AMBIGUITY IN SEMANTICALLY RELATED WORD SUBSTITUTIONS:

AN INVESTIGATION IN HISTORICAL BIBLE TRANSLATIONS

Maria Moritz and Marco Büchler

May 22 2017, *NoDaLiDa 2017 Workshop on Processing Historical Language*

INTRODUCTION

TEXT REUSE

Text Reuse:

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

For example:

- citations, allusions, and 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.

BACKGROUND

Detecting paraphrased and non-literal reuse is challenging

- See studies of reuse (Alzahrani et al., 2012) and plagiarism (Barrón-Cedeño et al., 2013) detection show that when reuse is modified (words changed) or paraphrased, most approaches are challenged.

Historical Text Reuse Detection is problematic as it comes with

- variants due to long transmission time, incomplete/erroneous witnesses, and diversity.

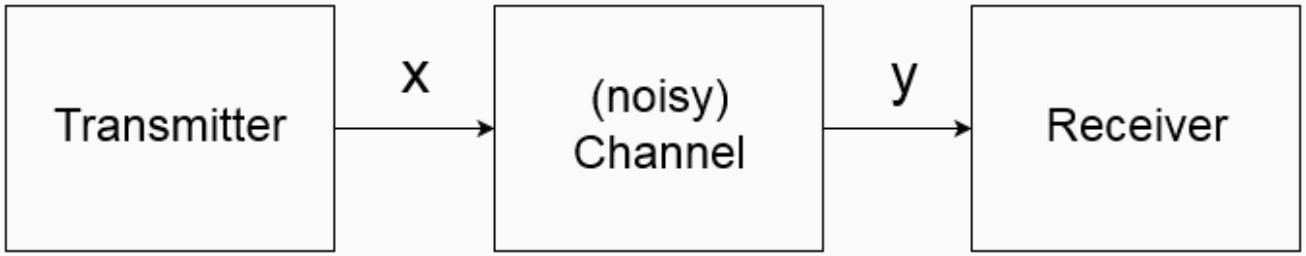
One solution: Reuse Style Investigation

- I.e., we need to learn how reuse is transferred, how literal it is, what kind of modification takes place, and further characteristics in reuse,
- To identify potential features that detection approaches can take into account.

UNAMBIGUOUS WORDS AS SUCH FEATURES?

Motivation

- Para-phrasal text reuse is a way to transfer knowledge.
- We are inspired by Shannon's (1949) conditional entropy (measures the information loss/ambiguity of a received message).
- We conjecture that ambiguous words are likewise less informative and no good substitution candidates for para-phrasal reuse (unsuitable as discriminating features).



