Fragmentary Texts and Digital Collections of Fragmentary Authors

Monica Berti
“I Frammenti degli Storici Greci”
Dipartimento di Antichità e Tradizione Classica
Università di Roma Tor Vergata, Italia

Marco Büchler
Natural Language Processing Group
Institute of Mathematics and Computer Science
University of Leipzig, Germany

Digital Classicist 2010 Summer Seminar Programme
Institute of Classical Studies, London, 30th July 2010
What is a fragment?  
(*Oxford English Dictionary, s.v. fragment*)

- a part broken off or otherwise detached from a whole
- a part remaining or still preserved when the whole is lost or destroyed
- an extant portion of a writing or composition which as a whole is lost
- a portion of a work left uncompleted by its author
Different kinds of fragments

- material fragments
- textual fragments
material fragments

= physical remains of ancient evidence

reconstruction of the monument
textual fragments (1)

textual fragments
= material fragments bearing textual evidence

→ surviving broken off pieces of ancient writings
In *The Foundings*, Hellanicus says that beer is made also of rye; he writes as follows: ‘They drink beer made of rye, as the Thracians drink it made of barley’. Hecataeus, in the second book of his *Description*, after saying of the Egyptians that they were bread-eaters, continues: ‘They grind up the barley to make the drink’. And in *The Description of Europe* he says that the Paeonians drink a beer made from barley, also *parabias*, made from millet, and even fleabane. ‘They also anoint themselves’, he says, ‘with an oil made from milk’. So much for that. (trans. Gulick)
print collections of fragmentary texts

- textual excerpts drawn from many different sources
- excerpts arranged according to various criteria
- length of the excerpts different from one edition to another
- when printed the excerpt gives a false illusion of materiality
- duplication of the same text in multiple editions

Berti – Büchler, Fragmentary Texts and Digital Collections of Fragmentary Authors
representing textual fragments

- construct truly hypertextual editions, including not only excerpts but links to the scholarly sources from which those excerpts are drawn
- create meta-information through an accurate and elaborate semantic markup
- produce meta-editions consisting not only of isolated quotations, but also of pointers to the original contexts from which the fragments have been extracted
- provide scholars with an interconnected corpus of primary and secondary sources of fragments that also includes critical apparatuses, commentaries, translations, and modern bibliography on ancient texts
• textual fragment as **hypertext**
  ◦ a text derived from another text and interconnected to many other different typologies of texts

• textual fragment as **multitext**
  ◦ the result of a work of stratification of manuscripts and scholarly conjectures
http://demo.fragmentarytexts.org

Berti - Büchner, Fragmentary Texts and Digital Collections of Fragmentary Authors
Bibliography


What is text mining?

„Process of deriving high-quality information from text“
(Feldman & Sanger 2006)
“Text Mining is 'big reading'.”
(Craig Bellamy on Twitter, Jul. 5th 2010)
Classes of text mining tools

- Unsupervised
- Supervised
- Bootstrapping
- Pattern
- Manual
Tasks of extracting and collecting fragmentary authors

- **Task 1**: Associations between person and work names
- **Task 2**: Extraction of fragments of an author
- **Task 3**: Finding new quotations and parallel texts
- **Task 4**: Expansion of the fragments' set
Task 1: Workflow person name extraction

- **Step 1**: Extraction of candidates by pattern such as VN VN, VN ETH, VN LOC
- **Step 2**: Resolving morphological dependencies using Morpheus
- **Step 3**: Statistical evidence criterion
- **Step 4**: Generating a similarity graph of those candidates and building valid concept classes
- **Step 5**: Applying validated patterns on text in order to extract less frequent occurrences
- **Step 6**: Iterating step 2 - 5
Task 1: Some results of the PN extractor

