

Ajalooliste tekstide tärgtuvastusest suurte keelemudelitega Anton Thor Helle 1732. a sõnastiku näitel

Madis Jürviste (EKI nooremteadur-leksikograaf, TÜ doktorant)
Joonatan Jakobson

Taust I

- “Klassikaline” OCR ajalooliste tekstidega:
ABBY, Transkribus, Kraken, eScriptorium jt
 - DIGAR MM 26.04.1919: “.. 5r.ihelkomia thecmajas
haamatud Eesli sõdurite heaks pidu-Zhtn”.
- Hea tulemus eeldab käsitsi segmenteerimist ja ulatuslikku *ground truth*’i
- LLMid? Kuni 2024 august “tavalised” Pythoni teegid, nt pypdf, tulemus pea loetamatu

Taust II

- Murrang: *vision*-funktsiooniga mudelid
- Senised katsed: Claude'i edumaa teiste ees
(GPT, Gemini, Mistral)
- Esmane küsimus: *vision*-LLMid + 18. saj fraktuur?
 - Katse: ATH 1732

Taust III

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

Viip (Claude 3.5 Sonnet *chat*):
“Vaata seda fraktuuris PDFi ja
kirjuta tulemused tabelisse.”

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

Prompti arendus

- API-päring (Python)
 - sisend: PDF (PNG)
 - väljund: CSV
- Väljakutse: Claude 3.5 segment. (lüngad: „tagurpidi L“)
 - lahendus 1: käsitsi segmenteerimine?
 - lahendus 2: PDF *chat*’ile 10 lk kaupa, väljund CSV?
 - lahendus 3: Claude 3.7 Sonnet

Claude 3.7 Sonnet ja OCR

- Esmane tulemus võrreldav *chat*'i-rakenduse väljundiga
- Pealtnäha suurepärase tulemus: illusioon või reaalsus?
- Mitte ainult OCR, vaid samaaegselt ka struktureerimine (CSV)
- Terve sõnastiku OCR + strukt ajakulu ja maksumus:
terve sõnastik ca 10 €, 1 lk < 1 min

Dictionary Pages

Sort by: Page Number ↑

Export Dictionary

Upload More

Anthonyus dictionary page 83. The page contains a list of entries in a two-column format. The entries are in a historical German dialect, likely from the 16th century. The text is written in a Gothic script. The page number 83 is visible in the bottom right corner.

Anthonyus dictionary page 84. The page contains a list of entries in a two-column format. The entries are in a historical German dialect, likely from the 16th century. The text is written in a Gothic script. The page number 84 is visible in the bottom right corner.

Anthonyus dictionary page 85. The page contains a list of entries in a two-column format. The entries are in a historical German dialect, likely from the 16th century. The text is written in a Gothic script. The page number 85 is visible in the bottom right corner.

Anthonyus dictionary page 86. The page contains a list of entries in a two-column format. The entries are in a historical German dialect, likely from the 16th century. The text is written in a Gothic script. The page number 86 is visible in the bottom right corner.

Page 83

Progress 100%

View Dictionary (47)

Page 84

Progress 100%

View Dictionary (53)

Page 85

Progress 100%

View Dictionary (48)

Page 86

Progress 100%

View Dictionary (50)

Prompt Details

[← Back to Prompts](#)

[✎ Edit](#)

[⚙ Use This Prompt](#)

Prompt Information

ID	3
Name	ATH Master 1 prompt NT
Model	claude-3-7-sonnet-20250219
Max Tokens	64000
Temperature	0.0
Thinking	Disabled
Created:	2025-03-06 08:29
Last Updated:	2025-03-06 08:51

 Delete Prompt

Usage History

Result #7	2025-03-06 08:53
Image: Helle-1732_DIGAR-ETERA_DICT_main_20250221_Page_001.png	
Entries extracted: 47	

Prompt Content

You are an expert in historical linguistics and lexicography analyzing a 1732 Estonian–German dictionary by Anton Thor Helle. Your task is to extract and structure ALL visible text entries from the provided page, ensuring no text is skipped. Be precise and follow these guidelines:

General Instructions:

