

Occupational terms and societal roles in early Estonian lexicography: applying large language models to historical lexicography research

Madis Jürviste, lexicographer, junior researcher
Institute of the Estonian Language & University of Tartu

Roadmap

- Manual collection of domain-specific words
- Custom Python scripts
- First API-requests
- OCR for ATH 1732 and AWH 1780
- Enriching lexicographical data with new content
 - e.g. *särjesapp* -> Buchgold
- Mapping vocabulary across sources

2022

- Manual study of six dictionaries:
 - Stahl 1637, Gutsclaff 1648, Göseken 1660, Vestring 17XX, Helle 1732, Hupel 1780
 - ~1600 professional titles and societal roles, 3 categories
- Nov 2022: first public version of “You Know Who”
 - How to harness the technology in old vocabulary research?

2023

- First attempts
 - Structuring and basic statistics of manually selected vocabulary
 - Local Python scripts, written with an LLM

Fail on genereeritud 25.03.2023 kell 19:49:05.

Leidsin loenditest järgmised sõnalõpud:

Lõpuga mees on 210 sõna.

Lõpuga r on 165 sõna.

Lõpuga ja on 492 sõna.

Lõpuga nna on 0 sõna.

Lõpuga nd on 31 sõna.

Lõpuga tüdruk on 7 sõna.

Kategooriapõhised esinemised:

K1 kategooria sõnu on kokku 678

K2 kategooria sõnu on kokku 542

K3 kategooria sõnu on kokku 378

Autorite sõnade esinemised:

Stahl sisaldab 76 ametisõna.

Gutslaff sisaldab 35 ametisõna.

Göseken sisaldab 432 ametisõna.

Vestring sisaldab 317 ametisõna.

Helle sisaldab 232 ametisõna.

Hupel sisaldab 620 ametisõna.

2024

- If LLMs cannot “read” PDFs, then we must “create” content for them?
- Several tests to see how well the LLMs can “understand” 17 and 18 century Estonian, matching word pairs
- Success? Accuracy: up to 90% (depending on the “difficultness” of the vocabulary)

Vana sõnakuju

Saksa vaste

Rackel	Hencker
Kassucksep	Kürßner
jütleja	Prediger
oppeja	Lehrer
sundija	Richter
Portt	Hure
Portopähline	Hurer
Kessineck	Bürge
Jeck	Hofnarr
Rahhajohataja	Münzmeister
kaffer	Ausgeber
kachtja	Färber
Kühna-Mees	Dieb
Kiwwi Tackar	Stein Metzger
Raamato Rutzoja	Buchdrucker
Pilli ajaja	Ein Sackpfeiffer

