Collating
diplomatic transcriptions of manuscripts

David J. Birnbaum
Open philology workshop
Leipzig, 2014-07-14
djbpitt@gmail.com
http://www.obdurodon.org
http://pvl.obdurodon.org

The Rus’ primary chronicle

- PVL (Повесть временных лет)
- Historical chronicle of the East Slavs
  - Russia, Ukraine, Belarus
- Beginning from the creation (Hamartolos)
- Historical period (since 852) arranged by year
- Repeated edits and recompositions
- Fixed 1116
  - Incorporated into later chronicles

PVL textual tradition

- α Archetype
- β Northern branch
  - δ Laurentian (Lav, 1377) [Trinity (Tro; lost)]
  - ε Radziwll (Rad, 1490s), Academy (Aka, late 15th)
- γ Southern branch
  - Hypatian (Ipa, ca. 1425), Xlebnikov (Xle, 16th)
  - [Pogodin (Pog, early 17th, to supplement Xle)]
- Novgorod first
  - Commission (Kom), Academy (NAk), Tolstoj (Tol)

PVL stemma

Why collate the PVL?

- Textual comparison
  - Relationships among the copies
  - Construction of a paradosis (alpha text)
  - History of transmission beyond alpha
- Why collate diplomatic transcriptions?
  - Linguistic comparison
  - Orthographic comparison
  - [More about diplomatic editions of manuscripts and critical editions of texts]

Practical issues

- No funding
  - Must reduce human effort
    - Especially repetitive human effort
- Edition is under constant development
  - Must be able to rerun collation
- Must be automated as much as possible
Print edition

- *The Povest’ vremennykh lět: An interlinear collation and paradosis*
- Donald Ostrowski, David J. Birnbaum, Horace G. Lunt
- Harvard UP, 2004
- 3 vv., 2368 pp.

Interlinear collation (print)

14:

Laur: передъя, катърь, тейкъ и до индикъя и домготъ
Tria: передъя ватъ, даже и до индикъя в домготъ
Ruth: передъя, катърь, даже и до индикъя, к домготъ
Acad: передъя, катърь, даже и до индикъя, к домготъ
Ипса: передъя, катърь, даже и до индикъя, к домготъ
Kłe: передъя, катърь, даже и до индикъя, к домготъ
Byck: Передъя, Ватъръ, даже и до Индикъя и Домготъ,
Shakl: Передъя, Ватъръ, даже и до Индикъя в Домготъ,
Lilj: Передъя, Ватъръ, даже и до Индикъя в Домготъ,
α: Передъя, Ватъръ, даже и до Индикъя в Домготъ,

Why interlinear?

- General
  - Variants are presented completely, not selectively
  - Ease of reading any individual copy
- Digital
  - Space, weight, cost are irrelevant
  - User can select witnesses
  - Searching on other than plain text
    - Lemma
    - Morphology

How interlinear?

- Alignment by line (per Karskii 1926 edition)
  - Ludolf Müller, *Handbuch zur Nestorchronik* (word index 1977)
  - Samuel Hazzard Cross, *The Russian primary chronicle Laurentian text* (English translation, 1930)
    - Performed manually for print edition
- Alignment by word
  - Too expensive to perform manually

Print edition workflow

- Typeset in troff
- Focus on producing print version
- Alignment is manual
  - Word-level alignment is impractical

Digital versions

- PDF of print edition
  http://hudce7.harvard.edu/~ostrowski/pvl/
- HTML edition
  http://pvl.obdurodon.org
Why is collation difficult?

- Exponential complexity
  - Worst case: compare every word in every witness to every word in every other witness
  - Complicated by repetitions and transpositions
- Diplomatic transcription
  - Efficient comparison algorithms require exact string matching, which is rare in diplomatic transcription
  - Finding closest match requires a completely different (more computationally expensive) method than finding exact match

Word-aligned version

A more challenging passage

CollateX

- http://collatex.net/
- Interedition (Huygens Institute, the Hague)
- Advantage
  - Use someone else’s collation algorithm and implementation
- Limitation
  - Requires exact string matching
  - Cannot find closest match
  - Cannot find logical matches that are not string matches
  - Digits vs words: 40000 ~ 40 тысяч
  - Synonymy: разумьн — съмысльн
Recent CollateX developments

- Ported to Python (module)
- New collation algorithm
  - Non-progressive
  - Suffix arrays and LCP arrays
  - Not subject to order effects
- Remaining limitations
  - Exact string matching
  - Repetition
  - Transposition

Wordflow (summary)

- Input is custom XML
- Convert to TEI, tokenize, add <w> tags
- Convert to JSON
- Preprocess: Enrich JSON with bespoke normalization
- Collate with CollateX, which creates variant graph
- Postprocess: Adjust rank in variant graph
- Generate JSON output
- Convert to custom XML
- Convert to HTML tables for rendering

Variant graph (excerpt)

Workflow (beginning)

- Input: Custom XML line blocks
- Convert to TEI with word (<w>) tags
  - Tokenizing mixed content
    - word1 <lb/>rd2 word3
    - <w>word1</w>
    - <w>wo</w> <lb/>rd2</w>
    - word2
  - Not all whitespace represents token break
    - word1 <lb/> word2
    - word3</w>
    - Convert to JSON, enrich with normalizations

