EESTI KEELE INSTITUUT

Large Language Models as Tools for Historical Lexicography: Automating Homonym Detection

Sven-Erik Soosaar, Institute of the Estonian Language
Madis Jürviste, Institute of the Estonian Language, University of Tartu
Tiina Paet, Institute of the Estonian Language

1st International Conference on Lexicology and Lexicography
Budapest, 30th September 2025

eki.ee

Project

R&D project “Applying large language models to lexicography: new opportunities and challenges”

(1.1.2024–31.12.2027)

Funded by the Ministry of Education and Research

Background I

- LLMs and lexicography: extensive research
- Our research topic: LLMs + 17th and 18th century Estonian dictionaries
- Main goals
  - ➢ providing old dictionary entries with modern Estonian equivalents
  - ➢ historical layers of the Estonian lexicon

Background II

- LLMs used in this experiment: GPT-4o, Claude 3 Opus, Gemini 2.0
- Material of present research: dictionary of Johannes Gutslaff (1648)
- Our previous research on Gutslaff but also Stahl (1637) and Göseken (1660) gave good results (Jürviste, Paet, Soosaar 2025)
- Automatic homonym detection with LLMs and comparison with contemporary Estonian

EESTI KEELE INSTITUUT

Background III

- History of written Estonian:
  - Two literary languages emerged in 17th century: North and South Estonian
  - South Estonian literary language faded away in the 19th century, speakers started to read materials in written North Estonian
  - Until recently South Estonian was considered a group of Estonian dialects (Tartu, Mulgi and Võro-Seto)
  - Since 2009 ISO 639-3 code vro