Vana sõnakuju

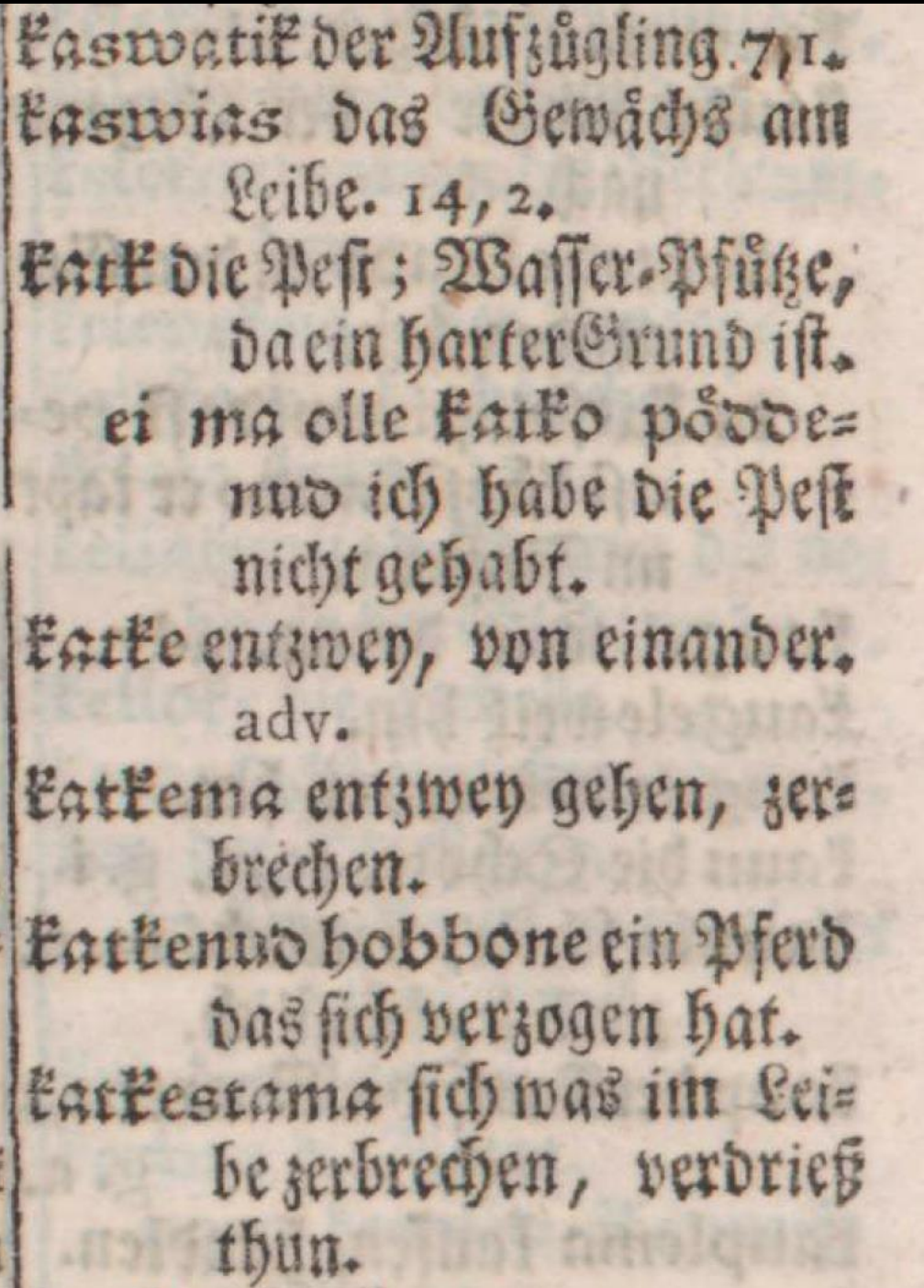
Saksa vaste

Laudamees	Ein unteutscher rechtsfinder
Mängimees	Ein Spielmann musicant
Lämmia	Ein Zauberer
kulaja	der Kundschaster
lambrine	der Schäfer
nikker	der Tischler
kolimees	der Schüler
siggur	der Schwein-Hirte
heikaja	Ausrufer, Herold
kandijanne	ein Tragender
kalla müja	Fischhändler
müütnik	der Zöllner
laiwa wallitseja	Steuermann
Jooks	Ein Curierer

2024

- *Vision-enabled models: Claude 3.5 Opus, followed by GPT and Gemini*
- New goal: digitisation of 3 dictionaries (Gutslaff 1648, Helle 1732, Hupel 1780) in addition to those already digitised (Stahl 1637, Göseken 1660, Vestring 17XX) by University of Tartu and the Estonian Literary Museum

2024: OCR, initial tests (ATH 1732)



Prompt (Claude 3.5 Sonnet *chat*):
“Take a look at this PDF in Gothic script (Fraktur), create a CSV.” (APPROX.)

kaswatik	subst	der Aufzügling		7,1
kaswias	subst	das Gewächs am Leibe		14,2
katk	subst	die Pest; Wasser-Pfüße, da ein harter Grund ist		
katke	adv	entzwey, von einander	adv.	
katkema	verb	entzwey gehen, zerbrechen		
katkenud				
hobbone	fraas	ein Pferd das sich verzogen hat		
katkestama	verb	sich was im Leibe zerbrechen, verdrieß thun		

2025: ATH?

- Write a thorough API-request to achieve a result good enough for manual review
- Creation of a custom web app for OCR: Anthonius (Joonatan Jakobson)

wöllo *eine alte Hexe.*

wöllw *die Welle im Mühl-rade*
Declension: 17,1.

wölwima *wölben.*

wölwitud *gewölbt*

wörk *das Netz*
Declension: 7,2.

2025 ATH key takeaways

- CER: 4 ... 120% (“traditional” goal < 5%)
 - Nevertheless ...
 - ... manual review was very fast
 - 41% of lines needed **no correction**
 - Cost of the operation: 10 €
- Some technical modifications for the next OCR operation (Hupel), esp. output format (TEI Lex-0 XML?)

Where to go from here?

