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