emnlp2016
Conference on Empirical Methods in Natural Language Processing
November 1–5, 2016, Austin, Texas, USA.

Non-Literal Text Reuse in Historical Texts:
An Approach to Identify Reuse Transformations and its Application to Bible Reuse

Maria Moritz[1], Andreas Wiederhold[1], Barbara Pavlek[2], Yuri Bizzoni[3], Marco Büchler[1]

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 resuse 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?

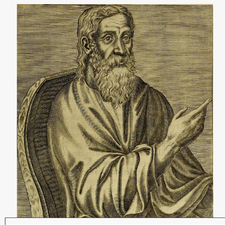
2 How is the non-literally reused text modified in our datasets?

2.1 How can linguistic resources support the discovery of non-literal reuse?

2.2 What are the limitations of an automated classification approach relying on linguistic resources?

Data

Clement of Alexandria



Christian theologian from the 2nd century.

Bernard of Clairvaux

French abbot from the 12th century.

We obtain Bible verse reuse pairs:
199 & 162

literal	Bible verse	Bernard reuse
Prov 18 3	impius cum in profundum venerit peccatorum contemnit 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)*	Impius , cum venerit in profundum malorum , contemnit (*When the wicked man is come into the depth of evil*)
less literal	**Bible verse**	**Clement reuse**
1Cor 13 13	νυνὶ δὲ μένει πίστις , ἐλπίς , ἀγάπη , τὰ τρία ταῦτα μείζων δὲ τούτων ἡ ἀγάπη (*And now remain faith, hope, love, these three; but the greatest of those is love.*)	ἀγάπην , πίστιν , ἐλπίδα (*love, faith, hope* – in accusative case)
		μένει δὲ τὰ τρία ταῦτα , πίστις , ἐλπίς , ἀγάπη · μείζων δὲ ἐν τούτοις ἡ ἀγάπη *(and remain these three, faith, hope, love; but the greatest among them is love)*
non-literal	**Bible verse**	**Clement reuse**
Mt 12 35	ὁ ἀγαθὸς ἄνθρωπος ἐκ τοῦ ἀγαθοῦ θησαυροῦ ἐκβάλλει ἀγαθά , 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.*)	Ψυχῆς , τὰ δὲ ἐκτός , κἂν μὲν ἡ ψυχὴ χρῆται καλῶς , καλὰ καὶ ταῦτα δοκεῖ , ἐὰν δὲ πονηρῶς , πονηρά , ὁ κελεύων ἀπαλλοτριοῦν τὰ ὑπάρχοντα (*[are in the] soul, and some are out, and if the soul uses them good, those things are also thought of as good, but if [used as] bad, [they are thought of] bad; he who commands the renouncement of possessions*)

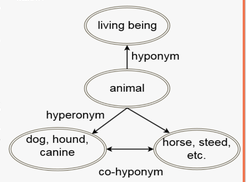
Methodology

Operation	Example
NOP(reuse_word, orig_word)	NOP(maledictus, maledictus)
lem(reuse_word, orig_word)	lem(penetrat, penetrabit)
repl_syn(reuse_word, orig_word)	repl_syn(magnificavit, glorificavit)
repl_hyper(reuse_word, orig_word)	hyper(cupit, habens)
repl_hypo(reuse_word, orig_word)	hypo(dederit, tollet)
repl_co-hypo(reuse_word, orig_word)	repl_co-hypo(magnificavit, fecit)
NOPmorph(reuse_tags, orig_tags)	NOPmorph(na,na)
repl_pos(reuse_tag, orig_tag)	repl_pos(n,a)
repl_case(reuse_tag, orig_tag)	repl_case(g,d)
lemma_missing(reuse_word, orig_word)	lemma_missing(tentari, inlectus)
no_rel_found(reuse_word, orig_word)	no_rel_found(gloria, arguitur)

i We define operations (OPs) reflecting literal reuse and semantic replacements (see above).

ii Our algorithm looks for identical & similar words and for morphological & semantic changes (see top right).

iii We apply both to our datasets using the Ancient Greek WordNet (see an example at the right).



iv We complement the automated approach by a manual analysis of a sample (60 & 100 instances resulting in 192 & 224 replacement operations) to find the limitations of our automated approach.

OPs used manually: `ins(word), del(word) and NOP, lem, repl_syn, repl_hyper, repl_hypo, repl_co-hypo`

Morphological information: from Perseus' tag set (Bamman & Crane 2011), e.g. `repl_num_s_p`

Future plans

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

A smarter automated approach for deriving an original text excerpt can be learning edit scripts, such as undetaken 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 relatons tha are presented in word nets.