- Continue OCR
 - Aligning and mapping the contents for *all* dictionaries
 - Diachronic word form developments across sources
 - Aligning and mapping the professional titles and societal roles in the dictionaries
 - Analysis of language change, concepts described
- Create a RAG-agent for analysing historical vocabulary across all available dictionaries and corpora (EKI, UT)

Aligning? (“AI paradox”)

<i>GUTSLAFF</i>	<i>STAHL</i>	<i>GÖSEKEN</i>	<i>VESTRING</i>
Angrias	Angrias	Angrias	Angrias
Raip	Raip	Raip	Raibe
Oddanck	öcht	öcht	Öhto
enge	agkas	aggas	Agga
Ohaya			
Telg	olla	telck	Tälg
Olk	ölgk	olgk	
Nurm	pöld	pöld	Kässa-Walli
Nastick	ader	Nastick	
Maria Jäh	Maria jeh	Marja jeh	Maarjad (Maarja Jä)
Sôn	ader	Soon	Aadrid
Pertike		pertick	Perdik.
Päh		peh	Pä
keick	keick	Keick	Keik
wanna	wanna	wanna	Wanna

Lummaja RAG-agent

Q: "Respond in English. Bring some examples of old professional titles and name the sources where you take this information."

R: "Thank you for your interest in these historical professional titles. Here is a list compiled from the provided documents:

1. (a) Ammet Mees
(b) Handwerck / Man
(c) Craftsman, artisan
(d) Göseken's "Manuductio ad Linguam Oesthonicam" (1660), p. 227
2. (a) AmmetiSeltz
(b) Ambts Gehülff
(c) Professional assistant, colleague
(d) Göseken's "Manuductio ad Linguam Oesthonicam" (1660), p. 105
3. (a) ettepajataja
(b) Anwalt
(c) Advocate, defender, lawyer
(d) Göseken's "Manuductio ad Linguam Oesthonicam" (1660), p. 107
4. (a) Juncker
(b) Meyer dem ein Fuhrwerck vertrauet ist
(c) Manager, steward (to whom transport is entrusted)
(d) Göseken's "Manuductio ad Linguam Oesthonicam" (1660), p. 294

5. (a) Juncker
(b) Schaffner
(c) Estate manager, steward
(d) Göseken's "Manuductio ad Linguam Oesthonicam" (1660), p. 349

6. (a) Juncker
(b) Ambtman
(c) Manor steward, bailiff
(d) Göseken's "Manuductio ad Linguam Oesthonicam" (1660), p. 105

7. (a) jütleia
(b) prediger
(c) Preacher
(d) Göseken's "Manuductio ad Linguam Oesthonicam" (1660), p. 326

8. (a) Amm
(b) Eine Amme
(c) Wet nurse
(d) Vestring's "Lexicon Esthónico Germanicum" (c. 1720), p. 5

9. (a) Teo Mees
(b) Arbeiter
(c) Worker, laborer
(d) Vestring's "Lexicon Esthónico Germanicum" (c. 1720), p. 5

Would you be interested in learning more about how these professional titles were used in daily life during the 17th and 18th centuries?"

Lummaja RAG-agent

EESTI
KEELE
INSTITUUT

EESTI RAHVA
MUUSEUM

INSTITUTE
OF THE ESTONIAN
LANGUAGE

LUMMAJA

Avasta vana aja sõnu ja esita oma küsimusi

Milliste vanade sõnade tähendus on tänapäevaks muutunud?

Roch

Wõlla

tebitz

Conclusion: why LLMs?

- Time efficiency
- Well prepared requests give reasonably good (if not excellent) outputs. Principle: *trash in* (poor prompt, wrong question) / *trash out* (AI slop). A good query needs preparation.
- Manual („wax-on, wax-off“) work can be focused on more specific and more interesting data.
 - ... e.g. analysing language change and cultural reflections illustrated by the professional titles and societal roles in those dictionaries

Thank you!

madis.jyrviste@eki.ee

Historical Estonian Word Form Search and Mapping Task

Context and Role

You are a historical lexicography expert specializing in 17th-18th century Estonian dictionaries. You will process a CSV file containing Estonian professional titles and societal roles, mapping them against historical dictionary entries from five major lexicographical works of the period.

Input Data Structure

The input CSV file contains approximately 1,600 lines with the following structure:

- ****Cat****: Category codes (K1, K2, K3 representing different types of professional/societal roles)
- ****Author****: Historical dictionary author (Stahl, Gutsclaff, Göseken, Vestring, Helle)
- ****Modern****: Modern Estonian word form

****IMPORTANT****: Hupel's data is MISSING from the database! Ignore all input rows where Author = "Hupel". Do not create a Hupel column in the output.

