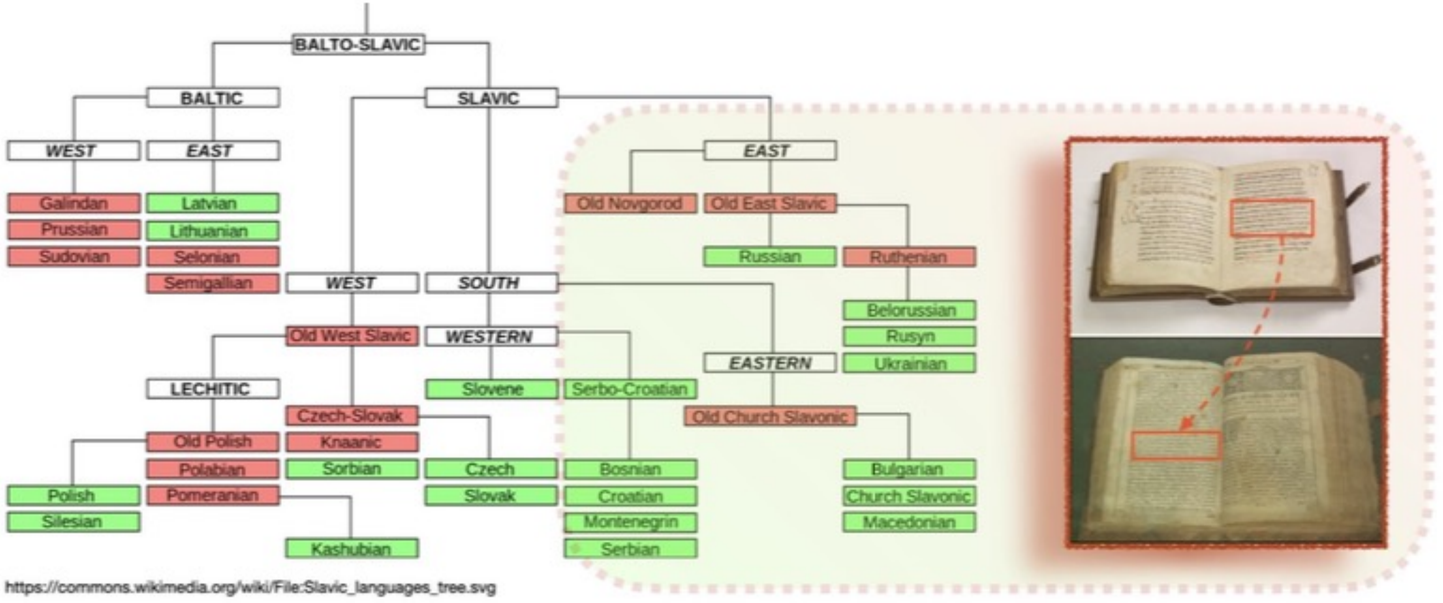
A Workflow for HTR-Postprocessing, Labeling and Classifying Diachronic and Regional Variation in Pre-Modern Slavic Texts

Piroska Lendvai[1], Maarten van Gompel[2], Anna Jouravel[3], Elena Renje[3], Uwe Reichel[4,5], Achim Rabus[3], Eckhart Arnold[1]

[1]Dept. of Digital Humanities, Bavarian Academy of Sciences, Munich, Germany
[2]Digital Infrastructure, Humanities Cluster, Royal Dutch Academy of Sciences, The Netherlands
[3]Dept. of Slavic Languages and Literatures, University of Freiburg, Germany
[4]audEERING GmbH, Germany
[5]Hungarian Research Centre for Linguistics, Budapest, Hungary
{piroska.lendvai, eckhart.arnold}@badw.de
proycon@anaproy.nl
{anna.jouravel, elena.renje, achim.rabus}@slavistik.uni-freiburg.de
ureichel@audeering.com

Historical NLP

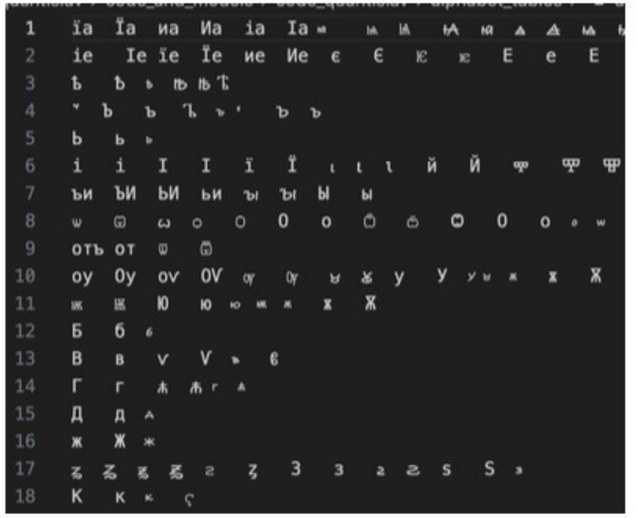
- Diachronic and variational historical linguistics, DH, NLP
- To track language use and language change
- Applied use case: chronological and geographic attribution of texts written in **Church Slavic** (its East and South Slavic recensions)