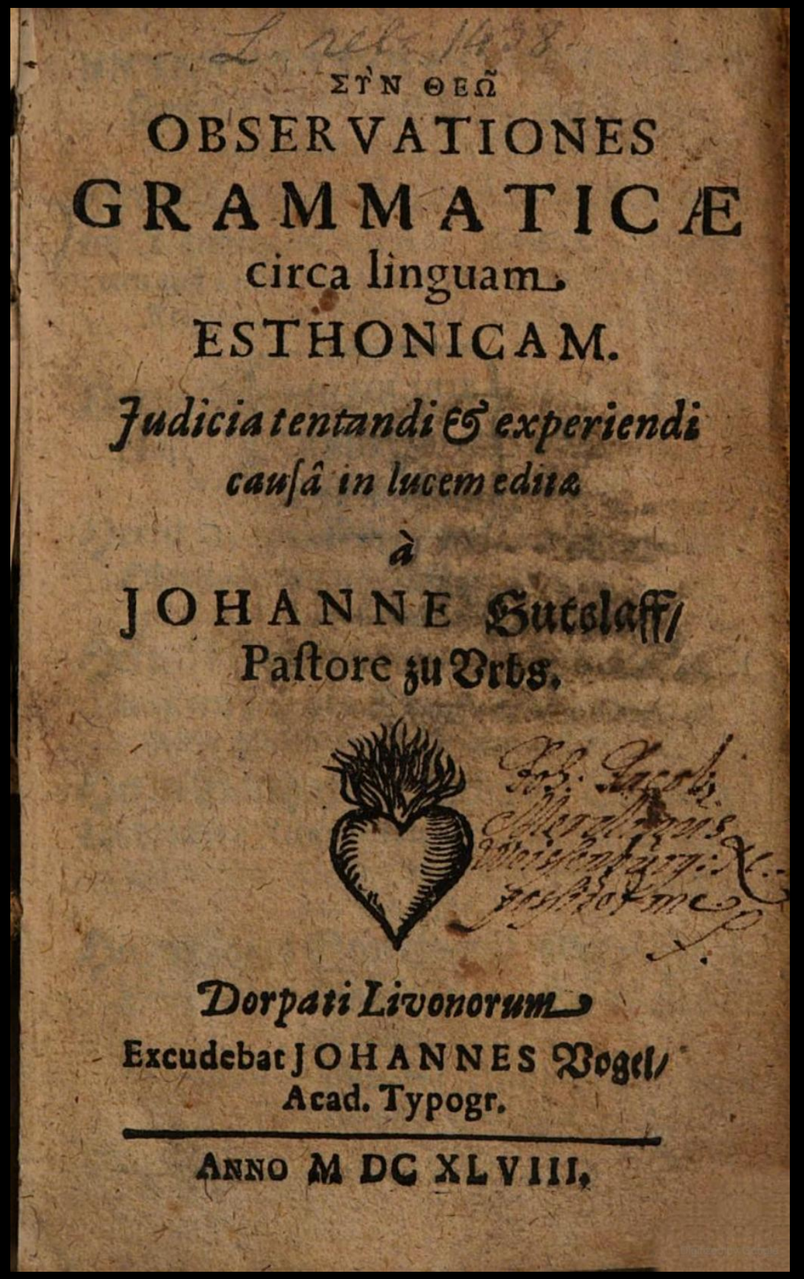
Background IV

- Johann(es) Gutslaff (? –1657)

  „Observationes grammaticæ circa linguam esthonicam“

