NON-LITERAL TEXT REUSE IN HISTORICAL TEXTS: AN APPROACH TO IDENTIFY REUSE TRANSFORMATIONS AND ITS APPLICATION TO BIBLICAL REUSE


Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, Marco Büchler

EMNLP 2016

October 13-14 2016, University of Hildesheim
• PhD student at the Georg-August University School of Science (GAUSS), Program in Computer Science (PCS)

• Topic: Historical Text Reuse Style (the way text is reused) Investigation

• Supervisors:
  • Prof. Dr. Ramin Yahyapour
  • Prof. Dr. Dieter Hogrefe
  • Dr. Marco Büchler
1. Introduction

2. Methodology

3. Results

4. Conclusion and Future Work
INTRODUCTION
Text Reuse:

- spoken and written repetition of text across time and space.

For example:

- citations, allusions, translations.

Detection methods are needed in different scholarly fields.

- They help to ensure clean libraries or identify fragmentary authors.

Text is often modified during the reuse process.
Paraphrasing and **non-literal reuse challenges** many approaches:

- **Alzahrani et al. (2012)**
  - study n-gram-, syntax-, and semantic-based detection approaches;
  - they find: as soon as reuse is slightly modified (words changed) most approaches fail.

- **Barrón-Cedeño et al. (2013)**
  - experiment with paraphrasing to improve plagiarism detection;
  - they found that complex paraphrasing with a high density challenges plagiarism detection, and
  - that lexical substitution is the most frequent plagiarism technique.
Not only Alzahrani et al.’s but most NLP research focuses on English, although many inflecting languages exist. Ancient languages are particularly problematic as they come with:

• **variants** due to transmission.
• **incomplete** testimonials.
• **diverse** reuse.

One solution: Reuse Style Investigation.

• i.e., how is reuse generally transferred and how literal is it?
Inspired by Shannon’s noisy-channel (Shannon, 1949) & the Kolmogorov Complexity (Li and Vitányi, 2008), we study Ancient Greek and Latin text reuse to understand how text is transferred.

- We identify operations that characterize word changes.
- We show how linguistic resources can help detect non-literal reuse.
- We complement the automated approach with a manual analysis.
METHODOLOGY
RESEARCH QUESTIONS

• RQ1. What is the extent of non-literal reuse in our datasets?
• RQ2. How is the non-literally reused text modified in our datasets?
  • RQ2.1. How can linguistic resources support the discovery of non-literal reuse?
  • RQ2.2. What are the limitations of an automated classification approach relying on linguistic resources?
We:

1. **define** operations reflecting literal reuse and semantic replacements.
2. **develop** an algorithm that looks for identical, similar words, morphological changes and semantic changes.
3. **apply** it to our two data-sets (next slide).
4. **manually analyze** a smaller sample of our reuse using further operations.
“ Salvation for the Rich”
Clement of Alexandria
Christian theologian, 2nd cent.
• Known for his retelling of biblical excerpts
• Reuse annotated by Biblindex team (Mellerin, 2014; Mellerin, 2016)
• We obtain 199 verse-reuse-pairs
• Pointing to 15 Bible books

Extracts from 12 works & 2 collections
Bernard of Clairvaux
French abbot, 12th cent.
• Known for his influence on the Cistercian order and his work in biblical studies
• Reuse extracted by Biblindex team (Mellerin, 2014; Mellerin, 2016)
• We obtain 162 verse-reuse-pairs
• Pointing to 31 Bible books
<table>
<thead>
<tr>
<th>more literal</th>
<th>Bible verse</th>
<th>Bernard reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proverbs 18:3</strong></td>
<td><em>impius cum in profundum venerit</em> peccatorum <em>contemnit</em> sed sequitur eum ignominia et obprobrium (When the wicked man is come into the depth of sins, also contempt comes but ignominy and reproach follow him)</td>
<td><em>Impius cum venerit in profundum</em> malorum, <em>contemnit</em> (When the wicked man is come into the depth of evil)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>less literal</th>
<th>Bible verse</th>
<th>Clement reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1Cor 13:13</strong></td>
<td>νυνὶ δὲ μένει πίστις, ἐλπίς, ἀγάπη, τὰ τρία ταῦτα μείζων δὲ τούτων ἢ ἀγάπη (And now remain faith, hope, love, these three; but the greatest of those is love.)</td>
<td>πίστει καὶ ἐλπίδι καὶ ἀγάπη (faith, and hope, and love - in dative case)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ἀγάπην, πίστιν, ἐλπίδα (love, faith, hope - in accusative case)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>μένει δὲ τὰ τρία ταῦτα, πίστις, ἐλπίς, ἀγάπη, μείζων δὲ ἐν τούτωι ἢ ἀγάπη (and remain these three, faith, hope, love; but the greatest among them is love)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>non-literal</th>
<th>Bible verse</th>
<th>Clement reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mt 12:35</strong></td>
<td>ὁ ἀγαθὸς ἄνθρωπος ἐκ τοῦ ἄγαθοῦ θησαυροῦ ἐκβάλλει ἀγαθὰ, καὶ ὁ πονηρὸς ἄνθρωπος ἐκ τοῦ πονηροῦ θησαυροῦ ἐκβάλλει πονηρά. (A good man out of good storage brings out good things, and an evil man out of the evil storage brings evil things.)</td>
<td>Ψυχῆς, τὰ δὲ ἐκτὸς, κἂν μὲν ἡ ψυχὴ χρῆται καλῶς, καλὰ καὶ ταῦτα δοκεῖ, ἐὰν δὲ πονηρῶς, πονηρὰ, ὁ κελεύων ἀπαλλατισθείην τὰ υπάρχοντα ([are whitin the] soul, and some are out, and if the soul uses them good, those things are also thought of as good, but if [they are used as] bad, [they are thought of as] bad; he who commands the renouncement of possessions)</td>
</tr>
</tbody>
</table>
We aggregate:

