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

**Maria Moritz[1], Andreas Wiederhold[2], Barbara Pavlek[3], Yuri Bizzoni[4],** and **Marco Büchler[1]**
[1]Institute of Computer Science, University of Göttingen
[2]Institute for Educational Science, University of Göttingen
[3]Minds and Traditions Research Group, Max Planck Institute for the Science of Human History, Jena
[4]Department of Philosophy, Linguistics, Theory of Science, University of Gothenburg

## Abstract

Text reuse refers to citing, copying or alluding text excerpts from a text resource to a new context. While detecting reuse in contemporary languages is well supported—given extensive research, techniques, and corpora—automatically detecting historical text reuse is much more difficult. Corpora of historical languages are less documented and often encompass various genres, linguistic varieties, and topics. In fact, historical text reuse detection is much less understood and empirical studies are necessary to enable and improve its automation. We present a linguistic analysis of text reuse in two ancient data sets. We contribute an automated approach to analyze how an original text was transformed into its reuse, taking linguistic resources into account to understand how they help characterizing the transformation. It is complemented by a manual analysis of a subset of the reuse. Our results show the limitations of approaches focusing on literal reuse detection. Yet, linguistic resources can effectively support understanding the non-literal text reuse transformation process. Our results support practitioners and researchers working on understanding and detecting historical reuse.

## 1 Introduction

The computational detection of historical text reuse—including citations, quotations or allusions —can be applied in many respects. It can help tracing down historical content (a.k.a., *lines of transmission*), which is essential to the field of textual criticism (Büchler et al., 2012). In the context of massive digitization projects, it can identify relationships between text excerpts referring to the same source. Specifically, detecting copies of the same historical text that have diverged over time (manuscript studies, a.k.a., *Stemma Codicum*) is an important task.

Although much work exists in the field of natural language processing (NLP), many new challenges arise when processing historical text. The most important challenges are the absence of supporting tools and methods, including an agreement on a common orthography, standardization of variants, and a wide range of clean, digitized text (Piotrowski, 2012; Geyken and Gloning, 2014; Zitouni, 2014). Typical statistical approaches from the field of NLP are difficult to apply to historically transferred texts, since these often cover a large timespan and, thus, comprise many different writing styles, text variants or even reuse styles (Büchler, 2013). Our long-term goal is to conceive robust text reuse detection techniques for historical texts. To this end, we need to improve the quantitative empirical understanding of such reuse accompanied by qualitative empirical studies. However, only few such works exist.

We study less- and non-literal text reuse of Bible verses in Ancient Greek and Latin texts. Our focus is on understanding how the reuse instances are transformed from the original verses. We identify operations that characterize how words are changed—e.g., synonymized, capitalized or part-of-speech (PoS) information changed. Since our approach uses external linguistic resources, including Ancient Greek WordNet (AGWN) (Bizzoni et al., 2014; Minozzi, 2009) and various lemma lists, we also show how such resources can help detecting reuse and where the limitations are. We complement the automated approach

*Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1859,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

with a qualitative manual analysis. We contribute:

- an automated approach to characterize how text is transformed between reuse and original,
- an application of the approach to two text datasets where reuse was manually identified,
- empirical data based on the automated approach, complemented by a manual identification.

Our resulting datasets[1] with rich information about the reuse transformation (e.g., PoS and morphology changes, and words becoming synonyms or hyperonyms, among others) can be used as a benchmark for future reuse detection and classification approaches.

## 2 Related Work

We first discuss why existing reuse detection approaches are not applicable to historical texts, and then present works trying to address this problem.

**Historical Text Reuse and Plagiarism Detection.** Büchler (2013) combines state-of-the-art NLP techniques to address reuse detection scenarios for historical texts, ranging from near copies to text excerpts with a minimum overlap. He uses the commonly used method fingerprinting, which selects n-grams from an upfront pre-segmentized corpus. While his approach can discover historical and modern text reuse language-independently, it requires a minimum text similarity—typically at least two common features.

Recognizing modified reuse is difficult in general. Alzahrani et al. (2012) study plagiarism detection techniques: n-gram-, syntax-, and semantics-based approaches. As soon as reused text is slightly modified (e.g., words changed) most systems fail. Barrón-Cedeño et al. (2013) conduct experiments on paraphrasing, observing that complex paraphrasing along with a high paraphrasing density challenges plagiarism detection, and that lexical substitution is the most frequent technique for plagiarizing. The AraPlagDet (Bensalem et al., 2015) initiative focuses on the evaluation of plagiarism detection methods for Arabic texts. Eight methods were submitted and turned out to work with a high accuracy on external plagiarism detection but did not achieve usable results for intrinsic plagiarism detection.

**Corpora.** Huge parallel corpora of modern languages are used in fields such as paraphrase generation and detection, typically used to train statistical models (Zhao et al., 2009; Madnani and Dorr, 2010). However, such corpora hardly exist for historical languages or are copyrighted, such as the TLG digital library (Pantelia, 2014). Especially in the field of modern reuse investigation, aligned corpora are often used, providing a rich source of paraphrasal sentence pairs in one, sometimes multiple languages. One of such is the Microsoft Research Paraphrase Corpus (MSRP), which contains 5801 manually evaluated, paraphrasal sentence pairs in English (Dolan and Brockett, 2005). Ganitkevitch et al. (2013) present a paraphrase database with over 200 million English paraphrase pairs and 196 million Spanish paraphrases. Each paraphrase pair comes with measures, such as a paraphrase probability score. In ancient literature, efforts are made to collect Biblical reuse. One of such is the collection of Ancinet Greek and Latin quotations based on the the Vetus Latina series and the Novum Testamentum Graecum Editio Critica Maior (Houghton, 2013a; Houghton, 2013b). It contains more than 150,000 Latin citations and about 87,000 Ancient Greek Bible references.

