

Ansätze zur Wort- und Satzsegmentierung in kirchenslavischen HTR-Transkriptionen

Jouravel, Anna

anna.jouravel@slavistik.uni-freiburg.de
Universität Freiburg, Deutschland
ORCID: 0000-0002-7767-4973

Rabus, Achim

achim.rabus@slavistik.uni-freiburg.de
Universität Freiburg, Deutschland
ORCID: 0000-0002-5366-1430

Scherrer, Yves

yves.scherrer@ifi.uio.no
Universität Oslo, Norwegen

Renje, Elena

elena.renje@slavistik.uni-freiburg.de
Universität Freiburg, Deutschland

Meindl, Martin

martin.meindl@slavistik.uni-freiburg.de
Universität Freiburg, Deutschland

Müller, Stefan

stefan.mueller@badw.de
Bayerische Akademie der Wissenschaften, Deutschland

Lendvai, Piroska

piroska.lendvai@badw.de
Bayerische Akademie der Wissenschaften, Deutschland

Mit zunehmender Entwicklung von Sprachanalysetools eröffnen sich den historisch arbeitenden Disziplinen neue Forschungsperspektiven (u.a. Polomac, 2014; Franzini et al., 2018; Camps et al., 2019; Rabus, 2019). Als *low-resource language* steht das Kirchenslavische vor typischen Herausforderungen der Textverarbeitung wie der begrenzten Verfügbarkeit annotierter Ground-Truth-Daten (GT) sowie einer erhöhten Fehleranfälligkeit bei der automatischen Handschriftenerkennung (HTR) und Segmentierung (Rabus et al., 2023).

Um eine Grundlage für eine präzisere Datierung und Lokalisierung slavischer mittelalterlicher Schriftzeugnisse zu schaffen, zielt unsere Forschung daher darauf ab, die Ausgabe verschiedener Tools durch automatisierte Abläufe zu

verbessern. Im Fokus stehen dabei zwei Ziele: die Verbesserung der Wortsegmentierung nach dem Transkriptionsprozess und die Verfeinerung der Satzsegmentierung.

In unserem Arbeitsablauf erfolgt die Erstellung der HTR durch *Transkribus*¹. Das verwendete Transkriptionsmodell für kirchenslavische Handschriften, die in *scriptura continua* abgefasst sind, erreicht eine Character Error Rate (CER) von 3,7%. Die für das Training verwendeten Daten sind jedoch nicht in *scriptura continua*, sondern enthalten bereits von Editor*innen vorgenommene Wortsegmentierungen. Evaluierungen der HTR-Ergebnisse ergeben, dass die Wortsegmentierung eine erhebliche Fehlerquelle darstellt. Daher besteht unsere erste Zielsetzung in der nachträglichen Verbesserung der **Wortsegmentierung**.

Außerdem zeigen qualitative Auswertungen erster Attribuierungsversuche kirchenslavischer Handschriften und Drucke aus drei Zeitabschnitten (10.–11. Jh., 15.–16. Jh., 18. Jh.) und zwei Regionen (Süd- und Ostslavia) durch Domain-Adaptation und Finetuning von BERT (Devlin et al., 2019), dass syntaktisch und semantisch kohärente Textteile verlässlicher zugeordnet werden (Lendvai et al., 2023). Daher besteht unsere zweite Zielsetzung in der Verbesserung der **Satzsegmentierung**, die unabhängig von der Wortsegmentierung betrachtet werden muss.

Hinsichtlich der Wortsegmentierung haben wir 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) einen *Church Slavonic Word Separator* trainiert. Dieser fügt in Zeichenketten, die in *scriptura continua* vorliegen, Leerzeichen hinzu. Ziel ist es, den unkorrigierten HTR-Output im Nachgang zu verbessern. Für die Aufbereitung des Testsets wurden alle Leerzeichen im HTR-Text entfernt, damit der Text in *scriptura continua* vorliegt. Das Trainingsmaterial umfasst neben öffentlich zugänglichen Datensets wie PROIEL² und TOROT³ ostslavisch-kirchenslavische Handschriften aus dem 16. Jh., wobei das beste Modell einen *validation loss* von 0.008 erreicht. Es zeigte sich eine sichtbare Verbesserung der CER zwischen (absolut) 0,5% und 1,4%. Allerdings sind trotz der deutlichen Verbesserung der CER in einer qualitativen Analyse Falschgenerierungen des Separators wie Textdopplungen oder Einfügungen von Leerzeilen aufgefallen, die zu ungewollter Wortsegmentierung führen.

