Ambiguity in Semantically Related Word Substitutions: an investigation in historical Bible translations

Maria Moritz **Marco Büchler**
Institute of Computer Science
University of Goettingen
mmoritz@gcdh.de mbuechler@etrap.eu

Abstract

Text reuse is a common way to transfer historical texts. It refers to the repetition of text in a new context and ranges from near-verbatim (literal) and para-phrasal reuse to completely non-literal reuse (e.g., allusions or translations). To improve the detection of reuse in historical texts, we need to better understand its characteristics. In this work, we investigate the relationship between para-phrasal reuse and word senses. Specifically, we investigate the conjecture that words with ambiguous word senses are less prone to replacement in para-phrasal text reuse. Our corpus comprises three historical English Bibles, one of which has previously been annotated with word senses. We perform an automated word-sense disambiguation based on supervised learning. By investigating our conjecture we strive to understand whether unambiguous words are rather used for word replacements when a text reuse happens, and consequently, could serve as a discriminating feature for reuse detection.

1 Introduction

Detecting text reuse is an important means for many scholarly analyses on historical texts. Nonetheless, the detection of para-phrasal reuse in historical texts is not yet well understood. Specifically, techniques borrowed from plagiarism detection (Alzahrani et al., 2012) are quickly challenged when words are substituted.

To improve historical text-reuse detection, we need to better understand the characteristics of reuse–such as the way and the ratio of word substitutions and modifications. We also need to learn about the characteristics of words that are often substituted to identify potential features that automated

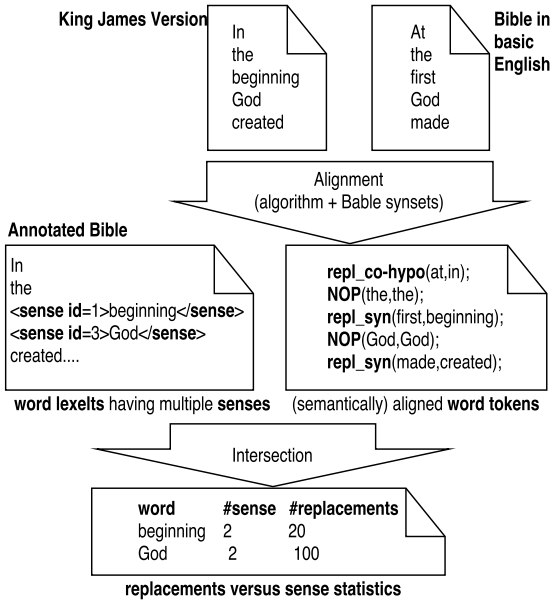

Figure 1: Methodology overview

reuse-detection techniques can take into account. In earlier work, we already investigated the ratios and modifications (morphological and semantic) in two smaller corpora of ancient text. In this paper, we investigate ambiguous words from an upfront word-sense annotated English Bible, and compare them with word substitutions that we find between the verses of this and two further English Bibles each. Since in historical text, text reuse is a way to transfer knowledge, we conjecture that words that are substituted in a para-phrasal, reused verse (of a para-phrasal, parallel corpus) are less likely ambiguous words and do not have multiple senses. We are inspired by Shannon's (1949) conditional entropy, which measures the ambiguity of a received message, i.e., the missing information of a message compared to what was actually sent (cf. Borgwaldt et al., 2005). We conjecture that ambiguous words are likely less specific (informative) and are no good candidates for a substitution (for a reused text in our case).

Fig. 1 illustrates our methodology. First, we determine the intersection of the ambiguous words from the first (word-sense-annotated) Bible and the replaced words between this Bible and the two other Bibles. Second, we disambiguate the two extra Bibles using a k-nearest neighbors classifier and a support vector machine classifier (based on the training data of the annotated Bible) and intersect the ambiguous words found that way (now knowing their numbers of senses as well) again with the replacements collected from the first step, to back-up our findings.

2 Related Work

Some works consider semantic information for detecting text smilarity. Sanchez-Perez et al. (2014) discover sentence similarity based on a tf-idf weighting during text alignment that allows them to keep stop words without increasing the rate of false positive hits in the result set. Their recursive algorithm allows to increase the alignment up to a maximal passage length. By using synset databases, Bär et al. (2012) consider semantic similarity in addition to structural features (e.g., n-grams of POS sequences) and stylistic characteristics (e.g., sentence and word length). They empirically show that taking their suggested wide variety of textual features into account works best to detect text reuse. Their method outperforms previous methods on every dataset on which their method was tested.

