Learning the CBGM by Design

Greek Paul Project Webinar
28 April 2022

Joey McCollum
Australian Catholic University
Institute for Religion and Critical Inquiry

james.mccollum@myacu.edu.au
@jamesjmccollum
jjmccollum
• Developed over thirty years by Gerd Mink, culminating in the latest updates to the *Editio Critica Maior (ECM)*

• **Recommended reading:**
  
  
  
  • *Tommy Wasserman and Peter J. Gurry*, *A New Approach to Textual Criticism: An Introduction to the Coherence-Based Genealogical Method*, *RBS 80* (Atlanta, GA: SBL Press, 2017)
  
About the CBGM

- Intended to solve *contamination*, or mixture across branches of the textual tradition
• Key assumption: *no hypothetical ancestors* (except the *Ausgangstext A*)

• Other important assumptions:
  1. Scribes typically copied their exemplars with fidelity.
  2. If a scribe introduced a variant, then it came from some other reading.
  3. Scribes typically used fewer sources rather than many.
  4. Scribes typically used closely related sources rather than distant ones.
• Key assumption: *no hypothetical ancestors* (except the *Ausgangstext A*)

• Other important assumptions:
  1. Scribes typically copied their exemplars with fidelity.
  2. If a scribe introduced a variant, then it came from some other reading.
  3. Scribes typically used fewer sources rather than many.
  4. Scribes typically used closely related sources rather than distant ones.
• *Not* a new methodology for evaluating variant readings, but a “meta-approach” to be used on top of existing methods

• *Not* a way to make computers do textual criticism, but a way for them to help us refine human judgments
About the CBGM

• *Not* a new methodology for evaluating variant readings, but a “meta-approach” to be used on top of existing methods
• *Not* a way to make computers do textual criticism, but a way for them to help us refine human judgments
• “Iterative workflow” highlighted in blue
Collation

- To compare manuscripts’ texts, we must first align them at independent *variation units*
- *Variant readings* occur at variation units

• Variation units serve as our points of comparison between witnesses in the CBGM
• Think of them as the columns of a table and the witnesses as rows

<table>
<thead>
<tr>
<th></th>
<th>3Jo 1:1/2</th>
<th>3Jo 1:1/6</th>
<th>3Jo 1:1/8</th>
<th>...</th>
<th>3Jo 1:15/23</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA 69</td>
<td>a</td>
<td>afl</td>
<td>a</td>
<td></td>
<td>a</td>
</tr>
<tr>
<td>GA 1739</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>...</td>
<td>a</td>
</tr>
<tr>
<td>GA 2243</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>...</td>
<td>a</td>
</tr>
</tbody>
</table>
Collation

- This is readily encoded in TEI XML format

```xml
<?xml version='1.0' encoding='UTF-8'?>
<TEI xmlns='http://www.tei-c.org/ns/1.0'>
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>A collation of Luke 10:2 in Swanson</title>
    </titleStmt>
    <publicationStmt>
      <p>Swanson, Reuben J., ed. <em>New Testament Greek Manuscripts: Variant</p>
    </publicationStmt>
    <sourceDesc>
      <listWit>
        <witness n="P75"/>
        <witness n="f1"/>
        <witness n="f13"/>
      </listWit>
    </sourceDesc>
  </fileDesc>
</teiHeader>
<text xml:lang="GRC">
  <body>
    <div type="book" n="B03">
      <div type="chapter" n="B03K10">
        <ab n="B03K10V2UZ">
          <app n="B03K10V2UZ">
            <rdg n="1" wit="f13">ἔλεγεν</rdg>
            <rdg n="1-f1" type="defective" wit="P75">ἔλεγεν</rdg>
            <gap/>λέγει</ab>
          </app>
        </div>
      </div>
    </body>
  </text>
</TEI>
```
Collation

reading_support = {
    "f13": "1",
    "P75": "1-f1",
    "f1": "2"
}
The Local Stemma

- The basic unit of comparison
- One for each variation unit
- A graphical representation of our judgments of readings
- Kurt Aland’s “local genealogical” principle

\[
\begin{array}{c}
3\text{Jo }1:1/2 \\
a: \text{o} \\
b: -
\end{array}
\]

\[
\begin{array}{c}
3\text{Jo }1:2/18 \\
a: \kappa\alpha\theta\omicron\varsigma \\
b: \kappa\alpha\iota \kappa\alpha\theta\omicron\varsigma \\
c: \kappa\alpha\theta\omicron\varsigma \kappa\alpha
\end{array}
\]
The Local Stemma

- Some are more complicated
  - *defective* readings (e.g., misspellings, reconstructions)
  - *orthographic* readings (e.g., regional differences)
  - *split* attestations of the same reading (coincidental emergence)
  - *ambiguous* readings (can be reconstructed as more than one reading)
- Some of these may be collapsed with other substantive readings

