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

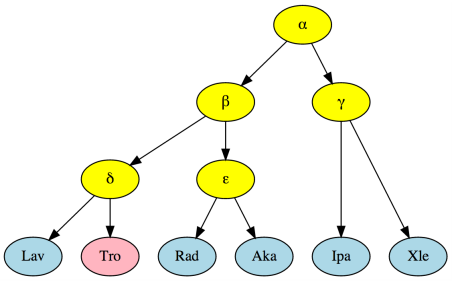
The *Rus' primary chronicle*

- *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 recompilations
- Fixed 1116
  - Incorporated into later chronicles

PVL textual tradition

- α Archetype
- β Northern branch
  - δ Laurentian (Lav, 1377) [Trinity (Tro; lost)]
  - ε Radziwiłł (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



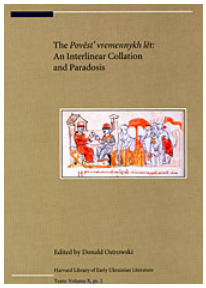
Why collate the PVL?

- 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)

1,4:

Laur: перснда. ватрь. тоже | н до ниднкниа в долготу
Trin: персида ватрь даже и до индикия в долготу
Radz: перснда. ватрь. доже и до ни҃кна. в до|лготоу
Acad: персїда. ватръ. до҃н н до ниднкїа. в долготѳ |
Hypa: перы|снда. ватрь. доже н до нны|днкниа. в долготу
Khle: перснда. ватръ. даже н до ниднкїа. въ | долготоу

Bych: Персида, Ватрь, доже и до Индикия в долготу,
Shakh: Персида, Ватрь доже и до Индикия въ дълготу,
Likh: Персида, Ватрь, доже и до Индикия в долготу,
α : Персида, Ватрь доже и до Индикия въ дълготу,

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

First digital version

1, 4

Lav	персида. ватрь. тоже | и до индикиꙗ в долготу
Tro	персида ватрь даже и до индикия в долготу
Rad	персида. ватрь. доже и до инⷣикиа. в до|лготоу
Aka	персїда. ватръ. доⷤи и до индикїа. в долготоу |
Ipa	перь|сида. ватрь. доже и до инь|дикиꙗ. в долготу
Xle	персида. ватръ. даже и до индикїа. въ | долготоу
Byč	Перснда, Ватрь, доже н до Ннднкня в долготу,
Šax	Перснда, Ватрь доже н до Ннднкня въ дълготу,
Lix	Перснда, Ватрь, доже н до Ннднкня в долготу,
α	Перснда, Ватрь доже н до Ннднкня въ дълготу,

First digital version

- Pro
  - Automated conversion from troff
  - Control over display
    - Fonts
    - Toggle individual witnesses on and off
  - Potential for annotation (lemma, morphology)
- Con
  - No support for word-level comparison

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

1,4

Lav	персида.	ватрь.	тоже |	и	до	индикиꙗ	в	долготу
Tro	персида	ватрь	даже	и	до	индикия	в	долготу
Rad	персида.	ватрь.	доже	и	до	инⷣикиа.	в	до|лготоу
Aka	персїда.	ватръ.	доⷤи	и	до	индикїа.	в	долготоу |
Ipa	перь|сида.	ватрь.	доже	и	до	инь|дикиꙗ.	в	долготу
Xle	персида.	ватръ.	даже	и	до	индикїа.	въ |	долготоу
Byč	Перснда,	Ватрь,	доже	н	до	Ннднкня	в	долготу,
Šax	Перснда,	Ватрь	доже	н	до	Ннднкня	въ	дълготу,
Lix	Перснда,	Ватрь,	доже	н	до	Ннднкня	в	долготу,
α	Перснда,	Ватрь	доже	н	до	Ннднкня	въ	дълготу,