- **Step 1**: Extraction of candidates by pattern such as Ἑλλάνικος Λέσβιος (VN ETH)
- **Step 2**: Resolving morphological dependencies
  - Removing candidates like Ἑλλάνικος Ἀκουσιλάω VN VN
- **Step 3**: Statistical evidence criterion like min freq is 4.
- **Step 4**: Generating a similarity graph of those candidates and building valid concept classes e.g.
  Ἑλλάνικος Λέσβιος (VN ETH)
  Ἑλλάνικος ὁ Λέσβιος (VN ZN ETH)
- **Step 5**: Applying validated patterns on text in order to extract less frequent occurrences
  - Ἑλλάνικος τε ὁ Λέσβιος
  - Ἑλλάνικος δὲ ὁ Λέσβιός
  - Λέσβιος Ἑλλάνικος
  - ...
  - Overall after 1 iteration 16 different versions of Hellanicus of Lesbos
### Task 3: Finding new quotations and parallel texts: pseudo algorithm

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( V = \text{segment} _ \text{corpus}(C) ) with ( v_1, v_2, ..., v_n \in V ), ( \cup v_i = C ) and ( v_i \neq v_j )</td>
</tr>
<tr>
<td>2</td>
<td>( \text{for each} \ v \in V )</td>
</tr>
<tr>
<td>3</td>
<td>( F_i = \text{train} _ \text{features}(v_i) )</td>
</tr>
<tr>
<td>4</td>
<td>( \text{for each} \ v \in V )</td>
</tr>
<tr>
<td>5</td>
<td>( \text{for each} \ f_k \in F_i )</td>
</tr>
<tr>
<td>6</td>
<td>( e_i = (v_i, v_j) \in E ) = select all ( v_j ) containing feature ( f_k )</td>
</tr>
<tr>
<td>7</td>
<td>( \text{for each} \ e_i \in E )</td>
</tr>
<tr>
<td>8</td>
<td>( s_i = \text{scoring}(e_i = (v_i, v_j) \in E; F_i; F_j) )</td>
</tr>
<tr>
<td>9</td>
<td>( \text{if} (s_i &lt; \text{threshold}) { E = E \setminus { e_i } } )</td>
</tr>
</tbody>
</table>

**Training**

**Linking**

**Scoring**
Task 3: Finding new quotations and parallel texts: Types of Completeness

Extraction of fragmentary authors

• *String approaches:*
  – GST
  – Letter n-grams
• *Syntactic approaches ((literal) quotations):*
  – N-gram expansion
  – Word n-grams
  – Distance based co-occurrences
• *Semantic approaches (parallel texts):*
  • Semantic clustering
  • Semantic graph based approach(es)
  • Relations of contrastive semantics
  • Radius retrieval
• *More complex approaches:*
  – DCT
  – Winnowing
## Task 2: Extraction of fragments: Role of named entities