<table>
<thead>
<tr>
<th>3Jo 1:4/22-26</th>
</tr>
</thead>
<tbody>
<tr>
<td>a: εν αληθεια περιπατουντα</td>
</tr>
<tr>
<td>af: εν αληθεια περιπατουντο</td>
</tr>
<tr>
<td>b: εν τη αληθεια περιπατουντα</td>
</tr>
<tr>
<td>c: περιπατουντα εν αληθεια</td>
</tr>
<tr>
<td>d: τη αληθεια περιπατουντα</td>
</tr>
</tbody>
</table>
Computationally, just a directed graph.

```
<graph type="directed">
  <node n="a" />
  <node n="af" />
  <node n="b" />
  <node n="c" />
  <node n="d" />
  <node n="d2" />
  <node n="zw-a/b" />
  <arc from="a" to="af" />
  <arc from="a" to="b" />
  <arc from="a" to="c" />
  <arc from="a" to="d" />
  <arc from="a" to="zw-a/b" />
  <arc from="b" to="zw-a/b" />
</graph>
```
The Local Stemma

- Relationships between readings are determined by checking for a path between them
  - $a = b$ (agreement): path of length 0
  - $a > b$ (prior): path of length $> 0$ from $a$ to $b$
  - $a < b$ (posterior): path of length $> 0$ from $b$ to $a$
  - NOREL (no relationship): no path from $a$ to $b$, but both have a common “ancestor” reading

\[
\begin{align*}
  &a \\
  &\downarrow \\
  &b \\
  &\downarrow \\
  &c \\
  &\downarrow \\
  &d
\end{align*}
\]

Acts 2:5/26-30

- $a$: υπο τον ουρανον
- $b$: υπο των ουρανων
- $c$: υπο του ουρανου
- $d$: υπ ουρανον
The Local Stemma

- UNCL (unclear): same as NOREL, but no common ancestor (reserved for when we don’t know where a reading fits in the stemma)
- We say that one reading explains another if
  - it is the same reading (explanation by agreement), or
  - there is a path of length 1 from it to the other reading
- Lacunae do not have to be explained, and they cannot explain readings

### 3Jo 1:13/24-26

- a: σοι γραφειν
- b: γραφειν σοι
- c: σοι γραψαι
- cf: σοι σοι γραψαι
- d: γραψαι σοι
- e: γραψαι
- f: –
• For the CBGM’s purposes, a witness is a sequence of readings
• Typically, the text of a known manuscript, minus the material baggage (date, provenance, etc.)
  • “How texts relate” ≠ “How manuscripts relate”

<table>
<thead>
<tr>
<th></th>
<th>3Jo 1:1/2</th>
<th>3Jo 1:1/6</th>
<th>3Jo 1:1/8</th>
<th>3Jo 1:15/23</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA 69</td>
<td>a</td>
<td>afl</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>GA 1739</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td>GA 2243</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
</tbody>
</table>
• Versions and fathers can also be treated as witnesses
• But back-translation may be ambiguous, and patristic citations may be “lacunose”
The relationship of two witnesses is the overall pattern of the relationships of their readings where both are extant.

The cost of a genealogical relationship is the number of explained readings that are not agreements (so the cost in the example below is 1).
It is convenient to encode genealogical relationships with bitmaps

\[
\begin{align*}
\text{agree} &= [1, 0, 0, 0, 0] \\
\text{prior} &= [0, 1, 0, 0, 0] \\
\text{posterior} &= [0, 0, 1, 0, 0] \\
\text{norel} &= [0, 0, 0, 1, 0] \\
\text{uncl} &= [0, 0, 0, 0, 1] \\
\text{expl} &= [1, 0, 1, 0, 0] \\
\text{cost} &= 1
\end{align*}
\]
For $d$ units and $n$ witnesses, $\sim n^2 d$ comparisons as one-time work

The `compare_witnesses` module (below) presents this computed data
• Potential ancestor = “more prior than posterior readings”
• *Textual flow* is a tool for helping us revise our judgments in a local stemma

• *Not* a global stemma (our ultimate goal), but still important
Textual Flow at a Variation Unit

• How do we find a given witness’s *textual flow ancestor*?
• We specify a *connectivity limit* $\kappa$ (i.e., a radius of “close-enough” neighbors)
• Then, for each witness:
  1. List its potential ancestors, sorted from most agreement to least
  2. If one of the first $\kappa$ has the same reading at this unit, then select it
  3. If not, then choose the first (non-lacunose) potential ancestor
• Core idea: use *general relationships* (between witnesses) to find *specific relationships* (between readings in a local stemma)
• Often, we just want to know the textual flow for witnesses with a specific reading

• (Numbers on edges represent the rank of the closest potential ancestor with the same reading, if it’s not 1)
• We can use it to evaluate alternate hypotheses about the initial text (A)
Textual Flow for a Variant Reading

3Jo 1:13/24-26a
Con = 10
• Or, we can look only at the parts of textual flow where a reading gets changed to find the most likely sources of unexplained readings (e and f)
Textual Flow for a Variant Reading

