

## 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
6. Etc.

## 1. Interlinear (synoptic) edition

```

1. A Mneus amigos maito me praz \vd amore//
   B Me9 amigos maito mj praz d amor
   V Me9 amigg maito mi praz d amor

2. A que entend ora que me quer matar
   B q entend [ ] que me auec matar
   V que entend ora que me quer matar

3. A pois mi a min deus non quis nen mia senhor
   B poys mh a mj de9 non quis ne mha senhor
   V poys mh a mi de9 non quis ne mha sen

4. A a que (e9) roquei de me del amparar
   B a que roquey simpex del emparar
   V a que o roquey de me del emparar

```

- 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

I O meu Senhor [Deus] me guisou  
 de sempre en a minha vida,  
 en quanto ao mundo viver,  
 e en TE, ai, dona meoira,  
 que me fez fillar por amor  
 e non II souso dizer: «jennora»!

II E, se Deus ouvo gran prazer  
 de me fazer outra vez,  
 que ben e end' E! soube guisar,  
 e non E!z tal dona veer,  
 que me fez fillar por [amor]  
 e non II souso dizer: «jennora»!

Se m'ou a Deus mal mereci,  
 non via que E! muito tardar  
 que se non quiseo virar  
 de mi, e eu tal dona vi  
 que me fez fillar por amor  
 [e non II souso dizer: «jennora»]!

Config. de amor de refon  
 Ms., d 225, E 60r, col. b; B 196, ff 88v, col. b - 89r, col. a, f  
 6, E 27v (A 6, E 1v, col. b).  
 I. Deus! ou. aDP. me] mi B! & mo] nou B T. ouo] ou-  
 ve E!z & me] mi DP. con]a F 1B. me] me] B! 1A. ou]

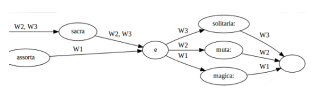
- Main text (reconstructed)
- Text type
- Traditio textus (witnesses and loci)
- Apparatus criticus (negative)
  - Location, lemma, reading, sigla

