Recent Developments of LT in Estonia

Kadri Muischnek

University of Tartu

Baltic HLT 2022

Based on
https://european-language-equality.eu/deliverables/

EUROPEAN
LANGUAGE
EQUALITY

D1.12

Report on the Estonian Language

Overview of my presentation

- Some general remarks
- Quick and incomplete overview of resources and tools for Estonian
- Easy-to-use tools created by ELG project

General Remarks

- Estonian has ca 1 million speakers → the market for LT products is also a small one
- Main force driving the development of Estonian LT is the public sector
  - → resources and tools developed by state-funded products are open source
  - → but the tend (or tended in the past) to be prototypes, not finished products
- What has changed during the last decade:
  - → the private sector has started to create more tools and solutions for Estonian
  - → there has been a considerable growth and development in language resources for Estonian

General Remarks

- The general attitude in Estonia towards digitization is positive
- People as well as government sector are ready to use AI and LT tools in their everyday life
- BUT despite various national and international programmes, initiatives and strategies, there is still a lack of continuity in research and development funding

A point copied from the Latvian ELE report: better synchronisation between national and international activities is necessary, especially regarding support for research infrastructures and for defining research priorities.

General Remarks

- During the last decade

→ some fields, especially machine translation and speech technology, have advanced significantly

→ we have better and bigger corpora of contemporary written language and bigger treebanks

- BUT several gaps that were identified by the Meta-Net White Paper in 2012 are still there, e.g.

→ we lack annotated semantic resources and tools for semantics

While talking about gaps, it is usually the case that we lack both annotated data and tools for certain tasks and, as annotating data is a time- and workforce-consuming process, it can be seen as an even bigger obstacle.

General Remarks

- BUT several gaps that were identified by the Meta-Net White Paper in 2012 are being filled, e.g.

→text generation: there are currently projects for summarization and text simplification
→ e.g. Henry Härm’s presentation about abstractive summarization yesterday

General Remarks

- The tools and resources for basic text processing and analysis work quite well if the input text follows the language and spelling norms, but ...

- → Estonian lacks annotated resources containing other language varieties: the written language variants used on social media sites, specialized languages used by professionals (healthcare, legal sphere)
- → Obviously there is a need for processing the variety of languages used on social media sites and also, with the growing popularity of Digital Humanities, the older written variants of Estonian.

Estonia’s LT strategies

- Since 2006 there has been a series of National Programmes for Language Technology
- → the current one in force until the year 2027
- The Estonian Language Development Plan 2021–2035
- → Development of Language Technology is stated as a priority
- *Bürokratt* - the vision of how digital public services should work in the age of AI: https://en.kratid.ee/burokratt
- → aims at an interoperable network of AI applications, which enable citizens to use public services with virtual assistants through voice-based interaction

LT/AL Specialists and Teaching

- We need more specialists!
- Teaching: University of Tartu
  - Computational Linguistics specialization at the Institute of Estonian and General Linguistics
  - Language technology courses (e.g. Natural Language Processing or Machine Translation) at the Institute for Computer Science

- Coming up Next: a quick overview of language resources and tools for Estonian

Corpora: large monolingual corpora

- Biggest monolingual corpus at the moment is Estonian National Corpus 2021, 2.9 billion tokens
  - → Large monolingual Estonian web corpora have been collected regularly and their size has been growing:
- → Estonian Web 2013 – 270 million tokens
- → Estonian Web 2019 – 1,5 billion tokens
- → Estonian Web 2021 – 2,4 billion tokens

Corpora: large monolingual corpora

- There is no data than more data, but ...
- → According to Koppel and Kallas (2021), 26% of the crawled Estonian texts were generated by machine translation
- → Also 36% of the tokens in this corpus belong to an “unknown” text type/genre
- → What is the future of web corpora? Can we trust the data crawled from the web?

Corpora: specialized corpora

- In specific domains, e g. court decisions or healthcare, large text collections exist, but they can be used only under very strict constraints.
- Obviously there is a need for processing the variety of languages used on social media sites, but the resources are scarce: Estonians don’t use Twitter much and using Facebook data is problematic

Bi- and multilingual corpora

- Estonian is included in the multilingual resources of the EU languages
- But there is little parallel Estonian-Finnish or Estonian-Latvian data that is a result of direct translation between these language pairs; most of the texts have been created by translating an English original into Estonian, Latvian or Finnish.

Audio resources

- We have at least a minimal necessary amount of audio resources for Estonian

- But more and/or bigger special corpora are needed:
  - children's and senior's speech, accented speech, and also speech of people having specific medical conditions (Parkinson's disease, Alzheimer's disease, dementia).
  - We also need more audio data for natural and noisy communication situations: spontaneous conversations, spontaneous meetings etc.

