About digital collation

David J. Birnbaum (University of Pittsburgh)
Helena Bermúdez Sabel (Universidade de Santiago de Compostela)
Text as process: Genetic and textual criticism in the digital age
University of Pittsburgh, 2016-04-05

Outline

- What is collation?
- The Gothenburg model
- Collation tools
  - CollateX
  - Juxta
  - Versioning machine

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. Stemma codicum
5. TEI XML

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)
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. Stemma codicum

- Hypothesis about textual transmission
- Nodes
  - Greek sigla, ochre: hypothetical
  - Upper-case Latin sigla, aqua: extant manuscripts
  - Lower-case Latin sigla, violet: lost manuscripts
- Edges
  - Solid line: antigraph → apograph
  - Dotted line: contamination

5. TEI parallel segmentation

```xml
<l>
  <app>
    <rdg wit="#one">atomentan</rdg>
  </app>
  <app>
    <rdg wit="#two">en el jardín</rdg>
  </app>
  <app>
    <rdg wit="#three">me</rdg>
  </app>
</l>
```

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)

Juxta

- Stand-alone desktop application
  - Input: XML and plain text
  - Output: Analytic visualizations
    - Side-by-side collation view
    - Heat map
    - Histogram
    - Critical apparatus
  - Annotation and image support
- Juxta Commons (online tool)

Juxta collation features

- Selection of base text
- Normalization
  - Punctuation
  - Case
  - Whitespace
- Accept/reject revisions

Juxta disadvantages

- Limited options for normalization
- The base text must be a single, specific witness
  - Does not support a dynamic base text
- Loss of mark-up information other than addition and deletion
  - E.g., abbreviation, editorial regularization, etc.
- Limited control over the publication style

Versioning machine

- Visualizes alignment; does not perform collation
- Input: TEI aligned texts (parallel segmentation method)
  - Developer determines collation to prepare input
- Visualization features and facsimile image support
Using the Versioning machine for publication

- Displays multiple layers of textual information
  - Preserves markup
- Selection of default behaviors (well documented)
- Code reuse: customization (programming experience required)

Summary

- CollateX
  - Benefit: Complete control over input, tokenization, normalization, collation, and visualization
  - Limitation: Requires user programming (Python, possibly others)
- Juxta
  - Benefit: Analytical tools (histogram, heat map, annotation)
  - Limitation: No control over output, loss of markup information
- Versioning machine
  - Benefit: User control over output
  - Limitation: Requires TEI parallel segmentation input, output control requires user programming (XSLT, CSS)

Thank you!

- David J. Birnbaum (University of Pittsburgh)
  - djbpitt@gmail.com
  - http://www.obdurodon.org
- Helena Bermúdez Sabel
  (Universidade de Santiago de Compostela)
  - helena.bermudez@usc.es