

The Estonian Reference Corpus: its composition and morphology-aware user interface

Heiki-Jaan Kaalep, Kadri Muischnek,
Kristel Uiboaed, Kaarel Veskis
University of Tartu

Composition of the Corpus

Overall size: ca 245 million words

non-balanced corpus; mostly standard written language

75% newspaper texts

2% fiction texts

2% science texts

5% legalese

5% parliament transcripts

9% texts of the “new media”

The Balanced Corpus 15 million words (newspapers, fiction, science)

Availability

Free for use for non-commercial purposes

www.cl.ut.ee/korpused

<http://www.cl.ut.ee/korpused/kasutajaliides/>

New, morphology-aware interface: www.keeleveeb.ee

Technical Annotation of the Corpus

Technical coding and annotation:

has been: TEI (Text Encoding Initiative) P3, SGML, ASCII + SGML entities

will be: TEI P5, XML, utf8

Migration from TEI P3 to TEI P5 currently under way

Annotated for text structure: paragraphs <p>, sentences <s>, headings <head>, authors <author> etc etc

Why does a corpus query interface need the knowledge about morphology?

Estonian **morphological system**: agglutinating, with fusional traits, e.g. ***tegema*** 'to do, make' :

1.sg.prs ***teen*** 'I do',

1.sg.pst ***tegin*** 'I did',

inf1 (*tahan*) ***teha*** 'I (want) to do',

inf2 (*hakkan*) ***tegema*** 'I (start) doing'

Morphological annotation of the Corpus enables:

searching by lemmas

searching by grammatical categories

searching by their combinations

Morphological annotation of the Corpus

Filosoft Ltd:

Morphological analyzer + guesser of out-of-dictionary words

HMM trigram disambiguator

Quality of the analysis + disambiguation:

10% tokens still ambiguous

mostly: participles; frequent verb form *on 'is, are'*

3-6% tokens (depending on text class) have got more or less incorrect analysis

Morphology-aware user interface

- 1) text fields + clickable boxes, or
- 2) query string

Features:

- the number of searched items is not limited
- by default, the order of searchable items in a sentence is not specified
- only well-formed grammatical tags (e.g. word class, case name) are allowed
- word forms, lemmas, 4-letter (and longer) substrings are allowed
- exclusion, i.e. „without“ is allowed
- option „immediately following“ is allowed

Näited

võta ainult Tasakaalus korpus, pane enne valmis kuni täpsema otsinguni

1: *tegema* lemma

Näita: klõpsatav allikaviide, klõpsatav sõnavorm

2. Otsimine gram kategooria järgi: Adjektiiv superlatiiv (U) sg translatiivis

Otsimine gram kategooriate kombinatsiooni järgi: Adjektiiv superlative in the translative case that is not used as an ??täiend. It means we should look for an adjective in superlative case in a sentence where there is no noun in a translative case

3. üks hea püsiühend: auku/augu/auk pähe rääkima

The idiomatic expression 'auku pähe rääkima' literally means 'to speak a hole in smb's head', and the metaphoric meaning of the expression is 'to convince somebody'??või parem tõlge.

While searching this expression from the corpus, one should take into account that the verb 'to talk', rääkima, inflects and should be retrieved using its lemma, the word-form 'pähe', meaning 'into the head'

Kas see või järgmine, mõlemat ilmselt ei jõua

Näited 2

4. impersonaal + N gen + poolt

Mati: Estonian has no indo-european-like proper passive, id est subjective action passive. The voice marked by special morphological form of a verb is called impersonal in Estonian grammar. The impersonal clause describes an action performed by an indefinite human agent. The main function of the impersonal voice in Estonian seems to be the ??varjamine of an agent, however, there is a limited possibility to add an agent phrase, like an english by-phrase to the clause. There has been discussion in estonian linguistics, whether it is a influence or even a borrowing from english and in what contexts the by-phrase can actually be used