Estonian Interlanguage Corpus

Collection of written texts produced by the learners of the Estonian as a second and foreign language

Tools for analysing texts:
Search, Statistics
Morpho-syntactic analysis
Syllables, lemmas
Reverse list
Frequency list
Error taxonomy

The research is supported by

Development of language software and language technology resource of the Estonian interlanguage corpus (2008–2010)

Digitalising of State Examination and Qualification Centre's manuscript materials (2009-2013)

Estonian Language and Cultural Memory (2009–2013),

Estonian Research Council Grants ETF8240, ETF8222. Cross-linguistic influence, Idiomatic Phrases (2010-2013)

Consistent treatment of the Estonian learner language (2015–2018)

Links to corpus

1. ICLE – The *International Corpus of Learner English,* Université catholique de Louvain (Louvain-la-Neuve, Belgium); Learner corpora around the world

https://www.uclouvain.be/en-cecl-lcworld.html

2. Centre of Estonian Language Resources, https://keeleressursid.ee/et/keeleressursid

3. Meta-Share

Publications

Written Estonian cluster analysis and it's perspectives for pedagogical purposes.

Use of the synonymous synthetic and analytical verbs

On the textual functions of adverbial structures in literary Estonian and Estonian learner language.

Constraints on morphosyntactic and lexical variability.

The comparative study of language use: morphological classes and clusters.

Learner language and Estonian Interlanguage Corpus.

On psycholinguistic reality of some word order patterns of Estonian.

The Prototype of word order assessment at the Estonian Interlanguage Corpus

The development of the prototype for an automatic word order error detector for Estonian.

... together ca 20

MA thesis

About the Rection of the Estonian Verbs (allative case). Kais Allikivi 2016

Usage patterns of analytic and synthetic verbal pairs based on 1990s Estonian literature texts. Heleriin Paeoja 2015

Thematic vocabulary and grammar in the textbook of the 8th grade Estonian studies as a second language. Marina Ševtšenko, 2014

Use of the Estonian ma- ja da-infinitives: Anastassia Smõteitšik (Kallikorm), 2009

On Possibilities of Research Estonian Learner Language and Development the Estonian Inerlanguage Corpus. Mare Kitsnik 2007

Bachelor thesis

Software for Morphosyntactic Cluster Extraction from Estonian texts. Sander Ots 2012

Error Displaying Module for EIC. Siim Medijainen, 2011

Word Grouping Tools for EIC. Kätlin Jõgisalu 2015

Error Marking Module for EIC Harry Nõmmann, 2015

R-language Use in Corpus Lingvistics. Magnus Kvell 2016

Use of Cluster Analysis for grouping texts by metadata. Virgo Hallik, 2016

Seminar works

- User Interface for Automatic Text Marking. Marko Sultsing, 2010
- Error Displaying Tree, Harry Nõmmann, 2014
- Text Similarities Comparing Web Application Virgo Hallik 2015
- Word Frequencies Comparison Web Application. Hanno Rudissaar 2016

Practical works

- Error Type Taxonomy 2014
- Reverse Dictionary for EIC 2015
- Language Level Suggestions for Texts 2015

Sub-corpora

L2 EIC core corpus
L2 State examination works
L2 Olympiad writings
L2 Academic learner language
L1 Academic Estonian
L1 Russian (referent corpus)
L3 Russian (referent corpus)

Together 3400000 tokens, 12000 texts

Similar researches in Estonia

Automatic text processing, Estonian Language Institute, Dictionaries

Lõo, Vajjala - Automatic CEFR level prediction for Estonian learner text.

Similar researches in World

- Cambridge Learning Language Corpus
- ASK Corpus Norway, Paul Meurer, Uni Computing
- ICLE, English as foreign language, Belgium, Louvaine
- Second Language Acquisition and Testing in Europe . Oulu
- ICLFI - International Corpus of Learner Finnish
- Tübingen University, Germany, Detmar Meurers

Problems

- Some inefficient algorithms and queries for big data
- Weak server for bigger use
- Relatively slow web access to text analysing tool Estmorf
- For bigger changes is good to rewrite code, user interface and data structures.

Plans

- Self learning and text analysing platform: statistics, similarities, groupings, word ordering, faulty text recognition, patterns finding
- Connections to learning environments and speller
- Training materials and exercises preparation and analyse
- Language level analyse, metadata guessing
- Interoperability with other language resources
- Student thesis in fields of EIC