**Historical Text Processing in General.** Efforts to automatically process ancient texts are made around the Perseus Digital Library project (Crane, 1985), among others. For example, Bamman (2008) presents the discovery of textual allusions in a collection of Classical poetry, using measures such as token similarity, n-grams or syntactic similarity. This allows finding at least the most similar candidates within a closed library. Some works have focused on text reuse in Biblical Greek text. Lee (2007) investigate reuse among the Gospels of the New Testament, aimed at aligning similar sentences. Using source alternation patterns, among others, the approach uses cosine similarity, source verse proximity, and source verse order. Focusing on high recall, the detection of Homeric quotations in Athenaeus' Deipnosophistai' was investigated by Büchler et al. (2012), searching for distinctive words within reuse.

While the approaches above rely on string or feature similarity, Bamman (2011b) attempts to process the semantic space using word-sense disambiguation (Patwardhan et al., 2003; Agirre and Edmonds, 2007). Using a bilingual sense inventory and training set, they classify up to 72 % of word senses correctly.

**Utilizing Linguistic Resources.** Word nets support

[1] https://bitbucket.org/mariamoritz/emnlp

identifying word relationships. Jing (1998) investigates issues that come with using WordNet (Miller et al., 1990) for language generation. Among others, these comprise issues arising from the adaption of a general lexicon to a specific domain. These were encountered by using a domain corpus and an ontology to prune WordNet to a certain domain.

In our work, we are interested in using linguistic resources (word nets and lemma lists) together with PoS information to model the transformation process of reuse, specifically on an ancient language text to find limitations when applied to non-literal text reuse.

## 3 Methodology

Our study addresses two main research questions:
**RQ1.** *What is the extent of non-literal reuse in our datasets?* This analysis provides a baseline for the following characterizations of the non-literal reuse.
**RQ2.** *How is the non-literally reused text modified in our datasets?* We study kinds and frequencies of semantic, lexical, and morphological changes. We develop an automated approach to identify the reuse transformation, and complement it with a manual, qualitative analysis. We formulate two sub-questions:
**RQ2.1.** *How can linguistic resources support the discovery of non-literal reuse?* We conjecture that non-literal reuse is difficult to capture automatically (especially due to domain- or author-specific words), but that taking linguistic resources into account helps. We analyze the coverage of words in lemma lists and a synset database, and investigate how useful they are for understanding the reuse transformations.
**RQ2.2.** *What are the limitations of an automated classification approach relying on linguistic resources?* Our manual analysis investigates the reuse in its full richness, to understand the limitations of the automated approach and identify further characteristics of the reuse in our datasets.

### 3.1 Study Design

Our study comprises the following main steps. First (**RQ1**), we identify and characterize the literal and non-literal overlap in reuse instances. Second (towards **RQ2**), we define operations reflecting literal reuse, replacements (inspired by semantic relationships, such as synonyms and hyperonyms, supported by AGWN), and morphological changes (e.g., when mapping words contain the same cognate). Our operations are based on a one-word-replacement to better quantify the results. Third (**RQ2.1**), we develop an algorithm that identifies operations by first looking for morphological changes between a word from the reuse and its corresponding candidate from the Bible verse and, in case of no success, by seeking for a semantic relation. We apply it to our two datasets and investigate the relationships of affected words and the literal share. We quantify occurrences of operations and calculate two measures $\text{sup}_{\text{lem}}$ (lemma support) and $\text{sup}_{\text{AGWN}}$ (AGWN support) to assess the resources' coverage for our approach. Fourth (**RQ2.2**), we manually analyze a smaller sample of our reuse datasets, using further operations, to understand the full richness of the reuse.

### 3.2 Datasets

We use the following two text sources, both reusing content from Bible verses. As a ground truth of the reuse, we use manually annotated versions of both, provided to us by Mellerin (2014) and the Biblindex project (Mellerin, 2016; Vinzent et al., 2013).

Our first dataset comes from the primary source text of "Salvation for the Rich" from the Ancient Greek writer Clement of Alexandria (Clément d'Alexandrie, 2011), a well-known author in Biblical literature (Cosaert, 2008). The Biblindex team annotated 128 text passages as Bible reuse instances, adding a footnote with Bible verse pointers to each. We select a total of 95 out of these 128, following four criteria: (i) reuse should not consist of an exact literal copy of a Bible verse (skipping six instances), (ii) reuse should be recognizable by our expert (skipping ten instances), (iii) the reference frame should be within five Bible verses (comparable with sentences) to avoid too much noise in our data to ensure a comparable length to the original Bible verse (skipping nine instances), and (iv) reuse instances should not exceed a length of 40 tokens (1–2 sentences), again to cut the long tail and avoid too much noise (skipping eight instances). Sometimes one reuse instance pointed to different Bible verses or one text passage contained more than one reuse instance, thus, we come up with 199 verse-reuse-pairs. The excerpts point to a total of 15 Bible books.

Our second dataset are extracts from a total of 14 volumes of twelve works and two work collec-

| | |
|---|---|
| **Jer 23 24** | si occultabitur vir in absconditis et ego non videbo eum dicit Dominus numquid non caelum **et terram ego impleo** ait Dominus (*Can anyone hide himself in secret places that I will not see him? Said the lord. Do not I fill heaven and earth? Said the Lord*) |
| literal | **et terram ego impleo** (*and I fill the earth*) |
| **Mk 10 30** | Ἤρξατο λέγειν ὁ Πέτρος αὐτῷ, Ἰδοὺ ἡμεῖς ἀφήκαμεν πάντα καὶ ἠκολουθήκαμέν σοι. (*Peter began to say to him: See, we left everything and followed you.*) |
| literal | ἡμεῖς ἀφήκαμεν πάντα καὶ ἠκολουθήσαμέν σοι (*we left everything and followed you*) |
| **Prv 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)* |
| more literal | **Impius** , **cum venerit in profundum** malorum , **contemnit** (*When the wicked man is come into the depth of evil*) |
| **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.*) |
| less literal | **πίστει** καὶ **ἐλπίδι** καὶ **ἀγάπῃ** (*faith, and hope, and love* - in dative case) |
| less literal | **ἀγάπην** , **πίστιν** , **ἐλπίδα** (*love, faith, hope* - in accusative case) |
| less literal | 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)* |
| **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 .*) |
| non-literal | Ψυχῆς , τὰ δὲ ἐκτός , κἂν μὲν ἡ ψυχὴ χρῆται καλῶς , καλὰ καὶ ταῦτα δοκεῖ , ἐὰν δὲ πονηρῶς , πονηρά , ὁ κελεύων ἀπαλλοτριοῦν τὰ 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*) |