A more challenging passage

9,2

Lav	и	то	творать		мовенье	собѣ	а	не	мученье.
Tro	и	то	творять		мовенье		а	не	мученье
Rad	и	тако	творать	не	мытвоу	собѣ		но	мⷹ |чение.
Aka	и	тако	творатъ	не	мытвоу	собѣ		но	моученье. |
Ipa	и		твораⷮ |	не	мытву	себѣ	а	не	му|ченье.
Xle	и		тво|раⷮ	не	мытвоу	себѣ		но	мѫченїе,
Byč	н	то	творять		мовенье	собѣ,	а	не	мученье”.
Šax	н	то	творять		мъвенне	собѣ,	а	не	мученне”.
Lix	н	то	творять		мовенье	собѣ,	а	не	мученье”.
α	н		творять	не	мытву	себѣ,		не	мученне”.

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 ~ ҂м҃ ~ 40 тысящь
    - Synonymy: разумьн- ~ съмысльн-

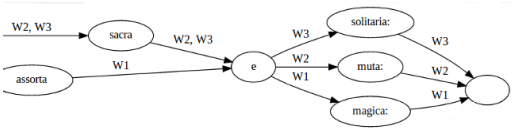
Recent CollateX developments

- 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 wo<lb/>rd2 word3
    - <w>word1</w> <w>wo<lb/>rd2</w> <w>word3</w>
  - Not all whitespace represents token break
    - word1 <lb/> word2
    - <w>word <lb/></w> <w>word2</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; degeminate
  - 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
  - оу, ѵ, ꙋ = ѵ
  - ѡ, ω, о, о = о
- Case fold, neutralize consonantal variants
  - Not always one-to-one, e.g., щ = шт
- Degeminate, delete other vowels, delete diacritics
  - Keep two letters of two-letter words
    - Higher information load
- Other conflations?
  - Knowledge based vs machine learning
- Expand abbreviations?
  - б҃га, бг҃а, б҃а = бога (бг)
- Truncate or zero-pad (to what length?)

Soundex sample (*Bdinski sbornik*)

Ch397	и	възвра|тить	дьщерьше	своѥѥ.
Ch384	и	възвратит҃	дьщершоу	свою.
Nbkm298	и	възвратити	братанитцꙋ |	своѭ
Berlin	и	възꙿвратити |	братаницꙋ	свою.

Ch397	и	взвр	дштр	св
Ch384	и	взвр	дштр	св
Nbkm298	и	взвр	бртн	св
Berlin	и	взвр	бртн	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
    - фраки ~ фраци (фрк ~ фрц)
  - CollateX recognizes only perfect matches
    - Unable to recognize *closest match*

3,5

3,5

Lav	гаръмати	тавр[і] |ани.	сируфыѩ.		фраци.
Tro	гаръмати	тавриани	скуфиа		фраки
Rad	сармати	таврилни	скоуфиа	и	фраци |
Aka	сармати.	тавріани	скоуфіа.	и	фра|ци
Ipa	сармати.	тавриани. |	скуфиѩ.		фраци.
Xle	сармати.	тавріани.	скѵфіа		фра|ци.
Byč	Саръматн,	Таврнанн,	Скуфна,		Фрацн,
Šax	Сарматн,	Таврнанн,	Скуфня,		Фрацн,
Lix	Саръматн,	Таврнанн,	Скуфна,		Фрацн,
α	Сарматн,	Таврнанн,	Скуфня,		Фрацн,

Numbers

18,4

Lav						.[до]	дв҃да.
Aka	.ѕ҃.л҃.	а		исхо|женїа	мѡисѣѡва	до	дв҃да.
Ipa	ѕ҃.л҃.		ѿ	исхожениѩ	мо|исѣѡва.	до	дв҃да.
Xle	.ѕ҃.л҃.	а	ѿ	исхо^ж^нїа |	мѡѵсеѡва	до	дв҃да
Byč	430;	а	отъ	исхожения	Моисѣова	до	Давида
Šax	430;	а	отъ	исхожения	Моисѣева	до	Давыда
Lix	430;	а	от	исхожения	Моисѣова	до	Давида
α	430;	а	отъ	исхожения	Моисѣова	до	Давыда