  Dorpat: Johannes Vogel, 1648

  Grammar + dictionary


ΣΤΗΝ ΘΕΩ

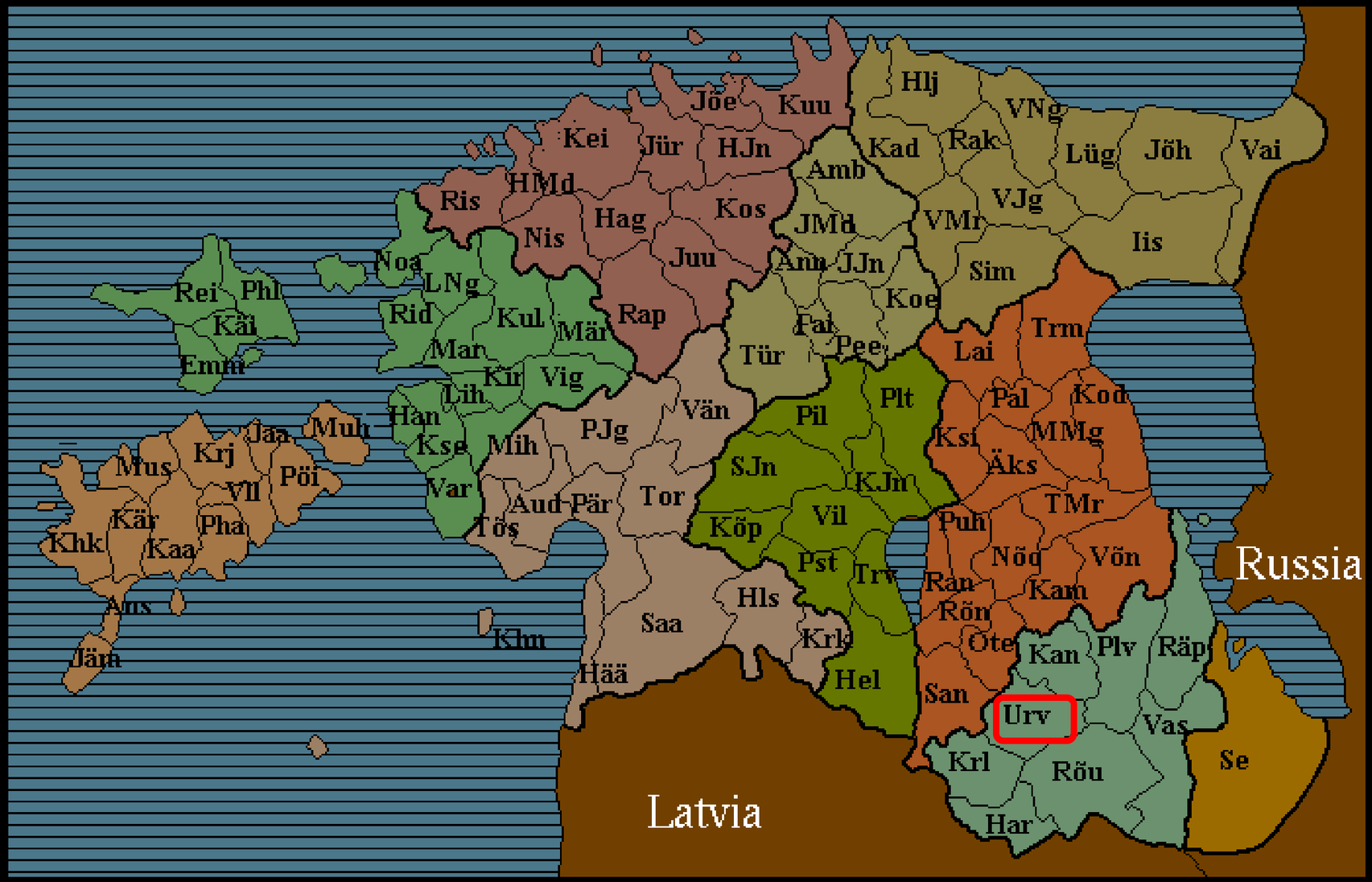
OBSERVATIONES
GRAMMATICÆ
circa linguam
ESTHONICAM.