- Process the page COLUMN BY COLUMN:
 - First extract all entries from the LEFT column (top to bottom)
 - Then extract all entries from the RIGHT column (top to bottom)
 - This is crucial for maintaining correct alphabetical order
- Process ALL text in each column, including:
 - Partially visible or incomplete entries at page boundaries
 - Entries spanning multiple lines
 - Marginal notes, footnotes, or annotations
- Preserve:
 - Original spellings, punctuation, and diacritical marks
 - Historical character usage (for example, "ö" instead of modern "õ")
 - All special characters, brackets, or nested parentheses
 - The exact order of entries as they appear in each column
- Pay special attention to:

Task Details

Results:

 View Page Dictionary

Project:	ATH Master 1	Image:	Helle-1732 DIGAR-ETERA DICT main 20250221 Page 001.png
Page:	#83	Prompt:	ATH Master 1 prompt NT
Start Time:	2025-03-06 08:52:39		
End Time:	2025-03-06 08:53:42		

Task Log

☒ Auto-scroll

2025-03-06 08:52:39

RUNNING

Initializing task processing...

2025-03-06 08:52:39

RUNNING

Starting processing of image 'Helle-1732_DIGAR-ETERA_DICT_main_20250221_Page_001.png' with prompt 'ATH Master 1 prompt NT'

2025-03-06 08:52:39

RUNNING

Initializing parser with model: claude-3-7-sonnet-20250219, thinking: False, temperature: 0.0

2025-03-06 08:52:39

RUNNING

Sending image to Claude API for processing...

