Modification Analysis in Historical Paraphrastical Parallel Text

An Empirical Work on Stable and Changing Elements in Historical Text Reuse

DISSERTATION

for the award of the degree

DOCTOR RERUM NATURALIUM

of the Georg-August-Universität Göttingen

within the Programme in Computer Science (PCS)

of the Georg-August University School of Science (GAUSS)

submitted by

Maria Berger (née Moritz)

from Leipzig

Göttingen, February 2019

Advisory Committee

Dr. Marco Büchler, eTRAP Early Career Research Group, Institute of Computer Science, University of Goettingen

Prof. Dr. Ramin Yahyapour, Practical Informatics, Institute of Computer Science, University of Goettingen, GWDG

Prof. Dr. Dieter Hogrefe, Telematics, Institute of Computer Science, University of Goettingen

Members of the Examination Board:
Reviewer:

Dr. Marco Büchler, eTRAP Early Career Research Group, Institute of Computer Science, University of Goettingen

2nd Reviewer:

Prof. Dr. Caroline Sporleder, Digital Humanities, Göttingen Centre for Digital Humanities, Institute of Computer Science, University of Goettingen

Further members of the Examination Board:

Prof. Dr. Ramin Yahyapour, Practical Informatics, Institute of Computer Science, University of Goettingen, GWDG

Prof. Dr. Dieter Hogrefe, Telematics, Institute of Computer Science, University of Goettingen

Prof. Dr.-Ing. Marcus Baum, Data Fusion, Institute of Computer Science, University of Goettingen

Prof. Dr. Carsten Damm, Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, University of Goettingen

Prof. Dr. Stephan Waack, Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, University of Goettingen

Date of oral examination: May 2nd 2019

Abstract

Clarifying the genesis of a passed down text is of outmost importance for many scholarly disciplines within the humanities such as history, literary studies, and Bible studies. The computational detection of such passed down texts in the form of historical text reuse, including citations, quotations or allusions, unintended reuse of a saying, or even of cross-linguistic reuse in the form of translations, 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 It can help tracing down historical content (a.k.a., lines of transmission), which is essential to the field of textual criticism. In modern literature it can help assigning text to authors. 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 is an important task. While detecting reuse in contemporary languages is well-understood—given the existence of extensive research, techniques, and corpora, automatically detecting historical text reuse is much more difficult. Corpora of historical languages often encompass various genres, linguistic varieties, and topics. In fact, the automated detection of historical text reuse is much less understood, requiring empirical work to improve its automation. Especially, the analysis of text reuse by quantitative methods is crucial to understand reuse in detail.

This work presents a technique for describing text reuse modification on a fine-grained level and collects empirical data based on the application of the technique to several datasets and use cases. In detail, this work presents a linguistic analysis of text reuse in two medieval datasets. In a more comprehensive analysis, it investigates modifications in a monolingual parallel corpus of English Bible translations and a parallel Corpus of German Bible translations. We design and implement an automated technique to analyze how a source text is modified compared to its reuse/parallel version, taking linguistic resources into account to understand how they help characterizing the transformation. Precisely, an operation set is designed considering operations based on morphological cognates and lexicon-based operations based on semantic relations to find a mapping between a source text and its reused/parallel version and apply it on top of a statistical alignment output to learn how precisely and to what extent text is modified. The work is complemented by a manual analysis of subsets of the medieval reuse datasets, and a manual evaluation of the alignment precision on subsets of the English Bible Corpus.

The results show the lack of resources for ancient texts, while lexical database for modern languages are widely available and can partially enhance the technique presented in this work. However, especially for a sufficiently preprocessed historical English text, linguistic resources can effectively support understanding the paraphrastical text reuse modification process. These results can support practitioners and researchers working on detecting historical reuse.

Zusammenfassung

Die Klärung der Entstehung eines überlieferten Textes ist für viele geisteswissenschaftliche Disziplinen wie beispielsweise der Geschichte, Literaturwissenschaft oder Bibelwissenschaft von größter Bedeutung. Die automatische Erkennung solcher überlieferten Texte in Form historischen Text Reuses—dies beinhaltet Zitationen, Zitate oder auch Andeutungen, sowie unbeabsichtigten Reuse eines Sprichworts oder sogar Fälle von sprachübergreifendem Reuse in Form von Übersetzungen—kann in vielerlei Hinsicht nützlich sein. Sie kann dabei helfen, historische Inhalte aufzuspüren, was zum Beispiel für das Forschungsgebiet der Textkritik von wesentlicher Bedeutung ist. In der modernen Literatur kann die Text-Reuse-Erkennung aber auch hilfreich sein, um Text Autoren zuzuordnen. Im Rahmen massiver Digitalisierungsprojekte können Beziehungen zwischen Textausschnitten identifiziert werden, die sich auf ein und dieselbe Quelle beziehen. Insbesondere das Erkennen von Kopien desselben historischen Textes, die im Laufe der Zeit voneinander abgewichen sind, ist eine wichtige Aufgabe der Text-Reuse-Erkennung. Während der Erkennung von Text Reuse in modernen Sprachen viel Aufmerksamkeint entgegen gebracht wird, und Studien aufgrund reichlich existierender Technologien und Text Korpora erleichtert werden, ist die automatische Erkennung von historischem Text Reuse viel schwieriger. Korpora historischer Sprachen umfassen oft verschiedene Gattungen, sprachliche Variationen und Themen. Tatsächlich ist die automatische Erkennung von Text Reuse in historischen Texten viel weniger bekannt, und empirische Studien sind notwendig um dessen Automatisierung zu ermöglichen und zu verbessern. Zu diesem Zweck ist die Analyse von Text Reuse mittels quantitativer Methoden unumgänglich. Dies hilft die Einzelheiten des Text Reuse zu verstehen, um schließlich exitierende Methoden zur Text Reuse Erkennung zu verbessern.

Diese Arbeit präsentiert eine Technik zur Beschreibung fein-granularer Veränderung von Text Reuse und erhebt empirische Daten, die auf der Andwendung dieser Technik auf verschiedenen Datensätzen und Use-Cases basieren. Im Detail präsentiert diese Arbeit eine sprachliche Analyse von Reuse in zwei kleineren Datensätzen mittelalterlichen Griechischs und Lateins. In einer umfassenderen Analyse wird Wortveränderung und -Ersetzung in einem parallelen Korpus englischer Bibelübersetzungen und einem parallelen Korpus deutscher Bibelübersetzungen untersucht. Es wird ein automatisierte Ansatz entworfen und implementiert, der hilft zu analy-

sieren wie ein Quelltext im Vergleich zu seinem Reuse beziehungsweise seiner parallelen Version verändert wurde. Dabei werden sprachlichen Ressourcen berücksichtigt, um zu verstehen was die Transformation charakterisiert. Es werden Operationen defininert, die auf morphologischen Veränderungen basieren, sowie Operationen, die auf semantischen Beziehungen basieren, um eine Zuordnung zwischen einem Quelltext und seiner wiederverwendeten Version zu finden. Diese Operationen werden im Nachgang eines statistischen Ansatzes zwischen potentiellen Wortpaaren modeliert. Dadurch werden Einsichten dazu erlangt, wie genau Text verändert wird. Ergänzt wird diese Arbeit durch eine manuelle Analyse von Teildatenbsätzen der mittelalterlichen Texte sowie einer manuellen Beurteilung der Alignmentgenauigkeit auf einem Teildatensatz des englischen Bibelkorpuses.

Die Ergebnisse zeigen den Mangel an Ressourcen für antike Texte, während lexikalische Datenbanken für moderne Sprachen reichlich vorhanden sind. Insbesondere für einen ausreichend vorverarbeiteten historischen englischen Text können Sprachressourcen jedoch das Verständnis des Modifikationsprozesses für paraphrastischen Text Reuse unterstützen. Diese Ergebnisse können Praktikern und Forschern dabei helfen die Erkennung historischen Text Reuses voranzutreiben.

Acknowledgments

I would like to thank Marco Büchler my supervisor, examiner and team leader for his constant guidance and availability. From the time I entered the field of Digital Humanities—when I wrote my Master's thesis—up until now, the completion of the PhD thesis he was not only my manager, but also my teacher in matters of research. I would also like to thank Caroline Sporleder for agreeing to review the thesis as well and for her valuable input whenever I had questions. Thanks also to the examination board which take care of the evaluation of this work. I thank the BMBF (FK/no. 01UG1509) for funding this work.

I thank all my great colleagues—from my Leipzig and Göttingen times, and the research stay in Jena—for the ongoing discussions and the support they gave me while growing as a researcher. Especially, I want to thank Greta Franzini who accompanied me the longest time of my research life for being available so many times as a consultant, a reflector on research ideas and a good friend.

Finally, I want to thank my family that never got tired to listen to my excessive explanations on the topics that occupied me from time to time, and my fiancé Thorsten Berger who always was ready to reassure me, to support me with answering countless questions on the whole topic of research, publishing and traveling, and many more.

List of publications included in the thesis

2018

Maria Moritz, Johannes Hellrich & Sven Büchel. A Human-Interpretable Method to predict Paraphrasticality. In: Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2018). Santa Fe, Aug. 20-26, 2018, pp. 113–118. ACL.
https://www.aclweb.org/anthology/W18-4513

Maria Moritz, Johannes Hellrich & Sven Büchel. Towards a Metric for Paraphrastic Modification. In: Proceedings of Digital Humanities (DH 2018). Mexico City, June 26-29, 2018, pp. 457–460. ADHO.
https://dh2018.adho.org/wp-content/uploads/2018/06/dh2018_abstracts.pdf

Maria Moritz. On the Impact of Time Proximity on the Alignment of Spelling Variants in Old English Bibles: A Case Study. In: Proceedings of Corpus-based Research in the Humanities (CRH-2). Vienna, Jan. 25-26, 2018, pp. 143-–151. Gerastree, Wien.
https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/CRH2.pdf

2017

Maria Moritz & Marco Büchler. An Automated Approach to Model the Transformation Process of the Reuse in Bernard de Clairvaux: How Do Lexical Resources help?. In: Proceedings of Digital Humanities (DH 2017). Montreal, Aug. 8-11, 2017, pp. 533—536. ADHO.
https://dh2017.adho.org/abstracts/142/142.pdf

Maria Moritz & Marco Büchler. Ambiguity in Semantically Related Word Substitutions: an investigation in historical Bible translations. In: Proceedings of the Workshop on Processing Historical Language at NODALIDA 2017 (ProcHistLang 2017). Gothenburg, May 22, 2017, pp. 18–23. Linköping University Electronic Press.
https://www.aclweb.org/anthology/W17-0505

2016

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, & Marco Büchler. Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and its Application to Bible Reuse. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, TX, Nov. 1-5, 2016, pp. 1849—1859. `https://www.aclweb.org/anthology/D16-1190`

Publications not included in the thesis

2018

Greta Franzini, Marco Passarotti, **Maria Moritz**, & Marco Büchler. Using and evaluating TRACER for an *Index fontium computatus* of the *Summa contra Gentiles* of Thomas Aquinas. In: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). Turin, Italy, December 10-12, 2018, No page information. AILC.

Maria Moritz & David Steding. Lexical and Semantic Features for Cross-lingual Text Reuse Classification: an Experiment in English and Latin Paraphrases. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan, May 7-12, 2018, pp. 1976–1980. ELRA.

2017

Markus Paluch, Gabriela Rotari, David Steding, Maximilian Weß, **Maria Moritz**, & Marco Büchler. Analysis of Part-Of-Speech Tagging of Historical German Texts. In: Proceedings of the International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2017). Göttingen, Germany, June 1-2, 2017, pp. 41–46.

2016

Maria Moritz, Barbara Pavlek, Greta Franzini, & Gregory Crane. Sentence Shortening via Morpho-Syntactic Annotated Data in Historical Language Learning. Journal on Computing Cultural Heritage (JOCCH). 9, 1, Article 3 (February 2016), 9 pages. ACM.

2014

Marco Büchler, Greta Franzini, Emily Franzini, & **Maria Moritz**. Scaling Historical Text Re-Use. In: Proceedings of the IEEE International Conference on Big Data 2014 (IEEE BigData 2014). Washington DC, Oct. 2014, pp. 27–30.

Maria Moritz, Monica Lent, Thomas Köntges, Emily Franzini, Maryam Foradi, & Gregory Crane. AncientGeek: Primary Sources Powering Historical Language Learning. In:

Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (ELEARN 2014). New Orleans, Louisiana, Oct. 2014, pp. 2167–2176. Chesapeake, VA: AACE.

Frederik Baumgardt, Monica Berti, Giuseppe Celano, Gregory R. Crane, Stella Dee, Maryam Foradi, Emily Franzini, Greta Franzini, Simon Frazier, Monica Lent, **Maria Moritz**, & Simona Stoyanova. Open Philology at the University of Leipzig. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik, Iceland, May 26-31, 2014, pp. 1682–1685. ELRA.

2012

Marco Büchler, Gregory Crane, **Maria Moritz** & Alison Babeu. Increasing Recall for Text Reuse in Historical Documents to Support Research in Humanities. In: Proceedings of Theory and Practise of Digital Libraries (TPDL 2012). Paphos, Cyprus, Sept. 23-27, 2012, pp. 95–100.

Contents

List of Tables

List of Figures

Abbreviations

AGWN Ancient Greek WordNet.

BLEU bilingual evaluation understudy.

DH digital humanities.

HMM hidden Markov model.

HTRD historical text reuse detection.

LW Latin WordNet.

MT machine translation.

NLP natural language processing.

NOP no operation.

OCR optical character recognition.

OP Operation: a modification happening to a word resulting in a modified version of that word.

POS part of speech.

TM text mining.

VSM vector space model.

Chapter 1

Introduction

Text reuse is the written repetition of text, sometimes in a new context. Clarifying the genesis of a passed down text is of outmost importance for many scholarly disciplines within—but not restricted to—the humanities, such as history, literary studies, and Bible studies. The computational detection of such passed, reused text in the form of historical text reuse—including (verbatim) quotations, allusions, the unintended reuse of a saying, or even cases of cross-linguistic reuse in the form of translations—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), or it can help assigning a text to an author (Gupta & Lehal, 2009; Steyvers* et al.*, 2004) if the original author is not clear. In the context of massive digitization projects, text reuse detection 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*)*[1] *is an important task. Finally, new insights from tasks that are originally motivated by the detection of historical text reuse, can be used to foster research in the field of plagiarism detection alike. This thesis' goal is analyzing historical text reuse to get deeper insights into how text changes when it is reused. Hence, it contributes to improve historical text reuse detection.*

1.1 Background

This section gives an overview of the background of this study. It introduces the role of natural language processing in the context of digital humanities, explains its challenges, and starts motivating the research work of this thesis.