Fernando and Stevenson (2008) present an algorithm that identifies paraphrases by using word-similarity information derived from WordNet (Fellbaum, 1998). They experiment with several measures for determining the similarity of two words represented by their distance in the WordNet's hierarchy. Their methods turned out to work slightly better than early works did—to which their methods are compared to.

Some works also consider the influence of word ambiguity for plagiarism detection. Ceska and Fox (2011) investigate whether ambiguous words impact the accuracy of their plagiarism-detection technique. Among others, they examine the removal of stop words, lemmatization, number replacement and synonym recognition, and how they affect accuracy. They find that number replacement, synonym recognition, and word generalization can slightly improve accuracy.

We want to find out about the role of ambiguous words in a reuse scenario to define new requirements for text-reuse detection methods in historical text as a long-term goal.

3 Study Design

We now describe our study design, including our research question, datasets, and tools that we used.

3.1 Research Question

We formulate one research question:

RQ1. *Is there a correlation between words that are often replaced during text reuse and words that are unambiguous (i.e., have one sense only)?*

In other words, we ask whether unambiguous words are more frequently substituted than ambiguous words in reused text. We think that unambiguous words are more likely replacement candidates in a text that is reused, because thy probably transport clearer information. This can depend on the reuse motivation (e.g., the reason to create an edition). However, we want to learn if we can find a trend that follows our conjecture.

3.2 Datasets

We use three English Bibles. The first is the King James Version (KJV) from 1611–1769. It has been annotated with word senses. The other two Bibles are the Bible in Basic English (BBE)—1941–1949—and Robert Young's Literal Translation (YLT). YLT from 1862 very literally follows the Hebrew and Greek language. Because these Bibles follow different linguistic criteria, they offer a greater lexical diversity. We consider both Bibles as the counterpart of the text reuse (target text), and the KJV as source text.

To obtain word senses for the latter two Bibles, we use the word senses of KJV as training data for a machine-learning task, which we then apply to both BBE and YLT.

3.3 Methodology

Our methodology comprises three steps.

1) We identify word substitutions pairwise between KJV and BBE, and between KJV and YLT. Therefore, we align words of a Bible verse hierarchically by first associating identical words and words wich have the same lemma in common, and then we look for synonym, hypernym, hyperonym, and co-hyponym relations between the words of two Bible verses, which we use BabelSenses (Navigli and Ponzetto, 2012), for.

2) We then compare the annotated words (multi- and single-sense words) of the sense-annotated

Bible	tokens	types
KJV	967,606	15,700
BBE	839,249	7,238
YLT	786,241	14,806

Table 1: General lexical information on the corpus

KJV with the substituted words from the former step (cf. Fig. 1).

3) Finally, we identify word senses in both BBE and YLT using a k-nearest neighbors classifier and a support vector machine classifier trained with the KJV annotations, and do the same comparison as in step 2 to see whether our conjecture still holds or not, or only holds for the new replacement words in BBE and YLT.

Step 2 and 3 rely on annotated training data that was created for KJV by Reganato et al. (2016).[1] They used BabelNet synsets (Navigli and Ponzetto, 2012) to identify semantic concepts and disambiguate words using the word sense disambiguation (WSD) system Babelfy (Moro et al., 2014). They performed semantic indexing on their Bible corpus after disambiguation and entity-linking. To evaluate the Babelfy output, they manually annotated two chapters of their Bible. The confidence score of the annotations is between 70%–100%.

4 Ambiguity in Replaced Words

Next, we investigate if words substituted between Bibles are rather unambiguous than ambiguous.

4.1 Data Preparation and Corpus Overview

Because of the age of KJV (18th century), we use MorphAdorner (Paetzold, 2015) for its lemmatization. We use the lemma output from TreeTagger (Schmid, 1999) for both BBL and YLT. We use the lemmas to query the BabelNet API to find synonyms, hypernyms, hyponyms and cohyponyms for a given word. We query BabelNet to find synonyms, hypernyms, hyperonyms, and cohyponyms presenting potential replacements when we compare the Bible verses. For orientation, Table 1 gives an overview of the Bible vocabulary, and Table 2 shows information on the annotation data. Both tables show raw information on the given corpora.