Figure 1: Examples of reuse

tions from the Latin writer Bernard of Clairvaux. We again use Biblindex' extracted Bible reuse, which offers over 1,100 reuse instances in alphabetical order. We follow the same selection criteria as for Greek and—starting top-down and dropping only two—we obtain 162 Bible-verse reuse-pairs, which is similar to the number of Greek reuse instances. Specifically, since those reuse instances come from several different primary source works, they point to a total of 31 Bible books. We use the Bible editions from Biblindex, specifically, the data based on *Septuagint* (Rahlfs, 1935b), *Greek New testament* (Aland and Aland, 1966), and *Biblia sacra juxta vulgatam versionem* (Weber R., 1969 1994 2007).

Fig. 1 shows reuse examples, illustrating the wide range of literalness in our data, comprising literal (all tokens overlap), less literal (important tokens overlap), and non-literal (no content word tokens overlap) reuse. For example, Clement's reuse ranges from introducing the overall topic by citing multiple verses, to supporting his argumentation. Specifically, Mk 10 30 is a fully literal reuse from a passage that discusses the problem of rich men in heaven. Clement uses this episode as a main point in his essay. Later he refers to 1Cor 13 13, he again refers to how hard it would be for rich men to enter heaven, explaining that salvation is independent of "external things," but depends on the "virtue of the soul," mentioning faith, hope, and love, the key words in the original verse.

**Algorithm 1:** Reuse classification algorithm

```
/* Executed for each reuse instance and its corresponding
   Bible verse.  morph(x) returns the part-of-speech
   and/or case of x.  repl_case and repl_pos are masked to
   repl_morph for clarity reasons.  checkm(x,y) returns
   NOPmorph(morph(x),morph(y)) if morph(x) equals morph(y)
   and repl_morph(morph(x),morph(y)) otherwise.            */
input  : $L \leftarrow$ set of word-lemma pairs obtained from the lemma resources
input  : $S \leftarrow$ set of synsets from AGWN; each synset contains an id and a parent id
input  : $T \leftarrow$ list of words of reuse instance (containing part-of-speech information)
input  : $B \leftarrow$ list of words of Bible verse (containing part-of-speech information)
output : $OP \leftarrow$ list of sets containing up to 3 parameterized operations
$s1, s2 \leftarrow$ any two synsets $\in S$.
$tmp\_op \leftarrow$ temporary variable which presents the absence of a relation but not of a lemma.
for $t$ in $T$ do
    for $b$ in $B$ do
        if $t$=$b$ then
            $OP \leftarrow OP \cup (NOP(t, b), checkm(morph(t), morph(b)))$
            break
        else if lowerCase($t$) = $b$ then
            $OP \leftarrow OP \cup (lower(t, b), checkm(morph(t), morph(b)))$
            break
        else if lowerCase($b$) = $t$ then
            $OP \leftarrow OP \cup (upper(t, b), checkm(morph(t), morph(b)))$
            break
        else if $t \in L$ and $b \in L$ then
            /* lemma found for original (b) and reuse word (t)  */
            if lemma($t$) = lemma($b$) then
                $OP \leftarrow OP \cup (lem(t, b), checkm(morph(t), morph(b)))$
                break
            else if $t \in s1$ and $b \in s2$ and $s1 \in S$ and $s2 \in S$ then
                if $s1$ = $s2$ then
                    /* t is synonym of b                                  */
                    $OP \leftarrow OP \cup (repl\_syn(t, b))$ break
                else if id($s1$) = parent_id($s2$) then
                    /* t is hyperonym of b                                */
                    $OP \leftarrow OP \cup (repl\_hypo(t, b))$ break
                else if parent_id($s1$) = id($s2$) then
                    /* t is hyperonym of b                                */
                    $OP \leftarrow OP \cup (hyper(t, b))$ break
                else if parent_id($s1$) = parent_id($s2$) then
                    /* synset of t and synset of b both have the same
                       synset as parent                                   */
                    $OP \leftarrow OP \cup (repl\_cohypo(t, b))$ break
            else
                $tmp\_op \leftarrow (no\_rel\_found(t, b))$
    end
    if $tmp\_op$ then
        $OP \leftarrow OP \cup tmp\_op$
    else
        $OP \leftarrow OP \cup (lemma\_missing(t))$
end
return $OP$
```

### 3.3 PoS Tagging

The automated and the manual approach also take PoS information into account to understand the reuse transformation. Following the Greek morphology tagging system of Perseus (Bamman and Crane, 2011a), which maps PoS and case information to single characters[2], we manually PoS-tag the 199 reuse instances of Ancient Greek and the 162 of Latin, as well as the original Bible verses. Since Latin and Ancinet Greek PoS-taggers lack available implementations, appropriate trained models or simply accu-

