**Non-Literal Text Reuse in Historical Texts: An Approach to Identify Transformation and Its Application to Bible Reuse**

Maria Moritz¹, Andreas Wiederhold¹, Barbara Pavlek², Yuri Bizzoni³, Marco Büchler¹

**Research Questions**

**Motivation**

Text reuse is the spoken and written repetition of text across time and space. It can be a quotation, an allusion or translation. Detection methods of historical text reuse are needed in different scholarly fields, e.g. to detect redundancies in digital libraries, to trace transmissions of historical thought or to identify fragmentary authors.

However, text is often modified during the reuse process, which makes the detection challenging. Therefore, we analyze the non-literal share in historical text reuse to obtain an understanding of the requirements for contemporary detection methods.

1. What is the extent of non-literal reuse in our datasets?
2. How is the non-literally reused text modified in our datasets?
3. How can linguistic resources support the discovery of non-literal reuse?
4. What are the limitations of an automated classification approach relying on linguistic resources?

**Data**

<table>
<thead>
<tr>
<th>Region</th>
<th>reseach programmes</th>
<th>reused</th>
<th>unclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pref 10:1</td>
<td>a) framework (textual similarities), b) word order and c) word-sense substitution</td>
<td>199</td>
<td>162</td>
</tr>
<tr>
<td>Pref 10:2</td>
<td>a) framework (textual similarities), b) word order and c) word-sense substitution</td>
<td>199</td>
<td>162</td>
</tr>
</tbody>
</table>

**Clement of Alexandria**

Christian theologian from the 2nd century.

**Bernard of Clairvaux**

French abbot from the 12th century.

**Methodology**

1. We define operations (OPs) reflecting literal reuse and semantic replacements (see above).
2. Our algorithm looks for identical & similar words and for morphological & semantic changes (see top right).
3. We apply both to our datasets using the Ancient Greek WordNet (see an example at the right).
4. We complement the automated approach by a manual analysis of a sample (60 & 100 instances resulting in 192 & 224 replacement operations) to find limitations of our automated approach.

**Results**

**Figure 1:** Ratios of operations in reuse instances, literal: NOP, lem, lower, etc.; nonlit: syn, hyper, etc.

The reuse is significantly non-literal and conceptualization might be preferred over stemming or semantic relations in the same POS category only.

**Research Questions**

**Future plans**

A more comprehensive study will strengthen the findings. For example using larger reuse datasets and additional languages, such as reflecting and non-infecting languages.

A smarter automated approach for deriving an original text excerpt can be learned edit scripts, such as undertaken by Kehrer (2014) also considering the movement of reuse except within the reuse or the syntactical tree.

Deeper analyses of reuse statistics might be supported by the semantic relations that are presented in word nets.

**We aggregate**

Bibliae Latine versiones. (65.5K Biblical Greek entries; 315K Latin entries)

Classical Latin Text Tool Kit (CLTK) (Ahnrich et al., 2014)

364K Ancient Greek & 270K Latin entries

Greek New Testament of the Society of Biblical Literature & St. Stevens

59.5K word lemma-pairs

Ancient Greek WordNet (Bizzoni et al., 2014; Mouzou; 2009)

Nine datasets of which 25K contain Ancient Greek and 27K contain Latin words.

**Greek Old Testament**


Classical Latin Text Tool Kit (CLTK) (Ahnrich et al., 2014)

364K Ancient Greek & 270K Latin entries

Greek New Testament


Latin Bible


**We acknowledge**

Laurence Mehler for providing the datasets and for valuable advice on their content, The German Federal Ministry of Education and Research for funding the work (grant 01UL15059).

**References**

1. Ling (2011). "How good is the Gale bloom" (1895).

Language resources support the identification of reuse components. In our datasets, co-hyponyms are often used to rephrase an idea. Many-to-many relationships show that meanings can be hidden in structural or expert knowledge.