[1] `http://homepages.inf.ed.ac.uk/s0787820/bible`

KJV annotated single-word lexelts	9,927
KJV annotated multi-word lexelts	2,794
total	12,721

Table 2: Information on annotated KJV Bible

4.2 Replacement Statistics

We first calculate the words that are substituted by another word, pairwise between each KJV and BBE, and between KJV and YLT. In Table 3 we list an overview of types and tokens of words containing relations such as synonyms, hyponyms, hypernyms, and co-hyponmys. In total, we find **4,172** lexelts (words that have one or multiple meanings) of the annotated KJV in the intersection with BBE and **3,312** lexelts in the intersection with YLT.

In the following, we show and explain diagrams of the results on these intersections. We relate the number of replacement operations of lexelts to the number of their senses. Note that the y-axis is logarithmic to compress the data points for clarity. In Fig. 2 we normalize the number of replacements between KJV and BBE by the number of senses, with the result that—judged by the box and median values—relatively above a sense number of four, the increase of the number of replacement operations stagnates a bit. This behavior is confirmed in Fig. 3, which shows the replacement operations between KJV and YLT by sense numbers of the replaced lexelts, again relative to the number of senses. Here, a strong increase is visible from four and six senses on (based on box and median).

5 Word Sense Disambiguation Task

Now, we investigate whether we obtain a similar result when we automatically disambiguate the word

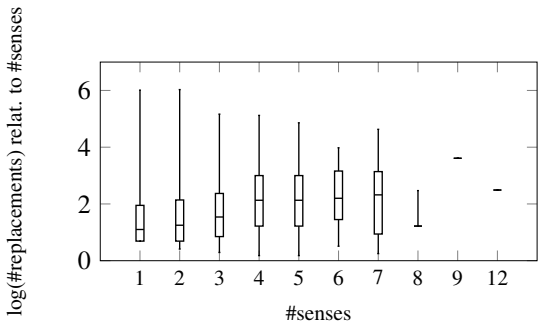

Figure 2: Relative numbers of replacement operations between KJV and BBE, per sense, normalized by number of senses (logarithmic quantities shown)

source Bible	target Bible	subst. types source B.	subst. types target B.	subst. tokens
KJV	BBE	4,947	2,048	150,938
KJV	YLT	3,915	4,094	74,851

Table 3: Substitution statistics between the Bibles

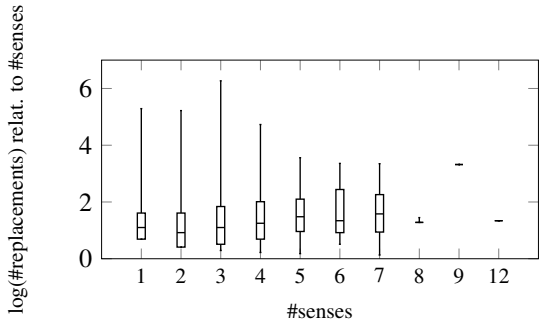

Figure 3: Relative numbers of replacement operations between KJV and YLT, per sense, normalized by the number of senses (logarithmic quantities displayed)

senses using two different machine learning classifiers.

5.1 Preparation of the Experiment

To obtain an understanding of the classifiers' accuracy, we first evaluate them using the given annotation data: we split the 66 files representing the Bible (one book per file) randomly into two thirds for training and one third for testing. We train and test two classifiers (explained shortly). We use three filter criteria for the testing data: i) all word and sense classes are only considered in the testing data if they also appear in the training data; ii) only words (lexelts) with at least two different senses are considered, and iv) only words with at least 30 instances per sense are considered. We choose 30 as the instance threshold, because we work with a 20-tokens-window feature space, thus feature matrices turn out sparse. On the other hand, we want to loose as few words as possible. Table 4 shows the baseline accuracy of this preparatory test, before we run the classifiers on our two other Bibles.

classifier	p	r	correct	attempted	total
KNN	.678	.670	8317	12266	12408
SVM	.679	.672	8334	12266	12408

Table 4: Performance—(p)recision and (r)ecall—of the KNN and SVM on the annotated test data

Classifiers Used: We use two classifiers from the sklearnpackage: the Linear Support Vector Classifier (SVM) and the KNeighbors Classifier (KNN). For the latter, we leave the number of neighbors and the weight at their default value. Table 4 shows the classifiers' ground performance on the training and testing data set from the annotated KJV Bible.
Error Rates per Sense Number: We further calculate the averaged error per sense number for both classifiers on the test data. Table 5 shows the results for the sense number 2, 3 and 4.