Judicia tentandi & experiendi
causâ in lucem edita
à
JOHANNE Gutslaff/
Paſtore zu Vrbs.

Dorpati Livonorum
Excudebat JOHANNES Vogel/
Acad. Typogr.

ANNO M DC XLVIII.

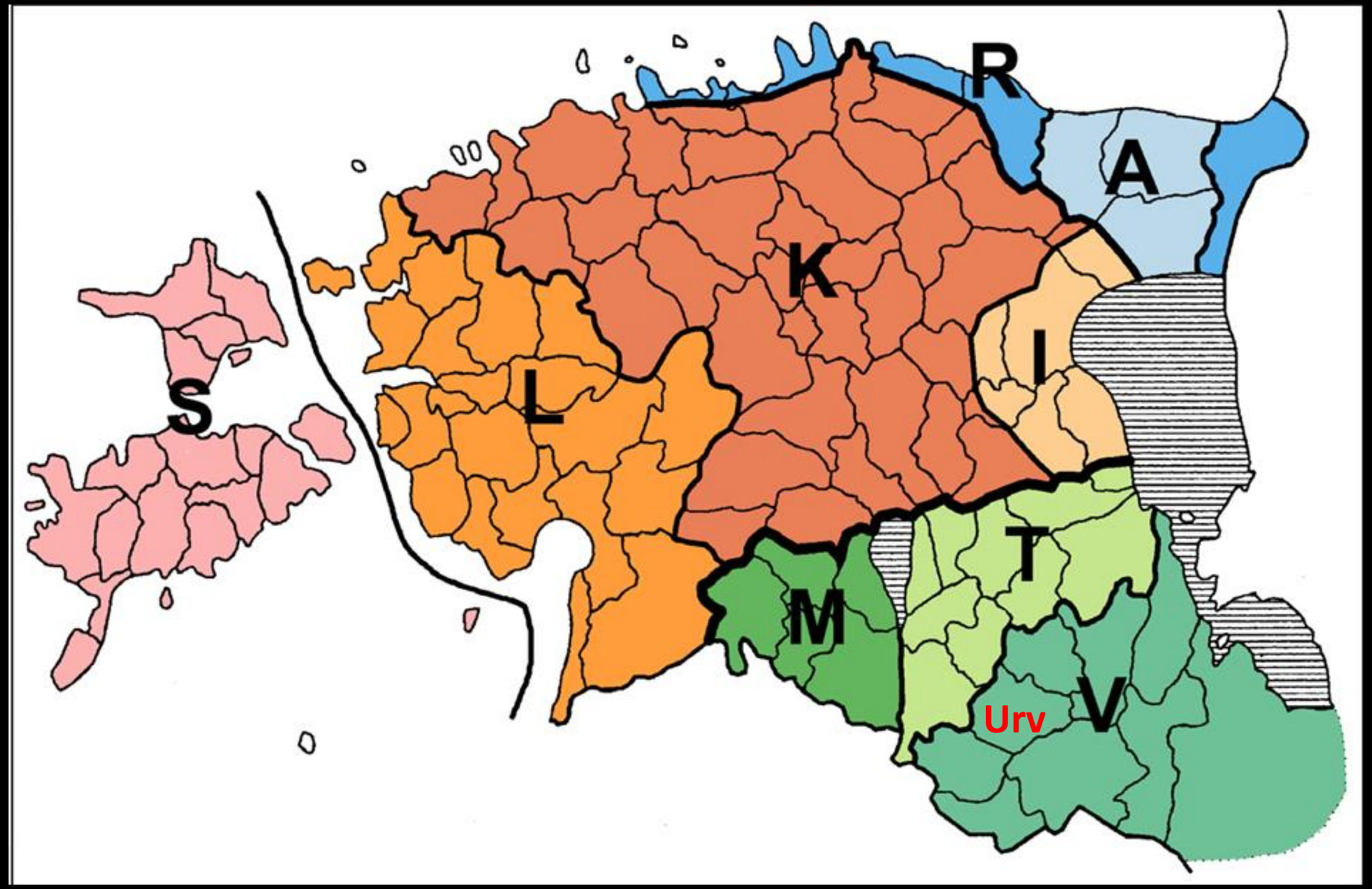
EESTI KEELE INSTITUUT



Urvaste

(Urv)

German:
Anzen
Latin: Urbs



EESTI KEELE INSTITUUT

Urvaste

(Urv)

Background V: Homonyms

Homonyms are words or word combinations that share identical spellings by coincidence but have completely unrelated meanings. Homonymy occurs less frequently in language than polysemy. While homonymy creates ambiguity like polysemy, it differs fundamentally in that homonymous words result from coincidental similarities in spelling and form. The core meanings of homonymous words bear no semantic relationship to one another. (Langemets 2022)

Homography in Estonian

In modern Estonian, homography can occur in the case of

- the difference between the 2nd and 3rd quantity of consonants, e.g. *linna* (*linna* ‘city.GEN’ ~ *linn´a* ‘city.PART’), *vaala* (*vaala* ‘whale.GEN’ ~ *vaa´la* ‘whale.PART’).
- in the case of palatalization e.g. *palk* (*palk* ‘salary’ ~ *pal’k* ‘beam’)