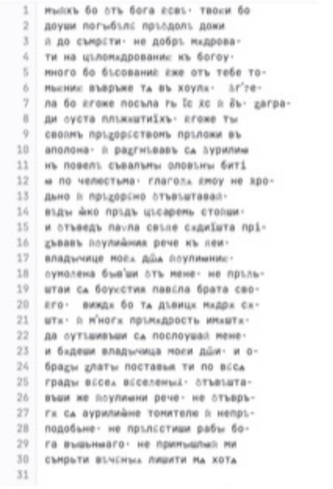


https://commons.wikimedia.org/wiki/File:Slavic_languages_tree.svg

Challenges

- Copying manuscripts in different geographic and cultural settings led to variation
- Exhibits on several linguistic levels: Grapheme variants, Morphological complexity, Dialectal variation
- Absence of gold boundary markers for words, sentences
- Presence of recognition errors (false word segments or characters)
- VERY few human domain experts for correction and labeling
- Little or no provenance labeled data

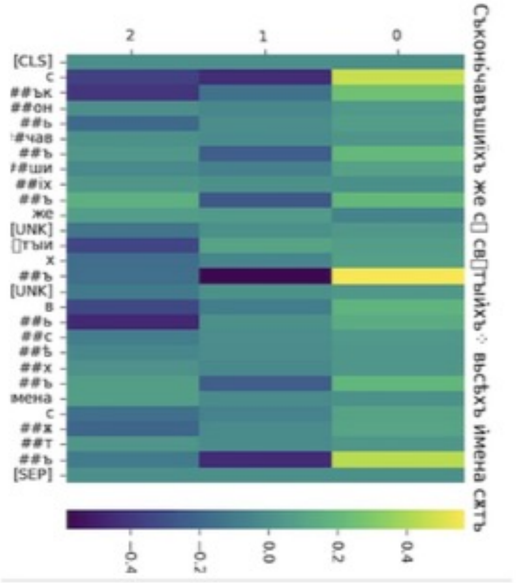


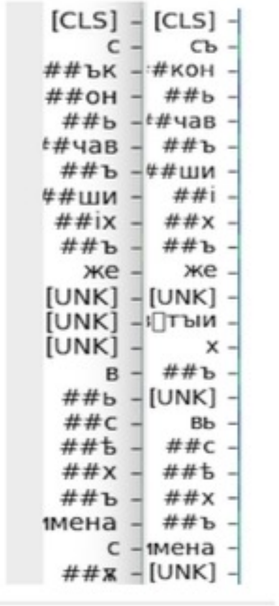


- With artefacts: word segmentation (whitespace, hyphenation)
- Sentence segmentation problematic (stanza, UDPipe) and not optimized per text temporality

Overarching research questions

- Can NLP perform downstream tasks and explain variation patterns in Church Slavic, across time and space?
- Can data-driven approaches identify or learn expert knowledge, and correct or be robust against HTR errors?





1. Workflow Start: After HTR

- Data acquisition: Manuscripts scans –> Handwritten Text recognition (HTR)
- NLP tools can be used (a) to diagnose HTR output as a feedback for recognition engines and (b) to utilize data-driven resources for correcting HTR errors
- Ecosystem: FoLiA XML `proycon.github.io/folia`
  - Data model and file format for linguistic annotation
  - Tooling: Converters, Tokenizer, Annotation tool, NLP tools, ...

2.1 Line-based alignment of GT and HTR

- For data diagnostics and cleaning: parallelized texts of ground truth (GT) and HTR. Currently we align texts on the manuscript line level
- Tool: **sesdiff** `github.com/proycon/sesdiff`
  - Alignment of GT-HTR in terms of Levenshtein + Character-level (abstract) edit notation
  - Enables search for editing patterns, e.g. false splits in HTR
- Tool: **TextAlign** `clarin.phonetik.uni-muenchen.de/BASWebServices`
  - Edit cost function is learned from smoothed conditional character co-occurrence probabilities
  - We can assess HTR misrecognitions and variation, motivated both on grapheme-level or phonetically