[1] `http://www.oxfordreference.com/view/10.1093/oi/authority.20110803100530975`

1.1.1 Overview of natural language processing

The field of natural language processing (NLP) focuses on the processing of natural language text to make it readable, mineable, and "understandable" by a machine to efficiently support a human's work with collections of textual data that is not manually tackleable anymore (e.g., Manning & Schütze, 1999; Manning *et al.*, 2014). In the context of digital humanities (DH) NLP plays the important role of a cross-sectional discipline, because most of DH's research questions circle around the preparation of textual data or textual description of non-textual data. One important goal that NLP in DH needs to address is that algorithmic results and output need to be interpretable, clear, and understandable to the humanist who uses NLP technologies to address her research questions. This is especially important, because applying tools to text causes modification, interpretation, and possibly the loss of information, which must be strongly traceable by the humanist (see e.g., Piotrowski, 2012).

Text mining (TM) is a sub-field of NLP that handles the process of extracting information from text. It contains sub-areas such as information extraction and retrieval, data mining, and lexical text analysis. Examples of NLP tasks are named-entity recognition, the querying of semantically similar text documents[2] or plagiarism detection (e.g., Heyer *et al.*, 2006). The context of this thesis is in improving plagiarism detection techniques and its adaptation in the field of DH. Precisely, historical text reuse detection (HTRD) differs from plagiarism detection, because of the characteristics that historical text ships with (e.g., strong spelling variations, absence of writing standards, fragmentary witness). That and the requirements stated above are the reasons why we address DH concerns by the use of TM techniques. That means that we use NLP and TM techniques to find, analyze, and visualize text, data and results that are collected in DH research.

One indispensable concept shall be introduced here already, because it is core to NLP and TM in the context of HTRD. That is Zipf's Law (Zipf, 1949), which states that the frequency of a word in a corpus of natural language is inversely proportional to its rank. The rank is the number of a word of a natural language text corpus when all words were ordered by its frequency decreasingly. The distribution of words of a corpus follows a power-law. Among the most frequent words are mainly so-called function or stop words, which cover a high ratio of all word tokens of a running text. In historical texts it is, however, critical to solely rely on this law, because words come with different writings, and inflection is stronger in historical English compared to contemporary English. This anomaly show-cases only one of the challenges that we encounter when working with historical text, because the words and their frequency are often used as components in a base measure to determine the similarity of two texts.

[2]Semantically similar text has the same meaning while using different vocabulary.

Further techniques to measure semantic similarities of texts are based on the distributions of words in a document. This means, different texts that have several words in common are to a certain degree similar. One established techniques to measure the similarity of texts is the vector space model (VSM), which represents the whole vocabulary of a text as a vector of the frequency of each word, and the cosine measure of the two vectors describes the degree of textual similarity (see Salton *et al.*, 1975). Again, remember that techniques relying solely on the vocabulary, and its frequency, that two text share, is not sufficient in the area of historical text reuse detection.

1.1.2 Challenges in reuse and plagiarism detection

Recognizing modified text—i.e., reuse or plagiarism—is difficult in general. Alzahrani *et al.* (2012) study plagiarism detection techniques: ngram-, syntax-, and semantics-based approaches. However, 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 density challenges plagiarism detection, and that lexical substitution and insertion is the most frequent technique of 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[3] plagiarism detection, but did not achieve usable results for intrinsic[4] plagiarism detection.

Further, also modern language text is affected by constant modification, for example, when meaning (i.e., polysemy) and use (see, e.g., Crossley *et al.*, 2010) changes in different domains. These challenges are caused especially by the change of language, which happens to historical text that is transported over centuries (see, e.g., Hellrich & Hahn, 2017).

1.1.3 Challenges in the detection of historical text reuse

Many more challenges arise when historical text needs to be processed for text reuse detection. These range from impaired digitization output to substantial differences in the research culture between humanities and computer science (Heyer & Büchler, 2010). Typical statistical approaches from the field of machine learning are difficult to apply to historically transferred texts, either because models do not exist, the critical mass of data for training does not exist, or the text data is too heterogeneous with respect to epoch and domain. Consequently, only sparse data is available for a certain period. Additionally, historical text has often been copied continuously over hundreds of years, being subject to constant modification. Hence, it comprises many different writing styles, text variants, paraphrasing,

[3] comparing a document to a set of reference documents for plagiarism
[4] finding writing style changes within one document

and other forms of non-literal reuse style (Büchler, 2013). The most important challenges, however, 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, or the tools for automatically processing such texts (see, e.g., Piotrowski, 2012; Geyken & Gloning, 2014; Zitouni, 2014).

To this end, we need to improve the quantitative empirical understanding of such reuse. However, only few works exist that started to empirically analyze modification between different text versions. These also have narrower focuses (see, e.g., Ketzan & Schöch, 2017), investigate the change of modification as a grammatical function (see e.g. Biber & Clark, 2002), and focus on the editorial life-cycle of a text (so called "fluid text", read Bryant, 2002).

Therefore, this thesis strives to investigate non-literal text reuse by means of qualitative and quantitative methods to improve the empirical understanding of historical, non-literal text reuse.

1.2 Research aim and significance of the study

Motivation for the research: The term text reuse refers to quoting, copying or alluding text excerpts from a text resource to a new context. Detecting such reuse is core to answering many important research questions in the humanities. Examples are the identification of Fragmentary Authors. These authors' thought only survived by other authors quoting, alluding, or copying them (Berti *et al.*, 2016). However, the resulting mixed texts need to be cleaned to reconstruct history.

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. These texts were not only transferred over a longer time, they were also modified to fit different contexts, time epochs or cultural backgrounds. Hence, a historical text is not simply copied and pasted to be reuse, it is culturally and linguistically adapted, continuously exposed to transformation errors due to the absence of any spelling and grammar standards. In fact, HTRD is much less understood, and empirical studies are necessary to enable and improve its automation.

Problem statement: Measures based on machine learning often are able to express some kind of similarity between two semantically equivalent text excerpts, but can not describe these similarity in detail and are not designed to record different degrees of modification, and what causes this modification. Hence, the analysis of text reuse by means of quantitative methods is important to understand the broader context of the process of reusing in order to improve reuse detection approaches. We think that the linguistic characteristics of a

reuse, compared to those of the original text, can help to understand the act of reusing and, consequently, help to discover reuse.

Research aim and significance of the study: This thesis investigates the text reuse process and contributes a technique to fit operations on each word of a reuse. This study defines an operation set to find a fine-grained mapping between a sentence or verse-aligned source text and its reused version. The operations follow the preprocessing steps that are applied on a text in preparation of a retrieval task such as normalization and lemmatization. Further operations reflect the semantic relationships two aligned/related words have according to their lexical classification. For this purpose, the study also takes linguistic resources into account to understand how they help characterizing the word transformations and modifications occurring during the reuse process. The operations are fitted using an algorithm conceived in this thesis, on several datasets. The empirical results show how text is reused in detail. Implications that affect the development of text reuse detection techniques that come with the empirical results are discussed. Analysis of text reuse in a range of different datasets is aggregated. The datasets comprise mainly Medieval Greek and Latin, as well as Early Modern English and Early New High German and New High German.

Impact of the study: The results show how and to what extend linguistic resources can support the task of reuse modification analysis especially for old text, and whether and how they can effectively support understanding the non-literal text reuse transformation process. The results can also support practitioners and researchers working on understanding and detecting historical text reuse. The results indicate the degrees of importance of i) several preprocessing steps—as modification is modeled using operations that are inspired by preprocessing steps—and ii) the consultation with lexical resource in order to capture the richness of historical reuse and to foster its detection capability. The long-term goal is to conceive robust text reuse detection techniques for historical texts.

1.3 Research hypotheses, questions, and methodology

This thesis addresses the analysis of non-literal/paraphrastic reuse in different datasets that come with different characteristics. Hence, some of the formulated research questions address similar golas and differ only slightly depending on the data to be investigated and the resources available. An overview of all research questions addressed follows in Ch. 5 to 7.

1.3.1 Research questions I

The main motivation is to study given reuse to learn about how reuse is performed in detail, and what specific changes are applied. The main research questions investigated in the **m**edieval texts (therefore, RQ **M**) in Ch. 5 are:

- **RQ M1** *What is the extent of non-literal reuse in our datasets?* We first, generally determine how much of the reuse is literal (no change) and how much is non-literal (morphological or lexical change).
- **RQ M2** *How is the non-literally reused text modified in the datasets when it was transported and reused?* We study frequencies of semantic, lexical, and morphological changes and develop an automated approach to identify the reuse transformation, and complement it with a qualitative analysis.

The chapter also investigates dictionary and database support of existing linguistic resources, refining the second question into three sub-questions:

- **RQ M2.1** *How can linguistic resources support the discovery of non-literal reuse?* The conjecture is 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 lexical database, and investigate how useful they are for understanding and defining the reuse transformations.
- **RQ M2.2a** *What are the limitations of an automated analysis relying on linguistic resources?* A 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 the datasets.

One more aim is the investigation of the database support of one lexical database created for Ancient Greek and Latin specifically, and one lexical database that is mainly built from modern language resources, some of which are also available in Latin. Hence, the third sub-question reads again:

- **RQ M2.2b** *What are the limitations of an automated approach to categorize modification relying on linguistic resources?* Both of the lexical databases are compared with regard to how well they support the categorization modification for modern languages that also supports Latin.

1.3.2 Research questions II

The next step is to run a larger analysis of reuse modification on a bigger dataset of parallel Bibles. The following research questions focus on historical **B**ible corpora (see Sec. 3.1.2 and Ch. 6):

Improving performance of writing variants

First, Ch. 6 shows whether time proximity of Bible editions can help to map historical word variants to modern writing using only a simple character-distance measure. The following research questions are formulated to this end:

- **RQ B1.1** *Does the use of temporally close Bibles improve the alignment of historical writing variants?*
- **RQ B1.2** *Whether and how does time proximity in historical texts (i.e., text that are published within short period) help to normalize old variants of text to modern spelling?*
- **RQ B1.3** *What are specific problems to align a historical Bible corpus?*

To address these questions the method that starts out with the study to address the RQ M block is applied on a selected subset of a Bible corpus. Operations are refined and added. Further, an evaluation of the method is presented and results are directly applied to the next steps of this study.

Empirical analysis of paraphrastic text reuse

Next, the modification is measured in two different ways: i) using a method to measure different modification levels in a prioritized order, ii) by analyzing part of speech (POS) changes between two verses of any two Bibles (within one language). The following questions guide the empirical analysis:

- **RQ B2.1** *How are the different types of modification distributed in paraphrastic text reuse and how does the use of different lexical resources affect theses distributions?*
- **RQ B2.2** *What does the number of POS changes tell when measured in the parallel Bible corpora?*

To address these questions modification of POS and the operations that are proposed are applied and empirically collected for both, the English and the German Bibles corpus. In parallel, two lexical databases are used to derive semantic relationships that then are applied between the words of two text excepts.

1.3.3 Research questions III

Towards a metric for paraphrastic modification

The last research question investigated based on the Bible corpus concerns the ability of the proposed method to measure distance in documents. For this purpose, a subset of

the English Bible corpus is divided into two groups. Literally and "normally"[5] translated Bibles, and a classifier is trained to estimate the importance of operations to distinguish similar and "distant" Bibles when they are aligned and modification is measured. The respective research question formulated is:

- **RQ B3** *How can the proposed method be used to measure distance between two Bibles with regard to both, the translation background and the time distance between them, and which of the operations designed in this thesis are important for this task?.*

The goal is to investigate whether the degree of modification measured based on operations—that are applied in a prioritized order as relations between the words of two sentences—serves as a good feature for paraphrase prediction. Scores such as Meteor (Denkowski & Lavie, 2011) make use of synonymy, but do not model other relationships. The method here, however, also integrates information on hypernymy, hyponymy, and co-hyponymy.

The following questions are investigated. *Compared to existing techniques, how does a human-interpretable method perform in* ***p****redicting semantic equivalency in:*

- **RQ P1** *a modern English parapahrase corpus,*
- **RQ P2** *a parallel Bible corpus,* and
- **RQ P3** *a Medieval Latin reuse dataset?*

All results on predicting paraphrases are compared with the performance of existing metrics borrowed from machine translation (MT) evaluation, such as BLEU (bilingual evaluation understudy) by Papineni *et al.* (2002) and Meteor.

1.3.4 Research hypotheses

The underlying research hypothesis of this thesis is that non-literal reuse does not necessarily have words in common with the original text and, thus, needs linguistic resources to be detected, even if we expect that not all of the reuse can be identified by the resources. Furthermore, we hypothesize, that—especially in historical text reuse—not only synonyms are used to preserve meaning when text was repeated or paraphrased, but also weaker semantic relations such as hypernymy or co-hyponymy are used.

1.3.5 Proposed method

The method proposed studies less literal and non-literal (a.k.a. parpahrastic) text reuse of Bible verses in Ancient Greek, Latin, historical English and historical German texts. The

[5] The difference is clarified in Sec. 7.1

focus is on understanding how reuse is modified and transformed with regard to the original Bible verse. To this end, operations are defined that characterize how words change—e.g., synonymized, capitalized or change the POS. Since the approach uses external linguistic resources, it also shows how such resources can help detecting reuse, where limitations are and how the recall changes when different resources are consulted. This automated analysis, which describes reuse changes using the operations, is complemented with a qualitative manual analysis.

The study comprises the following main steps. First, operations reflecting literal reuse, replacements (inspired by semantic relationships, such as synonyms and hypernyms, supported by Ancient Greek WordNet (AGWN) (Bizzoni *et al.*, 2014)), and morphological changes (e.g., when mapping words still share the same cognate) are identified. The operations are based on a one-word-replacement to better quantify the results. Second, an algorithm is developed that identifies operations by first looking for morphological changes between a word from the reuse/Bible verse and its corresponding candidate from the Bible verse and, in case of no success, by seeking for a semantic relation or recording a fallback operation. Third, the algorithm is applied to different datasets, and the relationships of affected words, and the literal share are investigated. Occurrences of operations are quantified and it is characterized to what extent the linguistic resources are helpful. Fourth, we compare a modern lexical (synset) database and one that is made to retrieve Latin and Greek with respect to their ability to identify semantic relationships among words. Smaller samples are manually analyzed using further operations to understand the full richness of the reuse. Fifth, the method is applied to test how alignment can be improved in a corpus of historical English Bibles. Afterwards, lessons are learned by refining the pre-processing and the operation set to improve the method. Last, the method is tested against other techniques in its capability to predict semantic equivalence. Empirical understanding about the characteristics of historical versus modern text reuse are summarized.

