO.

4.4 Coverage of the lexicon

A full-scale lexicon covering all parts of speech, including multi-word units.

4.5 Intended application of the lexicon

The intended use is language technology applications.

4.6 Reliability (automatically/manually constructed)

The resource has been manually constructed with the aid of automatic methods.

5.RELEVANT REFERENCES AND OTHER INFORMATION

Lars Borin, Markus Forsberg 2009. All in the family: A comparison of SALDO and WordNet *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series* [pdf]

8.3 SALDO examples

1. BASIC INFORMATION

1.1 Lexicon type

Computational lexicon

1.2 Representation of the lexicon

LMF

1.3 Character encoding

The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR

Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1Data structure of an entry

Structured according to the LMF specification (http://www.lexicalmarkupframework.org/).

3.2 Lexicon size

1 177 lexical entries.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Swedish

4. 2 Entry Type
LexicalEntry in LMF.
4.3 Attributes
The resource contains example sentences for SALDO senses.
4.4 Coverage of the lexicon
A small lexicon containing mostly verb examples.
4.5 Intended application of the lexicon
The intended use is language technology applications.
4.6 Reliability (automatically/manually constructed)
The resource has been manually constructed with the aid of automatic methods.

5.RELEVANT REFERENCES AND OTHER INFORMATION

Lars Borin, Markus Forsberg 2009. All in the family: A comparison of SALDO and WordNet *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series* [pdf]

8.4 Swedish Framenet

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon
1.2 Representation of the lexicon
LMF
1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)
Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se
2.2 Copyright statement and information on IPR
Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry
Structured according to the LMF specification (http://www.lexicalmarkupframework.org/).
3.2 Lexicon size
19 262 lexical entries, 538 frames

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Swedish
4. 2 Entry Type
LexicalEntry in LMF.
4.3 Attributes
The resource follows the description of the Berkeley Framenet (https://framenet.icsi.berkeley.edu) as long as it has a natural correspondence in the Swedish language.
4.4 Coverage of the lexicon
The Swedish Framenet currently covers around 50% of the frame set of the Berkeley Framenet.
4.5 Intended application of the lexicon
The intended use is language technology applications.
4.6 Reliability (automatically/manually constructed)
The resource has been manually constructed with the aid of automatic methods.

5.RELEVANT REFERENCES AND OTHER INFORMATION

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, Dimitrios Kokkinakis 2010. The Past Meets the Present in the Swedish FrameNet++ *14th EURALEX International Congress* [pdf]

8.5 Swesaurus

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon
1.2 Representation of the lexicon
LMF
1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION
2.1 Contact person (name, e-mail)
Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se
2.2 Copyright statement and information on IPR
Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION
3.1 Data structure of an entry
Structured according to the LMF specification (http://www.lexicalmarkupframework.org/).
3.2 Lexicon size
23 954 lexical entries

4. CONTENT INFORMATION
4.1 The natural language(s) of the lexicon
Swedish
4. 2 Entry Type
LexicalEntry in LMF.
4.3 Attributes
The resource currently contains graded synonymy relations, but will soon include other lexical-semantic relations and links to Core Princeton Wordnet.
4.4 Coverage of the lexicon
The goal of the resource is the creation of a Swedish Wordnet with graded synonymy.
4.5 Intended application of the lexicon
The intended use is language technology applications.
4.6 Reliability (automatically/manually constructed)
The resource has been semi-automatically constructed using information available in free lexical resources.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, Dimitrios Kokkinakis 2010. The Past Meets the Present in the Swedish FrameNet++ *14th EURALEX International Congress* [pdf]

Lars Borin, Markus Forsberg 2010. From the People's Synonym Dictionary to fuzzy synsets - first steps *Proceedings of the LREC 2010 workshop Semantic relations. Theory and Applications* [pdf]

8.6 Parole lexicon

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon
1.2 Representation of the lexicon
LMF

1.3 Character encoding

The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Name: Markus Forsberg

E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR

Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

Structured according to the LMF specification (http://www.lexicalmarkupframework.org/).

