Collating the *Rus' primary chronicle* (*Povest' vremennyx let*)

David J. Birnbaum
djbpitt@gmail.com
http://www.obdurodon.org

Varna, 2014-09-15

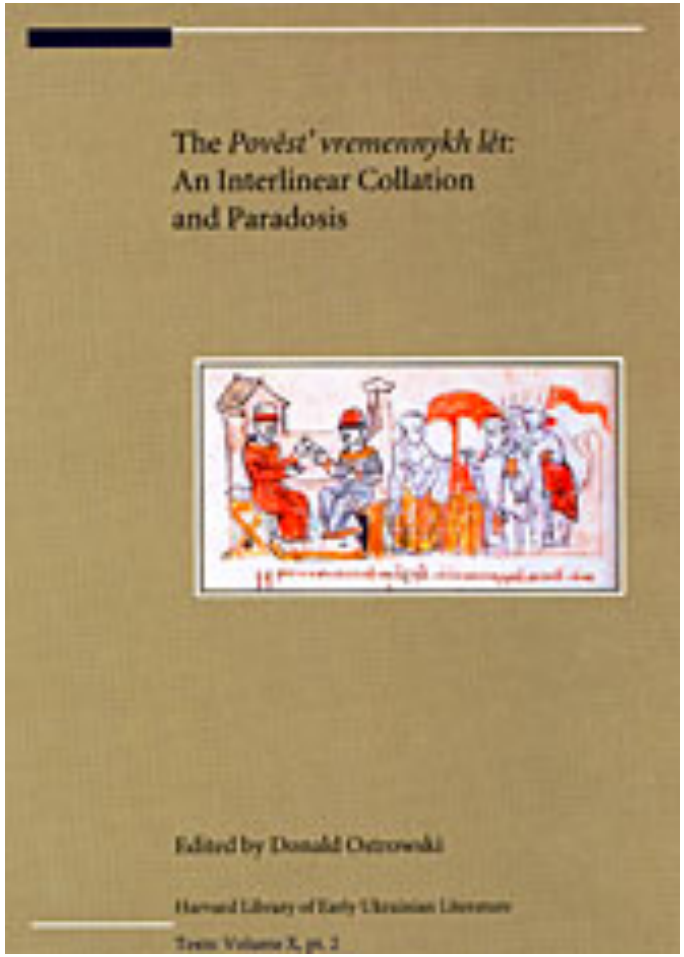
Why collate the *PVL*?

- Textual comparison
  - Relationships among the copies
  - Construction of a paradosis (alpha text)
  - History of transmission beyond alpha
- Linguistic comparison
- Orthographic comparison
  - Requires diplomatic transcription

Practical issues

- No funding
- Edition is under constant development
  - Donald Ostrowski, with David J. Birnbaum, Francis Butler, Inés García de la Puente
  - Must be able to rerun collation
  - Must be fully automated

Print edition



- *The Povest' vremennykh lĕt: 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 нндникнѧ в долготу
Trin: персида ватрь даже и до индикия в долготу
Radz: перснда. ватрь. доже н до нн҃никна. в до|лготоу
Acad: персїда. ватръ. дон҃ н до нндникїа. в долготѹ |
Hypa: перь|снда. ватрь. доже н до ннь|дикнѧ. в долготу
Khle: перснда. ватръ. даже н до нндникїа. 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
- If the interlinear edition were digital
  - Space, weight, cost are irrelevant
  - User can select witnesses
  - Searching on other than plain text
    - Lemma
    - Morphology

How interlinear?

- Initially
  - Alignment by line (per Karskii 1926 edition)
    - Used by Müller in the *Handbuch* and Cross in the English translation (1930)
- Target
  - Alignment by word

Digital versions

- PDF of print edition
  http://hudce7.harvard.edu/~ostrowski/pvl/
- HTML edition
  http://pvl.obdurodon.org

Print edition workflow

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

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 ин^{Д}икиа. в до|лготоу
Aka	персїда. ватръ. до^{ж}и 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
Vyč	Перснда, Ватрь, доже н до Ндникня в долготу,
Šax	Перснда, Ватрь доже н до Ндникня въ дълготу,
Lix	Перснда, Ватрь, доже н до Ндникня в долготу,
α	Перснда, Ватрь доже н до Ндникня 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
- Diplomatic transcription
  - Efficient comparison algorithms require exact string matching, which is rare in diplomatic transcription
  - Finding *closest* match requires a different—less computationally efficient—algorithm (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	ин^{д}икиа.	в	до|лготоу
Aka	персїда.	ватръ.	до^{ж}и	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	Персндa,	Ватрь,	доже	н	до	Ндднкня	в	долготу,
Šax	Персндa,	Ватрь	доже	н	до	Ндднкня	въ	дълготу,
Lix	Персндa,	Ватрь,	доже	н	до	Ндднкня	в	долготу,
α	Персндa,	Ватрь	доже	н	до	Ндднкня	въ	дълготу,

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
    - 40000 ~ ѧ҃м҃ ~ 40 тысящь
    - ρазумьнъ ~ съмысльнъ

Adapting CollateX

- Preprocessing
- Collation (CollateX)
- Postprocessing

Preprocessing

- Normalized “shadow” copy
- Soundex simplification
- Collate on normalization, render 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
  - Consonants carry more information than vowels
- Information load may be sensitive to position
  - Beginning of word carries more information than end
  - Especially true for lexical (not morphological) searching in inflected languages

Adapting Soundex to Church Slavonic

- Neutralize variant spellings of initial vowel
  - оу, у, ꙋ = у
  - ѡ, ѿ, ѻ, 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

- Ch397 и вьзвра|тить дьщерьше своюєе.
- Ch384 и вьзвратит͡ дьщершоу свою.
- Nbkm298 и вьзвратити дьщерьше своѫ
- Berlin и вьʒʿвратити | дьщерьше своѫ