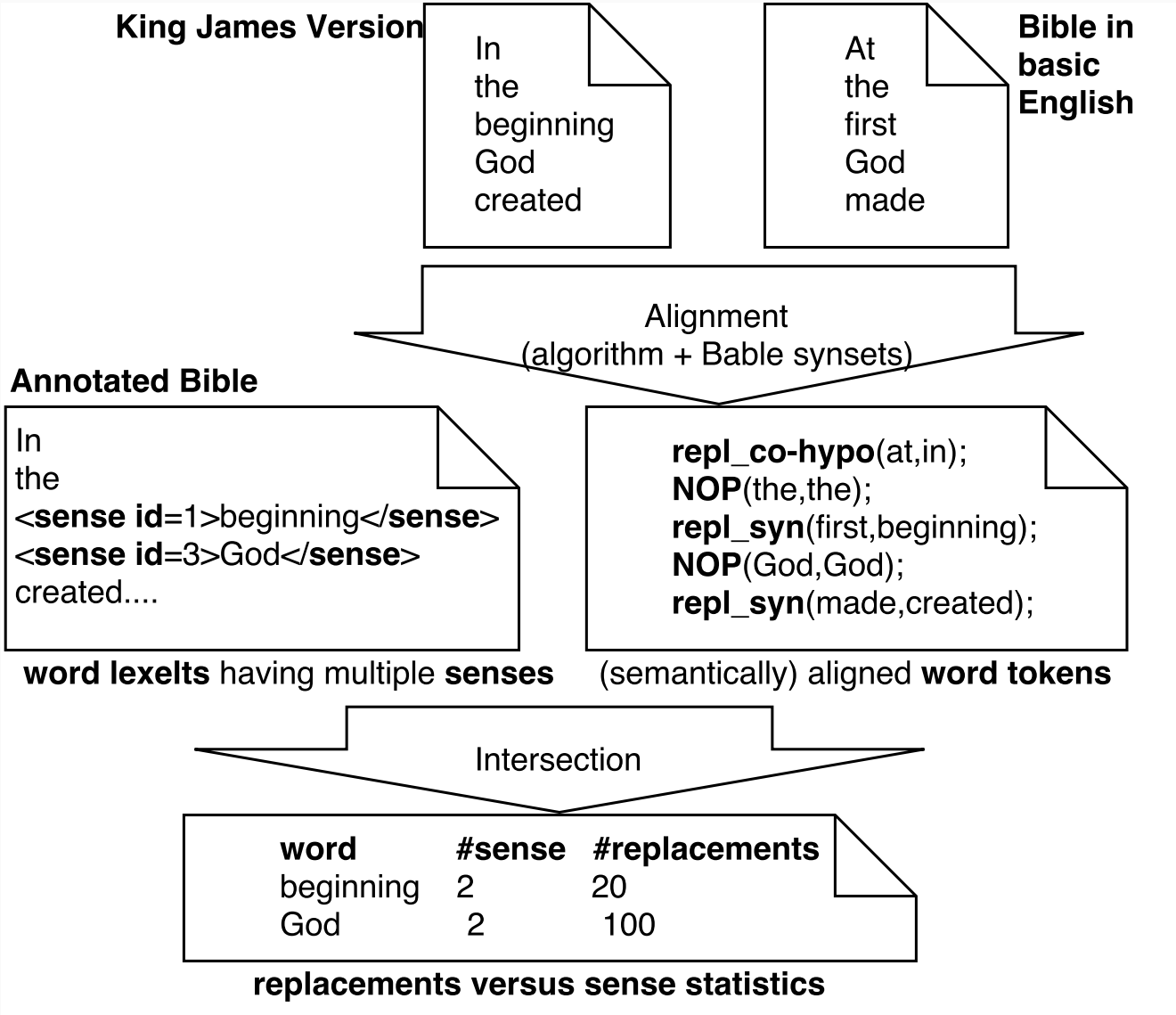
RESEARCH QUESTION

Is there a correlation between words that are often replaced during text reuse and words that are unambiguous?

METHODOLOGY

- We extract ambiguous words and their no. of senses from an upfront word-sense annotated English Bible.
- We identify word substitutions (e.g., synonyms, hypernyms, etc.) between two verses of this and two further Bibles.
- We intersect the words to compare substitution no. and no. of word senses.

METHODOLOGY



DATA AND DATA PREPARATION

DATA

We use

- King James Version[1](KJV, 1611–1769): word-sense-annotated by Raganato et al. (2016),
- The Bible in Basic English (BBE, 1941–1949), and
- Robert Young's Literal Translation[2](YLT, 1862), literally following Hebrew and Greek words and syntax.
- These Bibles follow different linguistic criteria, offering lexical diversity.
- We consider BBE and YLT the counterpart of the text reuse (target text), and the KJV the source text.

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

Table 1: Corpus figures

[1] http://www.biblestudytools.com/

[2] http://paralleltext.info/

DATA PREPARATION

- We lemmatize KJV (18th cnt.) using MorphAdorner (Paetzold, 2015), BBE, and YLT using Tree-Tagger (Schmid, 1999).
- We query the lemmas in BabelNet API to find synonym, hypernym, hyponym, and cohyponym relations between the words of two verses:

source B.	target B.	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 2: Substitutions between the Bibles

- We intersect those with the 9,927 single-word lexelts in KJV
- and find 4,172 lexelts in substitutions between KJV and BBE, and 3,312 between KJV and YLT.

RESULTS

RESULTS OF SUBSTITUTIONS I

A word's no. of replacements correlates to the no. of its senses.

