Estnltk — open source tools for Estonian natural language processing

Estnltk is a software library for analyzing and working with the Estonian language, with the goal of making *natural languge processing* (NLP) easy and productive for students and language researchers. It is similar to popular NLP libraries such as Natural Language Toolkit and TextBlob, which are mainly used for working with English [Bir06, tex]. Estnltk is meant to be used with Python programming language [VR+07], which is a popular tool among language researchers and is also taught in Estonian universities.

Esnltk ties together existing open source components into an application programming interface (API), which includes most common, but most required functionality: word and sentence tokenization, morphological analysis and generation, lemmatization / stemming. There are also API-s for more advanced features such as clause segmenter, temporal expression tagger, named entity recognition, verb chain detector, Estonian Wordnet integration, grammar based information extraction.

Main notion of Estnltk is a *corpus*, which is a blob textual data and functions that operate on it. You can load a corpus from either text files or create one programmatically from other sources such as databases. After that, you can do many NLP tasks just by calling the relevant function or using a particular property of the corpus instance. For example, property `corpus.lemmas` allows accessing the stemmed versions of the words. Many natural language processing use cases are covered in the documentation with examples.

Estnltk also has built-in pipeline for text clustering and evaluation. There are also graphical tools for building grammars with focus on information extraction. Estnltk is a relatively young software library, but is constantly being developed further with future goals to extend and integrate more tools. Estnltk is currently used for applied research and statistics by several major organizations in Estonia.

The aim of the presentation at Why Linguistics Conference is to introduce Estnltk library by showing how it can be used in practice for solving NLP tasks.

References

[Bir06] Steven Bird, *Nltk: the natural language toolkit*, Proceedings of the COLING/ACL on Interactive presentation sessions, Association for Computational Linguistics, 2006, pp. 69–72.

[tex] *Textblob: Simplified text processing*, http://textblob.readthedocs.org/, Accessed: 2015-02-22.

[VR+07] Guido Van Rossum et al., *Python programming language.*, USENIX Annual Technical Conference, vol. 41, 2007.