

# Leveraging Domain Corpora for Enhanced Terminology: The Case of Estonian-English Remote Sensing Termbase

Liisi Jakobson<sup>1</sup>, Jelena Kallas<sup>2</sup>,  
Erko Jakobson<sup>1</sup>

<sup>1</sup>Tartu Observatory of the University of Tartu  
<sup>2</sup>Institute of the Estonian Language



UNIVERSITY OF TARTU  
Tartu Observatory



EESTI  
KEELE  
INSTITUUT

# Introduction

Remote sensing is the process of detecting and monitoring the physical characteristics of an area by measuring its reflected and emitted radiation at a distance. It allows for collecting valuable information, such as monitoring forest fires, floodings, buildings etc.



Examples of remote sensing (pictures from a RITA Kaugseire project)



# Introduction

- Substantial increase in the adoption of remote sensing technology
- Lots of non-experts users
- Estonian remote sensing terminology has not been comprehensively analysed and standardised

# Terminological work in Estonia

- Institute of the Estonian Language ([eki.ee](http://eki.ee))
- In-house Dictionary Writing System named [Ekilex](#) (Tavast et al. 2018)
- Lench Classification
- Currently there are 130 termbases
- All termbases created using Ekilex are accessible to the public through dictionary portal [Sõnaveeb](#)
- Corpus Query System [Sketch Engine](#) (Kilgarriff et al. 2004)



# Estonian-English Remote Sensing Termbase

# Estonian Remote Sensing Corpus 2022

- Corpus Query System [Sketch Engine](#)
  - Corpus building from files and from the web
  - Terminology extraction function (Kilgarriff et al. 2014)
- **from files (57%) and the web (43%)**
- files (347 documents): BA and MA theses, handbooks, study materials
- **seed words:** 'kaugseire' (remote sensing), 'spektraalne lahutusvõime' (spectral resolution), 'spektraalmõõtmine' (spectral measurement), 'multispektraalne seade' (multispectral instrument), 'keskkonnasatelliit' (Earth observation satellite, environmental satellite), 'andmete ja teabe juurdepääsu teenus' (DIAS), and 'ülelennu sagedus' (revisit time, revisit period)
- goal: 5 million → 3 million

# The Terminology Extraction Module and the compilation of the term list

Focus corpus: the Estonian Remote Sensing Corpus 2022

Reference corpus: the Estonian National Corpus 2021 (Koppel & Kallas, 2022)

Estonian Term Grammar v 2.0: 37 distinct term patterns  
nouns, adjectives, indeclinable adjectives, adverbs, verbs, proper nouns,  
conjunctions, ordinal numbers, and acronyms

5-grams: 2 patterns

4-grams: 10 patterns

3-grams: 16 patterns

bigrams: 7 patterns

unigrams: 2 patterns

# The Terminology Extraction Module and the compilation of the term list

Evaluation: 500 top-ranking single-word term candidates and the 500 top-ranking multi-word terms

Results: 250 were identified as potential candidates for inclusion in the database (60% single-word terms and 40% multi-word terms)

Problems:

- the appearance of English terms in the list;
- terms from interconnected domains, primarily physics, biology, forestry, agriculture, metrology, meteorology, and climatology;
- general language items (e.g. 'majandus' (economy));
- mistakes in lemmatisation and morphological analysis.

Questionnaire in Tartu Observatory (ten remote sensing experts)

Final list: 100 terms

# Database elements in Ekilex

629476

Add term

Join

Duplicate

Delete

20.11.2023



Domain



009 - remote sensing (nt: the collection of information about an object without being in physical contact with the object)

Definition



et

- dünaamikavõrranditel põhinev atmosfääri liikumise matemaatiline mudel

[definitsioon](#) (Public)

[ [kaugseire ekspert](#) ]

Definition internal note

- Kasutatavad diferentsiaalvõrrandid on mittelineaarsed ning neid lahendatakse

numbriliselt. (Public)

[ [kaugseire ekspert](#) ]

en

- a mathematical model constructed around the full set of primitive dynamical equations which govern atmospheric motions [definitsioon](#) (Public)

[ [Wikipedia. Atmospheric model](#) ]

▼ et

**atmosfäärimudel**

Kaugseire terminibaas

Eemalda

Usage example



- Tugimõõtmiste süsteemi rakendamisel satelliidisensorite andmetele on oluline osa atmosfäärimudelil, mis arvestab kiirguse levimist maapinnale ja tagasi

satelliidisensorile. (Public)

[ [Kaugseire Eestis 2016](#) ]

▼ en

**atmospheric model**

Kaugseire terminibaas

Eemalda

# User interface in Sõnaveeb



atmospheric model

Language All languages ▾ Databases Kaugseire terminibaas ▾ Feeling lucky

## en atmospheric model

### TERMINOLOGICAL DATABASES

#### Kaugseire terminibaas

ID: 629476  20.11.2023

remote sensing (nt: the collection of information about an object without being in physical contact with the object)

et **dünaamikavõrranditel põhinev atmosfääri liikumise matemaatiline mudel** [kaugseire ekspert](#)

- Kasutatavad diferentsiaalvõrrandid on mittelineaarsed ning neid lahendatakse numbriliselt. [kaugseire ekspert](#)

en **a mathematical model constructed around the full set of primitive dynamical equations which govern atmospheric motions**  
[Wikipedia. Atmospheric model](#)

et **atmosfäärimudel**

#### Usage examples

Tugimõõtmiste süsteemi rakendamisel satelliidisensorite andmetele on oluline osa atmosfäärimudelil, mis arvestab kiirguse levimist maapinnale ja tagasi satelliidisensorile.  [Kaugseire Eestis 2016](#)

en **[atmospheric model](#)**

#### Usage examples

Atmospheric and meteorological data gathered at the time of the flight, at the surface, or using a radiosonde, can improve the atmospheric model.



# Practical advantages

- Choose relevant terms based on frequency
- Find possible variants
- Analyse the usage context
- Distinguish old variants
- Clarify the meaning based on context analyses
- Find contexts and definitions from trustworthy sources

# Conclusion

- Creating a corpus-based termbase for languages with a small number of native speakers is an important topic to be addressed.
- Our work demonstrates that adopting a corpus-based approach is viable even when dealing with relatively new topics.
- Ability to uncover various term variants along with their frequencies of occurrence.
- It is crucial to underscore that corpora do not replace expert knowledge in the termbase creation process.
- Corpus-based approach should complement an expert-based approach, as most terms still require expert consultation.



## Future plans

- to expand the Remote Sensing Termbase in Ekilex by adding new terms and revising existing ones based on user feedback
- to make the Estonian Remote Sensing Corpus 2022 publicly accessible through the Corpus Query System KORP, with future plans to provide end-users access through the KORP API also in Sõnaveeb

# References

- Kilgarriff, A.; Rychly, P.; Smrž, P. and Tugwell, D. (2004). The Sketch Engine. In Proceedings of the XI Euralex International Congress. Lorient: Université de Bretagne Sud, pp.105–116.
- Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P., Suchomel, V. (2014). Finding terms in corpora for many languages with the Sketch Engine. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, pp. 53–56.
- Koppel, K., Kallas, J. (2022). Eesti keele ühendkorpuste sari 2013-2021: mahukaim eestikeelsete digitekstide kogu. In Eesti Rakenduslingvistika Ühingu Aastaraamat 18, pp. 207–228. doi.org/doi:10.5128/ERYa18.12
- Tavast, A., Langemets, M., Kallas, J., Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts. Ljubljana, Slovenia, pp. 749–761.