Central Task Description

****Your primary task is to find OLD, HISTORICAL Estonian word forms from the five dictionary authors' works, NOT simply repeat modern forms!****

Each unique modern Estonian word form must be compared against your embedded knowledge database, which contains data about five 17th-18th century Estonian-German dictionaries. You **MUST** use your embedded knowledge database to find the actual historical spellings.

Required Output Format

Create a CSV with exactly 10 columns (WITHOUT Hupel column) in this order:

1. ****Category****: Consolidated category codes (K1, K2, K3) separated by commas if multiple categories apply to the same word
2. ****Modern form****: The modern Estonian word form from the input
3. ****Meaning****: Word meaning in Estonian language. Priority order for determining meaning:
 - If present in embedded database, use that definition
 - If not available, derive from German translations in the embedded database
 - If unavailable, use the Estonian word form structure and your lexicographical knowledge
 - As last resort, use general linguistic knowledge
4. ****Stahl word form****: Exact word form as recorded in Stahl's dictionary (preserve original orthography). ****IF NOT FOUND, LEAVE EMPTY****
5. ****Gutsclaff word form****: Exact word form as recorded in Gutsclaff's dictionary (preserve original orthography). ****IF NOT FOUND, LEAVE EMPTY****
6. ****Göseken word form****: Exact word form as recorded in Göseken's dictionary (preserve original orthography). ****IF NOT FOUND, LEAVE EMPTY****

7. **Vestring word form**: Exact word form as recorded in Vestring's dictionary (preserve original orthography). **IF NOT FOUND, LEAVE EMPTY**
8. **Helle word form**: Exact word form as recorded in Helle's dictionary (preserve original orthography). **IF NOT FOUND, LEAVE EMPTY**
9. **Comment**: Your analytical comment about the mapping process in Estonian, including:
 - Notes about orthographic variations
 - Difficulty in distinguishing between word variants vs. completely different words
 - Any semantic shifts or ambiguities observed
 - Missing entries in certain dictionaries
 - **DO NOT** include confidence level here - it goes in separate column
10. **Confidence**: Your confidence level in the mapping as a single number from 0 (minimum confidence) to 10 (maximum confidence)

Critical Requirements - AVOID HALLUCINATIONS

Absolute Prohibition Against Fabrication

- **NEVER INVENT OR GUESS** historical word forms
- **IF** a word form is not found in your embedded knowledge for a specific author, **LEAVE THAT CELL COMPLETELY EMPTY**
- **DO NOT** create plausible-looking historical variants
- **DO NOT** modify modern forms to "look historical"
- **ONLY** use word forms that you can definitively verify from your embedded knowledge

Data Consolidation

- Process only unique modern word forms (eliminate duplicates from input)
- Consolidate category information from all input rows containing the same modern word form
- **EXCLUDE** all rows where Author = "Hupel"
- Ensure all available author word forms are included for each unique modern word

Orthographic Integrity

- **NEVER** modify or "correct" the historical orthography - reproduce word forms exactly as they appear in the historical dictionaries
- Preserve all diacritical marks, archaic letters, and spelling variations
- **Use** empty cells for authors who do not have entries for specific words
- **Empty** cells are preferable to guessed or invented forms

Lexicographical Analysis

- **USE** your embedded knowledge database to find **ACTUAL HISTORICAL FORMS**
- Prioritize entries that are (a) orthographically and (b) semantically (in this priority order) related to the modern form
- Consider morphological variations typical of the historical period
- Account for German influence on Estonian orthography in these dictionaries
- Distinguish between genuine word variants and potentially different lexemes

- ****When uncertain, leave cells empty rather than guess****

Research Methodology

- Systematically search your embedded knowledge for each author's dictionary entries
- Cross-reference entries between different authors to identify relationships
- Note cases where historical forms significantly diverge from modern forms
- Comment on any lexicographical inconsistencies or gaps
- ****Be honest about missing data - empty cells are scientifically valuable****

Output Guidelines

- Begin your CSV output with the exact header row: Category;Modern form;Meaning;Stahl word form;Gutslaff word form;Göseken word form;Vestring word form;Helle word form;Comment;Confidence
- ****USE SEMICOLONS (;) as field separators, NOT commas****
- Process words in the order they first appear in the input file
- Ensure each output row represents a complete analysis
- Enclose fields containing commas or special characters in quotation marks
- ****Empty cells should contain no characters whatsoever between semicolons****

Temperature and Precision Instructions

- ****Use maximum precision and minimum creativity****
- ****Prioritize accuracy over completeness****
- ****When in doubt, leave cells empty****
- ****Historical accuracy is more important than filling all columns****

Final Instructions

Process the entire input file systematically, analyzing each unique word form against your embedded historical dictionary knowledge.