The noisy channel paradigm based on work by Shannon (1948) serves as a model to illustrate the overall approach. Conceptually, Shannon determines the degree of redundancy that an information flow must contain in order to ensure the successful transmission of the information. In this thesis, the model is used to illustrate that the channel itself is considered the place where modification happens. (Figure 4.1 displays this part in the middle rectangle.) The noisy channel hereby contains the minimal program (i.e., the minimal set of operations Kolmogorov (1963)).

In modeling reuse change in the form of modification, the aim is not only to apply operations that represent change, it is also desired to have a minimum operation set that closely follows the length of an input verse/sentence to calculate its output version. This task is inspired by the complexity of Kolmogorov (Kolmogorov, 1963; Li & Vitáni, 2008)—the minimal size of a program that computes a specific output.

1.4 Contribution, scope and limitations

1.4.1 Contribution

The main contribution of this thesis lies in the analysis of lexical modification in historical text reuse and the empirical data that result from this analysis. To achieve the goal, methods are developed to measure modification. These methods are then applied to different sorts of text. The contributions can be summarized by:

i A technique to measure modification in historical text reuse by formulating operations so that each represents a form of modification.

ii The application of the technique to two text data sets where reuse was manually identified, and the application to two bigger parallel Bible corpora of English and German.

iii Empirical data based on the automated approach that is applied to the data sets as well as the manual analysis as a complement, performed on samples of the data.

1.4.2 Scope and limitations

This study focuses on text in the languages English and German with a smaller analysis of two Medieval Greek and Latin datasets. Bibles in English and German are selected as research items, because they cover one strongly and one weakly inflecting language. The Bible and Biblical reuse is chosen due to its availability and representativeness also as a text that existed already centuries ago, and constitutes a good foundation for historical investigation in the context of modification of historical text. However, the techniques developed are also measured on a modern dataset, but did not show to have the same effect. Empirical figures of modification are only collected for the historical datasets. A further limitation is that for all languages analyzed lexical resources are necessary.

The scope of the thesis does not lie in improving or testing text reuse detection methods in historical text directly. However, the empirical insights and results of the work are supposed to eventually support the improvement of reuse detection techniques in historical texts.

1.5 Outline of the thesis

The remainder of this thesis is structured as follows. Chapter 2 gives an overview on related work of the field. Starting with preprocessing methods for historical English and German, it continues to introduce alignment techniques, the basics of word distributions in natural language, and it discusses text similarity based on string similarity. It finally introduces supervised techniques for paraphrase detection. Chapter 3 introduces the data on which this

research is conducted. It comprises mainly two smaller medieval datasets, a bigger parallel corpus of historical English Bible translations, and a corpus of German Bible translations. Chapter 4 proposes the method of this thesis, which has the goal to capture different degrees of modification between two texts. In Ch. 5 a small-scale analysis of modification is presented in two datasets of Medieval Greek and Latin. The procedure is considered a transformation step to understand how a reuse needs to be changed to obtain the primary text version that it was reused from. The degrees of support of linguistic resources for Latin and Greek are also denoted. Chapter 6 conducts a larger analysis on the two Bible corpora separately. It also investigates two lexical databases regarding their recall and support of identifying semantic relations among words in the parallel Bible corpus. In Ch. 7 the proposed method is used to classify dissimilar texts and it is compared against exiting techniques in sentence similarity prediction. Finally, Ch. 8 summarizes the contributions and discussed future work.

Chapter 2

Related work

This chapter gives an overview of research related to the topic of this thesis. It starts with an overview of work on the canonicalization of English and German text in Sec. 2.1. Thereafter, it gives an overview of sequence and statistical alignment in Sec. 2.2, followed by an outline of information retrieval methods for text reuse detection in Sec. 2.3. Further methods of text similarity and their diverse bases are discussed in Sec. 2.4, while Sec. 2.5 focuses on parallel corpora and sentence similarity scores borrowed from machine translation, which offers many evaluation strategies. Finally, in Sec. 2.6, the chapter closes with a summary, and it motivates the main contribution of this thesis, presented in the subsequent chapters.

2.1 Preprocessing historical texts

Canonicalization of text written in historical English and German can be achieved in many different ways, ranging from techniques based on dictionary knowledge via rule-based techniques to unsupervised learning. This section gives an overview of currently existing techniques and tools.

2.1.1 Lemmatizing historical English

Tools

VARD is probably the "goto" software in Early Modern English normalization. Baron & Rayson (2008) present the VARD tool, which combines a known variants lookup as well as replacement rules and phonetic matching to find a list of possible candidates for the normalized writing version of a word. The methods are combined in a confidence score, but candidates with a high Levenshtein distance measure (Levenshtein, 1965) are rejected. The candidates are then presented to the user via a graphical interface.

MorphAdorner (Burns, 2013), written in the programming language Java, performs morphological adornment of each word in a running text. It provides functions to assign normalization (standard spelling), POS, and the lemma to a word. It further provides tokenization,

sentence segmentation, and named entity identification. MorphAdorner was initially built to adorn text from the early Modern English period (late 14th to mid 16th century), however, it also works suitably well for the modern English language text. For lemmatization, MorphAdorner first looks up lemmas from the lexicon. For irregular forms, a mix of a list with associated forms and grammar rules, partially based on Martin Porter's suffix stripper (Porter, 1980) is used. MorphAdorner can be distinguished from VARD, because it is designed for longer datasets and texts. As such, it does not have a graphical user interface, but instead is used from the command line to process whole books at once. In this thesis, MorphAdorner is used to preprocess the historical English texts, since it can handle Archaic English well, is freely available, and is well documented.

Methods

Beyond the more established tools described above, Johnson (2009) investigates how a combined method of using the Levenshtein distance on the sorted vocabulary of a corpus of Old English can be used to lemmatize Old English words. Johnson's work shows that stemming by removing common endings homogenizes words that are related to each other and enables a more precise performance of the Levenshtein algorithm to determine the correct lemma. See Sec. 6.1 for more related work on methods to normalize and lemmatize historical English text.

2.1.2 German canonicalization

Dictionary and rule-based work

Bollmann *et al.* (2011) investigate the normalization of text written in Early New High German using context-aware rewrite rules, which map historical word forms to modern word forms. The rules are derived from an alignment of the original version of the Luther Bible and a version with modern spelling. Applying the normalization rules results in up to 93% of correct matches. Furthermore, using a threefold technique Hauser *et al.* (2007) relate modern-language writing with old word writing variants. They use a dictionary component covering new word forms, a rule-based component (e.g., Stockmann-Hovekamp, 1991), and a word distance that works with edit weights, which is a reimplemented version of the algorithm presented in (Brill & Moore, 2000), where edit operations are based on sequences of symbols instead of single symbols. Finally, candidates are ranked based on word similarity, frequency, and a heuristic finding possible candidates. With their approach, high recall values can be achieved with a precision that is still around 70%.

Supervised learning

Work on the canonicalization of text written in historical German is foremost done by researchers around Brian Jurish and the Deutsches Textarchiv (German Text Archive) (Jurish, 2008, 2010). Jurish (2008) presents work on mapping historical text to one or more canonical text types. To this end, Jurish uses phonetic conflation of word forms, and canonicalization based on lemma heuristics.[1] In another work, Jurish (2010) finds a trade-off between the precise transliteration approaches which are limited in coverage, and the highly recalling—even though comparably imprecise—phonetic conflation techniques: Jurish disambiguates words at the token level using textual context to find the most probable normalized word form for a given variant. A hidden Markov model (HMM) is dynamically computed based on conflation information from word tokens for every sentence. This disambiguation can be understood as the well-known tagging mechanism applied to normalized word forms representing the tags.

Gold standards

Scheible *et al.* (2011) describe a manually annotated gold standard corpus of Early Modern German, which is annotated with POS tags, lemmas, and normalized spelling of words. The corpus can be used as an evaluation test bed for NLP tasks adapted on historical texts.

2.2 Alignment

2.2.1 Statistical and word alignment

Methods used in word alignment often are based on language models and inspired by MT. For example, Vogel *et al.* (1996) use an HMM-based technique to consider the location distance of two words from two sentences of a bilingual corpus. IBM designed a series of alignment models, namely IBM Model 1 to IBM Model 6. These start with lexical translation, and increasingly consider further aspects and techniques, such as: reordering (Model 2), multi-word translation (Model 3), adding POS information of surrounding words to the probability distribution (Model 4), and language models combined with HMMs (Model 6) (Brown *et al.*, 1993; Och & Ney, 2000; Fernández, 2008; Schoenemann, 2010; Vulić, 2010). The Berkeley Word Aligner (Liang *et al.*, 2006; DeNero & Klein, 2007)—also designed for the purpose of MT—combines an HMM-based alignment model and takes the constituent structure (in German "Satzglieder," such as subject or object) of the target language explicitly into account. In this thesis, we use Berkeley Word Aligner to align parallel, monolingual corpora.

[1] Conflation here is the assignment of several words with highly similar sound to one canonical form, so covering many writing variants

2.2.2 Sequence alignment

Sequence alignment can be distinguished from word alignment by two major characteristics: i) it often uses dynamic programming rather than statistical methods, and ii) it is used to align rather long sequences, such as DNA which also come with a limited vocabulary. The algorithm by Needleman & Wunsch (1970) is designed to find the degree of similarity in two sequences. The principle is known from the edit steps of the Levenshtein distance (Levenshtein, 1965). The method uses dynamic programming, denoting the fact that distance scores (and paths) of earlier substrings/prefixes—of the two sequences that need to be aligned—are considered in later steps and dynamically updated. In contrast to the algorithm by Needleman & Wunsch (1970), which is considered to be global[2], the algorithm designed by Smith & Waterman (1981) is considered a local sequence alignment method. It especially finds several regions of very similar subsequences in two long sequences. Instead of adding a score for dissimilarity, it adds a score if matches are found. A trace-back step then finds multiple regions with high scores and returns the overall best alignment.

In the field of sequentially aligning historical text, the work by (Smith *et al.*, 2014) needs to be mentioned. The so called Passim method is three-fold. First, based on shingling, one relevant document pair is identified that contains significant overlap. Second, to increase the precision of the results, local alignment techniques are used to identify those passages that have a high chance to be reused. Last, making use of the links between passage pairs in the document collection (from the former step), clusters are built to remove duplicates. These duplicates appear when one passages is aligned multiple times. Furthermore, this steps helps to find connected (successive) reuse. Vesanto *et al.* (2017) apply BLAST[3] to text reuse detection in highly noisy Finnish newspapers and journals. These are digitized using optical character recognition (OCR) upfront. For Vesanto *et al.* (2017)'s purpose, BLAST vastly outperforms Passim (Smith *et al.*, 2014).

2.3 Information retrieval methods for text reuse detection

This section introduces some principles of natural language that are important to understand when frequency-based approaches are used to discover reuse. The section also gives a first overview on techniques for the discovery of text similarity based on the foundations of information retrieval.

[2]This means that it is especially useful when sequences are principally similar and differ only slightly.

[3]BLAST is a tool suite that combines several local and global alignment procedures and is widely used to analyze biological sequence data, see `https://blast.ncbi.nlm.nih.gov/Blast.cgi`

2.3.1 Zipf's law and frequencies of words

In (Zipf, 1949) George Kingsley Zipf states that the words of a language used to communicate follow the principle of least effort and economic efficiency. He introduces the famous Zipf's Law, which states that the frequency of a word in a corpus of natural language is inverse proportional to its rank (see Newman, 2005). The rank is the number of a word of a corpus when all words were ordered by its frequency decreasingly. A word's probability to appear in a corpus is then given by the simplified expression of Zipf's Law:

$$p_w(r) \sim \frac{1}{r}, \tag{2.1}$$

where r is the rank. The distribution of words of a corpus, thus, follows a power-law probability distribution, which means that the most frequent word ($r = 1$) will occur about twice as frequent as the second most frequent word ($r = 2$), and three times as often as the third word in the rank order, and so on. Among the most frequent words are mainly so-called function or stop words, which do not belong the the class of content words (nouns, verbs, adjectives and adverbs) and cover a high ratio of all word tokens of a running text.

Zipf's Law is important to consider when frequency-based techniques are used to measure similarity, because it has a direct effect on the results of a similarity measuring task, for example, when function words are kept in the processed text compared to when they are left out.

2.3.2 Frequency measures

The most obvious way to find repetition and semantic similarity in text collections is to search for words that are in common (see, e.g., Monostori *et al.*, 2000). However, using simply common (i.e., jointly appearing) words only enables the discovery of verbatim or near-verbatim reuse. So called fuzzy methods also consider tokenization, stemming and function word removal to reduce false positives (see Sec. 2.3.1). Alzahrani & Salim (2010) use these preprocessing steps in a task of extrinsic plagiarism detection[4] of the PAN 2010 challenge. Another way to allow fuzzy matching is to also consider ngram frequencies of characters and words. Potthast *et al.* (2011) use several different information retrieval methods. All of them represent documents as vectors of their word n-gram frequencies, too. Stemming and function word removal as well as term frequency weighting is deployed upfront. Stamatatos (2009) use profiles of character ngrams in the task of intrinsic plagiarism detection[5]. They further use a so-called style change function that was initially used for author identification

[4] I.e., plagiarism in a suspicious document that is compared to a collection of possible candidate documents.

[5] I.e., plagiarism must be found without a reference corpus, for example, by style inconsistency in the document of investigation (Zu Eissen & Stein, 2006).

to find variation in style. Finally, the tf-idf measure (term frequency – inverse document frequency) is a common way to find similarity between documents. It is denoted by the number of times that a word t occurs in a document d : $tf(t, d)$, and the log-scaled fraction of the total number of documents d in a document collection D divided by the number of documents that contain the word t : $idf(t, D) = \log(\frac{|D|}{d \in D : t \in d})$. tf-idf is then compute by:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{2.2}$$

This measure is not only used in reuse or plagiarism detection systems, instead, its principle is used for many retrieval algorithms of search engines in general (see, e.g., Baeza-Yates *et al.*, 1999).

2.4 Text similarity

This section first gives an overview of string-based similarity methods before it introduces work in text reuse detection of modern and historical language text.

