

## 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

```

1. A | Mneus amigos mauto me praz \d amor//
   B | Me9 amigos mauto mj praz d amor
   V | Me9 amigos mauto mi praz d amor

2. A | que estend ora que me quer mutar
   B | q estend [ ] que me acer mutar
   V | que estend ora que me quer mutar

3. A | pois mi a min deus non quis nen mia scnoor
   B | poyz mh a mj dey non quis ne mha scnoor
   V | poyz mh a mi dey non quis ne mha scn

4. A | a que (o) roquesy de me del amparar
   B | a que roquesy scayax del emparar
   V | a que o roquesy 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)

### 1. Sample interlinear collations

- *Povest' vremennykh let (Rus' primary chronicle)*
  - Donald Ostrowski (Harvard University), David J. Bimbaum (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

I O non Senzor [Dous me guison  
de sempre ou, ja coita soffer,  
en quanto te vindeu viver,  
a mi foz tal dous moiron  
que me fez illar por senzor  
e non li souso dizer: «jennora»].

II E, se Dous souso gra prazor  
de me fazer crida leve,  
que non a non foz crida quntar,  
a me fez tal dous veer  
que me fez illar por [senzor  
e non li souso dizer: «jennora»].

Se n' en a Dous mal merozi,  
non me quez E! moito tenar  
que se non quisesse virgar  
de mi, a se tal dous vi  
que me fez illar por senzor  
[e non li souso dizer: «jennora»].

Collaça de amor de refrao

Mss. A 223, E 89c, col. B, P 396, ff. 81v, col. b - 89r, col. a; F 6, E 2 v-v (A 6, E 1v, col. b)

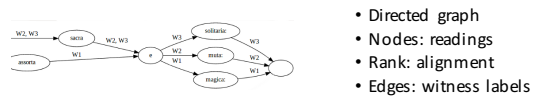
I. Dous] m. ADF, m] m] BF 6, m] non B 7, ouso] ou-  
ca BF 8, m] m] BF; con] F 10, m] m] BF 13, m]

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

### 2. Critical apparatus

<p>1. Poesía e arte medieval de Galicia</p> <p>2. Poesía medieval gallega</p> <p>3. Poesía medieval castelana</p> <p>4. Poesía medieval portuguesa</p> <p>5. Poesía medieval catalana</p> <p>6. Poesía medieval provenzal</p> <p>7. Poesía medieval francesa</p> <p>8. Poesía medieval italiana</p> <p>9. Poesía medieval española</p> <p>10. Poesía medieval alemana</p> <p>11. Poesía medieval inglesa</p> <p>12. Poesía medieval irlandesa</p> <p>13. Poesía medieval escocesa</p> <p>14. Poesía medieval nórdica</p> <p>15. Poesía medieval eslava</p> <p>16. Poesía medieval griega</p> <p>17. Poesía medieval árabe</p> <p>18. Poesía medieval hebraica</p> <p>19. Poesía medieval siríaca</p> <p>20. Poesía medieval copta</p> <p>21. Poesía medieval armenia</p> <p>22. Poesía medieval georgiana</p> <p>23. Poesía medieval india</p> <p>24. Poesía medieval china</p> <p>25. Poesía medieval japonesa</p> <p>26. Poesía medieval coreana</p> <p>27. Poesía medieval vietnamita</p> <p>28. Poesía medieval tailandesa</p> <p>29. Poesía medieval indonesia</p> <p>30. Poesía medieval filipina</p> <p>31. Poesía medieval vietnamita</p> <p>32. Poesía medieval birmana</p> <p>33. Poesía medieval nepalesa</p> <p>34. Poesía medieval tibetana</p> <p>35. Poesía medieval budista</p> <p>36. Poesía medieval hinduista</p> <p>37. Poesía medieval jainista</p> <p>38. Poesía medieval sijista</p> <p>39. Poesía medieval budista</p> <p>40. Poesía medieval hinduista</p> <p>41. Poesía medieval jainista</p> <p>42. Poesía medieval sijista</p> <p>43. Poesía medieval budista</p> <p>44. Poesía medieval hinduista</p> <p>45. Poesía medieval jainista</p> <p>46. Poesía medieval sijista</p> <p>47. Poesía medieval budista</p> <p>48. Poesía medieval hinduista</p> <p>49. Poesía medieval jainista</p> <p>50. Poesía medieval sijista</p>	<p>1. Poesía medieval gallega</p> <p>2. Poesía