Für die Satzsegmentierung wurden zwei Tools erprobt: *Stranza* (version 1.6.1; Qi et al., 2020) und *UDPipe* (version 2.12; Straka, 2018). Beide verfügen über Modelle für das Altkirchenslavische⁴ sowie für das Altostslavische⁵. Daneben haben wir ein eigenes regelbasiertes Python-Skript erstellt, das die Satzsegmentierung nach universellen syntaktischen und orthographischen Regeln forciert. Zu diesen Regeln gehört etwa die Wackernagelposition von Partikeln, i.e. ihre typische Zweitstellung im Satz, oder der Einsatz von Initialen, die einen Satzbeginn markieren. Beide Tools, das regelbasierte Skript sowie die Abfolge von Skript und Tool wurden auf drei Handschriftentexten unterschiedlicher Provenienz getestet: (1) die *Vita des Aninas* (11. Jh., südslavisch)⁶; (2) die *Katechesen Kyrills von Jerusalem* (11. Jh.,

ostslavisch)⁷; (3) der Traktat *De lepra* des Methodius von Olympus (16. Jh., ostslavisch)⁸. Die Ergebnisse wurden in Tabelle 1 anhand von ca. 350 GT-Sätzen pro Text quantitativ evaluiert (qualitativ in Jouravel et al., 2024)⁹.

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	

Tabelle 1: Vergleich F1-Score-Metrik der Satzsegmentierung mit regelbasiertem Skript, mit Tools und mit ihrer Kombination.

Insgesamt fallen die Ergebnisse auf unbekanntem und (ortho)graphisch heterogenem Textmaterial eher niedrig aus, was jedoch für historische Sprachstufen nicht ungewöhnlich ist. Dennoch spricht der höhere F1-Wert, der bei der Kombination aus regelbasiertem Skript und einem Tool in zwei von drei Fällen erzielt wurde, dafür, einen regelbasierten Ansatz in den Arbeitsablauf zu integrieren. Zugleich zeigt die Auswertung sowohl der Universal-Dependencies-benchmark-GT von Text (1) als auch der von uns erstellten GT von Text (2) und (3), dass eine allgemeingültige Satzgrenzdefinition für historische Sprachstufen ohne (satzanzeigende) Interpunktion nicht trivial ist und einer Diskussion innerhalb der Fachgemeinschaft bedarf.

Fußnoten

1. <https://readcoop.eu/de/transkribus/>
2. https://universaldependencies.org/treebanks/cu_proiel/index.html
3. https://universaldependencies.org/treebanks/orv_torot/index.html
4. *Stanza*: language code ‘cu’; *UDPipe*: old_church_slavonic-proiel-ud-2.12-230717
5. *Stanza*: language code ‘orv’, mit Python-Package ‘torot’; *UDPipe*: old_east_slavic-torot-ud-2.12-230717
6. aus *Codex Suprasliensis*; Sign. BN BOZ 201, Cod. Kop. 2Q, perg. I. 72; <http://suprasliensis.obdurodon.org/> Teil des Testsets der Universal Dependencies benchmark. vgl. https://github.com/UniversalDependencies/UD_Old_Church_Slavonic-PROIEL#data-splits
7. GIM, Sin. 478, ed. Weiher, 2017.
8. GIM, Sin. 995, Fol. 310r–315r, ed. Jouravel et. al., 2024.
9. vgl. das Evaluationsskript https://github.com/ufal/udpipeline/blob/udpipeline-2/udpipeline2_eval.py

Bibliographie

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, Piroška 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.

Jouravel, Anna, Janina Sieber und Katharina Bracht, hg. 2024. *Methodius Von Olympus: De Lepra. Griechischer Und Slavischer Text, Mit Einleitung Und Deutscher Übersetzung*. Die griechischen christlichen Schriftsteller der ersten drei Jahrhunderte NF, 31. Berlin: De Gruyter.

Lendvai, Piroška, 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.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton und Christopher D. Manning. 2020.

“Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, hg. von Asli Celikyilmaz und Tsung-Hsien Wen, 101–108. Stroudsburg, PA, USA: Association for Computational Linguistics.

Rabus, Achim. 2019. “Recognizing Handwritten Text in Slavic Manuscripts: A Neural - Network Approach Using Transkribus.” *Scripta & E - Scripta* 19: 9–32.

Rabus, Achim, Arnold Eckhart, Anna Jouravel, Piroška Lendvai, Martin Meindl, Vladimir Polomac und Elena Renje. 2023. “Developing a Pipeline for Automatic Linguistic Analysis of Historical Manuscripts and Early Printings: The Pre-Modern Slavic Case.” In *Proceedings of the Digital Humanities Conference*, Graz, Österreich, Juli 2023: 112–113.

Straka, Milan. 2018. “UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task.” In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, hg. von Daniel Zeman und Jan

Hajič, 197–207. Stroudsburg, PA, USA: Association for Computational Linguistics.

Weiher, Eckhard , hg. 2017. *Die altbulgarische Übersetzung der Katechesen Kyrills von Jerusalem*. Monumenta linguae Slavicae dialecti veteris 64. Freiburg i.Br.: Weiher.

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.