2025-03-06 08:52:43

RUNNING

Received content: [

2025-03-06 08:52:43

RUNNING

Received content: { "est

2025-03-06 08:52:44

RUNNING

Received content: onian_headword": "Ador

2025-03-06 08:52:44

RUNNING

Received content: it", "estonian_synonym

2025-03-06 08:52:44

RUNNING

Received content: s": "", "german

2025-03-06 08:52:44

RUNNING

Received content: _equivalent": "laß

2025-03-06 08:52:44

RUNNING

Received content: t ma zur Ader lassen", "german

Task log

Processing Runs

Available Runs

ATH Master 1

Selected

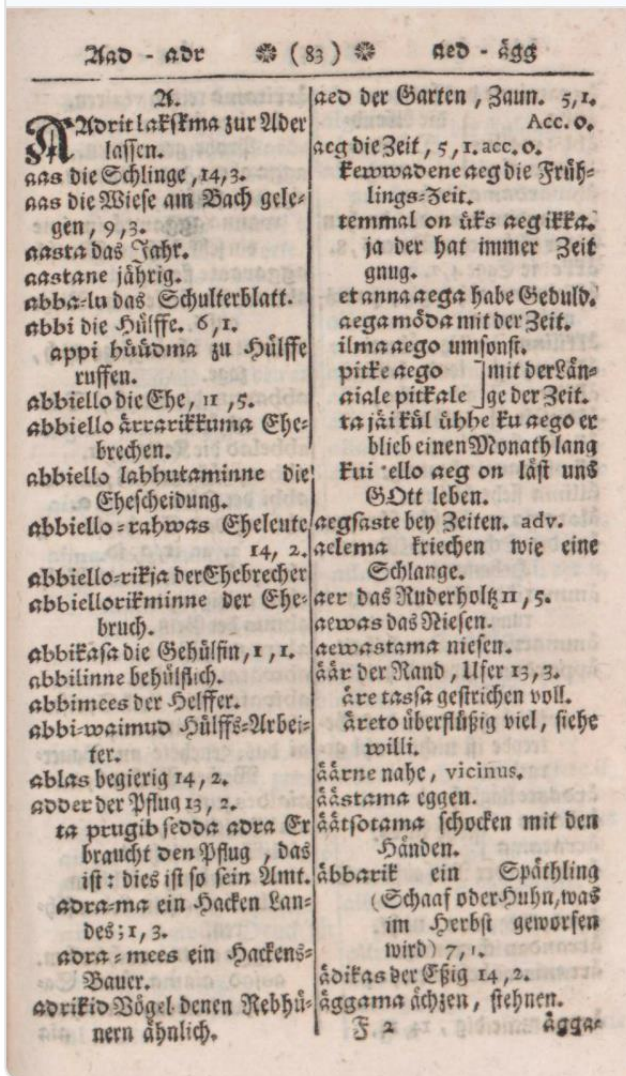


prompt NT

2025-03-06 08:52

Page Image

Page 83



⌵ aas

✓ Correct

—

✗ Incorrect

die Schlinge

Hide Details

Edit

Delete

```
{
  "estonian_headword": "aas",
  "estonian_mwu": "",
  "estonian_synonyms": "",
  "german_equivalent": "die Schlinge",
  "german_synonyms": "",
  "id": 302,
  "is_correct": null,
  "is_evaluated": false,
  "latin_explanation": "",
  "page_number": 83,
  "part_of_speech": "",
  "translated_mwu": ""
}
```

Copy JSON

⌵ aas

✓ Correct

—

✗ Incorrect

die Wiese am Bach gelegen

Show Details

Edit

Delete

⌵ aasta

✓ Correct

—

✗ Incorrect

das Jahr

Show Details

Edit

Delete

Tulemuse
kontroll ja
redigeerimine

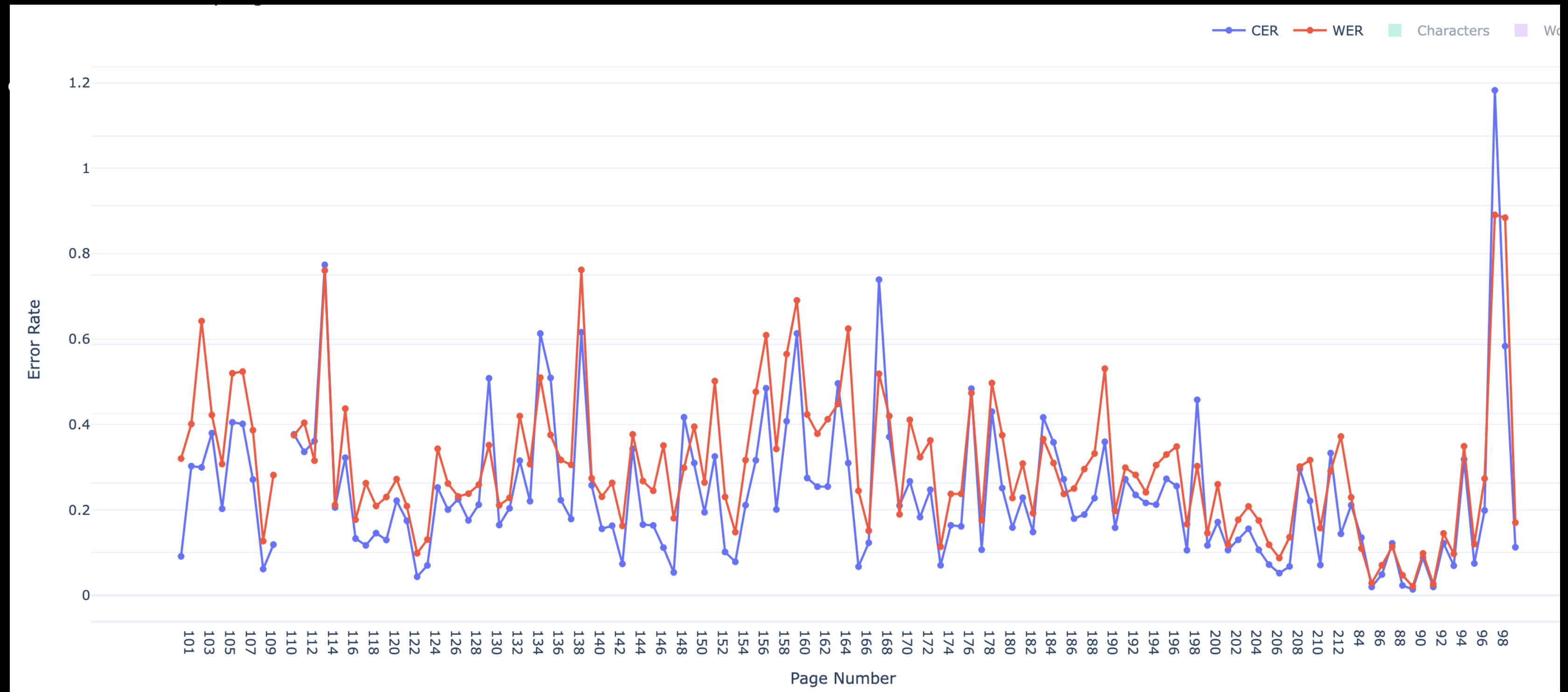
Claude 3.7 Sonnet: CER ja WER?