[2] http://nlp.perseus.tufts.edu/syntax/treebank/agdt/1.7/docs/README.txt

|  | **lemma coverage**[1] |  | **AGWN coverage**[2] |  |  |  | **total**[3] |
|---|---|---|---|---|---|---|---|
|  | corpus | lem. | syn. | hyper. | hypo. | co-hypo. |  |
| CLTK | Greek Bible[4] | 3238 | 1906 | 1422 | 1185 | 1422 | 4776 |
|  | Clement[5] | 739 | 326 | 231 | 175 | 231 | 2189 |
|  | Latin Bible[4] | 2473 | 1241 | 905 | 863 | 905 | 2618 |
|  | Bernard[5] | 1219 | 643 | 471 | 455 | 471 | 1335 |
| Biblindex | Greek Bible[4] | 752 | 103 | 58 | 67 | 58 | 4776 |
|  | Clement[5] | 455 | 54 | 24 | 33 | 24 | 2189 |
|  | Latin Bible[4] | 2473 | 1365 | 1057 | 1023 | 1057 | 2618 |
|  | Bernard[5] | 1219 | 701 | 531 | 520 | 531 | 1335 |
| SBLGNT & LXX | Greek Bible[4] | 4718 | 3385 | 2616 | 2092 | 2616 | 4776 |
|  | Clement[5] | 1297 | 824 | 582 | 421 | 582 | 2189 |
|  | Latin Bible[4,6] | n/a | n/a | n/a | n/a | n/a | 2618 |
|  | Bernard[5,6] | n/a | n/a | n/a | n/a | n/a | 1335 |
| combined | Greek Bible[4] | 4723 | 3449 | 2684 | 2156 | 2684 | 4776 |
|  | Clement[5] | 1548 | 899 | 653 | 495 | 653 | 2189 |
|  | Latin Bible[4] | 2473 | 1378 | 1057 | 1023 | 1057 | 2618 |
|  | Bernard[5] | 1219 | 706 | 531 | 520 | 531 | 1335 |

[1] number of tokens found by lemma resource
[2] number of lemmatized tokens covered by AGWN
[3] number of tokens in original and reuse
[4] original [5] reuse [6] no support for Latin

Table 1: Coverage of tokens by language resources

racy (Crane, 1991; vor der Brück et al., 2015), we perform this step manually to assure high accuracy. We also assign cases for the classes noun, article, adjective, and pronoun. We introduce *b* to represent the Latin ablative case, which does not exist in Greek.

### 3.4 Automated Approach

Our approach is to model the transformation process in terms of parameterized operations applied to the words in the reuse instance in order to obtain the original words. These operations use linguistic resources, such as lemma lists of classical Greek and Biblical Koine, and a synset database. For each transformation, we create the set of operations necessary to transform the reuse instance to its original.

**Linguistic Resources.** We investigate the following lemma lists to look up lemmatized forms of words—a prerequisite for looking up synsets: *Classical Language Tool Kit (CLTK)* (Johnson et al., 2014 2016) provides Ancient Greek and Latin lemma lists for 953,907 Greek and 270,228 Latin words. *Biblindex' Lemma Lists* contain entries for 65,537 Biblical Greek and 315,021 Latin words. *SBLGNT&LXX* refers to the Greek New Testament of the Society of Biblical Literature (SBLGNT)[3] and the Septuaginta (LXX), a translation of the Old Testament (Rahlfs, 1935a)[4] from the Center for Computer Analysis of Texts at UPenn. We acknowledge code-page corrections by M. Munson. SBLGNT&LXX provide 59,510 word-lemma-pairs.

We use *AGWN* (Bizzoni et al., 2014), which also contains Latin WordNet (Minozzi, 2009), to identify synsets (sets of synonyms) as well as hyperonyms, hyponyms, and co-hyponyms. From the wordnets' 98,950 synsets 33,910 synsets contain Ancient Greek and 27,126 synsets contain Latin words.

**Coverage.** Table 1 shows the coverage of each resource for our datasets. In the lower part of it we merge all lemma resources into one set of word-lemma pairs. The table shows that CLTK covers the Bible data better than the Hellenistic Greek as used in Clement of Alexandria, an author from 2nd century AD, writing in an archaic style with Biblical vocabulary, while also being influenced by Classical Greek. We also check the coverage of lemmata stemming from the same source (Biblindex) as our reuse. To increase the coverage for Greek, we consult SBLGNT&LXX, which in fact increases it. To not miss important information, we integrate all of the resources' data into our approach. For every lemma of a word we check the semantic relations in AGWN. We experimented with different ways of looking up lemmas and found that lower-casing all Latin tokens improved the success. For Greek, it had the opposite effect, which indicates that the Greek text contains more entities that are not available in lowercase in the lemma lists, so we did not change in that case.[5]

**Operations and Classification.** We define replacement operations using words and PoS as parameters, to transform a reuse instance *back into the Bible verse* it originates from. Table 2 lists the operations for the computational approach. We introduce the operations *NOPmorph*, *repl_pos*, and *repl_case* for words having the same cognate, and *lemma_missing(reuse_word)* when a word is not

[3] Logos Bible Software, Sbl new testament, 2014 `http://sblgnt.com/about/`

[4] CATSS LXX is prepared by the Thesaurus Linguae Graecae project directed by T. Brunner at UC Irvine, with further verification and adaptation by CATSS towards conformity with the individual Göttingen editions which appeared since 1935. LXXM is morphologically analyzed text of CATSS LXX prepared by CATSS led by R. Kraft (Philadelphia team)

[5] Often, the decision on whether to represent a word in upper or lower case letters is made by the editor, thus, our decision is affected by the edition we use for our research.

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

Table 2: Operation list for the automated approach

known to any of our lemma resources as well as *no_rel_found(reuse_word, orig_word)* when the relationship between a reuse word and each potential word from the original is not covered by AGWN.

Algorithm 1 shows our approach to classify the reuse transformation by identifying the operations. For each reuse token, we identify the first applicable operation matching the foremost Bible verse word (iterating the verse) in the following order: exact word match (*NOP: no operation*), case changed to *upper* or *lower*. Thereafter, we look up the lemma and return *lem* if the lemma of the reused word matches the lemma of the original. For these four, we also check the morphology, in addition returning whether the original has the same PoS and case (*NOPmorph*) or whether PoS changed (*repl_pos*), case changed (*repl_case*), or both. So up to three operations can be returned per word. Finally, we check for synonyms (*repl_syn*), hyperonyms (*hyper*), hyponyms (*hypo*), and co-hyponyms (*repl_co-hypo*), but do not check morphology. If a Bible verse word is used as a match, it is not used again for any other word from the reuse.