5.2 Substitutions in two Automatically Annotated Bibles

Now, we want to identify word senses in the two extra Bibles as well. For performing the WSD analysis on the BBE and the YLT, we use all Bible books of the annotated KJV Bible as training data (but again use only lexelts with at least 30 instances per sense to remain comparable), and the two classifiers already used before.

We find **88** lexelts contained in the intersection set. Next, we describe the results of the intersection. We intersect the words classified by SVM and KNN with the words that were replaced among BBE and KJV. Fig. 4 shows the results. The output of the classified word senses from both, KNN and SVM are intersected with the same replacement operations identified in the previous section. Fig. 4 shows the replacements for both classifiers' output. Again, the ratio of replacements seems to stagnate starting with a sense number of 5 (cf. Sec. 4.2 for information on replaced types and tokens between BBE and KJV, and YLT and KJV).

Next, we run the same procedure using substituted words between YLT and KLV. We find **138** lexelts in the intersection Fig. 5 interestingly shows

classifier	no. of senses		
	2	3	4
KNN	.47	.62	.74
SVM	.46	.60	.70

Table 5: Averaged classification error per sense number for the KNN and SVM classifier

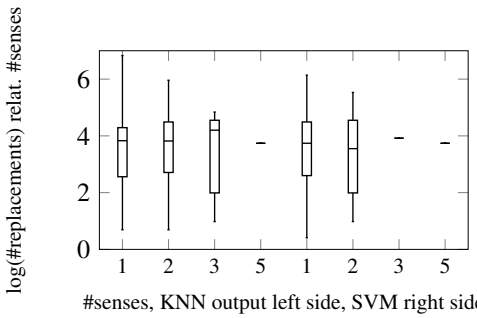

Figure 4: Relative numbers of replacement operations between BBE and KJV, per sense, normalized by number of senses (logarithmic quantities shown)

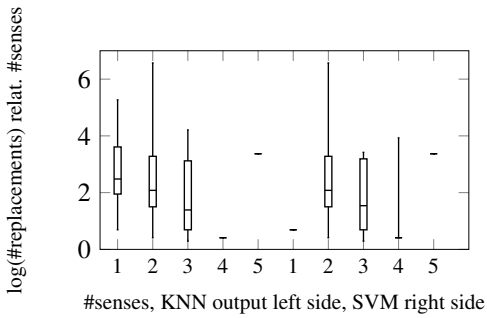

Figure 5: Relative numbers of replacement operations between YLT and KJV, per sense, normalized by number of senses (logarithmic quantities shown)

a decrease of replacements with an increase of the sense number of a word for results found using the KNN classifier. This can be explained by the closeness of YLT's language to the ancient, original text, and that its words in some contexts are less commonly used. Thus, words are substituted between YLT and KJV where none are substituted in between BBE and KJV, e.g.:

- repl_syn(sons,children) in [YLT,KJV], but NOP(children,children) in [BBE,KJV] (cf. Psalm 45:16)
- repl_syn(flames,fire) in [YLT,KJV], but NOP(fire,fire) in [BBE,KJV] (cf. Psalm 57:4)
- repl_syn(prepared,fixed) in [YLT,KJV], but NOP(fixed,fixed) in [BBE,KJV] (cf. Psalm 57:7)
- hypo(honour,glory) in [YLT,KJV], but NOP(glory,glory) in [BBE,KJV] (cf. Psalm 57:8)

Thus, they are good candidates for a replacement in a more common, even if older, translation as it is KJV. The calculated results using the SVM classifier, however, do not show statistically reliable data (too few data points for words with 1, 4 and 5 senses). Hence, we can not form an outcome based on them.

6 Threats to Validity

External Validity: A threat is that the word senses annotated in the King James Version of the Bible are generated from Babel Senses and the Word Sense Disambiguation system Bablelfy. Both use BabelNet synsets as the underlying knowledge base. Since we also use BabelNet to identify semantic relationships between two words of two Bible verses, we possibly find our conjecture influenced negatively from the beginning, because a unique word sense might never be given when its meaning is harvested by means of context vectors, which use a specific, surrounding context. This threat might be overcome in future work. A broader hand-annotated sense inventory together with a WSD classification task might be chosen instead of the given annotated Bible.