- Ei saa võrrelda võrreldamatut: struktuurielementide käsitlemine erineb “tavalisest” OCRist
- Ee mks vs sks vaste vs paigutus vs poolituskohad vs realõputähed jms trükikujunduse elemendid
- Eri struktuurielemente tuleb vaadelda eraldi
- Eelised ja väljakutsed?
 - Struktuur. Minimaalne ajaline ja rahaline kulu

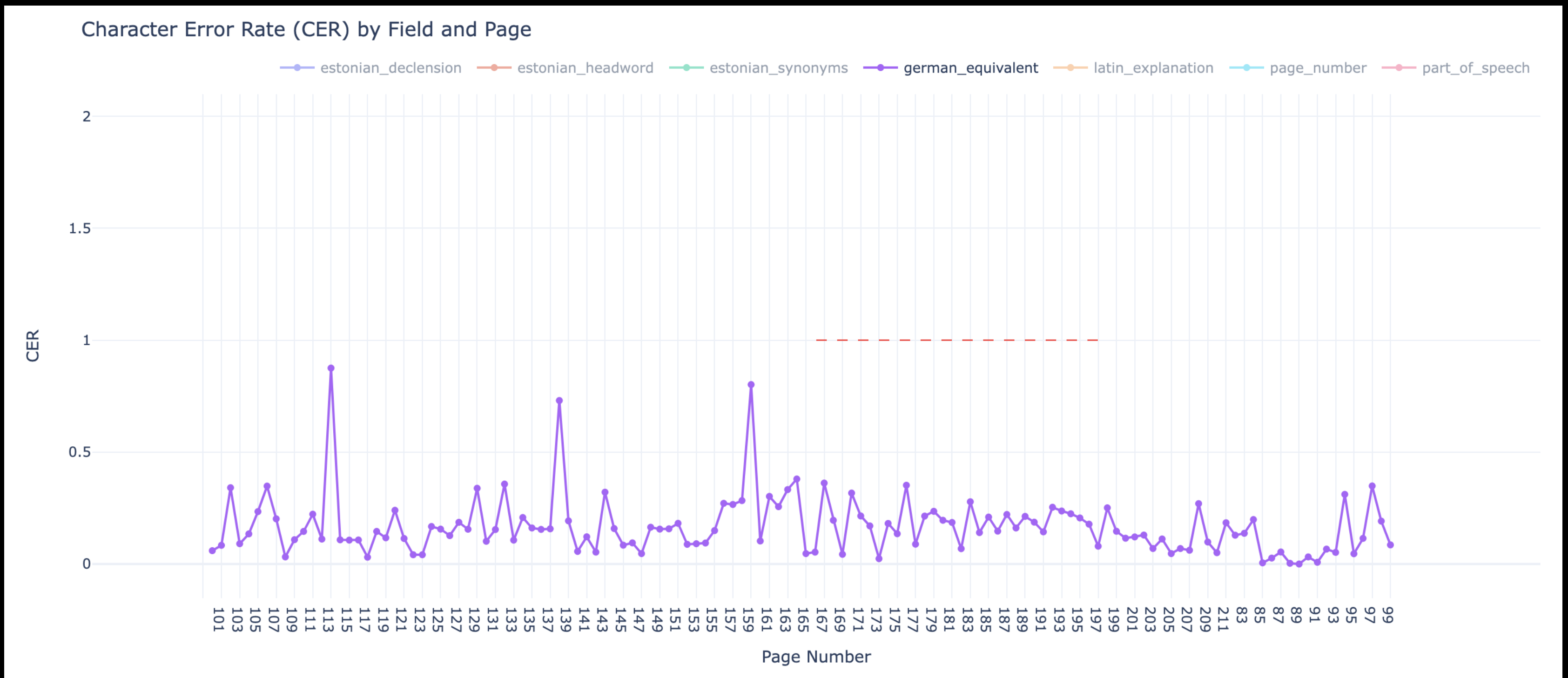
Claude 3.7 Sonnet: CER ja WER?

