


Leveraging Large Language Models for Historical Lexicography: Analysis and Enhancement of 17th-18th Century Estonian Dictionaries

Madis Jürviste (lexicographer, junior researcher, UT, EKI)

Tiina Paet (PhD, researcher, EKI)

Sven-Erik Soosaar (PhD, senior researcher, EKI)

Introduction

- Pilot project from 2018: ÕS 1918 → 
- Broader aim:
 - explore the possibilities of applying LLMs to identify word forms in older Estonian vocabulary
 - evaluate the identification of word forms by LLMs
- Long-term goal:
 - historical layer with LLMs

Experiments in 2024

- LLMs: GPT-4o, Claude 3 Opus, Gemini 1.5 Pro
 - **Sources:** Stahl (1637), Gutsclaff (1648), Göseken (1660), Vestring (17XX), Helle (1732), Hupel (1780)
 - **Exp 1:** modern forms > hist. forms (professional titles)
 - **Exp 2:** hist. S-Est. forms (Gutsclaff) <> modern forms
 - **Exp 3:** hist. loanwords (Stahl, Gutsclaff, Göseken) <> modern forms

Exp 1: professional titles / societal roles

- **Aim:** to find hist. professional titles / societal roles that correspond to modern terms: e.g. *kohtunik* (modern form) and *sundija* (hist. form) ‘judge’; *varas* (modern form) and *kühnamees* (hist. form) ‘thief’
- **Sources:** 6 dict-s
- **Input:** 30 words, incl. *Narr* vs *jeck* (‘jester’), *spioon* vs *kulaja* (hist.) (‘spy’), *nõid* vs *lämmia* (hist.) (‘witch’)

Exp 2: Gutsclaff's hist. <> modern forms

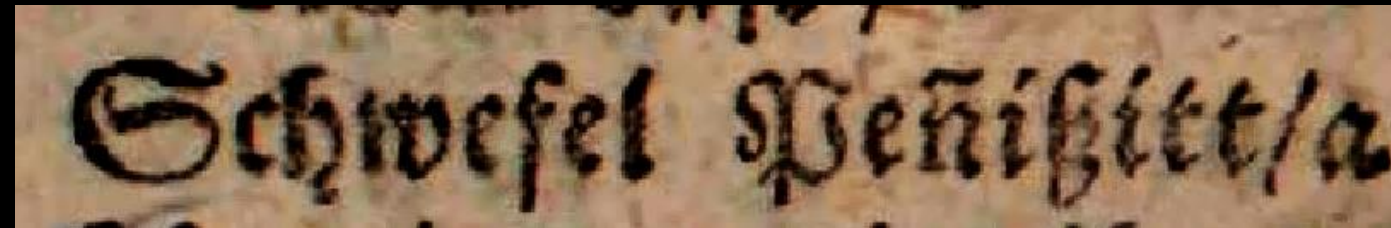
- **Aim:** find equivalents for hist. forms in modern Estonian and vice versa
- **Source:** Gutsclaff (1648) dialectal (S-Est.)
- **Input:** 54 words incl. **identical words** (e.g., *kirp* 'flea', *raha* 'money'); **minor orthographic variations** (e.g. *Hawd* (hist form) *haud* (modern Est) 'grave'); **completely different words** (e.g. *ramato nachk* and *paber* 'paper')

Exp 3: Loanwords

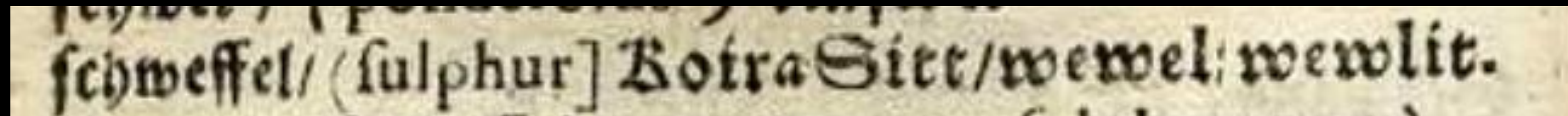
- **Aim:** to assess the ability of LLMs to link historical orthographic forms with each other and their corresponding modern forms
 - *mijrrit* (Göseken), *Mirra* (Gutslaff) *mürr* (modern form) ‘myrrh’
- **Sources:** Stahl 1637, Gutslaff 1648, Göseken 1660
 - No German equivalent given to the LLM.
- **Input:** 20 words
 - 1) minimally different: *pipar*, *pippar*, *pipper* ‘pepper’;
 - 2) somewhat different: *äädikas* and *Ettickas* ‘vinegar’;
 - 3) completely different: *väävel* and *koirasitt*, *pennisitt* ‘sulfur’, lit. ‘dog’s excrement’

Exp 3 example: sulfur

Gutslaff (1648): *Schwefel* (Germ.) and *Pennisitt* (Est.)



Göseken (1660): *schweffel* (Germ.) and *KoiraSitt / wewel* (Est.)