### 3.5 Qualitative Approach

To obtain a deeper understanding of the limitations of linguistic resources for our purpose, two graduate students (one Latinist, one Classical Archeologist) manually analyze 100 Greek and 60 Latin reuse instances with their expert knowledge, using an extended set of operations. It comprises *ins(word)* (insert a word) and *del(word)* (delete a word)—two operations we ignore in the automated approach where we focus on the coverage of the resources. It also has a richer set of replacement operations: those from the upper part of Table 2 (without *upper* and *lower*), and instead of only using *repl_case* when a cognate stays the same, we refine it and assign *all* changing morphological categories from Perseus' tag set for *any* "relativeness" between two words (e.g., repl_case_a_g).

## 4 Results

We now present the results for our research questions in Sec. 4.1–4.3, which are summarized and further interpreted in Sec. 4.4.

### 4.1 Literal Share of the Reuse (RQ1)

We obtain a first understanding of the reuse by looking at the percentage of overlapping words between reuse instance and original Bible verse. We measure the longest common substring based on word tokens. Fig. 2 shows the distributions, distinguishing between a *lemmatized* and *non-lemmatized* word comparison.

While lemmatizing words before comparison has only a small impact, we observe differences between the datasets. In our Latin dataset, the overlap is significantly higher than in the Greek dataset Sec. 3.2. 25 % (upper quartile) of Bernard's reuse instances have 50 % or more tokens overlap with their original, which is only the case for less than 25 % in Clement's Greek data. Still, large overlaps of up to 75 % (top

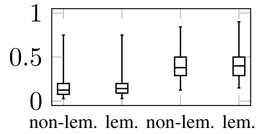


Figure 2: Ratios of literal overlaps between reuse instance and original (left: Greek, right: Latin)

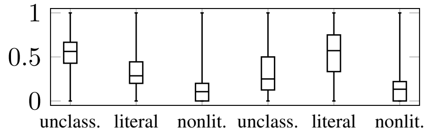


Figure 3: Ratios of unclassified, literal, and non-literal words in reuse instances (left: Greek, right: Latin)

whisker) in our Greek and up to around 90 % in our Latin dataset exist—so a small fraction of the reuse contains literal parts Sec. 3.2.

For a more precise understanding of the literalness, we group operations into literal (NOP, upper, lower, lem), non-literal (repl_syn, repl_hyper, repl_hypo, repl_co-hypo), and unclassified (no_rel_found and lemma_missing). Within each reuse instance, we calculate their relative occurrence using the results of the automated approach (explained shortly). Fig. 3 shows the distribution of these relative occurrences for all reuse instances. It confirms Fig. 2 by showing a higher rate of literalness for Latin compared to Greek. In summary, it also shows that the Latin reuse can be better classified by our approach, which takes the lemma lists and AGWN into account.

### 4.2 Automated Approach (RQ2.1)

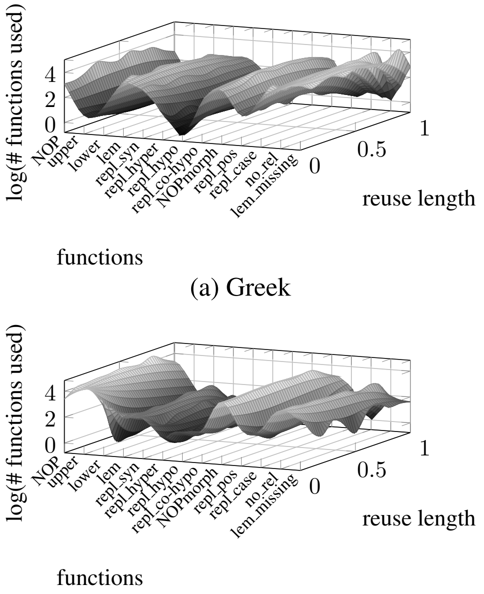
Table 3 shows the total number of operations identified for the transformation from reuse instances to the Greek and Latin originals. For 987 (45 %) out of 2189 words in the Greek instances and for 893 (67 %) out of 1335 words in the Latin instances, we were able to identify at least one operation, which already indicates to what extent the resources are helpful.

Fig. 4 visualizes the distribution of the frequencies (y-axis) of each operation (x-axis) together with the distribution of the operations' positions in the reuse instances (z-axis). The latter is calculated as the relative position $p \in [0..1]$ of an operation with respect to the length of the reuse instance. It indicates that most operation types are distributed over the whole reuse length without a particular trend in both datasets. We only encounter a frequent use of *upper* at the first position in Latin, which means that Bernard often starts his Biblical references with literal Bible words.

|  | NOP | upper | lower | lem | syn | hyper | hypo | co-hypo |
|---|---|---|---|---|---|---|---|---|
| Occ. Greek | 337 | 6 | 0 | 356 | 153 | 20 | 14 | 101 |
| Occ. Latin | 587 | 0 | 44 | 102 | 60 | 14 | 28 | 68 |

|  | NOPmorph | repl_pos | repl_case | no_rel_found | lem_missing |
|---|---|---|---|---|---|
| Occ. Greek | 420 | 49 | 258 | 563 | 639 |
| Occ. Latin | 617 | 46 | 75 | 347 | 85 |

Table 3: Absolute numbers of operations identified automatically



(a) Greek

(b) Latin

Figure 4: Occurrence of operations in reuse instances. X-axis: operations; Y-axis: relative position within reuse instances. Z-axis: natural logarithm of number of operations. Values are smoothed by spline interpolation. The order of operations is arbitrary.