2.2 HTR correction and variation analysis

Tool: **analiticcl**

- Approximate string matching or fuzzy-matching system that can be used for spelling correction or text normalization
- Texts can be checked against a validated or corpus-derived lexicon
- Diagnostics: retrieve spelling variants
- Correction: query false splits and generate suggestions

```
авьствьно
навьствьно 1.0 ['vkse2cyrilcathetical']
навьствьнѣ 0.89 ['vkse2cyrilcathetical']
авьствьнѣ 0.89 ['vkse2cyrilcathetical']
навьствено 0.77 ['vkse2apostolosscribe1']
навьственѣ 0.72 ['vkse2apostolosscribe3']
на҃вьств□нѡ̀ 0.56 ['dionisio2']
на҃вьстьвнь 0.51 ['dionisio2']
на҃вьстьвно 0.51 ['dionisio2']
```

```
мнѡ жаѥ <--
мнѡ́жаѐ 1.0 ['dionisio2']
мно́жае 0.86 ['dionisio2', 'vkse2elizabethbible']
мнѡ́жаѥ 0.86 ['dionisio2']
мнѡ́жаѣ 0.86 ['dionisio2']
мно́жаѥ 0.86 ['dionisio2']
===

и номоу <--
ѝномоу 1.0 ['vkse2suprasliensis']
и́номоу 1.0 ['vkse2cyrilcathetical']
ѥдино́моу 0.89 ['vkse2cyrilcathetical']
ѝномоу 0.86 ['vkse2suprasliensis']
о́номоу 0.86 ['vkse2cyrilcathetical']
```

Ground truth text bodies

Manuscript	Century	Region	Place of Copying	Language	Main genre	Tokens	Unique tokens	Text snippets
Codex Suprasliensis	10-11	South	Eastern Old Bulgaria	Old Church Slavic; South Slavic recension	hagiographical-homiletic	65,207	18,450	4,831
Cyril of Jerusalem's Cathechetical Lectures	11-12	East	Kyivan Rus'	Old Church Slavic; South Slavic recension; Transmitted version used: East Slavic recension	dogmatic	62,011	20,936	4,282
Dionisio corpus (printed)	15-16	South	Serbia, Macedonia	Serbian Church Slavic; South Slavic recension	liturgical	142,402	42,828	10,685
Apostolos (from the Uspensky version of the Great Menaion Reader)	16	East	Muscovy	Russian Church Slavic; East Slavic recension	gospel	230,660	50,302	14,058
Sluzhabnik	18	South	Serbia	Serbian Church Slavic; South Slavic recension	liturgical	56,785	13,197	3,350
Elizabeth Bible (printed)	18	East	Muscovy	Russian Church Slavic; East Slavic recension	Bible translation	204,322	21,335	11,796
Methodius of Olympus: De lepra ad Sistelium	16	East	Kyivan Rus'	Old Church Slavic; Transmitted version: East Slavic recension	exegetic treatise	3,743	2002	259

3. Text attribution on snippet level

BERT: Domain-adapted and finetuned

- Vocabulary extension with union of the 100 most frequent words of each manuscript to the tokenizers' vocabularies
- Masked language modeling with standard *BertForMaskedLM* head

Task	Model	From-Pretrained	From-Adapted
manuscript	KoichiYasuoka/bert-base-slavic-cyrillic-upos	0.922 (0.004)	0.941 (0.003)
manuscript	anon-submission/mk-bert-base-macedonian-bulgarian-cased	0.935 (0.002)	0.961 (0.001)
manuscript	bert-base-multilingual-uncased	0.945 (0.002)	**0.962** (0.003)
century	KoichiYasuoka/bert-base-slavic-cyrillic-upos	0.952 (0.002)	0.965 (0.001)
century	anon-submission/mk-bert-base-macedonian-bulgarian-cased	0.961 (0.001)	**0.977** (0.002)
century	bert-base-multilingual-uncased	0.959 (0.001)	0.976 (0.001)
region	KoichiYasuoka/bert-base-slavic-cyrillic-upos	0.96 (0.002)	0.976 (0.001)
region	anon-submission/mk-bert-base-macedonian-bulgarian-cased	0.968 (0.001)	0.984 (0.001)
region	bert-base-multilingual-uncased	0.979 (0.002)	**0.986** (0.001)

Table 2: Performance scores on the three downstream tasks on directly finetuned models (*From-Pretrained*) that we regard as baseline vs. domain-adapted and subsequently finetuned models (*From-Adapted*), in terms of Unweighted Average F-score arithmetic mean values and standard deviations (in brackets) obtained from five random seeds.

4. Labeling below text snippet level

- To make classification finer-grained: Annotation on token level
- Tool: **FLAT** annotation environment: Interfaces to FoLiA XML, CONLL-U, analiticcl corrections, text versions, ...

QuantiSlav Project Acknowledgments

Bavarian Academy of Sciences, Munich & University of Freiburg, Germany

Bundesministerium
für Bildung
und Forschung

Finanziert von der
Europäischen Union
NextGenerationEU