3.2 Lexicon size

35 006 lexical entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Swedish

4. 2 Entry Type

LexicalEntry in LMF.

4.3 Attributes

The resource contains sub-categorization frames for Swedish nouns, verbs, and adjectives.

4.4 Coverage of the lexicon

The resource covers nouns, verbs, and adjectives.

4.5 Intended application of the lexicon

The intended use is language technology applications.

4.6 Reliability (automatically/manually constructed)

The resource has been created in the PAROLE project, and linked with automatic methods to SALDO.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, Dimitrios Kokkinakis 2010. The Past Meets the Present in the Swedish FrameNet++ *14th EURALEX International Congress* [pdf]

PAROLE, Preparatory Action for Linguistic Resources Organization for Language Engineering, is an EC funded project aiming at generating LE resources.

8.7 Simple

1. BASIC INFORMATION

1.1 Lexicon type

Computational lexicon

1.2 Representation of the lexicon

LMF

1.3 Character encoding

The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Name: Markus Forsberg

E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR

Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

Structured according to the LMF specification (http://www.lexicalmarkupframework.org/).

3.2 Lexicon size

11 624 lexical entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Swedish

4. 2 Entry Type

LexicalEntry in LMF.

4.3 Attributes

The resource extends a subset of PAROLE with the semantic description developed in the SIMPLE project.

4.4 Coverage of the lexicon

The resource covers nouns, verbs, and adjectives.

4.5 Intended application of the lexicon

The intended use is language technology applications.

4.6 Reliability (automatically/manually constructed)

The resource has been created in the SIMPLE project, and linked with automatic methods to SALDO.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, Dimitrios Kokkinakis 2010. The Past Meets the Present in the Swedish FrameNet++ *14th EURALEX International Congress* [pdf]

SIMPLE, Semantic Information for Multifunctional Plurilingual Lexica is an EC funded project with focus on the encoding of semantic data relevant for NLP tasks.

8.8 Swedish LWT list

1. BASIC INFORMATION

1.1 Lexicon type

Computational lexicon

1.2 Representation of the lexicon

LMF

1.3 Character encoding

The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Name: Markus Forsberg

E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR

Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

Structured according to the LMF specification (http://www.lexicalmarkupframework.org/).

3.2 Lexicon size

1 460 lexical entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Swedish

4. 2 Entry Type

LexicalEntry in LMF.

4.3 Attributes

The resource links the Loan Word Typology list to SALDO.

4.4 Coverage of the lexicon

The resource covers the LWT list.

4.5 Intended application of the lexicon
The intended use is typological studies.
4.6 Reliability (automatically/manually constructed)
The resource has been manually linked using automatic support methods.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, Dimitrios Kokkinakis 2010. The Past Meets the Present in the Swedish FrameNet++ *14th EURALEX International Congress* [pdf]

8.9 Swedish Kelly list

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon
1.2 Representation of the lexicon
LMF
1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)
Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se
2.2 Copyright statement and information on IPR
Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry
*Structured according to the LMF specification (*http://www.lexicalmarkupframework.org/*).*
3.2 Lexicon size
8 425 lexical entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Swedish
4. 2 Entry Type
LexicalEntry in LMF.
4.3 Attributes
The resource links the Swedish Kelly list to SALDO.
4.4 Coverage of the lexicon
The resource covers a core vocabulary developed in the Kelly project.
4.5 Intended application of the lexicon
The intended use is language educations.
4.6 Reliability (automatically/manually constructed)
The resource has been created and linked using automatic methods.

8.10 Dalin

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon
1.2 Representation of the lexicon
LMF

1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)
Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR
Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry
*Structured according to the LMF specification (*http://www.lexicalmarkupframework.org/*).*

3.2 Lexicon size
62 973 lexical entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Swedish

4. 2 Entry Type
LexicalEntry in LMF.

4.3 Attributes
The resource is a digitized version of "Ordbok Öfver Svenska Språket", a paper dictionary of 19th century Swedish.

