

# 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

liisi.jakobson@ut.ee, jelena.kallas@eki.ee, erko.jakobson@ut.ee

## Introduction

Remote sensing is the process of detecting and monitoring the physical characteristics by measuring its reflected and emitted radiation at a distance. It allows for collecting valuable information, such as monitoring forest fires, floodings, heatwaves, snow, and various environmental phenomena (Figure 1).



Figure 1. Examples of remote sensing (Pictures from RITA Kaugseire project)

Over the past few decades, Estonia has seen a substantial increase in the adoption of remote sensing technology. What was once a topic primarily discussed in English and primarily by scientists and experts has now become a subject of widespread exploration among Estonian-speaking non-experts. Despite the growing interest in this field, Estonian remote sensing terminology has not been comprehensively analysed and standardised.

## Terminological work in Estonia

Traditionally, termbases in Estonia were compiled through expert knowledge. Nowadays, the process of termbase creation has become more automated, encompassing both the creation and management of databases as well as the compilation and analysis of domain corpora.

For compiling general language dictionaries and termbases, the Institute of the Estonian Language (eki.ee) has developed an in-house Dictionary Writing System named Ekilex (Tavast et al. 2018). The Lench classification system is used to cover subject fields within Ekilex.

For corpus data analysis, the Institute uses the Corpus Query Systems Sketch Engine (Kilgarriff et al., 2004).

All termbases created using Ekilex are accessible to the public through the dictionary portal Sõnaveeb (sonaveeb.ee), with each termbase having its dedicated homepage, such as the Estonian-English Remote Sensing Termbase. Currently, there are 130 databases spanning diverse fields, and this collection continues to grow.

## Estonian Remote Sensing Corpus 2022 compilation and Term Extraction using Sketch Engine

For corpora creation from scratch, we used a Corpus Query System Sketch Engine. The monolingual Estonian Remote Sensing Corpus 2022 (3 million words) was compiled from files (58%) and from the web (43%).

In total, 347 documents were uploaded. Documents were sourced from the University of Tartu Library database, encompassing BA and MA theses, handbooks, projects, and study materials. To obtain permission to add files to corpora, scientists/researchers were directly requested to provide their materials.

Texts from the web were crawled by providing seed words such as here: 'kaugseire' (remote sensing), 'spektraalne lahutusvõime' (spectral resolution), 'spektraalmõõtmine' (spectral measurement) etc. Most of the texts were from leading popular science journals, legal documents, materials related to remote sensing enterprises and remote sensing scientific projects.

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

Reference corpus: Estonian National Corpus 2021 (Koppel, Kallas 2021).

Focus corpus: Estonian Remote Sensing Corpus 2022

Term Grammar (Kilgarriff et al. 2014; Fišer et al. 2016)

Estonian term grammar version 2.0 enables term extraction for single and multi-word terms, with a maximum length of up to 5 words.

In total, the term grammar encompasses 37 distinct term patterns:

5-grams: 2 patterns, 4-grams: 10 patterns, 3-grams: 16 patterns, bigrams: 7 patterns, unigrams: 2 patterns.

## Evaluation I stage

500 top-ranking single-word term candidates and  
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 etc.;
- general language items (e.g. 'majandus' (economy));
- mistakes in lemmatisation and morphological analysis.

## Evaluation II stage

Questionnaire in Tartu Observatory (ten remote sensing experts)

**Results:** The final list of 100 terms served as a headword list for compiling the Estonian Remote Sensing Termbase using the Dictionary Writing System Ekilex. The work will continue.

## The Compilation of Remote Sensing Termbase using Dictionary Writing System Ekilex

Ekilex is the in-house Dictionary Writing System developed by the Institute of the Estonian Language. A term entry in Ekilex is concept-based, structured as a unit containing term variants, definitions, contexts, source references, notes, and domain information. Figure 2 illustrates an entry for the concept 'collection of equipment installed on the earth's surface that enables communications over one or more satellites'.