2.4.1 Stringology

String search

Crochemore & Rytter (2003) describe *stringology* as a term that unifies the field of string and text algorithms. In (Crochemore & Rytter, 2003) they give deep insights into all sorts of string-related techniques, such as search and sorting algorithms, compression algorithms, and pattern matching. However, in the context of this thesis, it is sufficient to introduce the following important search algorithm, namely that of Boyer & Moore (1977), who present an efficient way to search a shorter string (called the *pattern*) within a longer string. Compared to naive algorithms that try to first find a match of the first character of the pattern in the searchable text, this technique already initially jumps to the index that determines the end of the pattern in the searchable text, and proceeds only when a match is found. Their technique, a.k.a., Boyer-Moore string-search algorithm usually serves as a benchmark in the literature on string search.

Regular expressions

Regular expressions are a more abstract way to search regular patterns in searchable text. Thompson (1968) presents a method to locate character strings in text. He implements a compiler that accepts a regular expression as an input and returns a program with a searchable text that creates a signal when a regular expression is matched in the text.

Prefix, infix and suffix trees

A suffix tree is a fast way to lookup substrings after an index is created. Gusfield (1997) writes that suffix trees are constructed and stored in linear time and space according to the length of the string. Further operations that can be quickly and easily performed are regular expressions and longest common substring lookups. Jongejan & Dalianis (2009) use plain trees and directed acyclic graphs to store grammar rules that represent word formation by adding prefixes, infixes, and suffixes to an infinite word form. The lemma version of a word is then found by following these stored rules.

2.4.2 Detection of text reuse and plagiarism in modern language text

Recognizing modified reuse is difficult in general. Alzahrani *et al.* (2012) study plagiarism detection techniques based on ngram-, syntax-, and semantics. They find out that 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. Likewise, citation-analysis techniques (Gipp & Beel, 2010; Gipp & Meuschke, 2011) can often not be applied to historical texts due to the absence of references.

Lexicon-based approaches

Fernando & Stevenson (2008) present an algorithm that identifies paraphrases making use of word similarity information of English words, with information derived from WordNet Fellbaum (1998). While Lin (1998) define the semantic similarity of two words in a lexical database based on the fraction of the probability of their lowest common subsumer and the probability of the words themselves:

$$sim(w_1, w_2) = \frac{2 \log p(lowest\ common\ subsumer_{w_1,w_2})}{\log p(w_1) + \log p(w_2)} \tag{2.3}$$

Finally, synset databases support identifying word relationships based on their semantics. Jing (1998) investigates issues that come with using WordNet (Miller *et al.*, 1990) for language generation. Among others, these comprise issues arising from the adaptation of a general lexicon to a specific domain.

Machine and deep learning-based approaches

There is a vast number of machine learning-based techniques for text classification. In the following, some relevant techniques are exemplified. Rigutini *et al.* (2005) propose an algorithm based on expectation maximization. They train a classifier by means of a predefined set of text categories and a collection of labeled training data for a given language. A classifier for a different language is trained by translating the available labeled training set and tested on an additional set of unlabeled documents from the other language. The experiments are conducted in English and Italian. Osman *et al.* (2012) present a plagiarism detection technique based on the Semantic Role Labeling (SRL). They analyze text by identifying the semantic allocation/space of each term in a sentence and, consequently, find semantic arguments for each sentence. They also assign weights to the arguments and find that not all of them affect the detection of plagiarism. The work is conducted on the PAN-PC-09 datasets, which contain texts in English, German, and Spanish. Brlek *et al.* (2016) use word2vec to find and align semantic similar sentences and passages in a plagiarism detection task. They aggregate plagiarism cases from a seeding step to larger units using Duhaime (2015)'s sentence similarity measure. This work is conducted based on the PAN 2013 corpus, the PAN 2014 corpus and a corpus of web pages. Song & Roth (2015) study how the use of word2vec (Mikolov *et al.*, 2013) as a means of vector densification can help to improve the accuracy in detecting similarity in short English language texts. Zhang *et al.* (2017) present a framework that encodes sentences in the form of vectors. Sentences having semantic information in common are encoded as similar vector representations using an encoder-decoder model trained on a corpus of paraphrase sentence pairs. The technique is applied to the tasks of sentence paraphrasing and paragraph summarization. The results provide first insights into the usefulness of vector representations of sentences in advanced language embedding tasks.

2.4.3 Text reuse detection in historical text

The research field of automated detection of historical text reuse is still in its early stages. Up until now, Büchler (2013) combines information retrieval and language processing techniques to address a wide range of reuse detection scenarios for historical texts, covering near copies and also text excerpts with a minimum overlap. Specifically, he uses a fingerprinting approach by selecting certain ngrams from an upfront presegmentized corpus. Furthermore, focusing on high recall, the detection of Homeric quotations in Athenaeus' *Deipnosophistai* is investigated by Büchler *et al.* (2012), searching for distinctive words within reuse. Efforts to automatically process ancient texts are also made around the Perseus Digital Library project (Crane, 1985). For example, Bamman & Crane (2008) present the discovery of textual allusions in a collection of Classical poetry, using measures such as token similarity,

ngrams or syntactic similarity. This allows finding at least the most similar candidates within a closed library. In the Biblical context, Lee (2007) investigates reuse among the Gospels of the New Testament, aimed at aligning similar sentences. In this work, similar as in query retrieval, so-called alternation patterns are developed using the cosine similarity measure, a source verse proximity measure, and the source verse order. The research field of paraphrastic reuse detection in historical text is much more sparse. Bamman & Crane (2011b) process the semantic space of a word to be able to disambiguate word senses in historical text. They utilizing a bilingual sense inventory and achieve up to 72% of the word senses to be classified correctly. An example from the field of modification analysis is performed by Ketzan & Schöch (2017). They analyze modification, such as removals, insertions, substitutions, and minor token modifications in the re-edition of *The Martian*. They utilize computational methods, such as the *diff* algorithm (Hunt & MacIlroy, 1976).

In contrast, the present study, presented in the remainder of this thesis, focuses on the use of synset databases and POS information to model the transformation process of reuse, and to find limitations when applied to non-literal/paraphrastic reuse. The provided conceptual linkage supported by lexical databases and the abstraction level which comes with the POS information helps to identify the reuse transformation process on the Ancient Greek and Latin dataset used in the remainder.

2.5 Paraphrases and parallel text

2.5.1 Paraphrase identification in machine translation

The task of paraphrase identification is often used in the field of MT. The purpose of an MT system is to predict a semantically equivalent version of an input sentence. The purpose of an MT metric, however, is to determine how well the equivalency is achieved. Since this is even more difficult to apply cross-lingually, often, metrics are applied to MT output sentences and a human-generated reference translation (c.f., Finch *et al.*, 2005). The assessment of an MT metric with regard to its usefulness to semantical equivalency determination suggests the implication that MT metrics of the newer generations might be useful to measure paraphrasticality. In the following, some of the most common MT metrics are introduced.

Typically, metrics based on simple edit distance measures are used to evaluate MT systems, such as the Word Error Rate (WER) and the Position-independent Error Rate (PER), inititally defined by Tillmann *et al.* (1997). PER is similar to WER, but instead handles sentences as a bag of words. As such, only the words that occur in both sentences of interest are considered, all other overlapping words are counted as substitutions. BLEU (Papineni *et al.*, 2002) is probably the most famous MT evaluation metric, developed at the IBM

Watson Research Center. During the development of BLEU, IBM aimed at high correlation with human evaluation, language independence, and cheap processing costs. A simplified definition is:

$$BLEU = \exp(\sum_{n=1}^{N} \frac{1}{N} \log \frac{\sum_{i=1}^{I} \sum_{ngram \in s_i} Count_{sys \cap ref}(ngram)}{\sum_{i=1}^{I} \sum_{ngram \in s_i} Count_{sys}(ngram)}), \tag{2.4}$$

where N is the maximum ngram size, $Count_{sys \cap ref}(ngram)$ is the number of ngrams found in both sentences and $Count_{sys}(ngram)$ being the number of ngram found in the system output sentence. I is the length of the corpus in sentences.

Lavie & Denkowski (2009) present an MT evaluation metric called Meteor. Compared to IBM's BLEU, which only considers precision-based features, Meteor additionally incorporates measures for recall and supports a more flexible word matching by allowing morphological variation, and enabling synonym matching. Ch. 7.2 uses the metrics PER Tillmann *et al.* (1997), TER Snover *et al.* (2006), BLEU Doddington (2002), and Meteor Lavie & Denkowski (2009) to compare their performance in a task of paraphrase similarity.

Some work using translation metrics to measure equivalence in meaning is undertaken already in 2005. Finch *et al.* (2005) study the utility of the machine translation metrics BLEU, NIST, WER, and PER as features for classifiers that predict semantic equivalency. They also investigate the usefulness of POS information and of the Jiang-Conrath WordNet-based lexical relatedness measure (Jiang & Conrath, 1997) as part of their edit distance measure.

Madnani *et al.* (2012) present a more recent study on the usefulness of automated MT evaluation metrics for the task of paraphrase identification. In their experiments the authors train a meta classifier on three constituent classifiers—a logistic regression, a support vector machine, and an extension of nearest neighbor—using recent MT metrics as features. After testing their methodology on paraphrase benchmark corpora against known paraphrastic sentences, and a corpus created for the task of plagiarism detection, they find that they outperform existing methods in the former corpus, and obtain positive results for the latter corpus.

2.5.2 Gold standards and benchmark 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 & Dorr, 2010). Especially in the field of modern reuse investigation, aligned corpora can provide a rich source of paraphrastic sentence pairs in one, sometimes multiple languages. One of such is the Microsoft Research Paraphrase Corpus (MSRP), which contains 5801 manually evaluated, paraphrastic sentence pairs in English (Dolan & 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 Ancient Greek and Latin quotations based on the the Vetus Latina series and the Novum Testamentum Graecum Editio Critica Maior (Houghton, 2013a,b). It contains more than 150,000 Latin citations and about 87,000 Ancient Greek Bible references.

2.6 Summary and motivation

In summary, this chapter showed: i) that substantial research effort is put into plagiarism detection in modern languages, ii) that machine and deep learning techniques are used to find semantic equivalences in big collections of modern text, iii) but that comparable state-of-the-art techniques are difficult to apply in historical text—due to a richer variation in the vocabulary, the lack of resources or the shift of meaning (Hellrich *et al.*, 2018)—and even harder when the text is several hundreds of years old. Text collections of historical text of appropriate size are not necessarily eligible for the creation of stable language models, because these text collections are very heterogeneous in their genre and time of creation. However, preprocessing efforts are ongoing for historical languages, such as for Early New High German and Early Modern English. To drive research in historical text reuse detection, this thesis relies on valuable work in the field of text preprocessing. Notably, to meet the interests of practitioners and experts in the humanities, any technique used or built must be interpretable and clearly explain the processing steps. As such, before *collecting noisy, heterogeneous material, training a model based on context vectors and obtaining a highly unsatisfying accuracy in sentence similarity degree prediction, with results that are difficult to follow*, it is important to first understand how text changes, including what changes, how strongly words change, and how frequent these changes are. To this end, this thesis analyzes these questions in paraphrastic parallel texts and in reuse collections.

Chapter 3

Text data used

The automated processing of historical data is especially challenging due to reasons outlined in Ch. 1. To conduct this study on a representative sample of data available, we choose texts from different centuries and different languages. This chapter provides an introduction of different corpora containing examples of historical text reuse, and touches on the experiments that are executed based on those data. One smaller evaluation corpus of modern language English is also presented towards the end of this chapter. It is important for reasons of comparison of the proposed method against existing techniques.

3.1 Historical parallel corpora used

A parallel corpus is a form of bi-text that usually consist of verse or sentence-aligned text in two or more different languages. Yet, parallel corpora also exist mono-lingually. Then, the text usually is paraphrased to each other, such as different versions of one book, say, an original version and its simplified version. Throughout this thesis, we use three different parallel corpora of historical text reuse, i) two small Medieval Greek and Latin reuse datasets to test our main objective on first, ii) a parallel corpus of English Bibles, and iii) a parallel corpus of German Bibles. We choose to use parallel Bible corpora because they offer a sufficient amount of parallel text that covers many topics and offers a vast vocabulary. Using a diverse set of languages, we can better show the reliability of this studies' validity.

3.1.1 Medieval Greek and Latin reuse test dataset

To conduct a first test of the research hypothesis on how reuse is modified (see Sec. 1.3), we use two data sets form the medieval times. These datasets are especially well-suited because they are manually extracted by a team of biblical scholars and contain rather literal reuse, very allusive reuse, and several degrees of paraphrasticality in between. Following, this reuse data is shown in greater detail.

Clement of Alexandria

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

The first dataset comes from the primary source text of "Salvation for the Rich" from the Medieval Greek writer Clement of Alexandria (Clément d'Alexandrie, 2011), a well-known author in Biblical literature (Cosaert, 2008). The work contains a total of about 9600 words (punctuation excluded). It is unstructured and simply consists of verses, each of which comprising between one line (9 tokens) to a maximum of nine lines (95 tokens). Note that verses cross-cut sentences. The Biblindex team annotated 128 text passages as Bible reuse, adding a footnote with Bible verse pointers to each of them. Sometimes one reuse instance points to different Bible verses or one text passage contains more than one reuse instance. Thus, we come up with 199 verse-reuse pairs. The excerpts point to a total of 15 Bible books. The circles in Fig. 3.1 show these books (x-axis) and the number of pointers to each of them (y-axis), with Matthew (Mt) being the most frequently referenced one. Reuse instances in Clement's work are around 12 tokens, which is shorter than an average Bible verse (27 tokens). See Tab. 3.1 for type and token information on Clement's reuse.