Results

RQ1: What is the extent of non-literal reuse in our datasets?

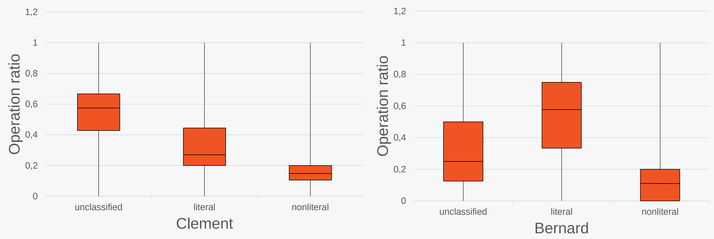


Figure 1: Ratios of operations in reuse instances. literal: NOP, lem, lower, etc.; nonlit: syn, hyper, etc.

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

RQ2.1 How can linguistic resources support the discovery of non-literal reuse?

	literal				non-literal				unclassified		
	NOP	upper	lower	lem	syn	hyper	hypo	co-hypo	no_rel_found	lem_mssing	total
Clement	337	6	0	356	153	20	14	101	563	639	2189
Bernard	587	0	44	102	60	14	28	68	347	85	1335

Consider operations that successfully look up a lemma as: `lem_success={lem,syn,repl_hyper,repl_hypo,repl_co-hypo, no_rel_found}`, with `lem_missing` holding tokens not found.

$$supp_{lem} = \frac{\sum_{Occ(o)} o \in \text{lem\_success}}{\sum_{Occ(o)} o \in \text{lem\_success} \cup \{\text{lem\_missing}\}}$$

$supp_{lem_{Clement}}$ = 0.65 and $supp_{AGWN_{Clement}}$ = 0.34
$supp_{lem_{Bernard}}$ = 0.88 and $supp_{AGWN_{Bernard}}$ = 0.33

RQ2.2 What are the limitations of an automated classification approach relying on linguistic resources?

exception	quantity	
	Clement	Bernard
Word changed to antonym	1[1]	0
Synonym and morphology changed	1	16
More than one morphological category changed	1	7
Synonym is multi-word expression	3	5
Many-to-many	0	1[2]

Exceptions preventing applying our OPs.

1) "**the** God, the good (**one**)" (Clement) vs. "**none** is good but the God" (Bible).
2) "judged calmly" (Bernard) vs. "fake friend" (Sal 12 18).

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.

1) {maria.moritz|andreas.wiederhold1}@stud.uni-goettingen.de
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

2) pavlek@shh.mpg.de
Max Planck Institute for the History of Human Genome
The Mint Minds and Traditions Research group

3) yuri.bizzoni@gu.se
UNIVERSITY OF GOTHENBURG

1) {mmoritz|mbuechler}@etrap.eu
eTRAP Electronic Text Reuse Acquisition Project

We aggregate

Biblindex' Lemmas
(65.5K Biblical Greek entries; 315K Latin entries)

Classical Language Tool Kit (CLTK) (Johnson et al., 2014)
954K Ancient Greek & 270K Latin entries

Greek New Testament of the Society of Biblical Literature[3] & Septuaginta[4]
59.5K word-lemma-pairs

Ancient Greek WordNet (Bizzoni et al., 2014; Minozzi, 2009)
99K synsets of which 33K contain Ancient Greek and 27K contain Latin words

Greek Old Testament
Alfred Rahlfs, editor. 1935. Septuaginta, id est Vetus Testamentum Graece juxta LXX interpretes. Rahlfs. 2 vol., 1950.

Greek New Testament
Kurt Aland and Barbara Aland, editors. 1966. The Greek New Testament. Deutsche Bibelgesellschaft-United Bible Societies, 27 edition.

Latin Bible
Gribomont J.Weber R., Fischer B., editor. 1969, 1994, 2007. Biblia sacra juxta vulgatam versionem. Deutsche Bibelgesellschaft.

Clément d'Alexandrie, Quel riche sera sauvé ?, Quis dives salvetur, P. A. O à Sources Chrétiennes, col. 537, p. 100 ff., 2011.

We acknowledge
Laurence Mellerin for providing the datasets and for valuable advice on their content. The German Federal Ministry of Education and Research for funding the work (grant 01UG1509).

3) Logos Bible Software, Sbl new testament, 2014 http://sblgnt.com/about/
4) Alfred Rahlfs, editor. 1935. Septuaginta. Württembergische Bibelanstalt, 9 edition. 1971.; UPenn

SPONSORED BY THE
Federal Ministry of Education and Research