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