Bible	**#tokens**	**#types**	**type-token ratio**	**tokens per verse**
Clement of Alexandria	3,721	826	4.5	19
Septuagint	4,779	1,230	4.0	24
Bernard of Clairvaux	9,588	2,705	3.5	9
Biblia Sacra Juxta Vulgatam	18,360	3,362	5.5	16

Table 3.1: Type and token figures of Clement's and Bernard's reuse, and the respective Bible verses (punctuation ignored)

Bernard of Clairvaux

The second dataset are extracts from a total of twelve works and two work collections from the Medieval Latin writer Bernard of Clairvaux who lived in the 12th century and also reused text from the Bible. Again manually extracted by the Biblindex team (Mellerin, 2016), the text excerpts forming the reuse are stored in alphabetical order summing up to over 1,100. Each of them again relates to a Bible verse. Typically, the reuse is about half as long as the verse. For the first experiment (see Ch. 5), we follow the same selection criteria as for the reuse of Clement and—starting top-down—we obtain 162 Bible-verse-/reuse pairs, which is similar to the number of Clement's reuse. Specifically, since Bernard's reuse comes from several different primary source works, it points to a total of 31 Bible books. The crosses in

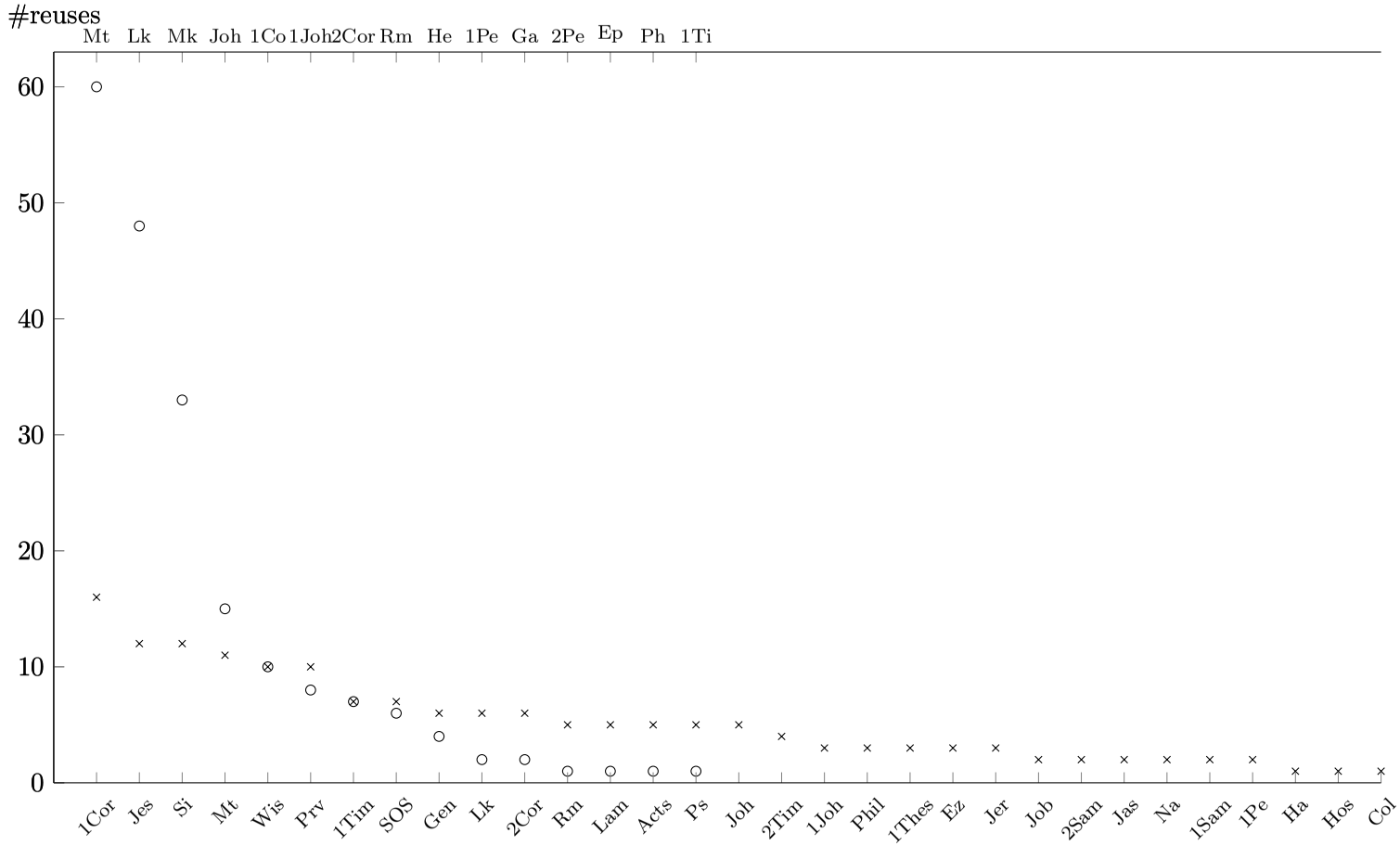


Figure 3.1: Bible verse reuse frequencies in the Latin work. X-axis: books of the Bible; Y-axis: total number a verse from each book was reused (top, circles: Clement, bottom, crosses: Bernard)

Fig. 3.1 show that Bernard's reuse is much stronger distributed over the books of the Bible.[1]

In another experiment, the whole reuse dataset of Bernard is used which sums up to exactly 1,127 reuse pairs. See Tab. 3.1 for type and token information on Bernard's reuse.

Latin and Greek Bibles

The Bible editions used to obtain the verses are the *Septuagint* (Rahlfs, 1935) (The Greek Old Testament), the Greek New Testament (Aland & Aland, 1966), and the *Biblia Sacra Juxta Vulgatam Versionem* (Weber R., 1969, 1994, 2007) (the Latin Bible). Again, circles and crosses in Fig. 3.1 show the distribution of Cement's and Bernard's reuse, respectively. (See Tab. 3.1 for type and token information on the Bible verses.)

Non-literal reuse

The literal reuse in both datasets consists of text excerpts that mainly contain words without inflection following the same order as the words from the Bible verses. Often reuse skips leading or following words from a Bible verse. Less literal reuse has important words in common with the Bible verse. Non-literal reuse has no content words in common with the original. For example, Clement's reuse is highly diverse. It ranges from introducing the overall topic of the relating Bible excerpt by citing multiple Bible verses, to simply supporting his argumentation by alluding to some key terms. 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, where 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. Examples are shown in Fig. 3.2 and 3.3.

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*)

Figure 3.2: Examples of literal reuse in the medieval datasets

Figure 3.4 shows reuse examples—starting by a Biblical verse followed beneath by its literal/ less literal/ more literal reuse. This illustrates the wide range of literalness in the data. It comprises literal (all tokens overlap), less literal (important tokens overlap), and non-literal (no content word tokens overlap) reuse.

[1] Bernard's works—from which the texts are extracted—were published between 1957 and 2010 in the Sources Chrétiennes edition.

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)

Figure 3.3: Examples of less literal reuse in the medieval datasets

Ancient Greek and Latin texts in the experiments

The two small reuse datasets of Medieval Greek and Latin are used in Ch. 5 where a first implementation of the proposed technique is tested. In the same chapter we also use the whole Latin dataset of Bernard of Clairvaux to compare the support of two lexical dictionaries. In Ch. 7.2, the big reuse dataset of Bernard is used again.

3.1.2 Historical English Bibles corpus

A major part of this thesis uses a parallel corpus of historical Bibles to conduct the research on. We choose the Bible because—as a historical source—it pictures a broad diversity of editions, a long history of transmission, and a rich vocabulary together with many domains. The publication history of Bibles in English covers a long time span in a reasonable density—especially in the 16th and 19th century. This work is performed on a total of twelve English language Bibles that were first published between the years 1500 and 1900, in several different use cases. We focus on Bibles that are at least 100 years old, because the goal is to investigate the phenomenon of change among historical text and its reuse. At the same time, we tried to select Bibles that are—even though written in English primarily—very diverse from each other in their evolutionary history and, hence, in their spelling and vocabulary. Following, an overview on the translation background of these Bibles is given so that the reader grasps an understanding on how their translation diversity.

Bible translations are downloaded from three different resources: i) Parallel Text Project (ptp, Mayer & Cysouw (2014)), ii) Mysword (mys, MySword (2011–2018)), and iii) Bible Study Tools (bst, Bible-Study-Tools (2018)) (see Tab. 3.2 for details). In this section we give an overview on the Bibles, their dating, and the preprocessing performed to suit the research purpose.

Bibles before 1600

Matthew Bible (MATT), 1537, mys: The MATT version was first published in 1537 by John Rogers using the pseudonym “Thomas Matthew”. MATT contains The New Testament, which was first published in 1526—and revised in 1534 and 1535—and more than

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 3.4: Examples of reuse in the medieval datasets

Bible	published	trans.	#tokens	#types	ttr[1]	#verses	tpv[2]
Matthew Bible (MATT)	1537	Anglican	781,894	24,362	32	31,102	25
Great Bible (GREAT)	1539	Anglican	793,722	22,439	35	31,102	26
Geneva Bible (GEN)	1560	Anglican	783,230	15,912	49	31,102	25
Douay-Rheims Catholic B. (RHE)	1582–1609	Catholic	898,143	18,414	49	35,891	25
Douay-Rheims Challoner R. (DRC)	1749–1752	standard	780,602	14,705	53	31,102	25
King James Version (KJV)	1769	standard	936,412	15,606	60	36,986	25
The Webster Bible (WBT)	1833	standard	785,493	14,045	56	30,999	25
English Septuagint (LXXE)	1851	literal	615,987	13,727	45	23,145	27
Young's Literal Translation (YLT)	1862	literal	784,192	14,469	54	30,999	25
Smith's Literal Translation (SLT)	1876	literal	777,955	13,500	58	31,102	25
Darby Bible (DBY)	1867–1890	standard	777,724	15,464	50	31,103	25
English Revised Version (ERV)	1881–1894	standard	792,389	13,801	57	31,102	25

[1] type-token ratio
[2] tokens per verse

Table 3.2: Overview of English Bible translations used

half of the Old Testament translated by William Tyndale. Tyndale translated directly from Hebrew and Greek sometimes consulting with the Vulgate and Erasmus' Latin version. He also used Luther's Bible (Tyndale, 1989). The Old Testament and the Apocrypha were later completed by Myles Coverdale using German and Latin texts (Tyndale, 1989). The Prayer of Manasseh, printed in 1535, was translated by John Rogers using a French translation.

Great Bible (GREAT), 1539, mys: The GREAT was published in 1539 and the first English Bible version that was authorized by King Henry VIII of England. Myles Coverdale compiled the GREAT, which includes many of the texts by Tyndale in which Coverdale made some revisions. Coverdale further translated the remaining books of the Old Testament and the Apocrypha from German and Latin translations (most probably Luther's translation and the Vulgate). Coverdale did not translate directly from the Ancient Greek, Hebrew and Aramaic primary texts (Pollard, 2003).

Geneva Bible (GEN), 1560, mys: During the governance of the Catholic Queen Mary I of England (1553-–1558), some protestants fled to Generva, Switzerland where John Calvin was a leading theological scholar. William Whittingham who was one of the scholars among Calvin became supervisor of the translation of the GEN together with Anthony Gilby, among others. Whittingham was responsible for the translation of the New Testament—which was published in 1557—while Gilby was responsible for the Old Testament. A first full edition of GEN was released in 1560 (Herbert & Darlow, 1968; Metzger, 1960).

Douay-Rheims Catholic Bible (RHE), 1582–1609, bst: The RHE Bible is a translation from the Latin Vulgate by the English College of Douai initiated by the Catholic

Church (Pope, 1910). The New Testament was published in Reims (France) in 1582, and preceded the publication of the Old Testament (1609, 1610) by the University of Douai. The RHE text makes strong use of Latin vocabulary, which makes reading difficult. Richard Challoner revised the RHE, which resulted in the Douay-Rheims Challoner Revision (DRC) (Newman, 1859).

Bibles from 1600–1800

Douay-Rheims Challoner Revision (DRC), 1749–1752, mys: Richard Challoner revised the Douay-Rheims Bible (RHE), which made strong use of the Latin vocabulary, because it was translated from the Vulgate, and made reading difficult. The New Testament was published in 1749, 1750, and 1752; the Old Testament in 1750. Challoner's revision is based on the text of the King James Bible (Newman, 1859) and is meant to be rigorously checked and extensively improved for readability.

King James Version (KJV), 1769, ptp: The story of the KJV begun in 1604 and its first edition was completed in 1611 (Dedicatorie, 1611). It was printed by Robert Baker and was the third translation that was approved by the authorities of the English Church. In the course of the 18th century, the KJV replaced the Vulgate as the default version for English scholars, and became—with the raise of type printing—the most frequently printed book in history. These prints are based on the edition of 1769, which was extensively reedited by Benjamin Blayney at Oxford (Daniell, 2003). We use the text of this last edition from 1769.

Bibles from 1800–1900

The Webster Bible (WBT), 1833, bst: The WBT by Noah Webster is a revision of the KJV. Webster mainly replaced words that became unusual or changed their meaning in the course of the centuries with better fitting contemporary words, eliminated archaic words and corrected and simplified Grammar. He also focused on socially acceptable language by eliminating words that are vulgar or offensive (Webster, 1833).

Darby Bible (DBY), 1867–1890, ptp: Darby wanted to create a highly literal version of the New Testament for study purposes. He used modern critical editions of the Greek primary text and augmented his text with critical and philological annotations. Darby also consulted with the translators of the New Testament of the English Revised Version, which was published in 1881 (Bruce, 1978). His New Testament was first published in 1867. Darby's translation of the Old Testament was published—after his death—by his students in 1890 and is based on Darby's German and French translations of the Old Testament. In 1890 Darby's Bible was published under the name "The Holy Scriptures. A New Translation from the Original Languages by J. N. Darby" by G. Morrish (Marlowe, 1867–1884). We primarily use Darby's Old Testament in the experiments.

English Revised Version (ERV), 1881–1894, mys: ERV is the most recent English Bible translation in our study. It is today's only officially authorized revised version of the King James Bible in Britain. Over fifty scholars were assigned to created this version. American researchers were invited to collaborate as well. The New Testament of the ERV was published in 1881, the Old Testament in 1885, the Apocrypha in 1894 (of Revised Version, 1989).

Bibles from 1800–1900 (literal translations)

English Septuagint (LXXE), 1851, mys: The LXXE is an English translation by Sir Lancelot Charles Lee Brenton. It is translated from the Codex Vaticanus version of the Greek Old Testament, which itself is a translation of the Hebrew Old Testament (Roger, 1958, 1959).

Young's Literal Translation (YLT), 1862, bst: Robert Young, the translator of YLT, created a highly literal translation of the original Hebrew and Greek texts. Young tried to be as consistent as possible in representing Greek tenses with English tenses. Among others, he used present tense where other translations used past tense (Young, 1898a,b). We see an example in the book Genesis:

"*In the beginning of God's preparing the heavens and the earth—*" (Genesis 1:1).

Smith's Literal Translation (SLT), 1876, mys: Upon its publication, Julia E. Smith Parker's Bible translation counted as the only English translation that was not only directly translated from the historical source texts (Hebrew and Ancient Greek), it also was the one that was written in a contemporary English. Smith aimed at complete literalness—what made her translation even seem flow-less—and consistently translated each original word with the exact same English word. For example, she consistently translated the Hebrew imperfect to English future tense (Malone, 2010).