4.4 Coverage of the lexicon
The resource covers the whole dictionary, including the supplement.

4.5 Intended application of the lexicon
The intended use is human consumption and language technology applications.

4.6 Reliability (automatically/manually constructed)
The resource has been manually digitized.

5. RELEVANT REFERENCES AND OTHER INFORMATION

DALIN, ANDERS FREDRIK *ORDBOK ÖFVER SVENSKA SPRÅKET*. VOL I–II. (STOCKHOLM 1850-1855).

Lars Borin, Markus Forsberg 2011. A diachronic computational lexical resource for 800 years of Swedish *Language technology for cultural heritage,* **Springer. 41-61**

8.11 Dalin morphology

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon

1.2 Representation of the lexicon
LMF

1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)
Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR
Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

Structured according to the LMF specification (http://www.lexicalmarkupframework.org/).

3.2 Lexicon size

62 797 lexical entries, 730 571 word forms

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Swedish

4. 2 Entry Type

LexicalEntry in LMF.

4.3 Attributes

The resource is a contains morphological information for the Dalin lexicon.

4.4 Coverage of the lexicon

The resource enrich a large part of the dictionary with full morphological information, and use the baseform for the rest of the entries.

4.5 Intended application of the lexicon

The intended use is language technology applications.

4.6 Reliability (automatically/manually constructed)

The morphological description is done manually, but the assigment of inflectional information to entries in the Dalin lexicon has been done automatically. The quality needs improvement.

9 RELEVANT REFERENCES AND OTHER INFORMATION

Lars Borin, Markus Forsberg 2011. A diachronic computational lexical resource for 800 years of Swedish *Language technology for cultural heritage,* **Springer. 41-61**

8.12 Schlyter

1. BASIC INFORMATION

1.1 Lexicon type

Computational lexicon

1.2 Representation of the lexicon

LMF

1.3 Character encoding

The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Name: Markus Forsberg

E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR

Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

Structured according to the LMF specification (http://www.lexicalmarkupframework.org/).

3.2 Lexicon size

10 067 lexical entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Swedish

4. 2 Entry Type

LexicalEntry in LMF.

4.3 Attributes
The resource is a digitized version of "Ordbok till Samlingen af Sweriges Gamla Lagar", a paper dictionary of Old Swedish.

4.4 Coverage of the lexicon
The resource covers the whole dictionary.

4.5 Intended application of the lexicon
The intended use is human consumption and language technology applications.

4.6 Reliability (automatically/manually constructed)
The resource has been manually digitized, but the conversion to LMF is not yet complete.

5. RELEVANT REFERENCES AND OTHER INFORMATION

SCHLYTER, C.J. *ORDBOK TILL SAMLINGEN AF SWERIGES GAMLA LAGAR*. (SAML. AF SWERIGES GAMLA LAGAR 13.) LUND 1877.

Lars Borin, Markus Forsberg 2011. A diachronic computational lexical resource for 800 years of Swedish *Language technology for cultural heritage*, **Springer. 41-61**

8.13 Söderwall

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon

1.2 Representation of the lexicon
LMF

1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)
Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se

2.2 Copyright statement and information on IPR
Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry
*Structured according to the LMF specification (*http://www.lexicalmarkupframework.org/*).*

3.2 Lexicon size
22 571 lexical entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Swedish

4. 2 Entry Type
LexicalEntry in LMF.

4.3 Attributes
The resource is a digitized version of "Ordbok Öfver svenska medeltids-språket", a paper dictionary of Old Swedish.

4.4 Coverage of the lexicon
The resource covers the whole dictionary.

4.5 Intended application of the lexicon
The intended use is human consumption and language technology applications.

4.1 Reliability (automatically/manually constructed)
The resource has been manually digitized, but the conversion to LMF is not yet complete.