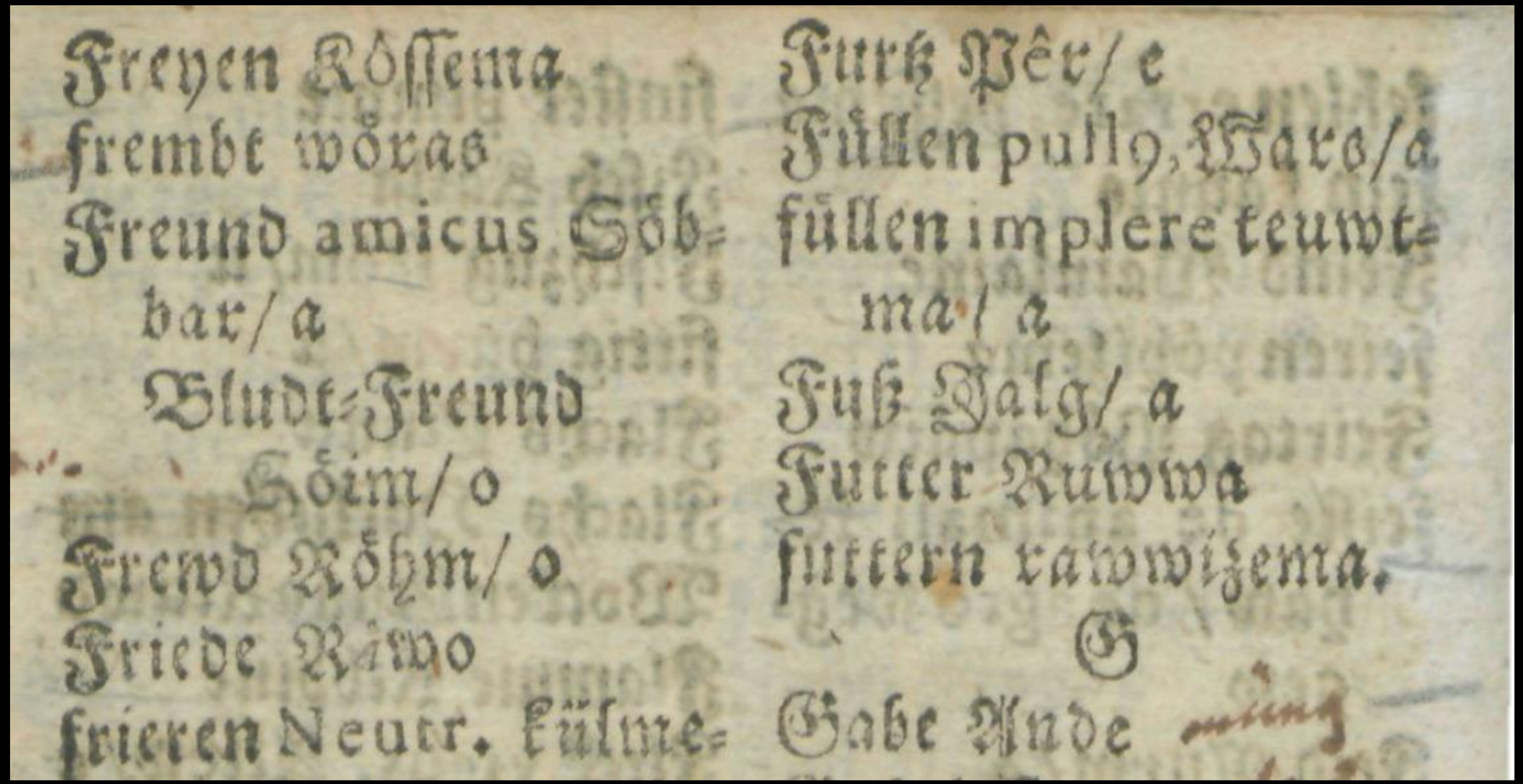
EESTI KEELE INSTITUUT

Orthography of Gutslaff

Gutslaff’s orthography varies. The same word may appear in different places with different spellings.

The main reasons for the differences are:

- Short vs. long *s*, e.g. *saisma ~ saißma*,
- Marking of velar stop *k/g* vs. *ck*, e.g. *Pösk ~ Pösck*,
- Marking of long vowel with 2 letters or with *h* like in German, e.g. *lohm ~ loom*,
- Marking of compound *ts*: *karritz ~ karritʃ* etc.