After having checked the overall coverage of the linguistic resources for all tokens (cf. Sec. 3.4), we now specifically investigate to what extent the resources support identifying the reuse transformation for the non-literal reuse using our approach. We introduce the measures $\text{sup}_{\text{lem}}$ and $\text{sup}_{\text{AGWN}}$ to calculate how often looking up a lemma or subsequently a synset element was successful. This is easy based on our operations. Let $\text{Occ}(o)$ be the number of occurrences of an operation o, obtained from Table 3. The operations that successfully looked up a lemma (before consulting AGWN) are lem_success={lem, syn, repl_hyper, repl_hypo, repl_co-hypo, no_rel_found}. Now recall that lem_missing represents the case when a reuse token was not found in the lemma resources. Then $\text{sup}_{\text{lem}} = \frac{\sum_{\text{Occ}(o)} o \in \text{lem\_success}}{\sum_{\text{Occ}(o)} o \in \text{lem\_success} \cup \{\text{lem\_missing}\}}$. We obtain a $\text{sup}_{\text{lem}}$ of 0.65 for the Greek reuse and 0.88 for the Latin reuse. Similarly, the operations that successfully looked up from AGWN are agwn_success={syn,

| operation | Greek | Latin | operation | Greek | Latin |
|---|---|---|---|---|---|
| repl_syn | 78 (40.6%) | 91 (40.4%) | repl_gender | 6 (3.1%) | 1 (0.4%) |
| repl_ant | 1 (0.5%) | 0 | repl_mood | 11 (5.7%) | 12 (5.3%) |
| repl_hyper | 3 (1.6%) | 0 | repl_number | 17 (8.9%) | 17 (7.6%) |
| repl_hypo | 11 (5.7%) | 0 | repl_person | 5 (2.6%) | 14 (6.2%) |
| lem | 1 (0.5%) | 2 (0.9%) | repl_pos | 18 (9.4%) | 33 (14.7%) |
| repl_co-hypo | 0 | 1 (0.4%) | repl_tense | 3 (1.6%) | 9 (4.0%) |
| repl_case | 38 (19.8%) | 36 (16%) | repl_voice | 0 | 8 (3.6%) |

Table 4: Numbers of replacement operations identified for the manual reuse transformation.

repl_hyper, repl_hypo, repl_co-hypo}, with no_rel_found representing a failed lookup. Then: $\text{sup}_{\text{AGWN}} = \frac{\sum_{\text{Occ}(o)} o \in \text{agwn\_success}}{\sum_{\text{Occ}(o)} o \in \text{agwn\_success} \cup \{\text{no\_rel\_found}\}}$. We obtain $\text{sup}_{\text{AGWN}}$ of 0.34 for Greek and 0.33 for Latin.

These values can be interpreted as follows. The lemma resources for genre- and time-specific text work well for less-literal reuse, but the resources for semantic relationships (synset databases) show a lack of support and need further development.

### 4.3 Qualitative Approach (RQ2.2)

We manually identify the transformation operations for 60 reuse instances of the Ancient Greek data and for 100 of the Latin data. Here, NOPs cover 9.3 %, insertions 49.8 %, and deletions cover 30.5 % in the Greek data. NOPs cover 26.1 %, insertions 49.7 %, and deletions 11.9 % in the Latin data.

Table 4 shows the ratios of the various *repl* operations based on the remaining 10.4 % and 12.2 %. Similar to the automated approach, we observe a strong use of synonyms and other semantic-level operations, and also a certain portion of switching morphological categories, which indicates para-phrasal reuse. In the Greek data, PoS changes cover about 9%, out of which a participle became a verb (7 times) and vice-versa (5 times). In our Latin data, PoS changes represent 15% of replacements: often a pronoun changed to a noun (6 times) and a participle became a verb (12 times). Case changes are shown in Table 5. Significantly often, an ablative became an accusative, because often changing prepositions expect different cases, or an accusative was replaced by an ablative or nominative, because para-phrasal expression changed.

| operation | Greek | Latin | operation | Greek | Latin |
|---|---|---|---|---|---|
| repl_case_a_b | 0 | 6 | repl_case_g_a | 5 | 2 |
| repl_case_a_n | 9 | 4 | repl_case_g_n | 4 | 2 |
| repl_case_b_a | 0 | 10 | repl_case_n_a | 7 | 5 |
| repl_case_d_a | 0 | 2 | repl_case_n_d | 3 | 0 |
| repl_case_d_g | 3 | 0 | repl_case_v_g | 0 | 2 |
| repl_case_d_n | 5 | 0 | | | |

Table 5: Numbers of case replacements

We encounter **exceptions** that prevent applying the operations. In the Greek data, one word is replaced with its antonym[6]; once, a synonym also changes its PoS. Four times, more than one morphological category changes, twice an auxiliary is deleted, and five times inserted. We find one writing variance (lem), and three times a synonym is replaced by a multi-word expression. In the Latin data, in 16 cases a synonym is replaced and morphological information changed. Seven times, more than one morphological parameter changes for the same cognate. Eight times, an auxiliary is inserted or deleted, and twice, a writing variance is encountered. A synonym is replaced by more than one word five times. In one case, a reuse is too paraphrasal for any word to match semantic relationships (e.g., *judged calmly*—Bernard vs. *fake friend* - Sal 12 18).

### 4.4 Summary and Discussion

**RQ1.** The reuse is significantly non-literal and only lemmatizing words does not help discovering it. Our results show that reuse in two substantial historical texts requires techniques beyond simple preprocessing (e.g., stemming or lemmatizing), which explains why plagiarism-detection systems fail when paraphrases are used (Alzahrani et al., 2012). Bible verses are often used to justify an author's claim, so only relevant parts of the Bible verse are reused. In the reuse the Bible verse is modified to better fit the syntactical and semantic context of an author's new text, as shown in Tables 4 and 5.

**RQ2.1.** The results from our automated approach are encouraging, showing the feasibility of extending reuse-detection techniques with linguistic resources. Yet, it is not clear which precision and recall could be achieved and how existing techniques need to be adapted and calibrated. This investigation is beyond the scope of this study and subject to our future work.