****CRITICAL RULES:****

1. You **MUST** use your embedded database to find **ACTUAL HISTORICAL** Estonian word forms
2. ****NEVER** invent, guess, or create plausible-looking historical forms**
3. ****IF** a word form is not found in your knowledge for a specific author, **LEAVE THAT CELL EMPTY****
4. Provide meanings in Estonian language
5. Write comments in Estonian language
6. Assign confidence scores from 0-10 based on how certain you are about the mappings

Use maximum available tokens for your response to ensure comprehensive coverage of all entries, making certain that your final analyzed line is complete and not truncated mid-analysis.

Category;Modern form;Meaning;Stahl word form;Gutslaff word form;Göseken word form;Vestring word form;Helle word form;Comment;Confidence

K3;aadler;aadlik, ülikklassi kuuluv isik;;;;Adel;;Leitud ainult Vestring'i sõnaraamatust saksa vastega 'Adel'. Teistel autoritel puudub;7

K1;abbati emand;abti naine, kloostri ülempreestri abikaasa;;;Abbati Emmand;;;Göseken'i sõnaraamat sisaldab täpse vaste. Teistel autoritel

K2;abielurikkuja;abielu rikkuv isik;;;;Abbiellorikkia;abbiellorikkuja;Vestring'i ja Helle vormid näitavad ortograafilisi erinevusi. Gösek

K2;abiline;abistaja, toetaja;;;;;Seda täpset vormi ei leidnud ühestki ajaloolisest sõnaraamatust. Võimalik, et kasutati teisi termineid;2

K2;abimees;abistav meesisik;;;Abbi Mees;;;Göseken'i sõnaraamatus esineb kaheosalisena. Teistel autoritel puudub;8

K2;abinaine;abistav naissoost isik;;;Abbi Naine;;;Göseken'i sõnaraamatus samuti kaheosalisena nagu abimees. Teistel puudub;8

K2;abivaim;abistav vaim;;;Abbi Waim;Abbi Waim;;Göseken ja Vestring kasutavad identset vormi. Helle sõnaraamatus ei leidunud;8

K2;abivaimud;abistavad vaimud;;;;;abbiwaimo;Helle sõnaraamatus esineb ainsuse vorm 'abbiwaimo'. Mitmuse vorm puudub;6

K1;abt;kloostri ülem;;;Abt;;;Göseken'i sõnaraamatus on saksa laensõna muutmata kujul. Teistel puudub;9

K1;abtiemand;abti naine;;;Abti Emmand;;;Göseken kasutab sama vormi nagu 'abbati emand' puhul. Teistel puudub;9

K3;adramees;põllumees, kündjaga töötav mees;;;Adra Mees;Addramees;addramees;Göseken kirjutab lahku, Vestring ja Helle kokku. Ortograafili

K1,K3;aednik;aedade eest hoolitseja;;;;;Ei leidnud ühtegi ajaloolist vastet. Võimalik, et kasutati teisi termineid nagu 'aiamees';1

K3;aganakott;aganate kott;;;;;Ei leidnud spetsiifilist vastet. Liitsõna, mille komponendid võivad eraldi esineda;2

K1,K3;aiamees;aia eest hoolitseja;;;Aia Mees;Aiamees;;Göseken kirjutab lahku, Vestring kokku. Helle sõnaraamatus puudub;8

K3;aidamees;aedade ehitaja või hoidja;;;Aida Mees;;;Göseken'i sõnaraamatus lahku kirjutatuna. Teistel puudub;7

K2;aitaja;abistaja, toetaja;;;Aitaja;Aitaja;aitaja;Kõigil kolmel autoril identne vorm. Stahl ja Gutslaff puuduvad;9

K2;äitsiline;ämmaemand, sünnitusabi andja;;;;;Ei leidnud ajaloolist vastet. Võib-olla kasutati teisi termineid;1

K2;ajaja;see, kes ajab, juht;;;Ajaja;Ajaja;;Göseken ja Vestring kasutavad sama vormi. Teistel puudub;8

K3;alam;madalamal positsioonil olev isik;;;;;Ei leidnud täpset vastet 'alam' kohta. Võimalik, et kasutati teisi termineid;2

K3;alamad;madalamal positsioonil olevad isikud;;;;Allamad;allamad;Vestring ja Helle kasutavad topelt-l-iga vormi. Teistel autoritel puudu

K3;alandik;alandlik isik;;;Allandik;;;Göseken'i vorm topelt-l-iga. Teistel autoritel ei leidunud;7