- CER ja WER määrad on väga varieeruvad:
 - CER 4% ... 120% sõltuvalt leheküljest (> 100 = lisandused)
 - Keskmine 24% (ee mks 42%, sks vaste 17%)
 - WER 2% ... 89% sõltuvalt leheküljest
 - Keskmine 30%
- Vrd “klassikaline” CER < 5%, uutel mudelitel pigem 1-2%
 - Eeldavad *ground truth*’i, mida ulatuslikum, seda parem
- *Exact match*: ee mks 64%, sks vaste 66%, terve rida 41%

Claude 3.7 Sonnet: CER ja WER?

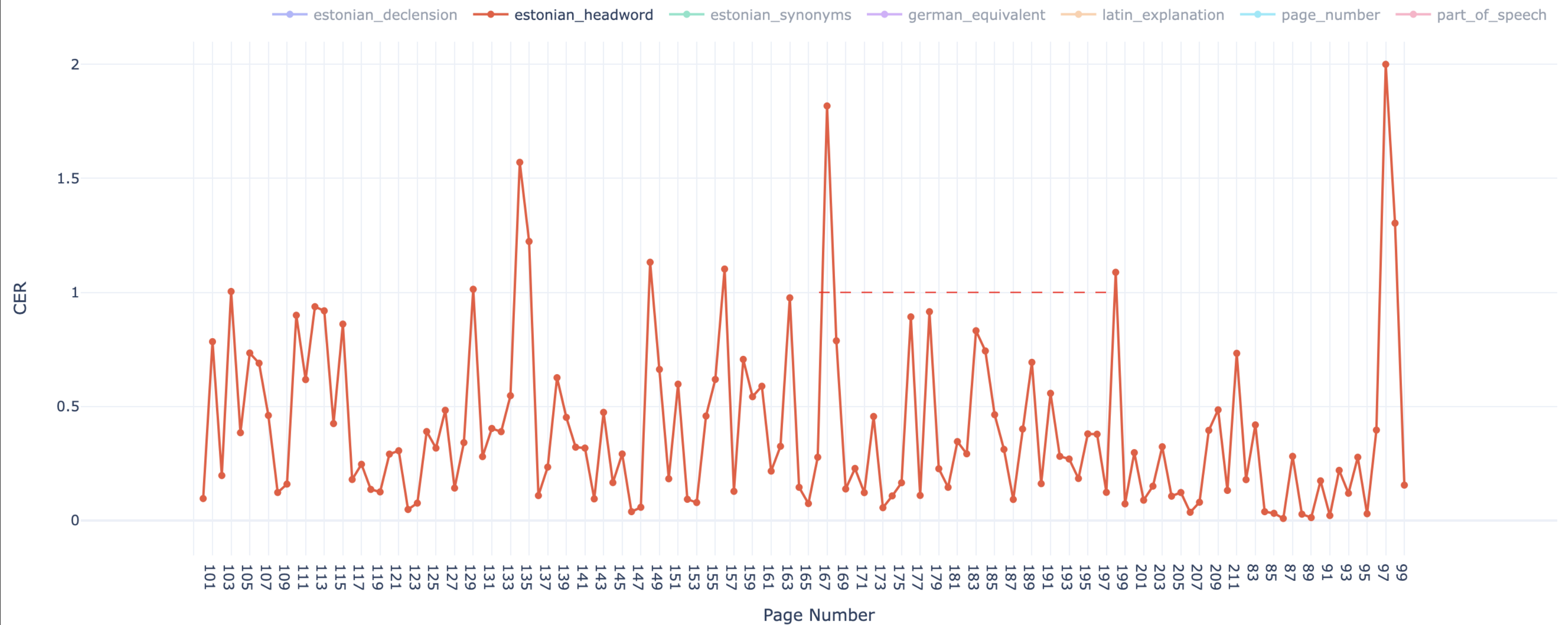


Claude: CER *saksa vaste*



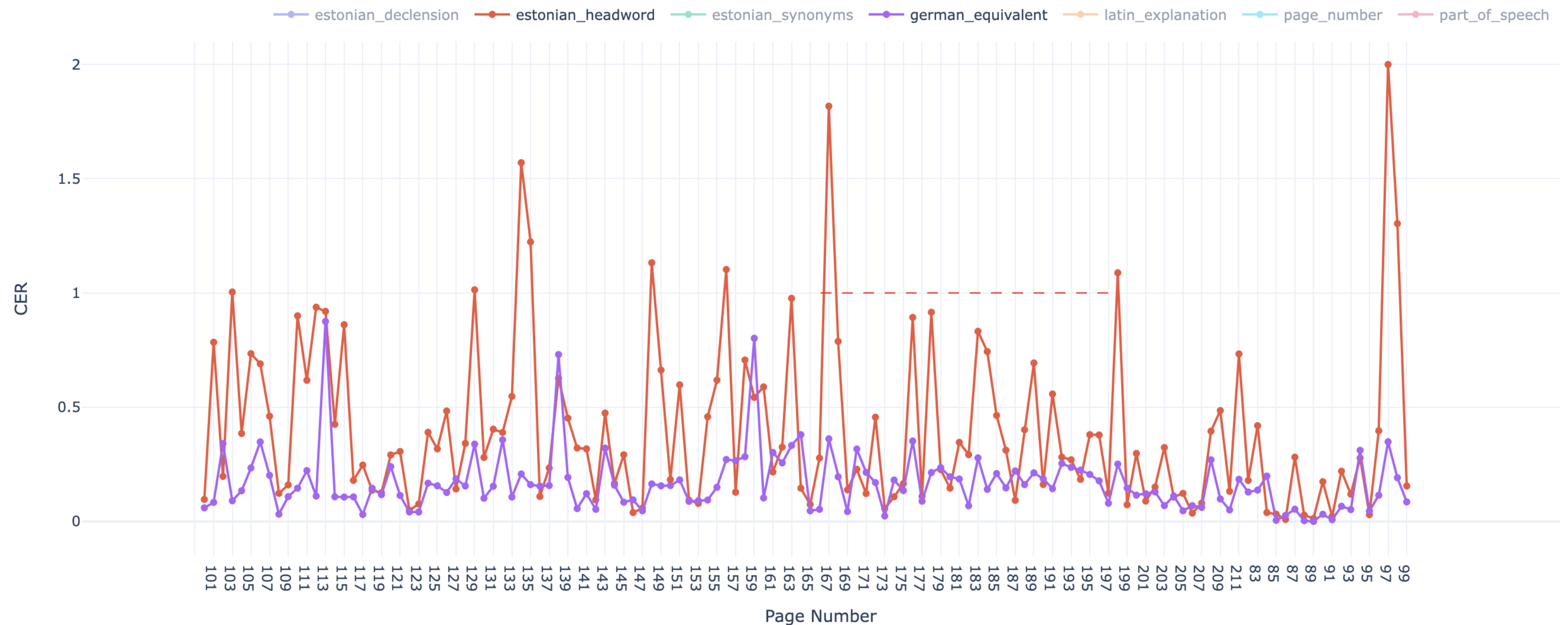
Claude: CER *eesti vaste*

Character Error Rate (CER) by Field and Page



Claude: CER *saksa* vs *eesti vaste*

Character Error Rate (CER) by Field and Page



Transkribus: zero-shot

Aad - adr ❀ (83) ❀ aed - ägg

A
Adrit lakskma zur Ader
lassen.
aas die Schlinge, 14, 3.
aas die Wiese am Bach gele-
gen, 9, 3.
aasta das Jahr.
aastane jährlich.
abba-lu das Schulterblatt.
abbi die Hülffe, 6, 1.
appi hüüdma zu Hülffe
ruffen.
abbiello die Ehe, 11, 5.
abbiello ärrarikkuma Ehe-
brechen.
abbiello labhutaminne die
Ehescheidung.
abbiello-rahwas Eheleute
14, 2.
abbiello-riika der Ehebrecher
abbiellorikminne der Ehe-
bruch.