Lexical-Conceptual resources

- Mostly lexicons, dictionaries and machine-readable dictionaries, as well as terminological databases
- Estonian Wordnet, 92 000 synsets at the moment
- But we lack a Framenet-type lexical resource for integrating syntax and semantics and for describing verb valency in Estonian
  - Or maybe we don’t need it anymore?

Computational Grammars

- One of full-coverage rule-based computational grammar for Estonian: Constraint Grammar, which contains rule-sets and lexicons for
  - → morphological disambiguation,
  - → clause segmentation,
  - → syntactic function labeling
  - → dependency structure.
- In both Grammatical Framework and Giellalt
  - → a rule- and lexicon-based morphological model, with the lexicon based on that of Vabamorf.

Language models

Monolingual

→ EstBERT https://huggingface.co/tartuNLP/EstBERT

→ Estonian RoBERTa https://huggingface.co/EMBEDDIA/est-roberta

→ ELMo https://www.clarin.si/repository/xmlui/handle/11356/1277

→ GPT2 model https://huggingface.co/tartuNLP/gpt-4-est-large

Multilingual

→ XLM-RoBERTa

→ FinEstBert

Tools: basic text analysis

- Sentence segmentation, tokenization: EstNLTK
- Morphological analysis and disambiguation: Vabamorf, also as a part of EstNLTK https://github.com/Filosoft/vabamorf
- Dependency parsing models trained on Estonian UD treebanks:
  - → Stanza https://stanfordnlp.github.io/stanza/available_models.html
  - → UDPipe https://ufal.mff.cuni.cz/udpipe/2/models
  - → SpaCy https://github.com/EstSyntax/EstSpaCy

Tools: EstNLTK

- EstNLTK Python Library – open-source tools for Estonian NLP
  https://github.com/estnltk/estnltk
  - text segmentation
  - morphological processing
  - parsing
  - information extraction (tools for extracting addresses, named entities and temporal expressions from texts)
  - embeddings: EstNLTK's pretrained language models
  - Estonian WordNet API
  - … and much more

Tools: Texta Toolkit

- Tools for text analytics and solutions based on the latter: https://docs.texta.ee/
- open-source
- performs e.g. anonymization, keyword extraction, entity extraction, summarization, among others

Tools: speech processing tools

Although the quality of speech processing tools and services is far from the quality of those for English, the situation for Estonian is quite good, at least for “ordinary” speech, i. e., while the speaker is speaking Estonian as their first language and has no specific health conditions and there is little background noise.

- Taltech’s speech recognition system https://tekstiks.ee
- subtitles for Estonian live broadcasts using ASR https://github.com/alumae/kiirkirjutaja
- a rich transcription system for the Estonian Parliament
- several models for speech synthesis http://www.eki.ee/heli/index.php
- also including a neural network-based one https://neurokone.ee

Tools: Machine Translation

- The EU’s translation tool eTranslation provides machine translation services for Estonian.
- Estonian is a featured language in Google Translate
- Microsoft Translator provides a text and speech translation service for Estonian as well as an offline translation pack.

Tools: Machine Translation

- independent MT services are important for government sector →
- → government has initiated the central translation platform project (*Tõlkevärav*) – a national platform to help public and private sector companies manage their translation jobs, translation memories, and use machine translation

→ part of this initiative was presented yesterday by Taido Purason „Open and Competitive Multilingual Neural Machine Translation in Production”

Tools: more MT projects at UT

Translation between Estonian, Latvian, Lithuanian, English, Finnish, German, Russian, Võro, Northern Sami and Southern Sami
https://translate.ut.ee/

Easy-to-use tools

- A change of view:
- ELG (European Language Grid)

View from aside

- IT system
- standard building blocks
- LT is a part ~ building block
- → need for simple uniform LT blocks
- → technically: microservice by a dockerised web server
- Alternative: use "language-independent" or create a simplistic system

View from outside

- Project 2021 – 2023
- https://www.lingsoft.fi/en/microservices-at-your-service-bridging-gap-between-nlp-research-and-industry
- Aim: create microservices, i.e. find, wrap & freeze in time
- Results after 1.5 years:
- "New" sources found mainly from github
- Finnish, Swedish, Icelandic, Norwegian, Spanish, Portuguese, Galician, North Sami, Komi
- Lithuanian 2 (SpaCy models) -- many tools already dockerised ?
- Latvian 3 -- LT developed & integrated in end-user systems (by Tilde?); => no public freestanding modules ?
- Estonian 11 -- indicates relative disconnect of academia from industry ?