CollateX and the formalization of textual Criticism
Bridging the gap between automated collation and edited critical texts

Gioele Barabucci, Franz Fischer
Cologne Center for eHumanities

Antwerpen 2016-10-04
Franz Fischer
Gioele Barabucci
Feedback wanted!

- Interesting for your research?
- Was it too complicated? Shallow?
- Something is deeply flawed?
- Do you know a good example of X?
- Should I absolutely cite paper Y?
Feedback wanted!
That's the way it goes

Lorem iBse
Loreri ipsum
Lorē ip.

Lorem <line>ipse</line>
Lorem ipsum
<abbr>Lorē</abbr> ip.

M1: Lorem (M1, M2, M3)
w2: ipse (M1, M3; M2: ipse [Spanish dialect])

M2: ipse, Spanish dialect

1 Lorem ipsum

1 M2: ipse, Spanish dialect
The problem
The problem

Lorem iBse
Loreri ipsum
Lorē ip.

Lorem <sic>ipse</sic>
Lorem ipsum
<abbr>Lorē</abbr> ipsum

w1: Lorem (M1, M2, M3)
w2: ipsum (M1, M3; M2: ipse [Spanish dialect])

1 M2: ipse, Spanish dialect
More problems

- Search?
- “Massive traditions”
- Guidelines on how to create volume 2?
- Third-party reviews?
The root cause

“The problem is not due to the implementation of the tools being incomplete.

The problem lies in the fragile theoretical foundations upon which these tools are realized. A tool cannot provide a list of meaningful variants if it does not know how to tell apart meaningful variants from non-meaningful variants. In turn, a tool cannot know how to identify a variant as meaningful if the editor has not explained it.

And, under the theoretical frameworks used by current tools, the editor has no way to explain the tool what are the rules to consider a variant meaningful and under which circumstances.
What is needed 1/3

A shared formalization

› define the basic elements they are working with
  • Letter? Hieroglyphs? Words? entire sentences? XML nodes?
› how to identify and classify or group these basic elements
  • which words are adjectives, which are names of people
› ...

What is needed 2/3

A shared formalization

› ... 

› what are the known editing phenomena and how they can be detected

• an homeoteleuton in B with respect to A is detected by the absence in B of words that happens to be included in A between two words that end with the same letters

› which classes of editing phenomena are important and which are not

• orthographic variations of the names of kings = IMPORTANT
• other orthographic variations in adjectives = NOT IMPORTANT
What is needed 3/3

A shared formalization

› ... 

› rules on how certain classes of editing phenomena influence the critical edition

• if document A contains a sentence similar to another found in B, but the sentence in A has been truncated due to homeoteleuton

  => then A must be a descendant of B in the stemma codicum

• all orthographic variants of the names of kings must appear in the critical apparatus
Formalization

- Thorpe, published by SHARP
- Machine-enforceable editorial guidelines
  lead to consistent editing style
- Requires explicit assumptions and methods
  allows document the process of creating the critical edition
- Requires identification of scholarly-relevant editing phenomena
  are based on
  - Semi-automatic generation of the critical apparatus
  - Semi-automatic generation of the stemma codicum
  - Advanced and extremely detailed queries in big corpora
- Together allow make methods comparable
Proposed formalization

- Based on
  - UniDM (Universal Delta model)
  - + “structured changes”
  - + detection rules
What is a document made of?

- CMV+P document model
- makes multiple abstraction levels coexist
- strings, trees, lattices, graphs... everything goes
- pointers to elements at specific levels of abstraction
Structured changes, deltas

“in excelsis Deo” → “ad excelsis dei”

DEL(“i”), DEL(“n”),
DEL(“D”), DEL(“o”),
ADD(“a”), ADD(“d”),
ADD(“d”), ADD(“i”)
The concept of structured changes

DEL(x, idx1)
ADD(y, idx2)  \[\text{with } idx1 == idx2\]

REPLACE(x, y, idx)

REPLACE(x, y, idx) \[\text{with } y == \text{lower-case}(x)\]

TO-LOWERCASE(x, idx)
The concept of structured changes

DEL(x, idx1)
ADD(y, idx2)
with idx1 == idx2

DEL('D', 13)
ADD('d', 13)

REPLACE(x, y, idx)
with y == lower-case(x)

REPLACE('D', 'd', 13)

TO-LOWERCASE(x, idx)

TO-LOWERCASE('D', 13)
Detection rules

\[ \exists c_1, c_2 | \]
\[ c_1 . time < c_2 . time \wedge \]
\[ c_1 . op = \text{DEL} \wedge c_2 . op = \text{ADD} \wedge \]
\[ c_1 . idx = c_2 . idx \]
\[ \Rightarrow \]
\[ c_n = \text{REPLACE} ( c_1 . data, c_2 . data, c_1 . idx ) \]
Possible operations

- CHANGE-OF-VALUE
- DELETION-FOR-SDMAM
- ORTOGRAPHIC-ADJUSTMENT-LTN-SPAIN-1400
- TRANSLITERATION-LTN-TO-GRC
- NAME-REPLACEMENT
- SYNONYM-REPLACEMENT
- ANTONYM-REPLACEMENT
- UNDO

Any classification that the editor can describe
Things that become possible

- Mode advanced collation algorithms
- Better collation output
- Feedback
- Automated generation of the critical apparatus (for the “boring” parts that the editor does not want to do manually)
- Test of conformance between stated methodology, used rules and produced apparatus
- Search for editing phenomena in big corpora
Thank you!

Q?