- **Biblindex’ Lemma Lists**
  - 65,537 Biblical Greek entries
  - 315,021 Latin entries

- **Classical Language Tool Kit (CLTK) (Johnson et al., 2014)**
  - 953,907 Ancient Greek words
  - 270,228 Latin words

- **Greek New Testament of the Society of Biblical Literature\(^1\) & Septuaginta (Rahlfs, 1935a; UPenn) 59,510 word-lemma-pairs**

\(^1\) Logos Bible Software [http://sblgnt.com/about/](http://sblgnt.com/about/)
99K synsets
of which 33K contain Ancient Greek and 27K Latin words
(Bizzoni et al., 2014; Minozzi, 2009)
## Table 1: Operation list for the automated approach

<table>
<thead>
<tr>
<th>operation</th>
<th>description</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOP(reuse_word, orig_word)</td>
<td>Original and reuse word are equal.</td>
<td>NOP(maledictus, maledictus)</td>
</tr>
<tr>
<td>upper(reuse_word, orig_word)</td>
<td>Word is lowercase in reuse and uppercase in original.</td>
<td>upper(kai, Kai) – in Greek</td>
</tr>
<tr>
<td>lower(reuse_word, orig_word)</td>
<td>Word is uppercase in reuse and lowercase in original.</td>
<td>lower(Gloriam, gloriam)</td>
</tr>
<tr>
<td>lem(reuse_word, orig_word)</td>
<td>Lemmatization leads to equality of reuse and original.</td>
<td>lem(penetrat, penetrabit)</td>
</tr>
<tr>
<td>repl_syn(reuse_word, orig_word)</td>
<td>Reuse word replaced with a synonym to match original word.</td>
<td>repl_syn(magnificavit, glorificavit)</td>
</tr>
<tr>
<td>repl_hyper(reuse_word, orig_word)</td>
<td>Word in Bible verse is a hyperonym of the reused word.</td>
<td>hyper(cupit, habens)</td>
</tr>
<tr>
<td>repl_hypo(reuse_word, orig_word)</td>
<td>Word in Bible verse is a hyponym of the reused word.</td>
<td>hypo(dederit, tolet)</td>
</tr>
<tr>
<td>repl_co-hypo(reuse_word, orig_word)</td>
<td>Reused word and original have the same hyperonym.</td>
<td>repl_co-hypo(magnificavit, fecit)</td>
</tr>
<tr>
<td>NOPmorph(reuse_tags, orig_tags)</td>
<td>Case or PoS did not change between reused and original word.</td>
<td>NOPmorph(na, na)</td>
</tr>
<tr>
<td>repl_pos(reuse_tag, orig_tag)</td>
<td>Reuse and original contain the same cognate, but PoS changed.</td>
<td>repl_pos(n, a)</td>
</tr>
<tr>
<td>repl_case(reuse_tag, orig_tag)</td>
<td>Reuse and original have the same cognate, but the case changed.</td>
<td>repl_case(g, d) – cases genitive, dative</td>
</tr>
<tr>
<td>lemma_missing(reuse_word, orig_word)</td>
<td>Lemma unknown for reuse or original word.</td>
<td>lemma_missing(tentari, inlectus)</td>
</tr>
<tr>
<td>no_rel_found(reuse_word, orig_word)</td>
<td>Relation for reuse or original word not found in AGWN.</td>
<td>no_rel_found(gloria, arguitur)</td>
</tr>
</tbody>
</table>
We manually analyze:

- **60 Ancient Greek & 100 Latin instances**
- **192 & 224 replacements**
- **Using** `ins(word), del(word)` and replacements:
  - NOP, lem, repl_syn,
    repl_hyper, repl_hypo,
    repl_co-hypo
- **We assign morphological categories from Perseus’ tag-set** (Bamman and Crane 2011)
  - **E.g., repl_case_a_g**
    repl_num_s_p