5. RELEVANT REFERENCES AND OTHER INFORMATION

SÖDERWALL, K.F. *ORDBOK ÖFVER SVENSKA MEDELTIDS-SPRÅKET*. VOL I-III. (LUND 1884-1918)

Lars Borin, Markus Forsberg 2011. A diachronic computational lexical resource for 800 years of Swedish *Language technology for cultural heritage*, **Springer. 41-61**

8.14 Söderwall Supplement

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon
1.2 Representation of the lexicon
LMF
1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)
Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se
2.2 Copyright statement and information on IPR
Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION

3.1 Data structure of an entry
Structured according to the LMF specification (http://www.lexicalmarkupframework.org/).
3.2 Lexicon size
19 172 lexical entries

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon
Swedish
4. 2 Entry Type
LexicalEntry in LMF.
4.3 Attributes
The resource is a digitized version of "Ordbok Öfver svenska medeltids-språket. Supplement", a paper dictionary of Old Swedish.
4.4 Coverage of the lexicon
The resource covers the whole dictionary.
4.5 Intended application of the lexicon
The intended use is human consumption and language technology applications.
4.6 Reliability (automatically/manually constructed)
The resource has been manually digitized, but the conversion to LMF is not yet complete.

5. RELEVANT REFERENCES AND OTHER INFORMATION

SÖDERWALL, K.F. *ORDBOK ÖFVER SVENSKA MEDELTIDS-SPRÅKET. SUPPLEMENT*. VOL IV—V. (LUND 1953-1973)

Lars Borin, Markus Forsberg 2011. A diachronic computational lexical resource for 800 years of Swedish *Language technology for cultural heritage*, **Springer. 41-61**

8.15 Old Swedish morphology

1. BASIC INFORMATION

1.1 Lexicon type
Computational lexicon

1.2 Representation of the lexicon
LMF
1.3 Character encoding
The character encoding is in UTF-8

2. ADMINISTRATIVE INFORMATION
2.1 Contact person (name, e-mail)
Name: Markus Forsberg
E-mail: sb-info@svenska.gu.se
2.2 Copyright statement and information on IPR
Open source: CC-BY-SA 3.0, LGPL 3.0

3. TECHNICAL INFORMATION
3.1 Data structure of an entry
*Structured according to the LMF specification (*http://www.lexicalmarkupframework.org/*).*
3.2 Lexicon size
3 022 lexical entries

4. CONTENT INFORMATION
4.1 The natural language(s) of the lexicon
Swedish
4. 2 Entry Type
LexicalEntry in LMF.
4.3 Attributes
The resource is a small morphological resource providing inflections for the three digitized paper dictionaries of Old Swedish.
4.4 Coverage of the lexicon
The resource includes nouns, verbs, adjectives and some of the closed word classes.
4.5 Intended application of the lexicon
The intended use is language technology applications.
4.6 Reliability (automatically/manually constructed)
The resource is manually created.

5. RELEVANT REFERENCES AND OTHER INFORMATION
Lars Borin, Markus Forsberg 2011. A diachronic computational lexical resource for 800 years of Swedish *Language technology for cultural heritage,* **Springer. 41-61**

9 Summing up

The following figures sum up the number of resources contained in first batch:

Resources in First Batch	
32	Lexical resources
12	Corpora
6	Treebanks
5	Resources for speech
3	WordNet
1	Tool
59	**Total**