- Ch397 и взвр дштр св
- Ch384 и взвр дштр св
- Nbkm298 и взвр брдр св
- Berlin и взвр брдр св

Two types of normalization

- Collation
  - Find alignment points
  - Coarse adjustments
  - No harm in conflating, e.g., imperfect and aorist or 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 results
- Utilize forced matches
  - A B C
  - A D C
- Misses
  - Gap in alignment (no forced match)
  - Imperfect match
    - фраки ~ фраци
  - CollateX recognizes only perfect matches
    - Unable to recognize *closest match*
    - Computational complexity

3,5

3,5

Lav	гаръмати	тавр[і] |ани.	сируфыа.		фраци.
Tro	гаръмати	тавриани	скуфиа		фраки
Rad	сармати	тавриани	скоуфиа	и	фраци |
Aka	сармати.	таврїани	скоуфїа.	и	фра|ци
Ipa	сармати.	тавриани. |	скуфиа.		фраци.
Xle	сармати.	таврїани.	скvфїа		фра|ци.
Byč	Саръматн,	Таврнаnн,	Скуфна,		Фрацн,
Šax	Сарматн,	Таврнаnн,	Скуфня,		Фрацн,
Lix	Саръматн,	Таврнаnн,	Скуфна,		Фрацн,
α	**Сарматн,**	**Таврнаnн,**	**Скуфня,**		**Фрацн,**

Numbers

18,4

Lav						.[до]	дв҃да.
Aka	.ѵ҃.л҃.	а		исхо|женїа	мѡисѣѡва	до	дв҃да.
Ipa	ѵ҃.л҃.		ѿ	исхожениꙗ	мо|исѣѡва.	до	дв҃да.
Xle	.ѵ҃.л҃.	а	ѿ	исхо^{ж}нїа |	мѡѵсеѡва	до	дв҃да
Byč	430;	а	отъ	нсхоженна	Монсѣова	до	Давнда
Šax	430;	а	отъ	нсхоження	Монсѣева	до	Давыда
Lix	430;	а	от	нсхоженна	Монсѣова	до	Давнда
α	430;	а	отъ	нсхоження	Монсѣова	до	Давыда

Problem areas

- Gaps in alignment
- No perfect match
- CollateX takes leftmost match
- 3,5
  - Orthography
    - скуфиа фракиⷪ (Tro)
    - скоуфиа и фраⷰциⷪ (Rad)
  - Soundex
    - скф фрк
    - скф и фрц

Postprocessing

- If there's a gap
  - If the column all matches, keep it
  - Else
    - Find all unique Soundex values in column and following
    - Move token to column with closest match
      - Edit distance (Damerau-Levenshtein)
        - » Insertion, deletion, substitution, transposition
      - Recursive

9,2

9,2

Lav	и	то	творѧть		мовенье	собѣ	а	не	мученье.
Tro	и	то	творѧть		мовенье		а	не	мученье
Rad	и	тако	творѧть	не	мытвоу	собѣ		но	моу | чение.
Aka	и	тако	творѧтъ	не	мытвоу	собѣ		но	моученье. |
Ipa	и		творѧ^{т} |	не	мытвѹ	себѣ	а	не	му | ченье.
Xle	и		тво | ря^{т}	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”.

In case of ties

- Thesaurus
- Most matches
- Length

Thesaurus

- Collect forced inexact matches
- Edit manually
- Use to break ties
- Close matches
  - ѡломиша ~ въѕломиша
  - плмш ~ взлм
- Non-matches
  - ѓазумьн¬ ~ съмысльн¬
  - ѓзмн¬ ~ смсл¬

Thesaurus

220,9

Lav	налегша	первоѭ	на	стополка	и	взломиша
Rad	налагоша	первие	на	стополка·	и	поломиша
Aka	налагоша	первое	на	стополка·	и	поломиша
Ipa	налагоша	пе|рвое	на	стополка·	и	възло|миша
Xle	налегоша	пръвое	на	стополка.	и	възло|миша
Byč	налегоша	первое	на	Святополка,	н	взломнша
Šax	налегоша	пьрвое	на	Святопълка,	н	възломнша
Lix	налегоша	первое	на	Святополка,	н	взломнша
α	**налегоша**	**пьрвое**	**на**	**Святопълка,**	**н**	**възломнша**

взломиша	поломиша
взлм	плмш

What's next: many-to-one

141,11

Lav	а	прочи^{х} |	вои	.м̄.		и	поиде	на	с̄тополка |	нарекъ
Rad	а	прочи^{х}.		ѯм̄.		и	поиде	[на]	с̄тополка.	на|рекъ
Aka	а	прочи^{х}.		.ѯм̄.		и	поиде	на	ст̄ополка.	нарекь |
Ipa	а	прочи|хъ	вои	.м̄.	тысаць.	и	поиде |	на	ст̄ополка.	нарекъ.
Xle	а	прочи^{х}	вои	.м̄.	тиса^{ш}. |		поиде	на	ст̄ополка	нарекь
Byč	а	прочнхъ	вой	40000,		н	понде	на	Святополка,	нарекъ
Šax	а	прочнхъ	вон	40	тысяць,	н	понде	на	Святопълка,	нарекъ
Lix	а	прочнх	вой	40,000,		н	понде	на	Святополка,	нарекъ
α	а	прочнхъ	вон	40	тысяць,	н	понде	на	Святопълка,	нарекъ

Recent developments

- CollateX has been ported from Java to Python
  - Python module
  - Easier to modify
  - Closer integration of preprocessing and postprocessing
  - Fivefold improvement in processing time (so far)
- Decision tree architecture under development for CollateX core
  - Closer integration of postprocessing with core