The linguistic resources support the automated

[6] Translation: "**the** God, the good (**one**)" (Clement) vs. "**none** is good but the God" (Bible).

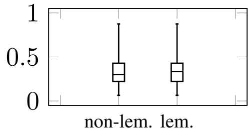


Figure 5: Ratios of literal overlaps in the whole Latin dataset

approach, but only for about one third of the lookups. The manually identified exceptions show that finding a connection between original verse and reuse can be difficult when there is only a vague semantic one.

**RQ2.2.** Our results show that the automated approach cannot capture the richness of the manual approach. Especially from the exceptions, it is clear that less-literal reuse does not only need information from a word's semantic environment, but also that it needs to be identified by looser relations, such as co-hyponyms, multi-to-multi-word associations or implicit meanings, which can be hidden in structural or more broader expert knowledge.

## 5 Threats to Validity

**External Validity.** We enhance the external validity of our work by focusing on Bible verses—one of the oldest, most conveyed, and cited sources of Ancient Greek, offering a vast amount of primary source text and also coming with a long history of scholars studying it. Clement of Alexandria is known for his retelling of biblical excerpts (Clemens, 1905 1909; Freppel, 1865), providing an interesting base for reuse investigation. The french abbot Bernard of Clairvaux (Smith, 2010) is equally known for his influence to the Cistercian order and his work in biblical studies. Furthermore, the chosen lemma resources are the most extensive ones existing for Ancient Greek and Latin. We chose the AGWN, since it is freely available, offering one of the largest synset database for Ancient Greek and Latin.

**Internal Validity.** A threat is that our ground truth has mistakes, as the PoS tagging was done by one author only and relied on a manual post-correction. The selection criteria in Sec. 3.2 were chosen to ensure quality and comparability. Extreme outliers in the length of the reuse instance or source (multiple Bible verses) are cut-off. For Greek, 33 are cut-off, as opposed to Latin, where our sample is significantly smaller than the whole population that we have. To automatically check whether the sample has similar characteristics with respect to the literal reuse, we create Fig. 5. It shows the overlap of the whole 1128 instances of Bernard's extracted reuse, which when compared to Fig. 2 (right) supports the representativeness of our sample. Last, we can only derive operation replacements when a word token was covered by the lemma sources, contained in AGWN, and when there actually exists a relation between two words. Also, our authors' vocabulary can differ in terms of domain knowledge, personal idiolect, and age of the Biblical vocabulary.

## 6 Conclusion

We presented a study of historical—and mostly non-literal—text reuse. We automatically and manually characterize the reuse and identify to what extent existing linguistic resources are able to cover non-literal text reuse. Our results show the potential as well as the necessity to develop robust techniques and to extend linguistic resources for analyzing and detecting such reuse. Our results can help to enhance paraphrase generation to model automatic ways on how small text portions can be rephrased. Considering the effects of syntactic rearrangement of reuse can also support such efforts. A smarter automated approach for deriving an original text excerpt would be learning so-called edit scripts (Kehrer, 2014; Chawathe et al., 1996), which more precisely identify operations an author performed on a text to transform it into another version. Whether learning edit scripts on such intricate transformations is possible is an open question and valuable future research. Finally, analyzing further languages and data sets helps to further complete our findings.

## Acknowledgments

We thank Laurence Mellerin for providing the datasets we used, and for valuable advice on its content. Our work is funded by the German Federal Ministry of Education and Research (grant 01UG1509).

## References

[Agirre and Edmonds2007] Eneko Agirre and Philip Edmonds. 2007. *Word Sense Disambiguation - Algorithms and Applications*. Springer Netherlands.

[Aland and Aland1966] Kurt Aland and Barbara Aland, editors. 1966. *The Greek New Testament*. Deutsche Bibelgesellschaft-United Bible Societies, 27 edition.

[Alzahrani et al.2012] Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *Trans. Sys. Man Cyber Part C*, 42(2):133–149.

[Bamman and Crane2008] David Bamman and Gregory Crane. 2008. The logic and discovery of textual allusion. In *LaTeCH (Language Technology for Cultural Heritage Data)*, Marrakech Morocco. LREC.

[Bamman and Crane2011a] David Bamman and Gregory Crane. 2011a. The ancient greek and latin dependency treebanks. In *Caroline Sporleder, Antal van den Bosch, & Kalliopi Zervanou (Eds) Language technology for cultural heritage: Selected papers from the LaTeCH Workshop Series*, pages 79–98, Berlin, Germany. Springer-Verlag.

[Bamman and Crane2011b] David Bamman and Gregory Crane. 2011b. Measuring historical word sense variation. In *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital libraries (JCDL 2011)*, pages 1–10. ACM Digital Library.

[Barrón-Cedeño et al.2013] Alberto Barrón-Cedeño, Marta Vila, M.Antònia Martí, and Paolo Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistic*, 39(4):917–947.

[Bensalem et al.2015] Imene Bensalem, Imene Boukhalfa, Paolo Rosso, Lahsen Abouenour, Kareem Darwish, and Salim Chikhi. 2015. Overview of the AraPlagDet PAN@FIRE2015 Shared Task on Arabic Plagiarism Detection. In *FIRE 2015 Working Notes Papers, 4-6 December, Gandhinagar, India*, December.

[Bizzoni et al.2014] Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The making of ancient greek wordnet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Büchler et al.2012] Marco Büchler, Gregory Crane, Maria Moritz, and Alison Babeu. 2012. Increasing recall for text re-use in historical documents to support research in the humanities. In *Theory and Practice of Digital Libraries, Lecture Notes in Computer Science*, volume 7489, pages 95–100. Springer, Berlin Heidelberg.

[Büchler2013] Marco Büchler. 2013. *Informationstechnische Aspekte des Historical Text Re-use (English: Computational Aspects of Historical Text Re-use*. Ph.D. thesis, Leipzig University, Germany.