Normalization

- Create a normalized “shadow” copy
- Normalization based on Soundex, adapted for early Cyrillic writing
- Collate on normalization, return original

Soundex

- English-language surnames, 1918
- Algorithm (simplified)
  - Retain first letter
  - Delete other vowels; degenerate
  - Conflate other letters according to phonetic similarity (e.g., t/d = 3; m/n = 5)
  - Truncate or zero-pad to four characters
- Examples
  - Birnbaum B-651 (also ✓ Barenboim; also ✗ Brumble)
Soundex assumptions

- Character differences are not all equivalent with respect to information load
- Information load may be sensitive to position
- Beginning of word carries more information than end
  - Especially inflected languages
- Consonants carry more information than vowels
  - Except in short words

Adapting Soundex to Church Slavonic

- Neutralize variant spellings of initial vowel
  - ay, y, ă ~ ā
  - w, u, o ~ ă
- Case fold, neutralize consonantal variants
  - Not always one-to-one, e.g., w ~ ă
- Degenerate, delete other vowels, delete diacritics
  - Keep two letters of two-letter words
  - Higher information load
- Other conflations?
  - Knowledge based vs machine learning
  - Expand abbreviations?
  - Truncate or zero-pad (to what length?)

Soundex sample (Bdinski sbornik)

- Ch397 и вьзвра тьдьщерьше своѥе
- Ch384 и вьзврати тьдьщершоу свою
- Nbkm298 и вьзвратити братанитцꙋ свꙋ
- Berlin и вьзвр братаницꙋ свꙋ
- Ch397 и вьзвр дышт св
- Ch384 и вьзвр дышт св
- Nbkm298 и вьзвр дышт св
- Berlin и вьзвр дышт св

Two types of normalization

- Collation
  - Find alignment points
  - Coarse adjustments
  - No harm in conflating grammatical forms
    - Imperfect and aorist; infinitive and supine
- Evaluation
  - Alignment points are already known
  - Finer comparisons
  - Many need to distinguish on the basis of small details

Collation after Soundex

- Greatly improved actual matches
- Forced matches
  - A B C
  - A D C
- Misses
  - Gap in alignment (no forced match)
  - Imperfect match
    - фра ~ фрац (фр ~ фрц)
  - CollateX recognizes only perfect matches
- Unable to recognize closest match

3.5

<table>
<thead>
<tr>
<th>3.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lav</td>
</tr>
<tr>
<td>Tro</td>
</tr>
<tr>
<td>Rad</td>
</tr>
<tr>
<td>Aka</td>
</tr>
<tr>
<td>Ipa</td>
</tr>
<tr>
<td>Xle</td>
</tr>
<tr>
<td>By İz</td>
</tr>
<tr>
<td>Šec</td>
</tr>
<tr>
<td>Lic</td>
</tr>
<tr>
<td>α</td>
</tr>
</tbody>
</table>
Numbers

<table>
<thead>
<tr>
<th>18.4</th>
<th>18.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/15/14</td>
<td>6</td>
</tr>
<tr>
<td>Numbers</td>
<td>Problem areas</td>
</tr>
<tr>
<td>• Gaps in alignment</td>
<td>• CollateX follows graph rank (leftmost match)</td>
</tr>
<tr>
<td>• No perfect match</td>
<td>• 3,5</td>
</tr>
<tr>
<td>• CollateX follows graph rank (leftmost match)</td>
<td>– Orthography</td>
</tr>
<tr>
<td>• 3,5</td>
<td></td>
</tr>
<tr>
<td>– Orthography</td>
<td>– Orthography</td>
</tr>
<tr>
<td>– Soundex</td>
<td>– Soundex</td>
</tr>
<tr>
<td>• Postprocessing</td>
<td>• In case of ties</td>
</tr>
<tr>
<td>• Thesaurus</td>
<td>• Thesaurus</td>
</tr>
<tr>
<td>• Most matches</td>
<td>• Most matches</td>
</tr>
<tr>
<td>• Length of match</td>
<td>• Length of match</td>
</tr>
</tbody>
</table>

Postprocessing

• Gap without perfect match on either side
  – Gap may span multiple columns
• Orthography
  – Soundex
  – Soundex
• Soundex
  – Soundex
  – Soundex
• If there’s a match, keep it
• Else
  – Find unique Soundex values in column and following
  – Move token to column with closest match

In case of ties

• Thesaurus
• Most matches
• Length of match

9,2

9,2

Thesaurus

• Collect forced inexact matches
  – A B C
  – A D C
• Edit manually
• Use to break ties
• Close matches
  – Close matches
• Non-matches
  – Non-matches
Acknowledgements

• Thanks to Ronald Dekker, lead developer of CollateX, for generous advice and consultation
• Thanks to Minas Abovyan, who implemented the PVL-specific Python code in our project

Thank you!

David J. Birnbaum
Open philology workshop
Leipzig, 2014-07-14

djbpitt@gmail.com
http://www.obdurodon.org