



Sentence Segmentation of Early New High German via Standard German Translation: Perplexity, Burstiness and Linguistic Analysis

HERZOG
AUGUST
BIBLIOTHEK

Fritz Kliche¹, Meryem Karalioglu², Kevin Röbler²

¹Akademie der Wissenschaften zu Göttingen, ²Herzog August Bibliothek Wolfenbüttel
fritz.kliche@adwgoe.de, karalioglu@hab.de, kroessler@hab.de

1. Sentence Segmentation

- Data: **16th-century texts from the Karlstadt edition** [1], written in Early New High German (ENHG)
- Their punctuation differs from modern sentence boundaries: virgules “/”, the periodus, capitalization at the beginning of subordinate clauses, and very long sentences characteristic of the writing style of the period
- The sentence boundaries are often optional: depending on personal decisions, a boundary can plausibly be assumed or rejected

Gold Standard

- 2,575 aligned transcript lines (about 21k tokens), 19 documents
- Two annotators + adjudication by a third annotator
- Labels: layout boundary, sentence boundary, potential sentence boundary

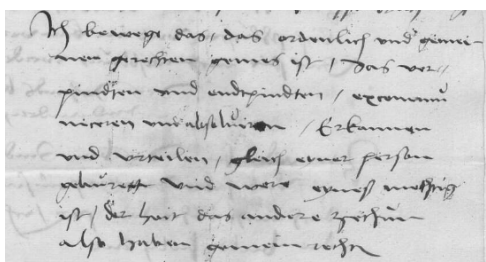
Research Questions

- Does a **translation into Standard German** improve sentence segmentation?
- How do prompts and human annotations differ in **perplexity, burstiness, and linguistic error patterns**?

2. Method and Results: Direct vs. Translation-Based Segmentation

- Prompting-based insertion of sentence boundary tags
- Run 1: Prompt (1) segments the ENHG text directly; prompt (2) translates the text into Standard German and then segments the translation
- Run 2: Prompts (3) and (4) are a second pass over the outputs.
- We use GPT 5.5 Thinking via ChatGPT
- We evaluate against all three labels, so including all optional boundaries
- **Translation maximizes precision**; the second-pass prompts maximize recall

Ich bewege das/ das ordenlich und gemeinen
gerechten gemes ist/ das verpindten
und endtpindten/ excommuniceren
und absolviren/ Erkennen
und urteilen/ gleich eyner person
geburett und wer eyneß mechtig
ist/ der hait das andere zethun
also haben gemein rechten.



Direct segmenta- tion in ENHG

und urteilen/ gleich eyner
person
geburett [GPT-SATZ?] und
wer eyneß mechtig
ist/ der hait das andere
zethun
also haben gemein rechten.
[SATZ]

Translate + segment in Standard German

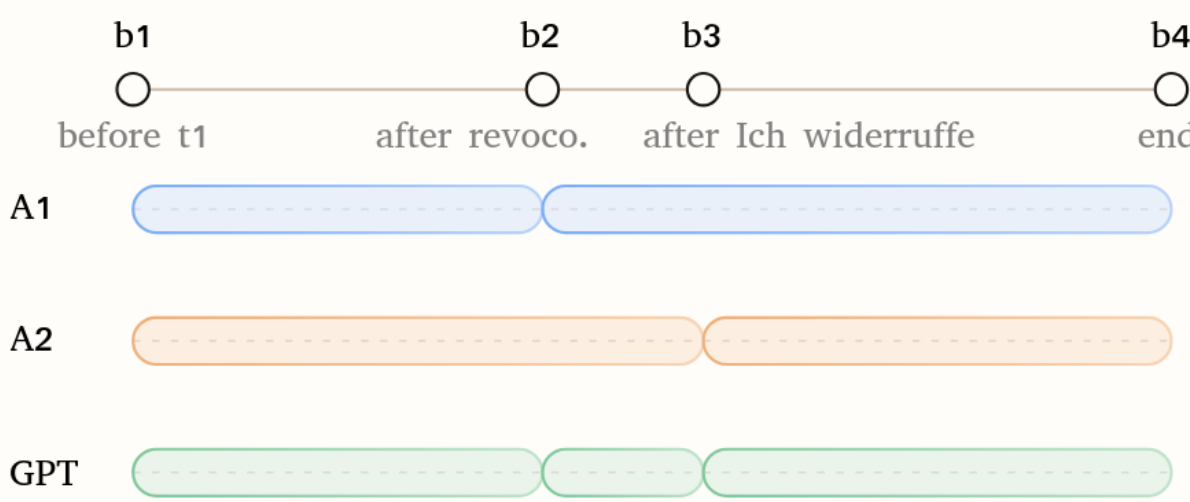
und Urteilen einer Person
gleich gebühren, und wer eines
mächtig
ist, der hat das andere zu tun,
also haben gemeine Rechte.
[SATZ]

Prompt	Precision	Recall	F1
(1) Direct insertion of sentence boundary tags in ENHG	0.828	0.887	0.857
(2) Translation into Standard German; then tag insertion	0.966	0.877	0.919
(3) Improve ENHG	0.831	0.941	0.883
(4) Improve translation	0.839	0.945	0.889