Exp 1–3 results

- Most accurate model: Anthropic's Claude 3 Opus
 - Exp 1: 90%
 - Exp 2: 74–81%
 - Exp 3: 60–90%
- Success depends on the choice of input words (the fewer differences from the modern form, the more accurate).
- Further experiments: OCR and enrichment of historical dict-s.

OCR: Anton Thor Helle 1732

- 2024: *vision*-enabled models (Claude 3.5 Sonnet, 3.7 Sonnet)
 - If LLMs can 'read' trees and cats, then ...
 - ... can they also read 18th century Fraktur?
 - Custom API script (Py), prompt development
 - Anthonius (Joonatan Jakobson)
 - Manual review

OCR: Anton Thor Helle 1732

Kaswatik der Aufzügling. 71.
Kaswias das Gewächs am
Leibe. 14, 2.
Katk die Pest; Wasser. Pfüge,
da ein harter Grund ist.
ei ma olle Katko pödde=
und ich habe die Pest
nicht gehabt.
Katke entzwen, von einander.
adv.
Katkema entzwen gehen, zer-
brechen.
Katkenu hobbone ein Pferd
das sich verzogen hat.
Katkestama sich was im Lei-
be zerbrechen, verdrieß
thun.

- Claude 3.7 Sonnet, API
- Thoughtful and thorough prompting
- Structured data extraction *in addition* to simple OCR
- Result: Near-perfect recognition?
... or ... not so perfect?
- Critical takeaway: even if key metrics seem catastrophic, manual post-editing can still be terribly efficient

OCR / 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 1780), esp. output format (TEI Lex-0 XML?)

Enrichment: why?

Let's decipher ...

- Gutsclaff 1648:
 - *karrakaro*
 - *padia*
 - *kicko, kantz*
- Hupe1 1780
 - *särke sap (r.) – Buchgold?* Modern form: särjesapp?
 - Roach bile?
 - What do books and gold have to do with fish?

Gu 1648 enrichment: workflow

- Gutsclaff 1648 + context (Stahl, Göseken, Vestring)
 - Stahl 1637 and Göseken 1660: with modern annotations
- Alignment (GPT + Claude)
 - 'AI paradox'
 - Manual review
- Claude 4 Opus (API Msty)
 - Manual review (incl. Viitkar 2005)

"Püsse", "püss", "püss", "Büchse", "Büchse", "'Püsse' on püss"
"Püssohand", "pisuhänd", "kratt, koletis", "Drache", "Drache", "'Püssohand' on pisuhänd, kratt"
"Putsaja", "sulg", "sulg", "Feder", "Feder", "'Putsaja' on sulg"
"Puttr", "puder", "puder", "Brey", "Brei", "'Puttr' on puder"
"Raip", "raibe", "surnukeha", "Aass", "Aas", "'Raip' on raibe"
"Ramat", "raamat", "kiri", "Briefff", "Brief", "'Ramat' on raamat kui kiri"
"Ramat", "raamat", "raamat", "Buch", "Buch", "'Ramat' on raamat"
"Ramm", "surnupaarid", "kanderaam", "Baare", "Bahre", "'Ramm' on surnupaarid"
"Rasa", "raas", "leivatükk", "Brocke Brods", "Brotstück", "'Rasa' on raas, leivatükk"
"Rasw", "rasv", "rasv", "Fett", "Fett", "'Rasw' on rasv"
"Reiss", "libejää", "sile jää", "glat Eiss", "glattes Eis", "'Reiss' on libejää"
"Rind", "rind", "rind", "Brust", "Brust", "'Rind' on rind"
"Risti", "rist", "rist", "Creutz", "Kreuz", "'Risti' on rist"
"Risti inniminni", "ristiinimene", "kristlane", "Christe", "Christ", "'Risti inniminni' on ristiinimene"
"röhmus", "rõõmus", "rõõmus", "anmutig", "anmutig/fröhlich", "'röhmus' on rõõmus"
"rüchkma", "ruttama", "kiirustama, jõudma", "eilen", "eilen", "'rüchkma' on ruttama"
"rummal", "rumal", "rumal", "dumm", "dumm", "'rummal' on rumal"
"Russick", "rusik", "rusikas", "Faust", "Faust", "'Russick' on rusik"
"saddama", "sadama", "sadama, langema", "fallen", "fallen/regnen", "'saddama' on sadama"
"salvama", "salvama", "hammustama", "beissen", "beißen", "'salvama' on salvama"
"Sanna", "saun", "saun", "Badstube", "Badstube", "'Sanna' on saun"
"Sant", "sant", "vaene", "Betler", "arm", "'Sant' on sant, vaene"
"Sarapuh", "sarapuu", "sarapuu", "Aschenbau(m)", "Haselstrauch", "'Sarapuh' on sarapuu"
"Savvissiwk", "vaskuss", "vaskuss", "Blindschleich", "Blindschleiche", "'Savvissiwk' on vaskuss"
"selge", "selge", "selge", "Clahr", "klar", "'selge' on selge"
"Seng", "säng", "voodi", "Bette", "Bett", "'Seng' on säng, voodi"
"sick", "sikk", "sokk", "Bock", "Bock/Ziegenbock", "'sick' on sikk"

Gutslaff 1648
enrichment:
API 1st result

Gu 1648 enrichment: initial results

- Claude 4 Opus (API) 03.06.2025
 - 278 out of 342 lines: all (or mostly) correct
 - 36 out of 342 lines: additional review needed
 - 28 out of 342 lines: manual override needed

>>> 81% -- 11% -- 8%

Perspectives?

