About digital collation

David J. Birnbaum (University of Pittsburgh)
Digital Humanities Literacy Workshop
Carnegie Mellon University
2016-05-18

Outline

- What is collation?
- The Gothenburg model
- Collation with CollateX

What is collation?

- *What:* Alignment and comparison of textual witnesses
- *Why:* Support text-critical analysis and edition
- *Input:* Multiple textual witnesses to the same work
- *Output:* Alignment of variants

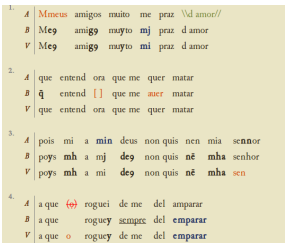
Types of variation

- Textual: insertion, deletion, mutation, transposition
- Substantive ~ non-substantive
  - Substantive: equipollent, linguistic, scribal error
  - Non-substantive: graphic
- Ignore non-substantive variation for comparison
  - Punctuation
  - Upper ~ lower case
  - Orthographic variation
    - Variant letterforms
    - Abbreviation

Types of output

1. Interlinear (synoptic) edition
   - Variant table
2. Critical apparatus
3. Variant graph
4. TEI XML
5. Stemma codicum
6. Etc.

1. Interlinear (synoptic) edition

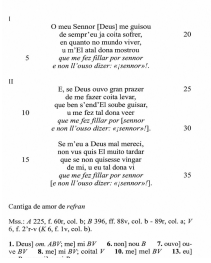


- Blocks: lines
- Rows: witnesses
- Columns: aligned tokens
- In this edition
  - Bold: graphic variation
  - Underline: equipollent reading
  - Orange: scribal error
  - Blue: linguistic variant
  - Other: deletions (red), insertions (green)

1. Sample interlinear collations

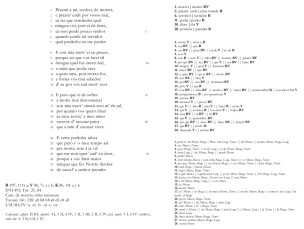
- *Povest' vremennykh let (Rus' primary chronicle)*
  - Donald Ostrowski (Harvard University), David J. Birnbaum (University of Pittsburgh), Horace G. Lunt (Harvard University)
  - http://pvl.obdurodon.org/browser.xhtml
- *Galician-Portuguese secular lyric: philology and historical linguistics*
  - Helena Bermúdez Sabel (Universidade de Santiago de Compostela)
  - http://gl-pt.obdurodon.org/index.xhtml

2. Critical apparatus



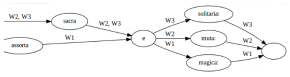
- Main text (reconstructed)
- Text type
- Traditio textus
  - Witnesses and loci
- Apparatus criticus (negative)
  - Location, lemma, reading, sigla

2. Critical apparatus



- Significant variants
  - — Equipollent (textual)
  - — Linguistic
  - — Scribal error
- Insignificant variants
  - — Graphic
- History of edition
  - — Critical annotations from prior editions (negative)

3. Variant graph



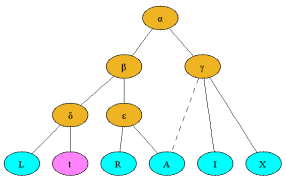
- Directed graph
- Nodes: readings
- Rank: alignment
- Edges: witness labels

4. TEI parallel segmentation

```
<l>
<app>
  <rdg wit="#one">se</rdg>
  <rdg wit="#two"><add>me</add></rdg>
  <rdg wit="#three">me</rdg>
</app>
atormentan
<app>
  <rdg wit="#one #two">en el jardín</rdg>
</app>
</l>
```

- Plain text: Shared textual reading
- <app>: Variation locus
- <rdg>: Textual variant
- @wit: Sigla of witnesses

5. Stemma codicum



- Hypothesis about textual transmission
- Nodes
  - Greek sigla, ocher
    - Hypothetical
  - Upper-case Latin sigla, aqua
    - Extant manuscripts
  - Lower-case Latin sigla, violet
    - Lost manuscripts
- Edges
  - Solid line
    - Antigraph → apograph
  - Dotted line
    - Contamination

6. Other output formats

- Plain text variation table
- HTML variation table
- XML variation table
- GraphViz DOT
- Etc.

The Gothenburg model

- History and goals
- Components
  1. Tokenization
  2. Normalization/regularization
  3. Alignment
  4. Analysis
  5. Visualization/output

The Gothenburg model: history and goals

- Developers of CollateX and Juxta
- Gothenburg 2009 joint workshop
- Sponsored by COST Action 32 and Interedition
- Identify core components of textual comparison at an abstract level

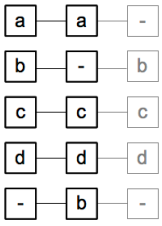
1. Tokenization

- (Presumes transcription and digitization)
- Divide the continuous text into units to be aligned (tokens)
- Typically whitespace-delimited words
  - May be at any level of granularity
  - "Syllables, words, lines, phrases, verses, paragraphs, or text nodes"
- Challenges
  - Ambiguity
  - Punctuation
  - Contraction, superscription, etc.
  - Markup

2. Normalization/regularization

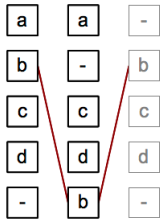
- Normalization during transcription ~ collation
- Ignore non-substantive variation for comparison
  - Punctuation
  - Upper ~ lower case
  - Orthographic variation
    - Variant letterforms
    - Abbreviation
- What goes into the output?

3. Alignment



- Alignment table
- Depth vs breadth
- Complications
  - Repetition
  - Transposition
  - Order effects
  - Computational complexity
  - Exact vs near (fuzzy) matching

4. Analysis/feedback



- Interpretation beyond linear alignment
- Manual intervention?

5. Visualization/output

- Markup, for further processing
  - XML, TEI, JSON, GraphViz DOT, LaTeX, etc.
- Textual alignment table, final form for edition
  - Plain text, HTML, PDF
- Textual visualization, for examination and analysis
  - Juxta
  - Versioning machine
- Graphic visualization, for examination and analysis
  - Variant graph

CollateX

- Java, Web app, and Python module
  - CollateX Java version:
    - http://collatex.net
  - CollateX Python package:
    - https://pypi.python.org/pypi/collatex
  - CollateX Python tutorial:
    - http://collatex.obdurodon.org
- Input: Anything at all (JSON)
- Output: Anything at all (JSON)

Line 1	**Draft**	\\Si// te atreves a ~~{comprender}~~ sorprender
	Published	Si te atreves a sorprender
Line 2	**Draft**	el sentido de esta vieja pared;
	Published	la verdad de esta vieja pared;
Line 3	**Draft**	y sus fisura\s,/~~{,}~~{,} desgarraduras~~{,}~~,
	Published	y sus fisuras, desgarraduras,
Line 4	**Draft**	formando rostros, esfinges,
	Published	formando rostros, esfinges,
Line 5	**Draft**	manos, clepsidras, ~~{...}~~
	Published	manos, clepsidras,

CollateX: Benefits and limitations

- Benefit
  - Complete control over input, tokenization, normalization, collation, and visualization (output)
- Limitation
  - Requires user programming (Python, possibly others)

Thank you!

- David J. Birnbaum (University of Pittsburgh)
  - djbpitt@gmail.com
  - http://www.obdurodon.org

Materials for this presentation were developed with the assistance of Helena Bermúdez Sabel (Universidade de Santiago de Compostela). An earlier version of this presentation was delivered at the *Text as process: Genetic and textual criticism in the digital age* conference (University of Pittsburgh, 2016-04-05).