aed der Garten, Zaun, 5, 1.
Acc. o.
aeg die Zeit, 5, 1. acc. o.
kewwadene aeg die Früh-
lings-Zeit.
temmal on üks aegikka.
ja der hat immer Zeit
gnug.
et anna aega habe Geduld.
aega möda mit der Zeit.
ilma aego umsonst.
pitke aego] mit der Län-
aiale pitkale] ge der Zeit.
ta jäi kül übbe ku aego er
blieb einen Monath lang
Fui ello aeg on läst uns
Gott leben.
aegfaste bey Zeiten, adv.
aelema kriechen wie eine
Schlange.
aer das Ruderholz, 11, 5.
aewas das Wiesen-

2 aed - ägg
3 10
4 Aad adr
5 (83.
6 Jaed der Garten, Zaun. 5, 1.
7 21.
8 Acc. o.
9 Adrit lakskma zur Ader
10 aeg die Zeit, 5, 1. acc. o.
11 lassen.
12 kewwadene aeg die Früh-
13 aas die Schlinge, 14, 3.
14 lings=Zeit.
15 aas die Wiese am Bach gele-
16 temmal on uks aegikka.
17 gen, 9, 3.
18 ja der hat immer Zeit
19 aasta das Jahr.
20 gnug.
21 aastane jährlich.
22 et anna aega habe Geduld.
23 abba-lu das Schulterblatt.

Mudel: The
Text Titan I
(*super model*)

Claude 3.7 Sonnet: *zero-shot*

Adorit	laßt ma zur Ader lassen					
aas	die Schlinge	14,3				
aas	die Wiese am Bach gelegen	9,3				
aasta	das Jahr					
aastane	jährig					
abba-lu	das Schulterblatt					
abbi	die Hülffe	6,1				
appi hüüdma	zu Hülffe rufen					
abbiello	die Ehe	11,5				
abbiello ärrarikkuma	Ehebrechen					
abbiello lahbutaminne	die Ehescheidung					
abbiello-rahwas	Eheleute	14,2				
abbiello-rikja	der Ehebrecher					
abbiellorikminne	der Ehebruch					
abbikasa	die Gehülfin	1,1				
abbilinne	behülflich					
abbimees	der Helffer					
abbi-waimud	Hülffs-Arbeiter					
ablas	begierig	14,2				
adder	der Pflug	13,2		ta prugib sedda adra	Er braucht den Pflug, das ist: dies ist so sein Amt	
adra-ma	ein Hacken Landes	1,3				
adra-mees	ein Hackens-Bauer					
adrikid	Vögel denen Rebhünern ähnlich					

API-päring,
Anthonius, *not*
thinking režiim

Kokkuvõte

- Põhiline eelis: aja ja raha kokkuhoid toimetatava faili saamiseks.
- Prompti täpsustamine: sünod, väljendid, ...
- Edasised katsed: väljundi formaat (nt yaml vs json?);
Transkribus (vmt) + LLM; eelmiste päringute arvestamine (...);
mudeli *fine-tune*'ing.
- Hupel 1780.

Helle, Anton Thor 1732. Kurtzgefaszte Anweisung Zur Ehstnischen Sprache. Halle: Stephan Orban. <http://www.digar.ee/id/nlib-digar:100071>

Jakobson, Joonatan; Jürviste, Madis 2025. Veebirakenduse Anthonius repositoorium. <https://github.com/joonatanjak/anthonius>

Jürviste, Madis; Paet, Tiina; Soosaar, Sven-Erik 2025. Eesti vanade sõnakujude tuvastamise võimalustest suurte keelemudelite abil. – Eesti Rakenduslingvistika Ühingu aastaraamat 21.

AITÄH