English Bibles in the experiments

The length of a Bible mostly is about 31,100 verses. But some Bibles contain books that are not contained in the canon, and these extra books are not persistent either. Exceptions are, for example, the RHE (ca. 36,000 verses), because it is a Catholic Bible and therefore, it also contains more Biblical books than a Protestant Bible. The KJV (ca. 37,000 verses) is longer, because back then the Biblical apocrypha were also read in the daily Old Testament liturgical lectionaries/readings. However, these books are not contained in the Masoretic text, hence, are not contained in many of the other Bibles either. Further, Biblical books that are not contained consistently throughout all the Bible editions—the Biblical Apocrypha and the Catholic Epistles for example—are ordered differently in a Protestant and a Catholic Bible,

hence, we can not always match these extra Books. Further, due to an error in the websites of biblestudytools' html-tree, we missed to download one book of the Bible that is about 100 verses in length out of about 31,000 verses. For minor differences in the Bibles' lengths, we refer to the sources that do not necessarily provide Bibles of homogeneous lengths, even though we checked, and can exclude major alignment inconsistencies. The reader may further be appointed to the fact that LXXE only contains the Old Testament, and for this fact it is shorter as well. In the specific experiments however, we only consider verses that all Bibles under investigation have in common. Table 3.2 also shows all Bibles next to their number of verses.

We use a subset of eight Bibles of the English Bible corpus in Sec. 6.1 for an *alignment* experiment, because here the aim is to investigate spelling changes over the centuries, and we ignore Bibles that build too closely on top each other. The specific Bibles are listed in Tab. 3.3. We use all English Bibles—except of the English Septuagint, because it only contains the Old Testament—in Sec. 6.2 where we calculate *empirical figures* of modifications that our procedural method collects. We use a different subset of six Bibles of the English Bible corpus in Sec. 7.1 for an experiment on how well the method confirms with scholarly knowledge on *morphological/lexical distance* of Bible editions. We again use eight Bibles in Ch. 7.2 to test the approach on its capability to predict *semantic equivalency*.

Bible	**published**	**trans.**	**6.1**[1]	**6.2**[2]	**7.1**[3]	**7.2**[4]
Matthew Bible (MATT)	1537	Anglican	x	x		
Great Bible (GREAT)	1539	Anglican	x	x		
Geneva Bible (GEN)	1560	Anglican	x	x		
Douay-Rheims Catholic Bible (RHE)	1582–1609	Catholic	x	x		
Douay-Rheims Challoner Rev. (DRC)	1749–1752	standard	x	x		x
King James Version (KJV)	1769	standard	x	x		x
The Webster Bible (WBT)	1833	standard	x	x	x	x
English Septuagint (LXXE)	1851	literal			x	x
Young's Literal Translation (YLT)	1862	literal		x	x	x
Smith's Literal Translation (SLT)	1876	literal		x	x	x
Darby Bible (DBY)	1867–1890	standard		x	x	x
English Revised Version (ERV)	1881–1894	standard	x	x	x	x

[1] alignment
[2] empirical figures
[3] lexical distance
[4] semantic equivalency

Table 3.3: English Bibles used in the experiments

3.1.3 German Bibles corpus

To expand the validity of the experimental results to another language, we also use a parallel corpus of German Bible translations, again focusing on historical text. However, exceptions are allowed, because the coverage of historical German Bibles available is not as dense as the coverage of English Bibles. To this end, we also use revisions of the old version of the Luther Bible and the Elberfelder Bible as well as three more. We downloaded the Bibles from the Parallel Text Project website (Mayer & Cysouw, 2014). We choose Luther's Bible in two versions, further the Elbersfelder Bible in its versions from 1871 and 1905, the Textbible—one Bible that was published around 1900—in its version from 1905 and two more recent Bibles. One of which is Gruenewalder Bible, and one is the New Evangelical Version in German. The latter two Bibles especially offer a different style as they follow more modern formulations and wording. An important property that also comes with German texts is the strong inflectional variance. In the following, we will introduce the Bibles in greater detail.

Bibles from 1500 - 1912

Luther Bible (LB1), 1545, Protestant: The Luther Bible is a translation from Hebrew and Ancient Greek by Martin Luther. The New Testament was first published in 1522 and the Old Testament including the Apocrypha, in 1534. Schaff (1858–1890) writes that Luther translated the New Testament from Koine Greek with the intention to make it accessible to the German people. He translated from the Greek New Testament that was written by Erasmus in 1519 and was known as the Textus Receptus. Luther did not primarily use the Latin translation, the Vulgate, which was the translation officially used by the Catholic Church. However, sometimes he oriented himself based on the Vulgate and conformed with it rather than with Erasmus' text. The Old Testament was translated by Luther from Hebrew. Among others, he owned a version of the Tanakh—the Hebrew Bible (Mackert, 2014). We use the final version from that period, i.e., the version from 1545.

Elberfelder Bible (ELB1), 1871, Catholic: The New Testament of the Elberfelder Bible was first published in 1855, the Old Testament in 1871. Its translation was initiated by Carl Brockhaus and John Nelson Darby. Against common use, the New Testament of the Elberfelder Bible is not based on the Textus Receptus, and instead makes use of new insights of the Bible textual criticism that arose in the 19th century. Hence, the translators used the new codices such as the Codex Sinaiticus (Lake, 1911) and the Codex Vaticanus (Vercellonis & Cozza, 1868) directly as they emerged. The Old Testament—with exceptions—is based on the Masoretic Text, which is the authorized Hebrew and Aramaic text (which is the Tanakh, the Hebrew Bible) for Rabbinic Judaism (e.g. of Elberfelder Bibel, 1985).

Elberfelder Bible (ELB2), 1905, Catholic: The revision from 1905 of the Elberfelder Bible was the first edition in Latin script, also known as "Perlbibel". It was published by R.

Brockhaus in Elberfeld, Wuppertal. Its full title reads "Die Heilige Schrift. Aus dem Urtext übersetzt" (e.g., Darby *et al.*, 1905).

Textbibel (TB), 1906, Protestant: Die Textbibel is a full Bible version that was published several times between 1899 and 1911 by the publisher J. C. B. Mohr. It is a collaboration of several German Protestant theologians, and was initially published by Emil Kautzsch in Hall in 1894. The Textbibel guarantees to follow the insights of textual criticism. We use a revision from 1906 (Kautzsch, 1894).

Luther Bible (LB2), 1912, Protestant: In 1858, the Bible society proposed to renew the Luther Bible mainly according to orthography and translation errors (Otte, 2014). In 1883, a test version—the product of a Halle-Stuttgart collaboration being a mix of the Cansteinsche Bible society together with core passages of the Württembergischen Bible society—was published in Halle (Otte, 2014). The final text was set in 1890. We use a revision of that Bible from 1912.

Bibles after 1912

Grünewalder Bible (GB), 1924–1934, Catholic: The Grünewalder Bible—also known as Riessler-Storr-Bibel or Mainzer Bibel—was translated by Paul Rießler (Old Testament) and Rupert Storr (New Testament) and published by Matthias Grünewald. It is a Catholic translation that uses the Hebrew and Aramaic (Old Testament) and Ancient Greek (New Testament) primary texts. The Vulgata (the Latin Bible translation) is neglected. The translations of the Wisdom of Salomon, the Psalms and the Prophets are written in metrical scheme (Rießler & Storr, 1934).

New Evangelistic Translation (NeÜ), 2010, Protestant: The New Evangelistic Translation was created by Karl-Heinz Vanheiden. Its New Testament was published in 2003, the Old Testament was finished in 2009. In 2010 the full version of the NeÜ was published by the Christliche Verlagsgesellschaft Dillenburg (Christian publishing company Dillenburg). The translator considered German and English language translations and commentaries as well as the Hebrew and Aramaic, and Ancient Greek primary texts. Vanheiden tried to keep the text clear and structured to also target people from outside the Biblical field. Linguistic clarity is prioritized over literal reproduction, and words are not consistently translated, instead they are fit into the semantic context and the German language feeling. The poetic texts follow a rhythmical speech. We especially choose this Bible because—even though close to the primary text—it adapts to contemporary language offering a broad diversity of parpahrasticality.[2]

[2]Excerpts translated from Vanheiden (2018)

German Bibles in the experiments

The lengths of the German Bibles is mostly around 31,000 verses (see Tab. 3.4). One exception is the Luther Bible from 1545 (LB1) which—as a Protestant Bible—also contains the Biblical Apocrypha. Yet, the revised Luther Bibles do not contain these Apocrypha normally. The Gruenewalder Bible also contains the Biblical Apocrypha.

Bible	published	#tokens	#types	ttr[1]	#verses	tpv[2]
Luther Bible (LB1)	1545	838,460	29,769	28	35,769	23
Elberfelder Bible (ELB1)	1855-1871	721,134	26,519	27	31,102	23
Elberfelder Bible (ELB2)	1905	721,754	26,410	27	31,102	23
Textbibel (TB)	1906	709,626	33,045	21	31,174	23
Luther Bible (LB2)	1912	696,970	22,572	31	31,102	22
Gruenewalder Bible (GB)	1924-1934	773,323	38,569	20	35,570	22
New Evangelistic Translation (NeÜ)	2010	678,604	30,060	23	30,955	22

[1] type-token ratio
[2] tokens per verse

Table 3.4: Overview of German Bible translations used

The whole German Bible corpus is used in Sec. 6.2 empirical figures of modifications from the proposed method are collected. In Sec. 7.1 we use the German Bibles to show how well one specific experiment adapts to other languages.

3.2 Modern corpus used

This thesis also considers a modern English reuse corpus. It is used to test the proposed method—which is based on operations that model modification between texts—on different texts, and compares them to existing techniques for predicting semantic equivalency.

As a Gold standard for paraphrase prediction, we use Madnani's paraphrase Gold corpus. It is a mono-lingual corpus of semantically equivalent sentences that origins from the PAN 2010 plagiarism detection challenge. Starting from text that was aligned on paraphrase level, Madnani *et al.* (2012) generated a set of aligned sentences by associating corresponding sentence pairs using a heuristic. Negative pairings are created by sampling sentences with an overlap of four words. The training set contains 10,000 sentence pairs, the test set contains 3,000 sentence pairs. Fifty percent of each are labeled as positive results, fifty as negative.

Madnani's Gold corpus of paraphrastic reuse is used in Ch. 7.2. In that chapter, the method proposed in this thesis is evaluated against other methods regarding its performance in paraphrastic reuse prediction.

Chapter 4

Proposed method

This chapter introduces the overall method to model modification in historical, paraphrastic text reuse using the example of monolingual, historical Bible translations. The first part of this chapter introduces a general model inspired by the noisy channel model. A detailed introduction of fine-grained operations follows, and the resources and tools that enable the use of these operations are presented. The method described here represents a central processing step of the text data in all experiments described in Ch. 5, 6 and 7. Some experiments use a sightly modified method. However, here the final state of the method is presented as it is used in most experiments of this study.

4.1 Modeling transformations inspired by the noisy channel

The proposed method is based on the noisy channel paradigm, which ground in work by Shannon (1948). The model considers a text source and its reuse as two sides of a noisy communication line—as it is known form information theory—in which the aim is to find a formal way to describe what happens to the data flow. Basically, Shannon determines the degree of redundancy that an information flow must contain in order to ensure the successful transmission of all the information. In this thesis the model is used to make clear that the channel itself is considered the place where modification happens. Figure 4.1 displays this part in the middle rectangle. The minimal program is a way to determine the minimal set of operations only (see Kolmogorov, 1963).

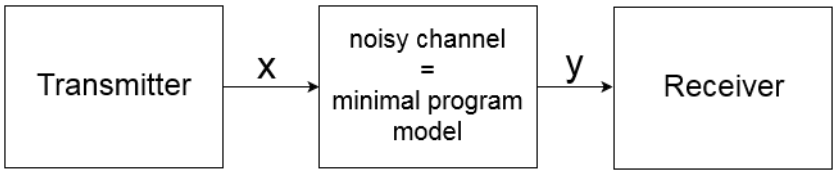


Figure 4.1: Principle of noisy channel model containing Kolmogorov's minimal program with in the noisy channel

4.1.1 Cheapest operations first

The noisy channel model first and foremost gives the reader an idea on what there actually is to model. In the field of humanities research—especially when the task is to collect scattered pages of a manuscript—scholars are often encountered with text that they possibly can date, but not necessarily attribute to a certain manuscript as these tent to be split and distributed (see an article by Shenton on reunification and preservation of manuscripts in Shenton, 2009). The noisy channel gives a means to possibly solve this issue by assuming an "average"[1] modification process between two versions of the same text that helps to consider what happened to a text during its transportation time. This work therefore focuses on identifying precisely and explicitly what modifications happen between two text excerpts.

The main target is to find a method by which modification in historical paraphrastic reuse can be described explicitly. Accordingly, it is obvious to first consider the basic modification operations as those introduced by Levenshtein (1965), namely *insert*, *delete*, and *replace*. Principally, these operations denote an action that needs to be performed on one item (word) of an input text in order to represent a related item (semantically equivalent word) in a target text. The proposed method especially focuses on modeling different versions of *replace*. In general, these versions come from two broader areas: i) operations that imitate a typical preprocessing stack as one would apply it to a document collection in a retrieval task, namely case-folding, normalization and lemmatization, and ii) operations that represent semantic operations such as synonymy, hypernymy, hyponymy and co-hyponymy. The latter are derived from a lexical database, which principally follows a tree-based structure. This structure is exemplified in Fig. 4.2.

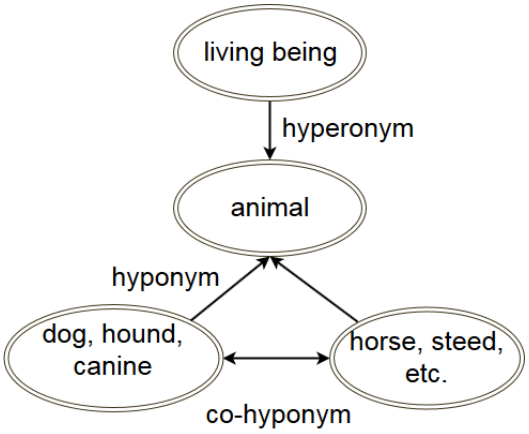


Figure 4.2: Principle architecture of a synset database

[1] This is usually represented by language models derived from huge text collection. However, calculating a language model is not in the focus of this work

The operations synonymy and co-hyponymy are considered symmetric relations. This means:

$$\forall x, y \in M : xRy \Rightarrow yRx \tag{4.1}$$

Where x and y are two words, which are synonym/co-hyponym to each other, and M is the amount of potential operations. Further, synonymy and co-hyponymy are equivalence relations. This means they fulfill—next to symmetry—also reflexivity:

$$\forall x \in M : xRx \tag{4.2}$$

and transitivity:

$$\forall x, y, z \in M : xRy \wedge yRz \Rightarrow xRz \tag{4.3}$$

Because, i) every word x is synonym and co-hyponym to itself—which fulfills reflexivity (we model this as no operation, NOP) and ii) if one word x is synonym/co-hyponym to one word y, and one word y is synonym/co-hyponym to one word z, then x is also synonym/co-hyponym to z—which fulfills transitivity. However, it can happen that one word has multiple senses meaning that it can appear in more than one synset. Hypernymy and hyponymy are considered unsymmetrical.