**Table 2: Excerpt from Perseus’ tag-set**

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>first person</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>second person</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>third person</td>
<td>3</td>
</tr>
<tr>
<td>number</td>
<td>singular</td>
<td>s</td>
</tr>
<tr>
<td></td>
<td>plural</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td>dual</td>
<td>d</td>
</tr>
<tr>
<td>tense</td>
<td>present</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td>imperfect</td>
<td>i</td>
</tr>
<tr>
<td></td>
<td>perfect</td>
<td>r</td>
</tr>
<tr>
<td></td>
<td>pluperfect</td>
<td>l</td>
</tr>
<tr>
<td></td>
<td>future perfect</td>
<td>t</td>
</tr>
<tr>
<td></td>
<td>future</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>aorist</td>
<td>a</td>
</tr>
</tbody>
</table>
RESULTS
What is the extent of non-literal reuse in our datasets?

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

**Figure 2:** Ratios of literal overlap between reuse instances and originals.
AUTOMATED APPROACH (RQ2.1)

How is the non-literally reused text modified in our datasets? (RQ2)
How can linguistic resources support the discovery of non-literal reuse? (RQ2.1)

Table 3: Absolute numbers of operations identified automatically.

<table>
<thead>
<tr>
<th></th>
<th>literal</th>
<th>non-literal</th>
<th>unclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NOP</td>
<td>upper</td>
<td>lower</td>
</tr>
<tr>
<td>Greek</td>
<td>337</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Latin</td>
<td>587</td>
<td>0</td>
<td>44</td>
</tr>
</tbody>
</table>
Operations that successfully looked up a lemma:
\[ \text{lem}_\text{success} = \{\text{lem}, \text{syn}, \text{repl}_\text{hyper}, \text{repl}_\text{hypo}, \text{repl}_\text{co-hypo}, \text{no}_\text{rel}_\text{found}\} \], with \text{lem}_\text{missing} representing not found tokens in the lemmata.

\[
\text{cov}_{\text{lem}} = \frac{\sum_{o \in \text{lem}_\text{success}} \text{Occ}(o)}{\sum_{o \in \text{lem}_\text{success} \cup \{\text{lem}_\text{missing}\}} \text{Occ}(o)}
\]

\[
\text{cov}_{\text{AGWN}} = \frac{\sum_{o \in \text{agwn}_\text{success}} \text{Occ}(o)}{\sum_{o \in \text{agwn}_\text{success} \cup \{\text{no}_\text{rel}_\text{found}\}} \text{Occ}(o)}
\]

We obtain a \text{cov}_{\text{lem}} of 0.65 for our Greek and 0.88 for the Latin data-set. And a \text{cov}_{\text{AGWN}} of 0.34 for our Greek and 0.33 for our Latin data-set.

Language resources help to get an idea of reuse components.
QUALITATIVE APPROACH (RQ2.2)

How is the non-literally reused text modified in our data-sets? (RQ2)
What are the limitations of an automated classification approach relying on linguistic resources? (RQ2.2)

Table 4: Exceptions that prevent applying the operations.

<table>
<thead>
<tr>
<th>Exception</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clement</td>
</tr>
<tr>
<td>Word changed to antonym</td>
<td>1²</td>
</tr>
<tr>
<td>Synonym and morphology changed</td>
<td>1</td>
</tr>
<tr>
<td>More than one morphological category changed</td>
<td>1</td>
</tr>
<tr>
<td>Synonym is multi-word expression</td>
<td>3</td>
</tr>
<tr>
<td>Many-to-many</td>
<td>0</td>
</tr>
</tbody>
</table>

²“the God, the good (one)” (Clement) vs. “none is good but the God” (Bible).
³“judged calmly” (Bernard) vs. “fake friend” (Sal 12 18).
• **RQ1.** *The reuse is significantly non-literal*
  
  - Techniques beyond stemming & even semantic closeness are required; suggestion: conceptualizing

• **RQ2.1.** *Our results show*
  
  - the possibility of supporting reuse detection with linguistic resources

• **RQ2.2.** *Qualitative complement*
  
  - especially the exceptions show that reuse detection needs looser relation (multi-to-multi-word) associations or implicit expert knowledge.
CONCLUSION AND FUTURE WORK
We contributed:

• an automated approach to characterize how text is transformed between reuse and original,

• an application of the approach to two text data-sets where reuse was manually identified,

• empirical data based on the automated approach, complemented by a manual identification.
A more comprehensive study could strengthen the findings:

- using larger reuse data-sets,
- additional languages (inflecting vs. non-inflecting),

An automated approach for deriving an original text excerpt would be learning edit scripts (Kehrer, 2014, Chawathe et al., 1996).
THANK YOU!


Old Testament:


New Testament:


Latin Bible:


Clement Edition:


The theme this presentation is based on is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Changes to the theme are the work of eTRAP.