[Chawathe et al.1996] Sudarshan S Chawathe, Anand Rajaraman, Hector Garcia-Molina, and Jennifer Widom. 1996. Change detection in hierarchically structured information. In *ACM SIGMOD Record*, volume 25, pages 493–504. ACM.

[Clemens1905 1909] Titus Flavius Clemens. 1905-1909. Werke (in greek). In Otto Stählin, editor, *Die Griechischen Christlichen Schrifteller, Berlin, v. 12, 15, 27*. Leipzig.

[Clément d'Alexandrie2011] Clément d'Alexandrie, 2011. *Quel riche peut-être sauvé*. éditions du Cerf, Paris, sources chrtiennes 537 edition.

[Cosaert2008] Carl P. Cosaert. 2008. *The Text of the Gospels in Clement of Alexandria*. New Testament in the Greek Fathers. Society of Biblical Literature.

[Crane1985] Gregory Crane. 1985. Perseus digital library. `http://www.perseus.tufts.edu/hopper/`.

[Crane1991] Gregory Crane. 1991. Generating and parsing classical greek. *Literary and Linguistic Computing*, 6(4):243–245.

[Dolan and Brockett2005] Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

[Freppel1865] Charles-Emile Freppel. 1865. *Clement d'Alexandrie*.

[Ganitkevitch et al.2013] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

[Geyken and Gloning2014] Alexander Geyken and Thomas Gloning. 2014. A living text archive of 15th-19th century german: Corpus strategies, technology, organization. In *Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP*. Narr Tbingen.

[Houghton2013a] H.A.G. Houghton. 2013a. Patristic evidence in the new edition of the vetus latina iohannes. In L. Mellerin and H.A.G. Houghton, editors, *Biblical Quotations in Patristic Texts (Studia Patristica 54)*, pages 69–85. Peeters, Leuven.

[Houghton2013b] H.A.G. Houghton. 2013b. The use of the latin fathers for new testament textual criticism. In B.D. Ehrman and M.W. Holmes, editors, *The Text of the New Testament in Contemporary Research. Essays on the Status Quaestionis second edition. NTTSD.*, pages 375–405. Brill, Leiden.

[Jing1998] Hongyan Jing. 1998. Usage of wordnet in natural language generation. In *Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems (COLING-ACL'98)*. Columbia University Academic Commons.

[Johnson et al.2014 2016] Kyle P. Johnson, Patrick J. Burns, Luke Hollis, Martín Pozzi, Amit Shilo, Stephen Margheim, Gitter Badger, and Eamonn Bell. 2014–2016. Cltk: The classical language toolkit. `https:`

//github.com/cltk/cltk. DOI 10.5281/zenodo.44555 v0.1.32.

[Kehrer2014] Timo Kehrer. 2014. Generierung konsistenzerhaltender editierskripte im kontext der modellversionierung. In Wilhelm Hasselbring and Nils Christian Ehmke, editors, *Software Engineering 2014, Fachtagung des GI-Fachbereichs Softwaretechnik, 25. Februar - 28. Februar 2014, Kiel, Deutschland*, volume 227 of *LNI*, pages 57–58. GI.

[Lee2007] John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic*, pages 472–479. Association for Computational Linguistics.

[Madnani and Dorr2010] Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguist.*, 36(3):341–387, September.

[Mellerin2014] Laurence Mellerin. 2014. New ways of searching with biblindex, the online index of biblical quotations in early christian literature. In Claire Clivaz, Andrew Gregory, and David Hamidovic, editors, *Digital Humanities in Biblical, Early Jewish and Early Christian Studies*, chapter 11, pages 175–192. Brill, Leiden.

[Mellerin2016] Laurence Mellerin. 2016. Biblindex. http://www.biblindex.mom.fr/.

[Miller et al.1990] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312.

[Minozzi2009] Stefano Minozzi, 2009. *Innsbrucker Beitrge zur Sprachwissenschaft*, volume 137, chapter The Latin WordNet Project, pages 707–716. Institut fr Sprachen und Literaturen der Universitt Innsbruck, Innsbruck.

[Pantelia2014] Maria Pantelia. 2014. Thesaurus linguae graecae. http://stephanus.tlg.uci.edu/index.php.

[Patwardhan et al.2003] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen, 2003. *Computational Linguistics and Intelligent Text Processing: 4th International Conference (CICLing)*, chapter Using Measures of Semantic Relatedness for Word Sense Disambiguation, pages 241–257. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Piotrowski2012] Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts (Synthesis Lectures on Human Language Technologies)*. Morgan & Claypool Publishers.

[Rahlfs1935a] Alfred Rahlfs, editor. 1935a. *Septuaginta*. Württembergische Bibelanstalt, 9 edition. 1971.

[Rahlfs1935b] Alfred Rahlfs, editor. 1935b. *Septuaginta, id est Vetus Testamentum Graece juxta LXX interpretes*. Rahlfs. 2 vol., 1950.

[Smith2010] William Smith. 2010. *Catholic Church Milestones: People and Events That Shaped the Institutional Church*. Indianapolis: Left Coast.

[Vinzent et al.2013] M. Vinzent, L. Mellerin, and H.A.G. Houghton, editors. 2013. *Biblical Quotations in Patristic Texts (Studia Patristica 54)*. Theory and Applications of Natural Language Processing. Peeters, Leuven.

[vor der Brück et al.2015] Tim vor der Brück, Steffen Eger, and Alexander Mehler. 2015. Lexicon-assisted tagging and lemmatization in latin: A comparison of six taggers and two lemmatization models. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 105–113, Beijing, China, July. Association for Computational Linguistics.

[Weber R.1969 1994 2007] Gribomont J. Weber R., Fischer B., editor. 1969, 1994, 2007. *Biblia sacra juxta vulgatam versionem*. Deutsche Bibelgesellschaft.

[Zhao et al.2009] Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 834–842, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Zitouni2014] Imed Zitouni. 2014. Natural language processing of semitic languages.