ICHLL15

BRIDGING 17TH AND 18TH CENTURY ESTONIAN AND LLMS: UNLOCKING HISTORICAL DICTIONARIES

Madis Jürviste – lexicographer, junior researcher – Institute of the Estonian Language & University of Tartu
Tiina Paet – researcher, Institute of the Estonian Language
Sven-Erik Soosaar – senior researcher, Institute of the Estonian Language

* This study was funded through the R&D project 'Applying large language models to lexicography: new opportunities and challenges' (EKKD-III1).

International Conference ICHLL15 Historical Lexicography Lexicology

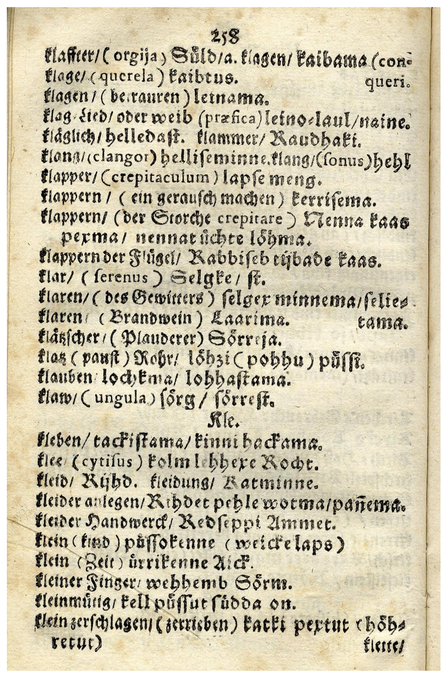
Dictionaries studied

- Stahl, Heinrich 1637. Anführung zu der Esthnischen Sprach.
- Gutslaff, Johannes 1648. Observationes Grammaticae circa linguam Esthonicam.
- Göseken, Heinrich 1660. Manuductio ad Linguam Oesthonicam. Anführung zur Öhstnischen Sprache.
- Vestring, Salomo Heinrich 2000 [~1720]. Lexicon Esthonico Germanicum.
- Helle, Anton Thor 1732. Kurtzgefaszte Anweisung Zur Ehstnischen Sprache.
- Hupel, August Wilhelm 1780. Ehstnische Sprachlehre für beide Hauptdialekte den revalschen und dörptschen; nebst einem vollständigen Wörterbuch.

Heinrich Stahl, "Anführung zu der Esthnischen Sprach ..." (1637)

***Riie* 'cloth, garment'**
Ger: Kleid

Stahl 1637: ***rihd***
Göseken 1660: ***Rijhd***
Vestring ~1720: ***Rie***
Helle 1732: ***rie***
Hupel 1780: ***rie***
Hupel 1818: ***riid, ride, rie***
Wiedemann 1893: ***rīe***
1918...2025: ***riie***



Heinrich Göseken, "Manuductio ...." (1660), p. 258

Research

Overview and initial experiments

Our research explores the significant potential of Large Language Models (LLMs) in historical lexicography. The larger aim of the project is to create a historical layer using LLM technology (see example on the left, *riie*.)

Material
Six German-Estonian and Estonian-German dictionaries, spanning from 1637-1780.

LLMs tested
GPT-4o, Claude 3 Opus, Gemini 1.5 Pro.

Aim
Explore the potential use of LLMs in studying 17th and 18th century lexicography, as well as finding the most suitable model for analysing historical Estonian lexical data.

Experiments
- Experiment 1: 30 professional titles (*kohtunik / sundija* 'judge', *varas / kühnames* 'thief') and societal roles across all the six sources.
- Experiment 2: 54 words from Gutslaff (1648) (South Estonian). Three categories tested: identical forms (*kirp* 'flea'), minor orthographic variations (*Hawd / haud* 'grave'), completely different lexemes (*ramato nachk* lit. 'book skin'/ *paber* 'paper').
- Experiment 3: 20 loanwords without German equivalents across 3 sources. Three categories tested: minimal differences (*pipar / pippar / pipper* 'pepper'); somewhat different forms with overlapping letters (*äädikas / Ettickas* 'vinegar'); radical divergences (väävel / koirasitt 'sulfur').

Success rates
- E1: 90% accuracy (Claude 3 Opus).
- E2: 76-81% accuracy (Gemini 1.5 Pro, Claude 3 Opus).
- E3: 60-90% accuracy rates (Claude 3 Opus).

Case-study: Gutslaff 1648. Enriching historical dictionaries with modern forms and meanings

Challenge and approach: Historical dictionaries contain entries (*karrakaro*, *padia*, *särge sap*) that are cryptic for modern readers, especially non-linguists. Envisaged solution: provide modern word forms and meanings for historical lexicographical data.

- **Aim**: Establishing replicable workflows applicable to other historical dictionaries and language pairs (Gutslaff 1648 has undergone manual analysis already: Lepajõe 1998, Viitkar 2005).
- **Example** : [et] *Ahwoneck* – [de] *Barß* > mod. *ahven*, 'perch'.

Method: multi-stage LLM-assisted analysis (cross-source mapping), followed by expert (human) validation and oversight.

Tested LLM: Claude 4 Opus (API).

Test sample: 342 entries out of 1714 (20%).

Success rate: 81% of entries (278 out of 342) required no correction. 11% needed additional review. 8% of entries required complete manual override.

"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"	"Brieff"	"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"

Excerpt from an API-request (CSV-output): enriching Gutslaff's 1648 dictionary with modern word forms and meanings. Data points:
>>> [original et], [modern et], [meaning], [original de], [modern de], [LLM comment]

Conclusions

- LLMs are capable of historical lexicographical analysis.
- Optimal resource efficiency in lexicographical analysis.
- Cross-linguistic potential: adaptability to other language pairs.
- Scalability: Gutslaff 1648 enrichment experiment is applicable to other historical dictionaries.
- Expert validation is essential: LLM analysis combined with human expert validation (manual review) yields best results.

Contacts

- madis.jyrviste@eki.ee
- tiina.paet@eki.ee
- svenerik.soosaar@eki.ee

INSTITUTE OF THE ESTONIAN LANGUAGE

UNIVERSITY OF TARTU · UNIVERSITAS TARTUENSIS · 1632

ACADEMIA DAS CIÊNCIAS DE LISBOA
FLUL FACULDADE DE LETRAS
CLUL
NOVAFCSH
CLUNL CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA
fct Fundação para a Ciência e a Tecnologia