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

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, Marco Büchler
(EMNLP16)
February 17 2017, Groningen, Netherlands
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 to support scholarly work.

- E.g. 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.
The most of current NLP research focuses on English, but many stronger inflecting languages exist. Their computational processing is more challenging. Historical, especially Ancient languages are particularly problematic as they come with:

- variants due to long transmission time,
- incomplete witnesses,
- diverse reuse.

One solution: Reuse Style Investigation.

- I.e., how is reuse transferred, how literal is it, what kind of modification takes place?
• Inspired by
  • Noisy channel model: given a “scrambled” word or sentence, guess the intended version of that sentence (Brill, 2000),
  • Kolmogorov Complexity: describes the length of the shortest program that produces an output string (Li and Vitáni, 2008),
• we study Ancient text reuse to understand how text is transferred.
  • Identify operations to characterize morphological & semantic changes
  • Design an algorithm which applies these OPs to our datasets
  • Transform one text excerpt into another by a minimum OP set
METHODOLOGY
• 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),
   - Identifying the literal overlap in reuse (RQ1),
   - Measuring occurrences of operations by covlem and covAGWN to assess the resources’ coverage for our approach (RQ2.1),
4. Manually analyze a smaller sample of our reuse using further operations. (RQ2.2).
“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

The data was tokenized and punctuation was kept but ignored in the analyses.

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
### Biblical Reuse Examples

<table>
<thead>
<tr>
<th>more literal</th>
<th>Bible verse</th>
<th>Bernard reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proverbs 18 3</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>1Cor 13 13</td>
<td>νυνὶ δὲ μένει πίστις, ἐλπίς, ἀγάπη, τὰ τρία ταῦτα μεῖζων δὲ τοῦτον ἢ ἀγάπη (And now remain faith, hope, love, these three; but the greatest of those is love.)</td>
<td>πίστει καὶ ἐλπίδα καὶ ἀγάπη (faith, and hope, and love - in dative case) ἀγάπην, πίστιν, ἐλπίδα (love, faith, hope - in accusative case) μένει δὲ τὰ τρία ταῦτα, πίστις, ἐλπίς, ἀγάπη· μεῖζων δὲ ἐν τοῦτοις 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>Mt 12 35</td>
<td>ὁ ἀγαθὸς ἄνθρωπος ἐκ τοῦ ἀγαθοῦ θησαυροῦ ἐκβάλλει ἄγαθά, καὶ ὁ πονηρός 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 καλὸς, καλὰ καὶ ταῦτα δοκεῖ, ἐὰν δὲ πονηρῶς, πονηρά, ὁ κελεύων ἀπαλλοτριοῦν τὰ ψυχῆν τα ὑπάρχοντα ([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,tollet)</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>

16/37
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.
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></th>
<th></th>
<th>non-literal</th>
<th></th>
<th></th>
<th>unclassified</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NOP</td>
<td>upper</td>
<td>lower</td>
<td>lem</td>
<td>syn</td>
<td>hyper</td>
<td>hypo</td>
<td>co-hypo</td>
</tr>
<tr>
<td>Greek</td>
<td>337</td>
<td>6</td>
<td>0</td>
<td>356</td>
<td>153</td>
<td>20</td>
<td>14</td>
<td>101</td>
</tr>
<tr>
<td>Latin</td>
<td>587</td>
<td>0</td>
<td>44</td>
<td>102</td>
<td>60</td>
<td>14</td>
<td>28</td>
<td>68</td>
</tr>
</tbody>
</table>
Operations that successfully looked up a lemma:
\[ \text{lem\_success} = \{\text{lem, syn, repl\_hyper, repl\_hypo, repl\_co-hypo, no\_rel\_found}\}, \text{ with } \text{lem\_missing} \text{ representing not found tokens in the lemmata.} \]

\[
\text{COV}_{\text{lem}} = \frac{\sum_{o \in \text{lem\_success}} \text{Occ}(o)}{\sum_{o \in \text{lem\_success} \cup \{\text{lem\_missing}\}} \text{Occ}(o)}
\]

\[
\text{COV}_{\text{AGWN}} = \frac{\sum_{o \in \text{agwn\_success}} \text{Occ}(o)}{\sum_{o \in \text{agwn\_success} \cup \{\text{no\_rel\_found}\}} \text{Occ}(o)}
\]