<table>
<thead>
<tr>
<th>Graph properties</th>
<th>Complete graph</th>
<th>w_id&gt;=100 &amp;&amp; freq(word)&gt;1</th>
<th>w_id&gt;=300 &amp;&amp; freq(word)&gt;1</th>
<th>w_id&gt;=500 &amp;&amp; freq(word)&gt;1</th>
<th>Named Entities</th>
<th>Normalised Named Entities</th>
<th>Normalised Text and Named Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>538,572</td>
<td>388,929</td>
<td>363,359</td>
<td>353,618</td>
<td>1,149</td>
<td>4,487</td>
<td>2,178</td>
</tr>
<tr>
<td>Number of co-occurrences</td>
<td>57,762,474</td>
<td>34,818,138</td>
<td>25,615,956</td>
<td>21,004,538</td>
<td>15,436</td>
<td>126,188</td>
<td>152,856</td>
</tr>
<tr>
<td>Number of significant co-occurrences</td>
<td>30,382,422</td>
<td>21,739,476</td>
<td>17,687,582</td>
<td>15,462,940</td>
<td>14,876</td>
<td>69,858</td>
<td>84,124</td>
</tr>
<tr>
<td>Percentage</td>
<td>0.53</td>
<td>0.62</td>
<td>0.69</td>
<td>0.74</td>
<td>0.96</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>Average degree</td>
<td>56.41</td>
<td>55.90</td>
<td>48.68</td>
<td>43.73</td>
<td>12.95</td>
<td>15.57</td>
<td>38.62</td>
</tr>
<tr>
<td>Argumentation trail properties</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of trails</td>
<td>&gt; 10⁹</td>
<td>&gt; 10⁹</td>
<td>&gt; 10⁹</td>
<td>&gt; 10⁹</td>
<td>361,094</td>
<td>7,958,240</td>
<td>3,087,581</td>
</tr>
<tr>
<td>Average degree</td>
<td>15.34</td>
<td>9.93</td>
<td>7.70</td>
<td>6.79</td>
<td>7.03</td>
<td>7.77</td>
<td>9.93</td>
</tr>
<tr>
<td>Average degree of internal node (trail length 2)</td>
<td>31.34</td>
<td>21.08</td>
<td>14.33</td>
<td>11.45</td>
<td>7.02</td>
<td>10.15</td>
<td>12.31</td>
</tr>
<tr>
<td>Average degree of internal node (trail length 3)</td>
<td>301.38</td>
<td>362.56</td>
<td>285.86</td>
<td>231.39</td>
<td>55.66</td>
<td>76.06</td>
<td>81.86</td>
</tr>
</tbody>
</table>

---

*Berti – Büchler, Fragmentary Texts and Digital Collections of Fragmentary Authors*
Task 2: Extraction of fragments: Possible ways?

- **Option 1**: Statistical based
  - Supervised

- **Option 2**: Pattern based
  - Pattern

- **Option 3**: Completely different?
  - Unsupervised
Hellanicus in *The Foundings* says that beer is made also of rye; he writes as follows: ‘They drink beer made of rye, as the Thracians drink it made of barley’. Hecataeus, in the second book of his *Description*, after saying of the Egyptians that they were bread-eaters, continues: ‘They grind up the barley to make the drink’. And in *The Description of Europe* he says that the Paeonians drink a beer made from barley, also *parabias*, made from millet, and even fleabane. ‘They also anoint themselves’, he says, ‘with an oil made from milk’. So much for that. (trans. Gulick)
'They drink beer made of rye, as the Thracians drink it made of barley'.

the Paeonians drink a beer made from barley, also *parabias*, made from millet, and even fleabane.

‘They also anoint themselves’, he says, ‘with an oil made from milk’.

Some „significance“ related properties:

- **tf.idf**: Except „Thracian“ and „Paeonians“ all other words have a term weight of 0 (function words) or are weak content words.
- **Difference analysis**: no discriminating words
- **Log–likelihood ratio**: no discriminating words


Is there any measurable content in this fragments?
Definition Co-occurrences:
- Common occurrence of at least two objects/events within a dedicated window
  » Possible windows in Classical Studies: line, sentence, paragraph, document, author, century

Motivation:
- Psycholinguistic experiments: Given a word: What is the first word test persons answer?

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Response Prob.</th>
<th># of Prob.'s</th>
<th>Co-occurrence</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>butter</td>
<td>Bread</td>
<td>60</td>
<td>Bread</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>40</td>
<td>Cheese</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Milk</td>
<td>32</td>
<td>Sugar</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Margarine</td>
<td>27</td>
<td>Milk</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Cheese</td>
<td>20</td>
<td>Margarine</td>
<td>22</td>
</tr>
<tr>
<td>Fat</td>
<td>Farina</td>
<td>16</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>yellow</td>
<td>Eggs</td>
<td>14</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>Bread and butter</td>
<td>Pound</td>
<td>8</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>Box / can</td>
<td>Meat</td>
<td>6</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>eat</td>
<td></td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overview of the approach (Humanists friendly)

- Observation:
  - Every word has a contextual profile (based on co-occurrences) in which it is typically used

Source: http://corpora.informatik.uni-leipzig.de
An example of Data Mining: Relation between beer and diapers I

There is a story that a large supermarket chain, usually Wal-Mart, did an analysis of customers' buying habits and found a statistically significant correlation between purchases of beer and purchases of nappies (diapers in the US). It was theorized that the reason for this was that fathers were stopping off at Wal-Mart to buy nappies for their babies, and since they could no longer go down to the pub as often, would buy beer as well. As a result of this finding, the supermarket chain is alleged to have the nappies next to the beer, resulting in increased sales of both.