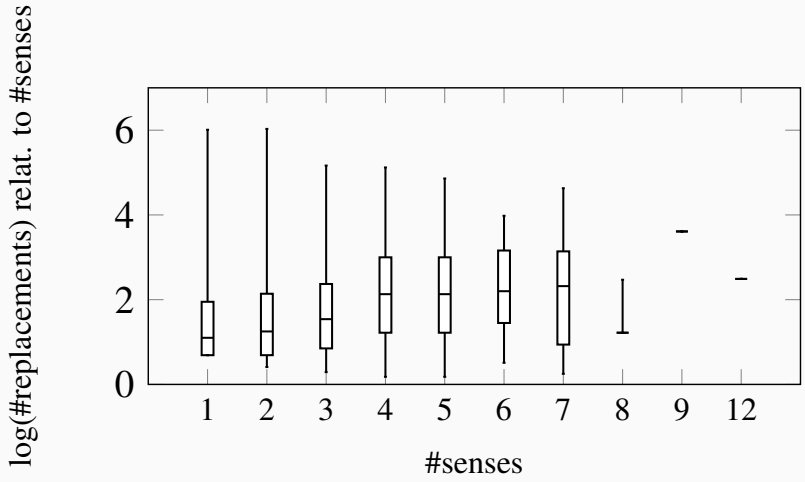


Figure 1: No. of replacements between KJV and BBE, per sense, normalized

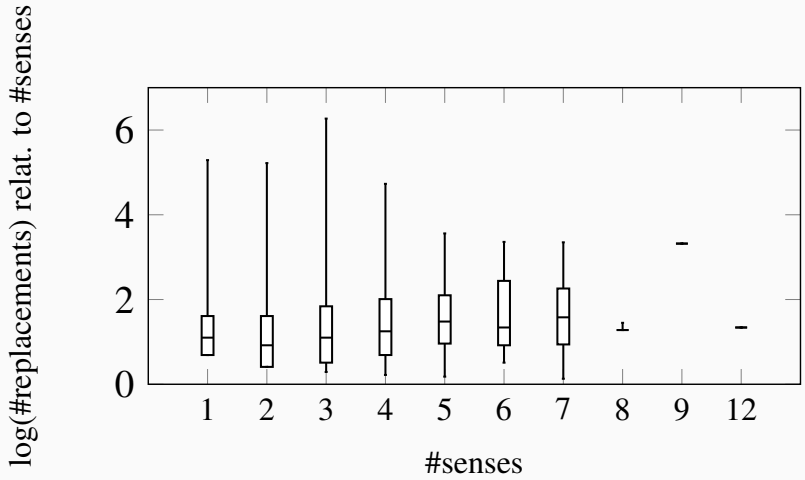


Figure 2: No. of replacements between KJV and YLT, per sense, normalized

SUBSTITUTIONS BETWEEN BBE (YLT) AND KJV

Identify senses using supervised learning

- Classifiers: SVM & KNN classifier
- Training data: KJV as training data
- Training criteria:
  - Words must have at least two different senses and 30 instances per sense to avoid a too sparse 20-tokens-window feature space, but still train with as many words as possible
- Again intersecting the classified words with those replaced among BBE (YLT) and KJV, we find 88 (138) lexelts in the intersection set.

RESULTS OF SUBSTITUTIONS II

Between YLT and KJV the no. of a word's replacements decreases with the increase of its sense.