Problem areas

- Gaps in alignment
- No perfect match
- CollateX follows graph rank (leftmost match)
- 3,5
  - Orthography
    - скуфиа фраки (Tro)
    - скоуфиа и фраци (Rad)
  - Soundex
    - скф фрк
    - скф и фрц

Postprocessing

- Gap without perfect match on either side
  - Gap may span multiple columns
- Orthography
  - скуфиа фраки (Tro)
  - скоуфиа и фраци (Rad)
- Soundex
  - скф фрк
  - скф и фрц
- If there's a match, keep it
- Else
  - Find unique Soundex values in column and following
  - Move token to column with closest match
    - Damerau-Levenshtein edit distance
      - Insertion, deletion, substitution, transposition

9,2

9,2

Lav	и	то	творать		мовенье	собѣ	а	не	мучenье.
Tro	и	то	творять		мовенье		а	не	мученье
Rad	и	тако	творать	не	мытвоу	собѣ		но	м^оу^ |чение.
Aka	и	тако	творатъ	не	мытвоу	собѣ		но	моученье. |
Ipa	и		твора^т^ |	не	мытву	себѣ	а	не	му|ченье.
Xle	и		тво|ра^т^	не	мытвоу	себѣ		но	мѫченїе,
Byč	и	то	творять		мовенье	собѣ,	а	не	мученье”.
Šax	и	то	творять		мъвение	собѣ,	а	не	мучение”.
Lix	и	то	творять		мовенье	собѣ,	а	не	мученье”.
α	и		творять	не	мытву	себѣ,		не	мучение”.

In case of ties

- Thesaurus
- Most matches
- Length of match

Thesaurus

- Collect forced inexact matches
  - A B C
  - A D C
- Edit manually
- Use to break ties
- Close matches
  - поломиша ~ възломиша
  - плмш ~ взлм
- Non-matches
  - разумьн- ~ съмысльн-
  - рзмн- ~ смсл-

Thesaurus

220,9

Lav	налегша	первоє |	на	с҃тополка	и	взломиша
Rad	налагоша	первие	на	с҃тополка·	и	поломиша
Aka	налагоша	первое	на	с҃тополка·	и	поломиша |
Ipa	налагоша	пе|рвое	на	с҃тополка·	и	възло|миша
Xle	налегоша	пръвое	на	с҃тополка.	и	възло|миша
Byč	налегоша	первое	на	Святополка,	и	взломиша
Šax	налегоша	пьрвое	на	Святопълка,	и	възломиша
Lix	налегоша	первое	на	Святополка,	и	взломиша
α	налегоша	пьрвое	на	Святопълка,	и	възломиша

взломиша поломиша
взлм плмш

What's next: many-to-one

141,11

Lav	а	прочи^{х} |	вои	.м҃.		и	поиде	на	с҃тополка |	нарекъ
Rad	а	прочи^{х}.		,м҃.		и	поиде	[на]	с҃тополка.	на|рекъ
Aka	а	прочи^{х}.		.,м҃.		и	поиде	на	с҃тополка.	нарекь |
Ipa	а	прочи|хъ	вои	.м҃.	тысаць.	и	поиде |	на	с҃тополка.	нарекъ.
Xle	а	прочи^{х}	вои	.м҃.	тиса^{ѱ}. |		поиде	на	с҃тополка	нарекь
Byč	а	прочихъ	вой	40000,		и	поиде	на	Святополка,	нарекъ
Šax	а	прочихъ	вои	40	тысящь,	и	поиде	на	Святопълка,	нарекъ
Lix	а	прочих	вой	40,000,		и	поиде	на	Святополка,	нарекъ
α	а	прочихъ	вои	40	тысящь,	и	поиде	на	Святопълка,	нарекъ

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