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.