Freyen Rõssema
frembt wõras
Freund amicus Söb-
bar/a
Bludt-Freund
Hõim/o
Frewd Röhm/o
Friede Rahho
frieren Neutr. külme-
Furtz Pêr/e
Füllen pullo, Waro/a
füllen implere teuwt-
ma/a
Fuß Jalg/a
Futter Ruhwa
futtern rahwitzema.
G
Gabe Ande

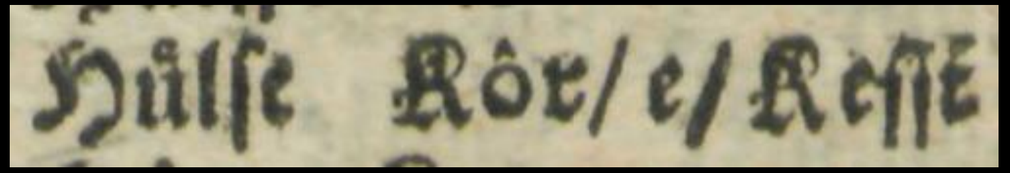
frembt wöras
Freund amicus Söbbar
…
Frewd Röhm

Fremd võõras
Freund amicus sõber
…
Freude rõõm

Homography

In the case of Gutslaff', there is more homography than in contemporary Estonian:

- ä can be written as e in the first syllable, e.g. *kesck*, which is used for *kesk* and *käsk* and *kesi* and *käsi* (*kessi*) become homographs,
- the weak plosive of a consonant cluster at the end of a word is marked with a strong one: *palck* (*palk* and *palg*).
- He does not distinguish the close-mid back unrounded vowel ɤ, the letter *õ* for it was introduced only at the beginning of the 19th century by O. W. Masing.



EESTI KEELE INSTITUUT

Input

The input for the experiment were words identified from Gutslaff's German-Estonian dictionary that were identical in form or had minor spelling differences due to orthographical fluctuations.

Prompt 1

```
prompt = (
"You are an expert in Estonian spelling. „
"You are given a list of modern Estonian words. „
"Please group them only based on spelling and phonetic similarity. „
"Ignore meanings – only the spelling of the words matters!\n\n„
"**Grouping rules:**\n„
"- If the words are the same or differ by only one letter, put them in one group.\n„
"- If the words look different, put them in separate groups.\n„
…)
```

Prompt 2

```
prompt = (
                    " You are an expert in 17th-century Estonian lexicography. Here is Gutslaff's dictionary from 1648. "
"Below is a group of modern Estonian words that are spelled similarly. "
"Your task is to decide whether these words are homonyms or not.\n\n"
"Words are homonyms if they have the same or very similar spelling, but different meanings. "
"If words have the same or very similar spelling and the same meaning, they are synonymous."
"**Criteria:**\n"
"- Homonyms: same spelling, but completely different meanings. For example, ‘silm' can mean both the organ of vision and a fish.\n"
"- Not homonyms: same spelling, meanings are either the same or very similar.\n\n"
"Use the following Gutslaff dataset (Estonian word of Gutslaff, German word of Gutslaff, Latin specification, Estonian synonym of Gutslaff, German synonym of Gutslaff):\n\n"
+ table_header + table_rows +
"\n\n### Answer exactly in the following format (do not add any other information!)\n"
"Yes/No; Brief explanation why these words are or are not homonyms. |\n "
```

Results

The surviving part of Gutslaff's dictionary contains 1,866 German headwords with Estonian equivalents.
A total of 473 identical or almost identical words were found.

Of these, there were

- 172 pairs
- 33 triplets
- 4 quadruplets
- 3 five-word groups

A total of 212 word groups

Results

Five-word groups:

- *pessema* (dreschen, geisseln, klopffen, questen, staupen)
- *komb~kombe* (Geberde, Gewohnheit, Arth, Sitte, Weise)
- *Wigga* (Fehl, Gebrechen, Mangel, Plage, Seuche)

Quadruplets:

- *kastma* (feuchten, netzen, tauchen, tuncken)
- *keelma* (hindern, stewren, verbieten, versagen)
- *ossa* (Ast/Zweig, Knast, Part, Theil)
- *helle* (helle, Klang, Laute, Schall)

Results

- Claude 3.5 Sonnet did well with the first part of the test, identifying homonyms even in cases where they are not homonyms in modern Estonian. For example, Gutslaff's *selg* is *selg* and *selge* in modern Estonian. However, there were also mistakes, for example when the difference was only in one letter and sound, e.g. *haud* and *haug*, where Claude ignored the meanings of the word *haud* and read it as an orthographic variant of *haug*. Some words with the same form were not found, e.g. *astma* (steigen, treten), *kütsma* (backen, braten), *kabal* (Band, Reiff).