- Continue OCR tests: Hupe 1780 (1818)
 - Improve format (XML), prompt, workflow
- Continue enrichment / commenting of old sources
 - Gu 1648, Ve ~1720, ATH 1732, ...
- Preparations for a comprehensive “Dictionary of Dictionaries”
 - 17th–19th century, Stahl 1637 ... Hupe 1818

Conclusions

- 2024 experiments: LLMs *can* analyse historical lex data
 - Surprising results (*pärdik X ahv*, versus *pennisitt <> väävel*), sometimes unpredictable / incoherent
- Anton Thor Helle 1732 OCR: “solved” case for LLMs
 - If the case is “solved”, then let’s remove the quotes
- Gutsclaff 1648 enrichment: initial results very promising
 - Cryptic content > accessible content

References

- Beliga, S., & Filipović Petrović, I. 2024.** Large language models supporting lexicography: Conceptual organization of Croatian idioms. Conference on Language Technologies and Digital Humanities. Ljubljana, Slovenia.
- De Schryver, Gilles-Maurice; Joffe, David 2023.** The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks. 20th CODH Seminar, Center for Open Data in the Humanities. Research Organization of Information and Systems, National Institute of Informatics.
- Gutslaff, Johannes 1648.** Observationes Grammaticae circa linguam Esthonicam. Dorpat: Johannes Vogel. <http://www.digar.ee/id/nlib-digar:100419> (10.09.2024).
- Göseken, Heinrich 1660.** Manuductio ad Linguam Oesthonicam. Anführung zur Öhstnischen Sprache. Reval: Adolph Simon. <https://kivike.kirmus.ee/meta/AR-11170-72005-62344> (10.09.2024)
- Helle, Anton Thor 1732.** Kurtzgefaszte Anweisung Zur Ehstnischen Sprache. Halle: Stephan Orban. <http://www.digar.ee/id/nlib-digar:100071> (10.09.2024)
- Hupel, August Wilhelm 1780.** Ehstnische Sprachlehre für beide Hauptdialekte den revalschen und dörptschen; nebst einem vollständigen Wörterbuch. Riga–Leipzig: Johann Friedrich Hartknoch. <http://www.digar.ee/id/nlib-digar:100926> (10.9.2024)
- Jürviste, Madis; Paet, Tiina; Soosaar, Sven-Erik (2025).** Eesti vanade sõnakujude tuvastamisest suurte keelemudelitega. Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics, 21, 63–83.
- Keem, Hella 1998.** Johannes Gutslaffi eesti keel ja Urvaste murrak, teoses Gutslaff 1998, lk 317–332.
- Kikas, Kristel 2002.** Mida sisaldab Heinrich Stahli *Vocabula*? Toim. Valve-Liivi Kingisepp. Tartu Ülikooli eesti keele õppetooli toimetised 11. Tartu: Tartu Ülikool.
- Kingisepp, Valve-Liivi; Ress, Kristel; Tafenau, Kai 2010.** Heinrich Gösekeni grammatika ja sõnastik 350. Tartu: Tartu Ülikool.
- Stahl, 1637.** Stahl, Heinrich 1637. Anführung zu der Esthnischen Sprach. Reval: Chr. Reusner der älter. <http://www.digar.ee/id/nlib-digar:101060> (10.09.2024)
- Tafenau, Kai 2011.** *Heinrich Gösekeni sõnaraamatu seni märkamata eeskujud*. – Keel ja Kirjandus, 6, pp. 425–439.
- Vestring 2000** = Vestring, Salomo Heinrich 2000 [17XX]. Lexicon Esthónico Germanicum. Võrguteavik. Toim. Ellen Kaldjärv, Krista Aru, Arvo Krikmann. Tartu: Eesti Kirjandusmuuseum. <https://www.folklore.ee/~kriku/VESTRING/index.htm> (17.09.2024)
- Viitkar, Urve 2005.** Johannes Gutslaffi “Observationes grammaticae circa linguam Esthonicam” (1648) sõnastiku leksika. Bakalaureusetöö. Tartu: Tartu Ülikooli eesti keele õppetool. Käsikiri Tartu Ülikooli eesti keele osakonna raamatukogus.
- Wild, Kate 2024.** Using Large Language Models for a Large Historical Dictionary: Challenges and Opportunities for the Oxford English Dictionary.

Thank you!

madis.jyrviste@eki.ee

tiina.paet@eki.ee