Internal Validity: A threat is that we can only find intersections with words that were successfully lemmatized upfront and for which we can find an entry in BabelNet. A lemma lookup failed in 6,210 cases for the BBE Bible and in 11,312 cases for the YLT Bible. No corresponding counterpart for a token was found 139,565 times for the intersection of KJV with BBE, and 83,285 times for YLT. Lemma lookups often failed when words contained special characters (such as "'s") due to a lemma-list cleaning we performed, or when a named entity was not used in both verses, and a lowercase version could not be found. Especially in the automated annotated data we encounter low data points. In the future we want to experiment with different thresholds to find a good setting between recall and precision.

Finally, we intentionally do not call our conjecture hypothesis, since we do not perform hypothesis testing using statistical tests, mainly since the results do not indicate that our conjecture holds. We are currently exploring other potential, discriminating features. Upon indications that they hold, we will perform statistical hypothesis testing.

7 Discussion

Our results show that—against the initial conjecture—the likeliness of a word being replaced correlates to its number of senses (shown by

the fact that—even though normalized—boxes in Fig. 2 to Fig. 5 tend to raise instead of fall). There is no conspicuousness in the use of unambiguous words as potential substitution candidates in a parallel para-phrasal corpus, such as the one used in this paper. Thus, if a word is unambiguous, it is no discriminating criteria for a word to be a potential candidate for replacements in a reuse situation. As mentioned in Sec. 6, this possibly relies on the selection of the resources we use to find semantic relatives (e.g., synonyms) for the words in our parallel Bible corpus.

However, we found an interesting discrimination in the second part of our experiment. It turned out that between the YLT and the KJV indeed more unambiguous words are in the replacement set. This might be influenced by the fact that YLT contains much more types when much fewer tokes were replaced at the same time (cf. Table 3).

Moreover, we only tested the conjecture on one genre (the Bible), whereas it might be possible that other sorts of text reuse behave differently, which also might be a further aspect to investigate.

8 Conclusion

We showed whether and how (ambiguous) words—when substituted—correlate to the number of their senses. In contrast to our initial conjecture, there is no significance in the use of unambiguous words as replacements candidates. Instead, the use of a word as a substitution candidate for para-phrasal reuse increases with the number of the senses of a word. In future work, we strive to compare word substitutions to another sense annotated dataset and to define the ambiguity by a word's appearance in only one or multiple synonym sets directly. In any case, we will further investigate the characteristics of words from reused text to derive more understanding on how text is constituted when reused.

Acknowledgments

Our work is funded by the German Federal Ministry of Education and Research (grant 01UG1509).

References

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.

Daniel Baer, Torsten Zesch, and Iryna Gurevych. 2012. Text reuse detection using a composition of text similarity measures. In *Proceedings of COLING 2012*, pages 167–184, Mumbai, India. The COLING 2012 Organizing Committee.

Susanne R Borgwaldt, Frauke M Hellwig, and Annette M B De Groot. 2005. Onset entropy matters–letter-to-phoneme mappings in seven languages. *Reading and Writing*, 18(3):211–229.

Zdenek Ceska and Chris Fox. 2011. The influence of text pre-processing on plagiarism detection. Association for Computational Linguistics.

Christine Fellbaum. 1998. *WordNet An Electronic Lexical Database*. MIT Press.

Samuel Fernando and Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. *Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloqium*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, December.

GH Paetzold. 2015. Morph adorner toolkit: Morph adorner made simple.

Alessandro Raganato, Jose Camacho-Collados, Antonio Raganato, and Yunseo Joung. 2016. Semantic indexing of multilingual corpora and its application on the history domain. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 140–147, Osaka, Japan. The COLING 2016 Organizing Committee.

Miguel A Sanchez-Perez, Grigori Sidorov, and Alexander F Gelbukh. 2014. A winning approach to text alignment for text reuse detection at pan 2014. In *CLEF (Working Notes)*, pages 1004–1011.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Claude E Shannon. 1949. Communication theory of secrecy systems. *Bell Labs Technical Journal*, 28(4):656–715.