Results

- In the second part of the test, the LLM was asked to find which words in Gutslaff's dictionary were homonyms, using the German equivalents and the synonyms provided by Gutslaff. Claude found 212 word groups, of which 56 were considered homonyms. In some word pairs, the model was unable to identify the word pair despite the spelling coincidence, for example, the word *Wilja*, which occurs as the equivalent of the German words *Frucht* and *Haab*, the model found that *Wilja* (=Frucht) occurs only once and therefore the homonymy cannot be identified. Gutslaff has presented the words *Warra, Nöw* as synonyms for *Wilja* (=Haab), from which the model considered the word *Nöw* to be a homonym of the word *Nöuw*, which occurs earlier in the table.

Results

- Claude identified almost all homonyms, but also considered several pairs of words as homonyms that are more correctly considered polysemous words, e.g. *vöras* in the meanings ‘unknown’ and ‘guest’. Some words were left unanalyzed for unknown reasons, e.g. Claude considered the Gutslaff’s word *Otza* to be homonyms with the meanings *Brodt Kante* and *Ende*, but left the word *Stirn* unanalyzed.
- 18 pairs of homonyms were correctly identified, and 5 pairs were not identified by Claude 3.5 Sonnet. In addition, it considered 38 pairs to be homonyms, although they are not.
- Other models got much worse results in the second part of the test:
  - GPT-4o: 10 correct-13 not identified-49 wrong
  - Gemini 2.0: 9 correct-14 not identified -84 wrong.

