Ansätze zur Wort- und Satzsegmentierung in kirchenslavischen HTR-Transkriptionen

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Anna Jouravel[1], Achim Rabus[1], Yves Scherrer[2], Elena Renje[1], Martin Meindl[1], Stefan Müller[3], Piroska Lendvai[3]
[1]Albert-Ludwigs-Universität Freiburg, Deutschland, [2]Universität Oslo, Norwegen, [3]Bayerische Akademie der Wissenschaften, Deutschland

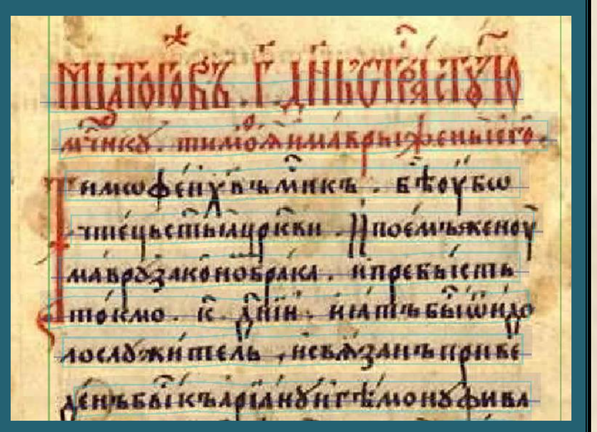
https://quantislav.badw.de/

1. Ausgangslage

- Neue Analysewege für die historisch arbeitenden Disziplinen dank zunehmender Verbreitung digitaler Tools (u.a. Camps et al. 2019, Franzini et al. 2018, Polomac 2014, Rabus 2019).
- Tools zugleich weiterhin fehleranfällig: **geringe Anzahl annotierter Ground Truth** (GT) Daten bei flexionsreichen, historischen Sprachstufen (etwa **Kirchenslavisch**), mit ausgeprägter orthographischer Variabilität und *scriptura continua*

2. Zielsetzung und Tools

- **Datierung und Lokalisierung** slavischer Schriftzeugnisse mit Hilfe digitaler Werkzeuge.
- Verarbeitung **unkorrigierter** Handwritten-Text-Recognition (HTR) Daten.



- HTR mit **Transkribus** erstellt.
- Fehlerhafte Wortsegmentierung als Problem für Analysetools
  → Verbesserung der **Wortsegmentierung**.
- Erste Attribuierungsversuche durch **domain adaptation und finetuning von BERT** (Devlin et al. 2019): kohärente Textteile lassen sich zuverlässiger zuordnen.
  → Verbesserung der **Satzsegmentierung** (Jouravel et al. 2024, Lendvai et al. 2023).

3. Wortsegmentierung

Training eines *Church Slavonic Word Separators* auf Basis eines multilingualen Text-to-Text-Transfer-Transformers (**mT5-Modell**) mit Byte-to-Byte-Erweiterung (**ByT5**) (Xue et al., 2020; Xue et al., 2022)

- Fügt Zeichenketten, die in *scriptura continua* vorliegen, Leerzeichen hinzu.
- Trainingsmaterial: PROIEL und TOROT Treebanks sowie eigene Transkriptionen

Ergebnisse Wortsegmentierung

- Validation loss: **0.008**
- CER: Verbesserung **zwischen 0,5% und 1,4%** (absolut)
- **Falschgenerierungen:** Textdopplungen, Einfügungen von Leerzeilen

4. Satzsegmentierung

- Zwei Tools: ***Stanza*** und ***UDPipe***
- Jeweils zwei Modelle: Altkirchenslavisch (**cu**) und Altostslavisch (**orv**)
- Zusätzlich: regelbasiertes **Python-Skript** zur groben Vorsegmentierung nach universellen und orthographischen Regeln
- Vergleich der **F1-Score-Metrik** auf drei Handschriftentexten unterschiedlicher Provenienz (11./14./16. Jh.; ost- und südslavisch): Tools, Skript, Kombination

Ergebnisse Satzsegmentierung

Text	⌀ Tokenanzahl pro Satz in GT	UDPipe		Stanza		Regeln	Regeln + Stanza		Regeln + UDPipe	
		proiel	torot	cu	orv		Regeln + cu	Regeln + orv	Regeln + proiel	Regeln + torot
(1) Aninas	17	32.2	23.3	**38.9**	28.9	29.8	32.8	29.6	30.7	25.3
(2) Kyrill	10	11.9	10.8	11.1	8.0	3.9	9.6	8.7	**12.6**	12.0
(3) Lepra	11	18.4	18.1	13.3	18.7	17.2	10.7	**20.1**	19.2	19.6

https://github.com/ufal/udpipe/blob/udpipe-2/udpipe2_eval.py

5. Diskussion und Zusammenfassung

- Trotz der signifikanten Verbesserungen bei der **Wortsegmentierung** weiterhin qualitative Fehler, insbesondere in Form von Halluzinationen.
  → Notwendigkeit weiterer Trainingsdurchläufe sowie qualitativer Analyse.
- Erzielte F1-Werte bei der **Satzsegmentierung** auf unbekannten Texten erwartungsgemäß niedriger.
  → Herausforderungen bei der Verarbeitung historischer Sprachstufen ohne formal klare Satzgrenzen.
- Verbesserung durch die Implementierung eines regelbasierten Skripts.
  → Hybride Ansätze vielversprechend.

6. Referenzen

Camps, Jean-Baptiste, Thibault Clérice und Ariane Pinche. 2019. „Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis." arXiv preprint, arXiv: 2012.03845.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North*, hg. von Jill Burstein, Christy Doran und Thamar Solorio, 4171–4186. Stroudsburg, PA, USA: Association for Computational Linguistics.
Franzini, Greta, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, Joanna Byszuk und Jan Rybicki. 2018. "Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm." Frontiers in Digital Humanities, 5.4: 1–15.
Jouravel, Anna, Elena Renje, Piroska Lendvai und Achim Rabus. 2024. "Assessing Automatic Sentence Segmentation in Medieval Slavic Texts." In *Proceedings of the Digital Humanities Conference*, Arlington, VA, USA, August 2024.
Lendvai, Piroska, Uwe Reichel, Anna Jouravel, Achim Rabus und Elena Renje. 2023. "Domain-Adapting BERT for Attributing Manuscript, Century and Region in Pre-Modern Slavic Texts." In *Proceedings of the 4th International Workshop on Computational Approaches to Historical Language Change 2023 (LChange'23)* co-located mit EMNLP2023, Singapur, Dezember 2023: 15–21.
Polomac, Vladimir. 2024. "Macarius a HTR Model for Romanian Slavonic Early Printed Books." *Slavistica Vilnensis* 68.2: 10–23.
Rabus, Achim. 2019. "Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus." *Scripta & E-Scripta* 19: 9–32.
Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua und Colin Raffel. 2020. „mT5: A massively multilingual pre-trained text-to-text transformer." *arXiv preprint arXiv:2010.11934*.
Xue, Linting, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts und Colin Raffel. 2022. „Byt5: Towards a token-free future with pre-trained byte-to-byte models." *Transactions of the Association for Computational Linguistics* 10: 291–306.