In the course of this work, we found that two more operations help to improve the alignment performance. These are *deriv*, representing relations between words that are derivations from each other (e.g., hold, holder), and words that can be aligned to each other by a strict character distance-based similarity score, *editdist*. The operations are applied between two text excerpts—verses or sentences—of a parallel corpus. Every operation takes parameters, which are either the words themselves or the POS tag, or both. For example, in *lower(LORD,Lord) lower* is the operation, and *LORD* and *Lord* are the parameters. Table 4.1[2] lists the operations following the prioritized order that they are applied in.

4.1.2 Transformation of minimum costs and length

In modeling modification, the aim is not only to apply operations that represent change explicitly and in detail. It is also demanded to have a minimum operation set that closely follows the length of an input verse/sentence to calculate its output version. This task is inspired by the complexity of Kolmogorov (Kolmogorov, 1963; Li & Vitáni, 2008)—the minimal size of a program that computes a specific output—and by so-called edit scripts (Chawathe

[2]Note that we distinguish operations with and without changes in POS, hence we work with up to twenty-one different operations. NOPs are displayed to set ratios of operations into relation.

no.	operation	description	example
1	NOP(word1,word2)	no operation necessary	NOP(above,above)
2	lower(word1,word2)	lower-casing matches	lower(LORD,Lord)
3	norm(word1,word2)	normalizing matches	norm(desireth,desires)
4	lem(word1,word2)	lemmatizing matches	lem(mine,my)
5	deriv(word1,word2)	derivation match	deriv(help,helper)
6	editdist(word1,word2)	short edit distance match	editdist(Phinehas,Phinees)
7	syn(word1,word2)	words are synonyms	syn(went,departed)
8	hyper(word1,word2)	word1 is hypernym of word2	hyper(coat,doublet)
9	hypo(word1,word2)	word1 is hyponym of word2	hypo(spears,arms)
10	co-hypo(word1,word2)	words are co-hyponyms	co-hypo(steps,feet)
11	fallback	unaligned	-

Table 4.1: Overview of transformation operations; The upper part presents first order operations, the lower part presents second order operations. MorphAdorner's tag set distinguishes POS tags in detail, e.g., verbs are distinguished in 2nd and 3rd person present, in infinitive and past tense and past participle, and conjunction of wh-words are distinguished from adverbs. Operations are applied in the order of their running number.

et al., 1996), which transform documents (e.g., program code) by applying a minimum number of operations. For this purpose we stop the operation alignment when tokens from the shorter version of two text excerpts finished iterating. Figure 4.1 contains the minimum program in the center of the graphic.

4.1.3 Alignment

The operations introduced previously are modeled on top of the word-aligned verses from the parallel corpora. After testing the alignment performance of an iterative approach to associate words from one verse with words from a counter verse, results showed that performance increases significantly when the text is statistically prealigned. Hence, Berkeley Word Aligner (DeNero & Klein, 2007) is applied on the tokenized versions of two Bibles each from the parallel corpora[3] before we model the operations—introduced earlier—on top of the associated verse indexes that are the output of Berkeley Aligner.

Figure 4.3 shows the overall procedure of the method starting in the upper left part with the preprocessing of parallel corpora used, which is performed by tools such as Berkeley Word Aligner and MorphAdorner (Burns, 2013). The following sections in this chapter introduce in a more detailed manner which sources and tools are used to create normalized,

[3]In German inflection is much more complex, especially for the old Luther Bible. Hence, for German Berkeley Aligner operates on the normalized text version, not the tokenized versions.

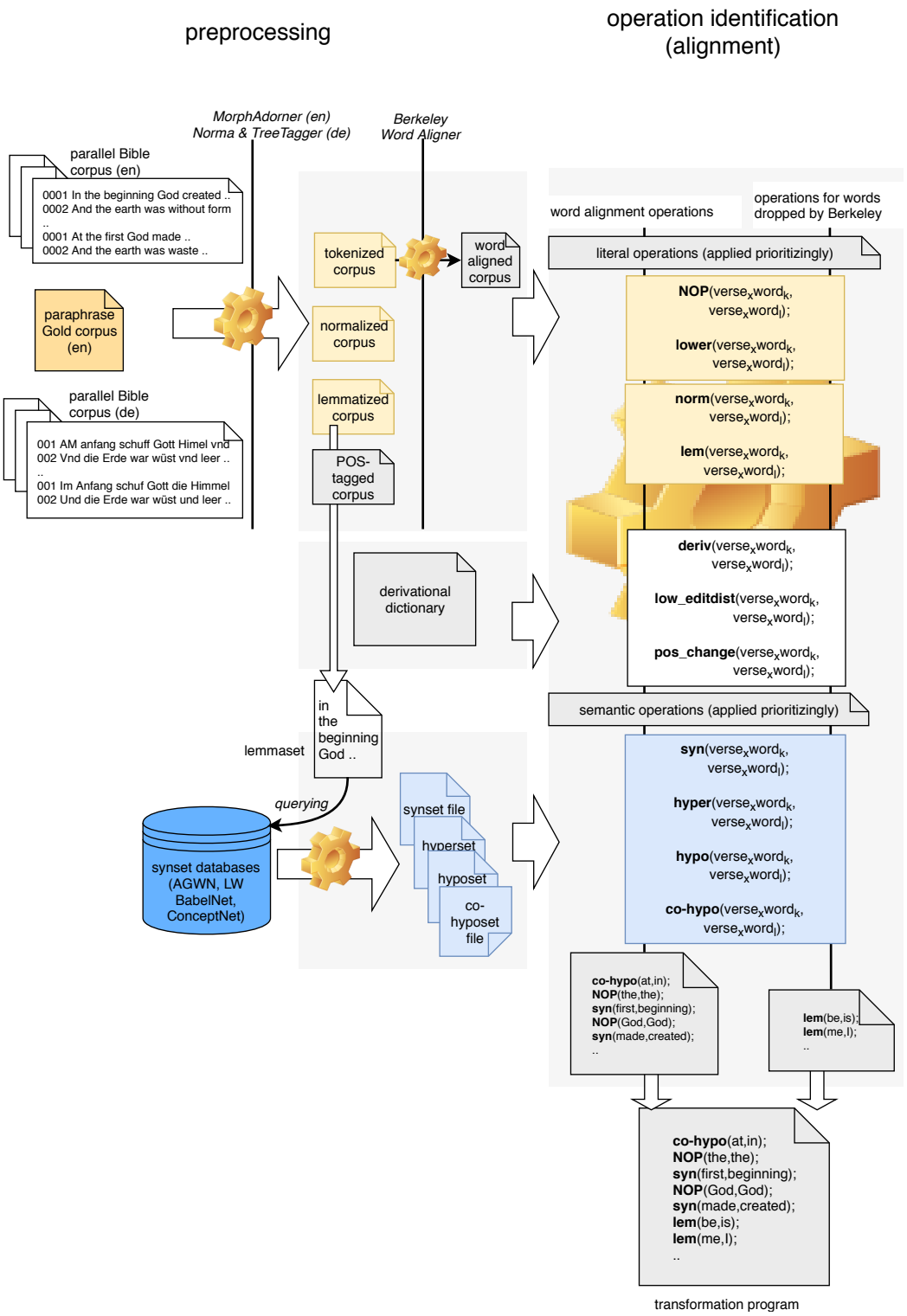


Figure 4.3: Overview of the data processing workflow

lemmatized, and POS tagged versions of the corpus' texts. For this purpose, we break down Fig. 4.3 to explain parts of it in the relating sections.

4.2 First-order operations - Morphological modification

First-order operations are operations that are applied to associate two words that still share the same cognate, i.e., words that are variants or inflections of each other. They are called first-order operations, because during the modification analysis these operations are preferred above operations that represent semantic equivalency such as synonyms. These operations are *NOP*, *lower*, *norm* and *lem*. *NOP* is applied when two words are not modified at all, *lower* we apply when the case-folded versions of two words are equal. To find *norm* and *lem* representations of words, we use tools that are specifically designed to work with historical text. Namely MorphAdorner (Burns, 2013) for the English texts, and Norma (Bollmann, 2012) and TreeTagger (Schmid, 1999) for the German text[4].

MorphAdorner initially was built to adorn text from the Early Modern English period. However, it also works suitably well for the more modern English language Bibles. For lemmatization MorphAdorner first looks-up lemmas from the lexicon. For irregular forms a mix of a list containing associated word forms, and grammar rules partially based on Martin Porter's suffix stripper (Porter, 1980) are used. According to Wilkens (2008) MorphAdorner achieves an error rate of 1.9% on historical texts.[5]

Norma was constructed within the Anselm project[6]. Its aim is to normalize Early New High German to modern German spelling. For example, "vnse lybe vrouwe" to "unsere liebe Frau".[7] For this purpose, Norma uses a combination of a look-up list containing manually created normalized word forms associated with a word variant, context-aware character rewrite rules that are based on a modified edit-distance algorithm, and a so called "Weighted Levenshtein Distance (WLD)" normalizer that uses a weighted edit distance to choose the most probable alignment for one word and its variant (Bollmann *et al.*, 2011). According to Bollmann *et al.* (2011) Norma achieves up to 90% of accuracy[8] on the Luther Bible.

The Norma output is used as the TreeTagger input. TreeTagger then determines the lemmatized word forms. Overcoming the issue of sparse data transitions that Markov Model-based taggers encounter, TreeTagger uses a decision tree to calculate estimates for

[4] See Ch. 5 to learn about preprocessing for the Ancient Greek and Latin reuse data

[5] The test corpus in Wilkens (2008)'s study consisted of a hand-tagged corpus of mostly nineteenth century English fiction which sums up to about 3.8 million tokens.

[6] `https://www.linguistics.ruhr-uni-bochum.de/comphist/projects/anselm/index.html`

[7] Example from `https://www.linguistics.ruhr-uni-bochum.de/comphist/resources/norma/index.html`

[8] Accuracy is defined by: (TP+TN)/(TP+FP+TN+FN).

tag sequences for given input sequences. Thus, the algorithm operating on a tree-based structure can handle a short memory of preceding tokens together with the associated tags more reliably, and can then easier decide which tag is the most probable one to be assigned in the current step. According to Schmid (1999) TreeTagger achieves an accuracy of up to 96.81%.[9] All tools require a running text input that provides one word per line or offer scripts to tokenize the text based on simple heuristics.

POS tags are provided by both, MorphAdorner and TreeTagger. Later, in the experiments, the operations are distinguished into operations with unaltered POS tag and with altered POS tag. Figure 4.4 shows an overview of the preprocessing procedure. After the texts are run through MorphAdorner and TreeTagger respectively, the tokenized, normalized, lemmatized and POS-tagged versions of all Bibles are available, out of which the tokenized versions are inserted (pairwise) into Berkeley Word Aligner.

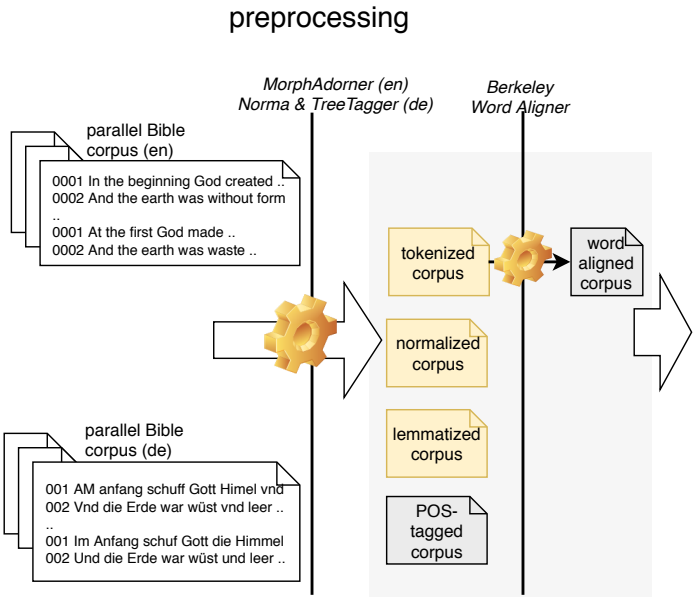


Figure 4.4: Preprocessing overview

[9] Recall figures are not provided.

4.3 Derivational information and resources used

4.3.1 Categorial variation database for English

To further improve the alignment accuracy, operations that are represented by a word's derivations are also enabled. This means, two associated words are linked because one word is build by the diverseness of derivational morphology of the other word, hence changing the part of speech—as opposed to inflections that are covered by *lem*. For aligning English texts, the Categorial Variation Database (CatVar) published by Habash & Dorr (2003) is used. CatVar contains in version 2.0 62,232 clusters covering 96,368 unique lexemes. These belong to the four POS classes (Noun 62%, Adjective 24%, Verb 10% and Adverb 4%). CatVar combines a range of resources and algorithms such as—among others—the Brown Corpus section of the Penn Treebank (Marcus *et al.*, 1993), the English normalization lexicon NOMLEX (Macleod *et al.*, 1998), WordNet 1.6 (Fellbaum, 1998), and the Porter stemmer (Porter, 1980).

4.3.2 Derivation dictionary for German

For German, we use the derivation dictionary DErivBase by Zeller *et al.* (2013). DErivBase is a rule-based framework for inducing derivational families, i.e., word lemmas that go beyond POS boundaries. Applied on the SdeWaC corpus (Faaß & Eckart, 2013) it finds over 280,000 single lemmas distributed over 235,288 derivational families out of which 17,000 are non singleton clusters. We use this knowledge to find word associations beyond lemmatisation, but before semantic relations as delivered by the synset databases are used.