We obtain a \( \text{cov}_{\text{lem}} \) of \textbf{0.65} for our Greek and \textbf{0.88} for the Latin data-set. And a \( \text{cov}_{\text{AGWN}} \) of \textbf{0.34} for our Greek and \textbf{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>Word changed to antonym</td>
<td>1(^2)  0</td>
</tr>
<tr>
<td>Synonym and morphology changed</td>
<td>1        16</td>
</tr>
<tr>
<td>More than one morphological category changed</td>
<td>1        7</td>
</tr>
<tr>
<td>Synonym is multi-word expression</td>
<td>3        5</td>
</tr>
<tr>
<td>Many-to-many</td>
<td>0        1(^3)</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, such as Bible translations:\(^4\):
  - 27 English Bible translations, e.g., the Tyndale Bible (16th cent.), King James V. (17th cent.), American Standard V. (19th cent.), Revised Standard V. (20th century),
  - 25 German Bible translations, Luther Bibel (16th cent.), Elberfelder V. (19th cent.), Elberfelder V. (20th cent.).
- Using additional languages, such as
  - Historical German—strongly inflecting—and English—less inflecting.
- Improving the alignment priority of the algorithm and support nested operations in the form of trees.

\(^4\)www.paralleltext.info/ and www.biblestudytools.com/
FUTURE WORK - QUESTIONS

Research Questions:

• RQ1 How is the reuse changed? Do word stems remain the same? What remains stable?
  • Obtain a comprehensive analysis of modifications and stable elements
• RQ2 How high is the ratio of changes in a reused text? How often are words substituted and how often does morphology change?
  • Consider a high ratio of part-of-speech changes, not supported by a word net, then, better cluster together cognates as common concept
FUTURE WORK - EXPERIMENTS

- Identify **nested operations** in the form of **trees** by help of word nets
  - to gather reliable information on the share of semantic and morphological changes
- Visualizing morphological changes as **part-of-speech matrix**
  - to find possible dependencies between them and the context
- Back-up the word net coverage with word embedding approaches, e.g. **distributional semantics**
- Identify the support of a Semitic language as **pivoting element**
- Calculate a reuse model (probabilities by OP parameters)
- Publish a data set of historical writing variants (at best generate a BibleNet)
Apply the essential approach to bigger data sets

• 27 English Bible translations
  • e.g., Tyndale Bible (inspired by Luther) (ca. 1526–1536),
  • King James V. (1611–1769),
  • Webster’s Revision (of the King James V.) (1833),
  • Darby Bible (1890),
  • American Standard V. (19th century).

• 25 German Bible translations
  • e.g., Luther Bibel (1545), Elberfelder V. (1871 and 1905).

• van Dyck’s Arabic Bible from 1865 to use it as a pivot language

• The verse-based structure of the Bible enables a reliable mapping of reuse by the unique book, chapter, verse IDs. This also ensures a specific window in which the modification is measured.

---

5www.paralleltext.info/ and www.biblestudytools.com/
Most Arabic words consist in a root of three consonants

- This root represents (carries) the meaning of a word, the topic
- Affixes and infixes between these consonants determine the part of speech or the morphological categories (case, number, count, etc.)
  - e.g., K-T-B carries to write
  - KaTaBa (he wrote, he did write)
  - KāTiB (writing, writer)
  - KiTāB (book)
  - maKTūB (written, letter)
  - maKTaBa (library)
- Simplified: the root consonants have semantic function and the vocals morphological function
Another example *to fly*\(^6\)

\(^6\)http://aratools.com/
Hypothesis: Using a root stemmed Arabic text possibly helps to find/evaluate semantically related words in an Indo-German language (given an aligned Bible Corpus).

- Due to the “root concept” semantic relationships can be verified or identified
- A 1st step is to prove the distribution of roots over the potential topic clusters in a DS task
THANK YOU!
APPENDIX
Old Testament:


New Testament:


Latin Bible:


Clement Edition:

• Bernard de Clairvaux, Amour de Dieu, Liber de diligendo Deo, P. A. O à Sources Chrétienes, col. 393, 1993.
• Bernard de Clairvaux, Grâce et le Libre Arbitre, Liber de gratia et de libero arbitrio, P. A. O à Sources Chrétienes, col. 393, 1993.
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.