Homonyms in contemporary Estonian

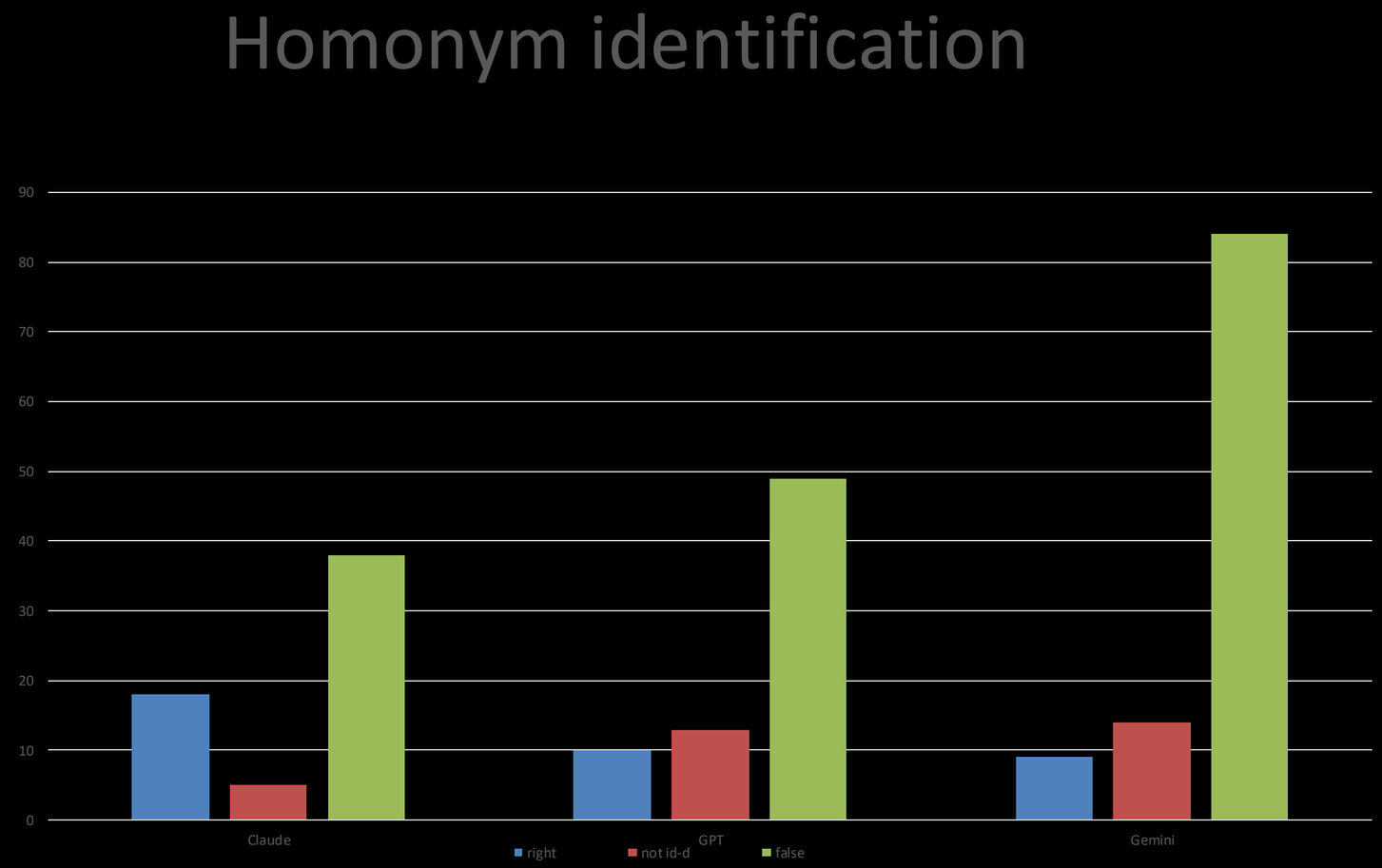
EESTI KEELE INSTITUUT

- In the Unified Dictionary of Estonian there are 845 groups of words marked as homonyms out of almost 180 000 entries.
- Of these, 97 are homonym groups of foreign words, 73 are letters, 36 are abbreviations, 28 are place names, and 14 are compound words. If these 248 are excluded, 597 groups remain.
- Of those 597 still many are instances of false homonymy (interjections used as nouns, adjectives used as adverbs or nouns etc (e.g. *tänu* – noun, preposition and interjection)

EESTI KEELE INSTITUUT

SUMMARY

- Best model: Claude 3 Opus, which outperformed other LLMs by a large margin
- There are significant differences in LLMs
- Some LLMs are useful for homonym detection with subsequent human checking



References

Estevez-Rams E, Mesa-Rodriguez A, Estevez-Moya D 2019. Complexity-entropy analysis at different levels of organisation in written language. PLoS ONE 14(5): e0214863. https://doi.org/10.1371/journal.pone.0214863

Gutslaff, Johannes 1648. Observationes grammaticae circa linguam esthonicam. Dorpat: Johannes Vogel. http://www.digar.ee/id/nlib-digar:100419 (10.09.2024)

Jakubíček, Miloš; Rundell, Michael 2023. The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography?. In Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century. Brno: Lexical Computing, pp. 508-523.

Jürviste, Madis; Paet, Tiina; Soosaar, Sven-Erik (2025). Eesti vanade sõnakujude tuvastamise võimalustest suurte keelemudelite abil. [Identifying Old Estonian Word Forms Using Large Language Models.] Eesti Rakenduslingvistika Ühingu aastaraamat 21. [Estonian Papers in Applied Linguistics 21.]

Langemets, Margit 2010. Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus keelevaras. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.

Langemets, Margit 2022. EKI teatmik: Homonüümid. https://teatmik.eki.ee/teatmik/homonuumid/

Lew, Robert 2023. ChatGPT as a COBUILD lexicographer. Humanit Soc Sci Commun 10, 704 (2023). https://doi.org/10.1057/s41599-023-02119-6