Contrastive relation: (beer, diapers)

Source: http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html
An example of Data Mining: Relation between beer and diapers II

There is a story that a large supermarket chain, usually Wal-Mart, did an analysis of customers' buying habits and found a statistically significant correlation between purchases of beer and purchases of nappies (diapers in the US). It was theorized that the reason for this was that fathers were stopping off at Wal-Mart to buy nappies for their babies, and since they could no longer go down to the pub as often, would buy beer as well. As a result of this finding, the supermarket chain is alleged to have the nappies next to the beer, resulting in increased sales of both.

**Latent relation:** (beer, diapers)

**Context:** fathers, stopping off, Wal-Mart, could no longer go down to the pub as often

**Result of this relation:** nappies next to the beer

What should be the result?

father

could no longer go
down to the pub as
often

stop off

Unexpected & contrastive relation

Relevant context

beer

diaper

Berti – Büchler, Fragmentary Texts and Digital Collections of Fragmentary Authors
Some examples (if the original text is still existent)

• Relation of (Ὀδόμαντοι, πέος)
  – engl.: (Odomastai (a folk in Thrace), penis)
  – Context: Found in an Ancient comedy (Aristophanes, 5th c. BC)

• Relation of (κοπρολόγος, ψάλτρια) – engl.: (shit collector, dancing girl)
  – Context: ἀστυνόμοι – engl.: (protecting the city, public festivals)
  – Found in Aristotle (4th c. BC)
Results

• Lots of *contrastive semantic relations* can be found (manual evaluation is still in progress)

• But depending on text sort:
  – Other clusters can be found additionally
    • As shown in examples *comedy*
    • Sarcasm
    • Cynicism
    • *Artificial ambiguity* like „Michael Schumacher the red king“ (translated from a German corpus)
  • *Scope to gnomology & philosophical texts*
Contrastive semantic relations from a bird's eye view

• What did I do with the example of beer and diaper?
  – If I would write it down: A semantically textual reference.
• Is there a relation between contrastive relations and textual reuse?
  – Clearly, yes.
  – First evaluation results: More than 90% of the latent relations (Settings: minimum frequency: 2, Except the contrastive relation itself not more than 2 additional associations)

Focus:
___ Here: Why is knowledge reused?
Nobody would reuse something like: „Milk is white and good for you“.
Why: It's well-known.
Why was it reused?

‘They drink beer made of rye, as the Thracians drink it made of barley’.

the Paeonians drink a beer made from barley, also parabias, made from millet, and even fleabane.

‘They also anoint themselves’, he says, ‘with an oil made from milk’.

Dissimilarities in the contextual usage (TLG):
• (milk,oil): 72%

• (fleabane, millet): 92%, (parabias, millet): 97%, (fleabane, parabias): 94%, (barley, fleabane): 94%, ...

• (rye, barley): 80%
Further work: Semantic spaces
How can Marco benefit from work with Monica?

- **NEW QUESTION:** Shannon's Noisy Channel Theorem (witness):
  - Transmitter $x$ (noisy) Channel $y$ Receiver

- **NEW QUESTION:** Not HOW but why is something quoted?
  - Contrastive semantics

- **EVALUATION:** How to evaluate text reuse & knowledge transfer?
  - Collection of fragmentary authors as highly reviewed *Gold Standard*
Summary

To be, or not to be, that is the question

Hamlet, Shakespeare