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.