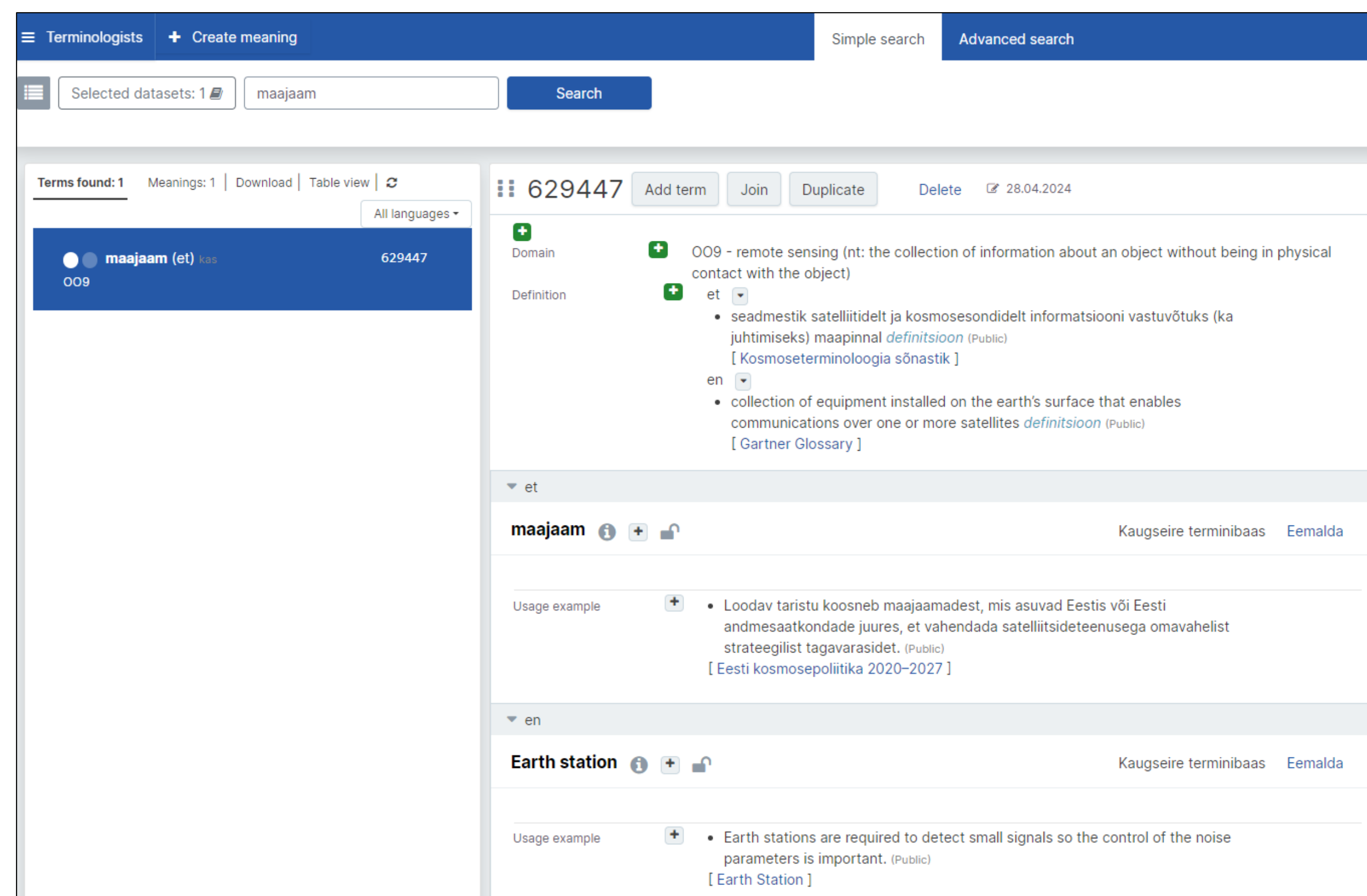


Figure 2. A concept entry in a Dictionary Writing System Ekilex

## Discussion

The practical advantages that corpora offered during the creation of the Estonian Remote Sensing Termbase include the ability to:

- choose relevant terms based on frequency through term extraction;
- detect possible variants of the same term;
- analyse the usage context of variants, helping to understand in which domains the term is used. For instance, the English term 'reflectance' might be used as 'heleduskordaja' in the water remote sensing domain, while in vegetation remote sensing domain, the term 'peegeldustegur' is preferred;
- identify preferred variants based on statistical data;
- distinguish old variants based on data from different time. By dividing the data into two periods: 1993-2010 and 2011-2022, we found that Estonian equivalent for English term 'reflectance' 'peegeldustegur' is significantly more used after 2010, while 'heleduskordaja' is significantly less used after 2010. This analysis indicates that 'peegeldustegur' is substituting the term 'heleduskordaja';
- clarify the meaning based on context analyses;
- find contexts and definitions from trustworthy sources.

## Conclusions

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. One of the primary advantages of corpus data lies in its ability to uncover various term variants along with their frequencies of occurrence. However, it is crucial to underscore that corpora do not replace expert knowledge in the termbase creation process. Instead, a corpus-based approach should complement an expert-based approach, as most terms still require expert consultation.

## Future plans

Our future plans involve expanding the Remote Sensing Termbase in Ekilex by adding new terms and revising existing ones based on user feedback. We aim to make the Estonian Remote Sensing Corpus 2022 publicly accessible through the Corpus Query System Korp.

## References

- Fišer, D., Suchomel, V., Jakubíček, M. (2016). Terminology Extraction for Academic Slovene Using Sketch Engine. In Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN). Karlova Studánka, Czech Republic, pp.135–141
- 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.