medieval castelana</p> <p>3. Poesía medieval portuguesa</p> <p>4. Poesía medieval catalana</p> <p>5. Poesía medieval provenzal</p> <p>6. Poesía medieval francesa</p> <p>7. Poesía medieval italiana</p> <p>8. Poesía medieval española</p> <p>9. Poesía medieval alemana</p> <p>10. Poesía medieval inglesa</p> <p>11. Poesía medieval irlandesa</p> <p>12. Poesía medieval escocesa</p> <p>13. Poesía medieval nórdica</p> <p>14. Poesía medieval eslava</p> <p>15. Poesía medieval griega</p> <p>16. Poesía medieval árabe</p> <p>17. Poesía medieval hebraica</p> <p>18. Poesía medieval siríaca</p> <p>19. Poesía medieval copta</p> <p>20. Poesía medieval armenia</p> <p>21. Poesía medieval georgiana</p> <p>22. Poesía medieval india</p> <p>23. Poesía medieval china</p> <p>24. Poesía medieval japonesa</p> <p>25. Poesía medieval coreana</p> <p>26. Poesía medieval vietnamita</p> <p>27. Poesía medieval tailandesa</p> <p>28. Poesía medieval indonesia</p> <p>29. Poesía medieval filipina</p> <p>30. Poesía medieval vietnamita</p> <p>31. Poesía medieval birmana</p> <p>32. Poesía medieval nepalesa</p> <p>33. Poesía medieval tibetana</p> <p>34. Poesía medieval budista</p> <p>35. Poesía medieval hinduista</p> <p>36. Poesía medieval jainista</p> <p>37. Poesía medieval sijista</p> <p>38. Poesía medieval budista</p> <p>39. Poesía medieval hinduista</p> <p>40. Poesía medieval jainista</p> <p>41. Poesía medieval sijista</p> <p>42. Poesía medieval budista</p> <p>43. Poesía medieval hinduista</p> <p>44. Poesía medieval jainista</p> <p>45. Poesía medieval sijista</p> <p>46. Poesía medieval budista</p> <p>47. Poesía medieval hinduista</p> <p>48. Poesía medieval jainista</p> <p>49. Poesía medieval sijista</p> <p>50. Poesía medieval budista</p>	<ul style="list-style-type: none"> <li>• Significant variants           <ul style="list-style-type: none"> <li>– Equipollent (textual)</li> <li>– Linguistic</li> <li>– Scribal error</li> </ul> </li> <li>• Insignificant variants           <ul style="list-style-type: none"> <li>– Graphic</li> </ul> </li> <li>• History of edition           <ul style="list-style-type: none"> <li>– Critical annotations from prior editions (negative)</li> </ul> </li> </ul>
--	--	--

### 3. Variant graph



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

### 4. TEI parallel segmentation

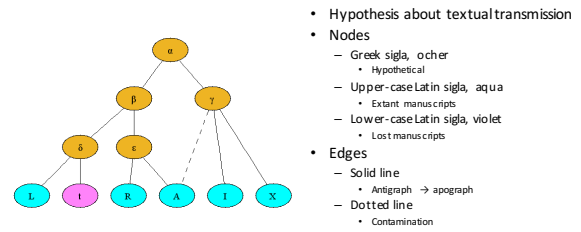
```

<|>
<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

### 5. Stemma codicum



## 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

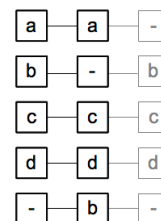
## 1. Tokenization

- (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

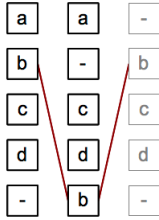
- Normalization during transcription ~ collation
- Ignore non-substantive variation for comparison
  - Punctuation
  - Upper ~ lower case
  - Orthographic variation
    - Variant letterforms
    - Abbreviation
- What goes into the output?

## 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	<code>\\Si// te atreves a <del>(comer)</del> sorprender</code>
	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) <del>(e)</del> desgarraduras <del>(s)</del> ,
	Published	y sus fisuras, desgarraduras,
Line 4	Draft	formando rostros, esfinges,
	Published	formando rostros, esfinges,
Line 5	Draft	manos, clepsidras, <del>(=)</del>
	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. Bimbaum (University of Pittsburgh)
  - [djbpitt@gmail.com](mailto: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 EnBicSm in the digital age* conference (University of Pittsburgh, 2016-04-05).