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.