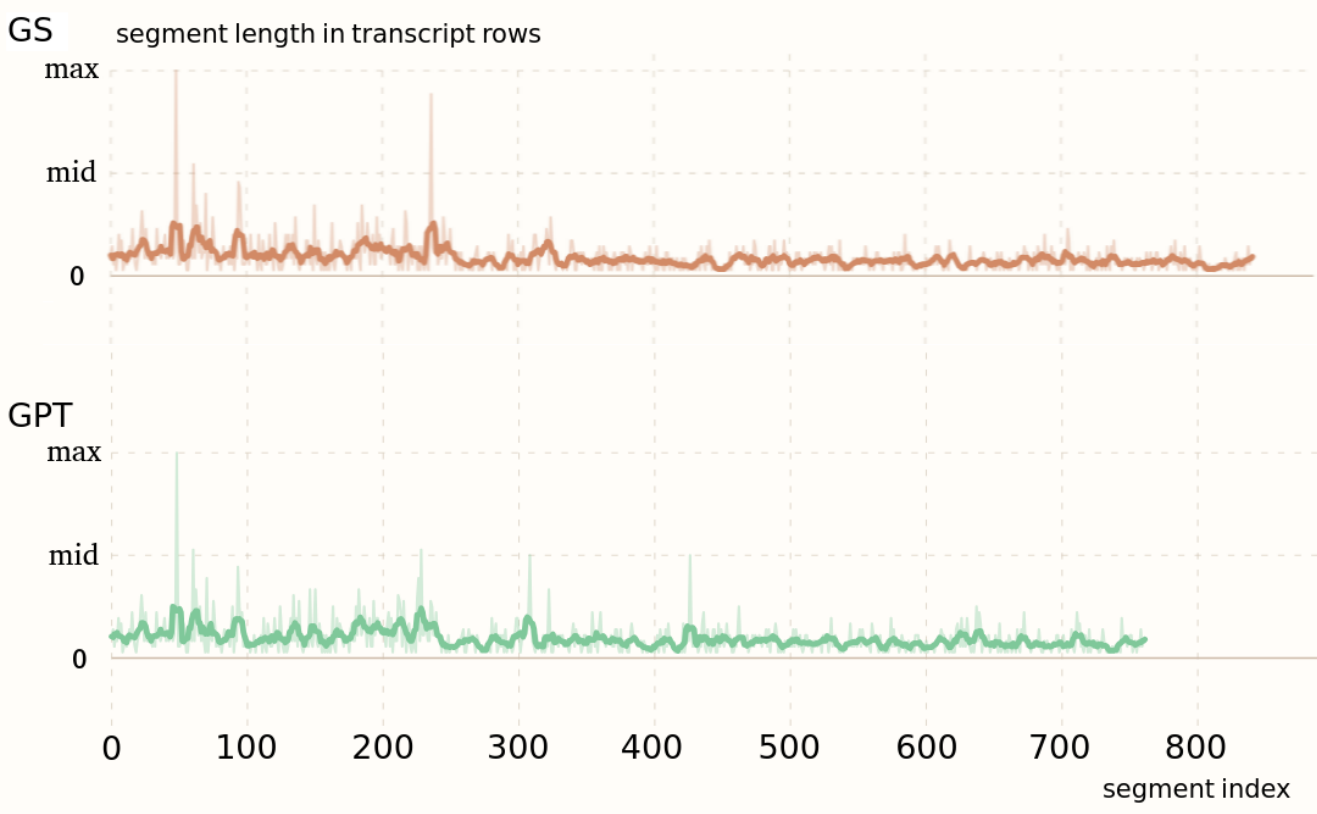
3. Boundary Rhythm: Perplexity and Burstiness

- Perplexity and burstiness are metrics which have been used for detecting AI texts
- We use them to compare the sentence lengths of the human and AI segmentations (We measure the sentence length in terms of lines in the edition between two sentence boundary markers)
- Gold standard segments are denser/finer-grained; prompt 2 often keeps larger units together
- All four prompts have lower burstiness than the human annotations. Prompt 2 combines **high perplexity** with **low burstiness**

KGK transcript 072, letter from Martin Luther to Andreas Karlstadt
"Das weis ich/ das ich der allerangenemst und liebste were/ wenn ich dis
einig wort spreche/ revoco. Das ist/ Ich widerruffe. Aber ich wil nicht zu
einem Ketzer werden [...]",
"I know/ that I would be the most acceptable and beloved one/ if I
uttered this one word/ revoco. That is/ I revoke. But I do not want to
become a heretic [...]"



Segment Length Curve
GS [sx:] vs. GPT [GPT-SATZ]/[GPT-SATZ?]/[SATZ], prompt 2 "insert in translation"



Annotation	Perplexity	Burstiness	Labels
Prompt 2: Translation into Standard German	7.284	-0.118	786
Other AI prompts (range)	6.105–6.492	-0.102–(-0.098)	926–979
Humans (range)	6.468–6.887	-0.082–(-0.062)	822–865

4. Linguistic Analysis

- False GPT annotations are often connected with **ENHG-specific ambiguity**: conjunctions such as *und*, *das/dass*, capitalization, and the historical *periodus*
- Several problems are **reduced by translation**. GPT also adapts less clearly than humans to a shift toward more paratactic prose

5. Conclusions

- **Pivot translation improves precision**; a second pass creates **high-recall candidates** for later refinement
- Specialized-domain evaluation should capture **granularity and coherent alternatives**, not only “right vs. wrong” boundaries