3Jo 1:13/24-26

a: σοι γραφεῖν
b: γραφεῖν σοι
c: σοι γράψαι
cf: σοι σοι γράψαι
d: γράψαι σοι
e: γράψαι
f: –
• Between coherence (a form of external evidence) and internal evidence, we can attempt to explain previous unexplained readings

• A necessary step for our ultimate goal of constructing a global stemma
• The *substemma* of a witness is the portion of the global stemma consisting of the witness and its ancestors in the stemma

• Requirement: *every* extant reading in the witness must be explained by a reading in at least one of its ancestors
• A witness may have multiple valid substemma (i.e., ones that explain all of its readings), but some are better than others
• Two of the CBGM’s methodological assumptions are important here:
  3. Scribes typically used fewer sources rather than many.
  4. Scribes typically used closely related sources rather than distant ones.
• A balancing act: the substemma \{L938\} is more parsimonious, but may not explain as many readings by agreement
Finding a (Good) Substemma

- Also called *substemma optimization*
- For *n* potential ancestors, a *weighted set cover* problem with *n* sets (and *2^n* − 1 combinations to check!)

<table>
<thead>
<tr>
<th>Substemma</th>
<th>Variation Units Explained</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A}</td>
<td>✓  ✓  ✓  ✓  ✓</td>
<td>4</td>
</tr>
<tr>
<td>{B}</td>
<td>✓  ✓  ✗  ✗</td>
<td>1</td>
</tr>
<tr>
<td>{C}</td>
<td>✗  ✓  ✓  ✓  ✓</td>
<td>2</td>
</tr>
<tr>
<td>{A, B}</td>
<td>✓  ✓  ✓  ✓  ✓</td>
<td>4+1=5</td>
</tr>
<tr>
<td>{A, C}</td>
<td>✓  ✓  ✓  ✓  ✓</td>
<td>4+2=6</td>
</tr>
<tr>
<td>{B, C}</td>
<td>✓  ✓  ✓  ✓  ✓</td>
<td>1+2=3</td>
</tr>
<tr>
<td>{A, B, C}</td>
<td>✓  ✓  ✓  ✓  ✓</td>
<td>1+2+4=7</td>
</tr>
</tbody>
</table>
Finding a (Good) Substemma

- If a witness has many potential ancestors, then checking all $2^n - 1$ possible substemmata by brute force is prohibitive.

- The *branch-and-bound* heuristic (pictured left) finds all minimum-cost substemmata quickly in practice.

- Easily adapted to find all substemmata within a given cost.
• In practice, the optimize_substemmata module finds all desired substemmata in seconds.

Substemmata for witness W1 = 5 (116 extant passages):

<table>
<thead>
<tr>
<th>ANCESTORS</th>
<th>COST</th>
<th>AGREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>623, 2344</td>
<td>9</td>
<td>113</td>
</tr>
<tr>
<td>623, 69</td>
<td>9</td>
<td>112</td>
</tr>
<tr>
<td>623, 33</td>
<td>9</td>
<td>112</td>
</tr>
<tr>
<td>623, 326</td>
<td>10</td>
<td>113</td>
</tr>
<tr>
<td>623, 2541</td>
<td>10</td>
<td>113</td>
</tr>
<tr>
<td>623, 1890</td>
<td>10</td>
<td>112</td>
</tr>
<tr>
<td>623, 049</td>
<td>10</td>
<td>112</td>
</tr>
<tr>
<td>623, 2186</td>
<td>10</td>
<td>112</td>
</tr>
<tr>
<td>623, 307</td>
<td>10</td>
<td>112</td>
</tr>
<tr>
<td>623, 400</td>
<td>10</td>
<td>112</td>
</tr>
<tr>
<td>623, 429</td>
<td>10</td>
<td>112</td>
</tr>
<tr>
<td>623, 453</td>
<td>10</td>
<td>112</td>
</tr>
<tr>
<td>623, 918</td>
<td>10</td>
<td>112</td>
</tr>
<tr>
<td>623, 180</td>
<td>10</td>
<td>112</td>
</tr>
</tbody>
</table>
The Global Stemma

- Just as the local stemma relates readings, the *global stemma* relates witnesses
- Combination of all substemmata into a single graph
• But every reading in every local stemma except the initial one must be explained by another reading
• Otherwise...
• If we “complete” every local stemma (and ignore or manually account for super fragmentary witnesses) ...
• If we “complete” every local stemma (and ignore or manually account for super fragmentary witnesses) ...
• How is this different than a textual flow diagram?
  • A witness can have more than one ancestor
  • All readings in a witness must be explained by readings in its ancestor(s)
  • More computationally intensive, so takes a bit longer to produce
• The `open-cbgm` library (my implementation of the CBGM, based on these principles) is freely available at https://github.com/jjmccollum/open-cbgm, and the standalone command-line utility is available at https://github.com/jjmccollum/open-cbgm-standalone
  • Supported on Windows, Mac, and Linux
• The INTF’s official implementation (using a Docker container) is now also available (download and instructions at http://ntvmr.uni-muenster.de/intfblog/-/blogs/download-the-cbgm-docker-container)