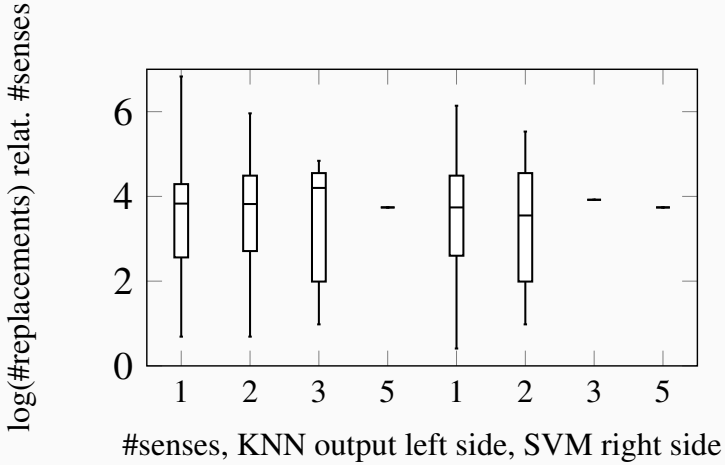


Figure 3: No. of replacements between BBE and KJV, per sense, normalized

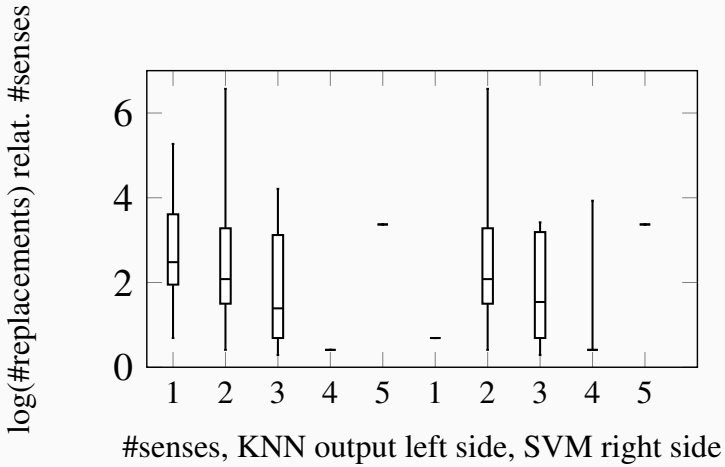


Figure 4: No. of replacements between YLT and KJV, per sense, normalized

RESULTS OF SUBSTITUTIONS II

Decrease of replacements; potential explanation: words in some contexts are less commonly used. E.g. words substituted between YLT and KJV, but not 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 (interesting/discriminating features) in a more common, even if older, translation as it is KJV.

CONCLUSION

SUMMARY

UNAMBIGUOUS WORDS AS SUCH FEATURES?

Motivation

- Para-phrasal text reuse is a way to transfer knowledge.
- We are inspired by Shannon's (1949) conditional entropy (measures the information loss/ambiguity of a received message).
- We conjecture that ambiguous words are likewise less informative and no good substitution candidates for para-phrasal reuse (unsuitable as discriminating features).

5/22

RESEARCH QUESTION

Is there a correlation between words that are often replaced during text reuse and words that are unambiguous?

6/22

METHODOLOGY

- We extract ambiguous words and their no. of senses from an upfront word-sense annotated English Bible.
- We identify word substitutions (e.g., synonyms, hypernyms, etc.) between two verses of this and two further Bibles.
- We intersect the words to compare substitution no. and no. of word senses.

7/22

RESULTS OF SUBSTITUTIONS I

A word's no. of replacements correlates to the no. of its senses.

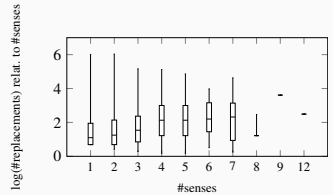


Figure 1: No. of replacements between KJV and BBE, per sense, normalized

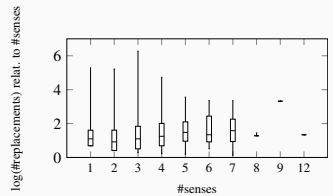


Figure 2: No. of replacements between KJV and YLT, per sense, normalized

13/22

We could not find a conspicuous pattern in the preferred use of unambiguous words as substitution candidates, instead this depends on the direction and intention of the

FUTURE WORK

Clarify the ambiguity and refine the research question

- Possibly use another sense-annotated dataset or define ambiguity by a word's appearance in multiple synsets.
- Refine the research question and use a uniform data format (i.e., investigate which words are replaced with which others in more detail).

Thank you!

http://www.etrap.eu/

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.
- Barrón-Cedeño and Marta Vila and 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.
- G. H. 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), 140–147, Osaka, Japan. The COLING 2016 Organizing Committee.
- Helmut Schmid (1999) Improvements in part-of speech tagging with an application to german. In Natural language processing using very large corpora, 13–25. Springer.
- Claude E Shannon (1949) Communication theory of secrecy systems. Bell Labs Technical Journal Vol. 28, No. 4:656–715.

LICENCE