Table 1: *The total sum of resource types in batch 1*

TYPE	RESSOURCE NAME	ORG.	COUNTRY
Corpora	1.2 Corpus of Latvian literature	Tilde	Latvia
	3.7 Estonian reference corpus, ESTREFCORP	UT	Estonia
	5.3 Corpus of Old Literary Finnish	UHEL	Finland
	5.4 Corpus of Early Modern Finnish	UHEL	Finland
	5.5 Classics of Finnish Literature	UHEL	Finland
	6.2 Icelandic Frequency Dictionary Corpus, IFD Corpus	HI	Iceland
	6.6 The Saga Corpus	HI	Iceland
Annotated corpora	3.3 Estonian corpus with morphological annotation, ESTMORFCORP	UT	Estonia
	3.4 Estonian corpus with shallow syntactic annotation, ESTSYNCORP	UT	Estonia
Parallel corpora	1.7. Legislation Corpus of the Republic of Latvia	Tilde	Latvia
	3.5 Estonian-English parallel corpus, ESTENGPARCORP	UT	Estonia
TMX	5.11 TRIS corpus	UIB	Norway
Computational lexicons	2.4 STO-LMF	UCPH	Denmark
	4.1 Lexical database for Norwegian	UIB	Norway
	4.2 Lexical database for Swedish	UIB	Norway
	4.3 Lexical database for Danish	UIB	Norway
	4.7 Norsk ordbank	UIB	Norway
	4.8 Scarrie Lexicon	UIB	Norway
	8.1 SALDO	UGOT	Sweden
	8.2 SALDO morphology	UGOT	Sweden
	8.3 SALDO examples	UGOT	Sweden
	8.4 Swedish Framenet	UGOT	Sweden
	8.5 Swesaurus	UGOT	Sweden
	8.6 Parole lexicon	UGOT	Sweden
	8.7 Simple	UGOT	Sweden
	8.8 Swedish LWT list	UGOT	Sweden
	8.9 Swedish Kelly list	UGOT	Sweden
	8.10 Dalin	UGOT	Sweden
	8.11 Dalin morphology	UGOT	Sweden
	8.12 Schlyter	UGOT	Sweden
	8.13 Söderwall	UGOT	Sweden
	8.14 Söderwall Supplement	UGOT	Sweden
	8.15 Old Swedish morphology	UGOT	Sweden
Lexical Databases	3.6 Database of Estonian multi-word expressions, ESTMWE	UT	Estonia
	7.2 Database of Neologisms	LKI	
Lexical lists	5.6 Modern Finnish Word List	UHEL	Finland
	5.7 Frequency List of Written Finnish Word Forms	UHEL	Finland
Dictionaries	1.1 EuroTermBank	Tilde	Latvia
	1.3 The Lithuanian-Latvian dictionary	Tilde	Latvia
	1.4 The Latvian-Lithuanian dictionary	Tilde	Latvia
	1.5 The Estonian-Latvian dictionary	Tilde	Latvia
	1.6 Multilingual dictionary of person names	Tilde	Latvia
	6.5 Pronunciation Dictionary for Icelandic	HI	Iceland
	7.1 Modern lithuanian dictionary (DŽ)	LKI	Lithuania
WordNet	2.1 DanNet	UCPH	Denmark
	3.1 Estonian Wordnet	UT	Estonia
	5.1 Finnish WordNet	UHEL	Finland
Treebanks	2.2 Danish Treebank - CDT1	UCPH	Denmark
	2.3 Danish English Treebank – CDT2	UCPH	Denmark
	3.2 ESTONIAN TREEBANK, ESTTRE	UT	Estonia
	4.9 Sofie Treebank	UIB	Norway
	5.2 Finnish TreeBank: Grammar Definition Corpus	UHEL	Finland
	6.1 Icelandic Parsed Historical Corpus, IcePaHC	HI	Iceland

TYPE	RESSOURCE NAME	ORG.	COUNTRY
Acoustic databases	4.4 Acoustic database for Norwegian	UIB	Norway
	4.5 Acoustic database for Swedish	UIB	Norway
	4.6 Acoustic database for Danish	UIB	Norway
Speech corpora	6.3 Parliament Speech Corpus	HI	Iceland
	6.4 Hjal Speech Corpus, HJAL corpus	HI	Iceland
Tools	4.10 Oslo-Bergen tagger	UIB	Norway

Table 2: *All resources in batch 1 divided into resource type*

Summary of the language resources show that most focus has been on the data resources in contrast to the language tools in this first batch. Particularly well represented are lexical resources. The work on language resources continues and more resources will be provided in the second and third batch on M18 and M24 respectively.

It is further our intention to migrate all of the metadata and resources to the META-SHARE node as soon as the necessary META-SHARE software functionality will been implemented and tested.