## 2. Critical apparatus

1. Prouer e, ouder, de amor,  
 2. Prouer e, ouder, de amor,  
 3. Prouer e, ouder, de amor,  
 4. Prouer e, ouder, de amor,  
 5. Prouer e, ouder, de amor,  
 6. Prouer e, ouder, de amor,  
 7. Prouer e, ouder, de amor,  
 8. Prouer e, ouder, de amor,  
 9. Prouer e, ouder, de amor,  
 10. Prouer e, ouder, de amor,  
 11. Prouer e, ouder, de amor,  
 12. Prouer e, ouder, de amor,  
 13. Prouer e, ouder, de amor,  
 14. Prouer e, ouder, de amor,  
 15. Prouer e, ouder, de amor,  
 16. Prouer e, ouder, de amor,  
 17. Prouer e, ouder, de amor,  
 18. Prouer e, ouder, de amor,  
 19. Prouer e, ouder, de amor,  
 20. Prouer e, ouder, de amor,  
 21. Prouer e, ouder, de amor,  
 22. Prouer e, ouder, de amor,  
 23. Prouer e, ouder, de amor,  
 24. Prouer e, ouder, de amor,  
 25. Prouer e, ouder, de amor,  
 26. Prouer e, ouder, de amor,  
 27. Prouer e, ouder, de amor,  
 28. Prouer e, ouder, de amor,  
 29. Prouer e, ouder, de amor,  
 30. Prouer e, ouder, de amor,  
 31. Prouer e, ouder, de amor,  
 32. Prouer e, ouder, de amor,  
 33. Prouer e, ouder, de amor,  
 34. Prouer e, ouder, de amor,  
 35. Prouer e, ouder, de amor,  
 36. Prouer e, ouder, de amor,  
 37. Prouer e, ouder, de amor,  
 38. Prouer e, ouder, de amor,  
 39. Prouer e, ouder, de amor,  
 40. Prouer e, ouder, de amor,  
 41. Prouer e, ouder, de amor,  
 42. Prouer e, ouder, de amor,  
 43. Prouer e, ouder, de amor,  
 44. Prouer e, ouder, de amor,  
 45. Prouer e, ouder, de amor,  
 46. Prouer e, ouder, de amor,  
 47. Prouer e, ouder, de amor,  
 48. Prouer e, ouder, de amor,  
 49. Prouer e, ouder, de amor,  
 50. Prouer e, ouder, de amor,  
 51. Prouer e, ouder, de amor,  
 52. Prouer e, ouder, de amor,  
 53. Prouer e, ouder, de amor,  
 54. Prouer e, ouder, de amor,  
 55. Prouer e, ouder, de amor,  
 56. Prouer e, ouder, de amor,  
 57. Prouer e, ouder, de amor,  
 58. Prouer e, ouder, de amor,  
 59. Prouer e, ouder, de amor,  
 60. Prouer e, ouder, de amor,  
 61. Prouer e, ouder, de amor,  
 62. Prouer e, ouder, de amor,  
 63. Prouer e, ouder, de amor,  
 64. Prouer e, ouder, de amor,  
 65. Prouer e, ouder, de amor,  
 66. Prouer e, ouder, de amor,  
 67. Prouer e, ouder, de amor,  
 68. Prouer e, ouder, de amor,  
 69. Prouer e, ouder, de amor,  
 70. Prouer e, ouder, de amor,  
 71. Prouer e, ouder, de amor,  
 72. Prouer e, ouder, de amor,  
 73. Prouer e, ouder, de amor,  
 74. Prouer e, ouder, de amor,  
 75. Prouer e, ouder, de amor,  
 76. Prouer e, ouder, de amor,  
 77. Prouer e, ouder, de amor,  
 78. Prouer e, ouder, de amor,  
 79. Prouer e, ouder, de amor,  
 80. Prouer e, ouder, de amor,  
 81. Prouer e, ouder, de amor,  
 82. Prouer e, ouder, de amor,  
 83. Prouer e, ouder, de amor,  
 84. Prouer e, ouder, de amor,  
 85. Prouer e, ouder, de amor,  
 86. Prouer e, ouder, de amor,  
 87. Prouer e, ouder, de amor,  
 88. Prouer e, ouder, de amor,  
 89. Prouer e, ouder, de amor,  
 90. Prouer e, ouder, de amor,  
 91. Prouer e, ouder, de amor,  
 92. Prouer e, ouder, de amor,  
 93. Prouer e, ouder, de amor,  
 94. Prouer e, ouder, de amor,  
 95. Prouer e, ouder, de amor,  
 96. Prouer e, ouder, de amor,  
 97. Prouer e, ouder, de amor,  
 98. Prouer e, ouder, de amor,  
 99. Prouer e, ouder, de amor,  
 100. Prouer e, ouder, de amor,

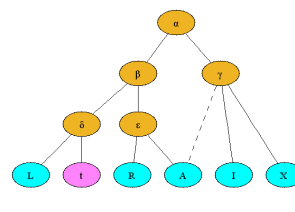
- Significant variants
  - Equipollent (textual)
  - Linguistic
  - Scribal error
- Insignificant variants
  - Graphic
- History of edition
  - Critical annotations from prior editions (negative)

## 3. Variant graph



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

## 4. Stemma codicum



- Hypothesis about textual transmission
- Nodes
  - Greek sigla, ocher: 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

```
<|>
<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 jardin</rdg>
</app>
</|>
```

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

## 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

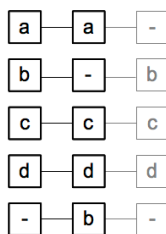
### 1. Tokenization

- (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

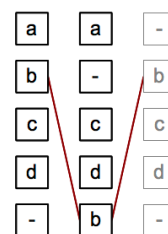
- Normalization during transcription ~ collation
- Ignore non-substantive variation for comparison
  - Punctuation
  - Upper ~ lower case
  - Orthographic variation
    - Variant letterforms
    - Abbreviation
- What goes into the output?

### 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	\\\$!// te atreves a (====der) 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,

## 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](mailto:djbpitt@gmail.com)
  - <http://www.obdurodon.org>
- Helena Bermúdez Sabel (Universidade de Santiago de Compostela)
  - [helena.bermudez@usc.es](mailto:helena.bermudez@usc.es)
  - <http://gl-pt.obdurodon.org>