4.4 A strict edit distance-based operation

Following the hierarchy of operations that are applied to record modification in the reuse data, the next priority is an own development based on Levenshtein's (Levenshtein, 1965) edit distance. This operation, called *editdist*, is considered as a trade-off between the derivational and morphological modifications (in the upper part of Tab. 4.1), and the operations representing semantic relations (i.e., complete word substitutions, lower part of Tab. 4.1). The *editdist* operation determines two words as related if their edit distance is not higher than 2/7 of the length of the shorter word, and is only applied when the shorter word is at least six characters in length. This measure was found after experiments, and is especially useful to align writing variants of named entities that have a certain length and cannot be captured by the preprocessing tools. See Ch. 6.1 for more details and improvement of the alignment.

4.5 Second-order operations - Semantic relations

4.5.1 BabelNet as primary resource for semantic relations

The multilingual lexico-encyclopedic dictionary BabelNet (Navigli & Ponzetto, 2012) comes with both, lexicographic and encyclopedic coverage of about 16 mio. entries, while, at the same time, it is a semantic network, which stores concepts and entities together with the semantic relations among them. BabelNet integrates resources such as—among others—WordNet (Miller & Fellbaum, 2007) and Wikipedia[10].[11] We call the operations representing semantic relations stored in such a lexicographic database are second-order operations, because they are applied after first-order operations.

Synsets

First and foremost, BabelNet is a synset database that stores words with the same meanings (senses) together in a synset. One word can have different senses which makes them occur in different synsets. Synsets have edges which are either labeled with hypernym or hyponym relations. BabelNet's core version was build by automatically integrating lexical knowledge from WordNet and encyclopedic knowledge from Wikipedia to form a multilingual, widely covering network. Machine translation techniques provide compensation for underrepresented languages in Wikipedia. The principle approach determines the mapping of Wikipedia pages to WordNet senses considering techniques based on simple bag-of-words representations and more advanced graph representations—based on WordNet itself. BableNet maps tens of thousands of Wikipedia pages to their corresponding WordNet senses with an F1 measure of about 78% (Navigli & Ponzetto, 2012). Later, more resources were added to the BabelNet core, such as VerbNet (Schuler, 2005), WikiData (Vrandečić & Krötzsch, 2014), and FrameNet (Baker *et al.*, 1998). This study makes especially use of the lexical information stored for the English and German language. We use the BabelNet API 3.7 to query for synonym, hypernym, hyponym and co-hyponym relations to all our word lemmas in all experiments that make use of BabelNet data. Recall Fig. 4.2 for the principle architecture of a synset database.

Going beyond synsets

Linking words with the same meaning (synonyms) together is important to find semantic equivalency in text (see also the definition of equivalency in Sec. 4.1.1). However, the method proposed here, goes beyond this standard relations and further considers hypernyms,

[10] `https://www.wikipedia.org/`
[11] `https://babelnet.org/about`

hyponyms and co-hyponyms. The following example shows why considering these extra relations—unlike existing similarity metrics[12]—is important:

1. the *dog* eats the bone & the *hound* eats the bone → 5% distance
2. the *poodle* eats the bone & the *dachshund* eats the bone → 65% distance
3. the elephant eats the *orange* & the elephant eats the *pear* → 60% distance
4. the elephant eats the *peanut* & the elephant eats the *nut* → 60% distance
5. the elephant eats the *peanut* & the elephant eats the *groundnut* → 5% distance

These examples are calculated using Meteor (Denkowski & Lavie, 2011), a machine translation evaluation score that tests for semantic equivalency for machine translation output compared to a reference translation. Next to character ngrams that sentences have in common, Meteor also considers synonyms utilizing WordNet for this task. Here, the Meteor similarity score is simply applied to measure the similarity of two short sentences to show how it operates. The figures especially show that even with semantically very similar lexicalization, exiting techniques tend to calculate unrealistic similarity. Figure 4.5 shows an overview of retrieving semantic relations: First, a lemma is looked up in a synset database (such as BabelNet), then, all relevant synsets with their related hypersets/hyposets are downloaded for further use.

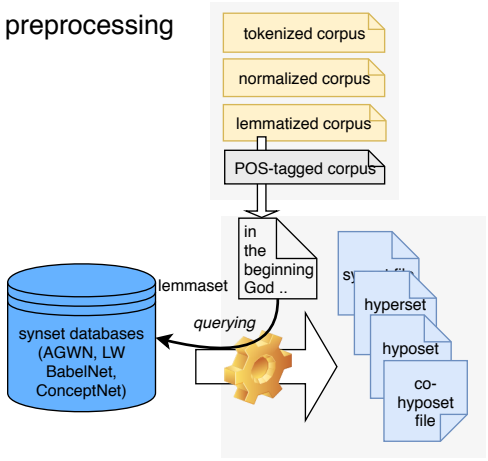


Figure 4.5: Overview of the querying and the download of synsets and their related hyper and hyposets

[12]To better fit these metrics to scope of this work, the *distance* is denoted next to the following examples: $distance = 1 - similarity$

Replacement frequency of words with multiple senses

To grasp an understanding of words with multiple senses i.e., words that appear in multiple synsets, we now take a small detour. One extra part of the presented study is to investigate the frequency/appearance of words replaced during paraphrastic reuse and the number of their different meanings. Precisely, Moritz & Büchler (2017) investigate the distribution of ambiguous words that are substituted between i) a Bible written in basic English and the King James Bible, and ii) a translation following the primary source wording literally and the King James Bible. The work investigates whether and how theses substituted words correlate to the number of their senses. It turned out that, against initial conjecture, there is no significance in the frequency of use between less ambiguous and more ambiguous words that are replaced between one Bible translations and the other. Instead, the likelihood of a word to be replaced with a synonym, hypernym, hyponym or co-hyponym tends to be increased correlating to the number of its senses. Figure 4.6 shows the distribution of replaced words (y-axis) grouped by the number of senses of those words (x-axis) for all words from the King James Bible that are replaced between the King James Bible and the Bible written in basic English. The values of the y-axis are normalized by the number of senses displayed on the x-axis. For the purpose of this study, we learn that the number of meanings of a word and the use of a sense of a word depends on the richness of the vocabulary of a Bible and the time in which a words was common.

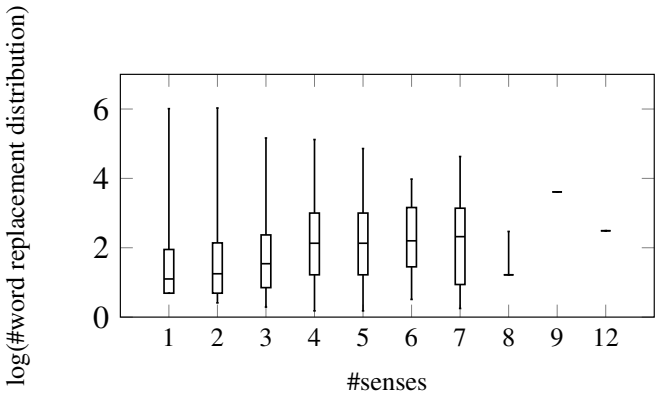


Figure 4.6: Distribution of replacement frequency of lexelts (y-axis), no. of senses of lexelts (x-axis)

4.5.2 BabelNet versus ConceptNet

We also consider ConceptNet (Speer & Havasi, 2012). ConceptNet is a semantic network that has the purpose to represent the meaning of words. It was created starting 1999 at

the MIT Media Lab by crowd sourcing, later expert knowledge, and knowledge generated by games that users can solve. It stores 37 relations among words out of which "Synonym" and "isA" (hyponymy) are only two. Many of the relations are word-class specific such as "Entails" (verbs) and "CreatedBy" (nouns). From a freely available database dump—that basically stores two words together with their relations—we form the synsets, hypersets, hyposets and co-hyposets to be used in the overall approach.

ConceptNet is used in Sec. 6.2 in which empirical insights of the proposed method are described. Compared to BabelNet ConceptNet is slightly smaller. As such—in a preliminary test—it identifies more synonyms, but not hypernyms and hyponyms, which is owed to the way its taxonomy is stored. This means, e.g., hyponyms are string specializations of their hypernyms (for example, "tank" is a hyponym of "weapon". Biblical vocabulary is supported rarely. It also finds only about 10% of the co-hyponyms compared to considering BabelNet alone, when enabled in the overall transformation alignment set-up. Table 4.2 shows a detailed overview of the pure numbers calculated. It turns out that BabelNet rather identifies relations as co-hyponyms where ConceptNet supposes them to be synonyms. This is a question of how an architect of such a database defines semantic equivalence. This can differ between several projects. A detailed analysis follows in Sec. 6.2.

operation	**BabelNet**	**BabelNet & ConceptNet**
syn	697,234	732,021
hyper	219,846	0
hypo	218,726	0
co-hypo	337,217	39,600

Table 4.2: Overview of recall in semantic relations considering BabelNet and BabelNet plus ConceptNet for the alignment of eight historical Bibles. The resolution of multi-word alignment by Berkeley Aligner results in an unordered operation assignment: typically the cheapest operation is preferred.

One further lexical database to mention here is EuroWordNet (Vossen, 1998), a resource that combines WordNets of European languages and follows the structure of Princeton WordNet, which is based on synonyms. However, it is not freely available—an academic license is between 200 and 400 Euro per language. The Open Thesaurus (Naber, 2005) contains lexical knowledge in Spanish and German and some East European languages. We exclude it, because we want to have one database that supports information for both our languages under investigation, English and German, for reasons of comparability.

4.6 The recall versus precision trade-off

In this work, we primarily use lexical databases to determine semantic relations between words. That is, especially because we are interested in learning about the nature of these relations. These relations, e.g., synonyms, etc., and the frequency they appear in, help to understand their contribution in the reuse process. This again enables a twofold information gain, i) it tells what paraphrastic reuse looks like in detail and explicitly described, and ii) it gives a snapshot of available resources and resource support to perform the task. A further advantage of using lexicon-based databases—instead of pure statistics[13]—is precision. Princeton WordNet (Fellbaum, 1998), the first lexical database that stored semantic relations, is handmade. Relations between words are manually defined. Meanwhile a lot of such synset databases exist. They typically are generated manually such as FrameNet or using semi-automated approaches such as BabelNet.

A counter approach to the lexicon-based method primarily used, is the identification of similar text excerpts and words based on statistical techniques that range from simple probability distributions of words and characters (unigrams, bigrams and ngrams) in a text, up to Markov models (see e.g., Brill, 1995), flat-layered graphs—such as conditional random fields (Lafferty *et al.*, 2001, see e.g.,) and deeply connected networks (see e.g., Socher *et al.*, 2011). These techniques possibly can determine similarity for more terms then a lexical database can ever contain, and they can tell how semantically similar words are. However, they do not provide the explicit lexical knowledge that we are interested in. Further, they require a critical mass of training data for word contexts, which makes their use likewise difficult for low-frequent words or exotic writing variants (see Chap. 6.1). Finally, word embeddings—nowadays a wide-spread method to perform deep learning for linguistic analysis—are successfully applied to digital humanities research questions. One example is the computational investigation of the variance of word meaning in diachronic text corpora using distributional semantics and human emotion dictionaries (Hellrich *et al.*, 2018). Kestemont *et al.* (2017) perform a detailed analysis of computational approaches in Paleography using primarily methods based on bag-of-words and deep learning, as well as stochastic neighbor embedding for the visualization of the data. Under the assumption that word embeddings are specifically trained for text of a certain domain, genre, and time epoch, those possibly can help to foster research in areas in which hand-made resources are still preferred.

[13] Measures based on machine learning often are able to tell some type of relevance between two semantically equivalent text excerpts, but can not describe these similarity explicitly.

Chapter 5

A small-scale reuse analysis in two Medieval Greek and Latin datasets

This chapter is an expansion of the following papers:

- Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, & Marco Büchler. Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and its Application to Bible Reuse. In: Proceedings of EMNLP 2016. ACL.[1]
- Maria Moritz & Marco Büchler. An Automated Approach to Model the Transformation Process of the Reuse in Bernard de Clairvaux: How Do Lexical Resources help?. DH 2017. ADHO. *priced with ADHO Tavel Award '17 of €850*

Automated HTRD is not much understood yet, hence, empirical studies are necessary to enable and improve its automation. This chapter presents a linguistic analysis of text reuse in two medieval datasets. Precisely, it gives a deeper understanding how historical text reuse is constituted. To this end, the operations introduced in Sec. 4.1 are applied to find a mapping between a source text and its reused version. An algorithm decides when which operation is applied. This algorithm is processed on reuse data by Clement of Alexandria and Bernard of Clairvaux and empirical results on how text is reused in detail are gathered. We formulate implications that come with the empirical results from this task. The automated approach is complemented by a manual analysis of a subset of the reuse.

[1] Contributions: 1) Research and methodological design (Büchler & Moritz), 2) Coordination, contacting Bible experts, data processing, algorithm implementation, automated transformation measurement, transformation guidelines design for manual transformations, results presentation and interpretation, paper writing, rebuttal (Moritz), 3) Manual transformations (Pavlek and Wiederhold), 4) Ancient Greek WordNet consulting (Bizzoni)

5.1 Overview

This section contains a small-scale case study of the overall study of this thesis. To this end, it follows the overall motivation, research questions and hypotheses. Because the overall motivation and research questions are already defined in Ch. 1, the purpose of the next section is to give an idea of how the research questions (Sec. 1.3.1) are addressed and to refresh information about the data used (Sec. 3.1.1).

5.1.1 Method to measure reuse in two medieval datasets

The proposed automated approach is used to characterize transformations from a reuse instance back to its original, and it is applied on two medieval datasets. The automated experiment is complemented with a manual analysis of a smaller sample. The study, hence, comprises the following main steps:

- First (**RQ M1**), we identify/distinguish and characterize the literal and non-literal overlap in the reuse instances by grouping the operations.
- Second (**RQ M2**), we consider the operations defined earlier that reflect literal reuse, morphological modification, and replacements represented by semantic relationships (as defined in Sec. 4.1).
- Third (**RQ M2.1**), we apply an algorithm that identifies operations by first looking for exact matches, 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 (as defined in Sec. 4.1).

The proposed procedure is applied to two datasets and the relationships of affected words and the literal share are investigated. Occurrences of operations are quantified and we characterize to what extent the linguistic resources are helpful. Two measures sup_{lem} (lemma support) and sup_{AGWN} (support by the lexical database) are calculated to assess the resources' coverage for the study.

- Finally (**RQ M2.2**), a modern synset database (see Sec. 4.5) and one that is made to retrieve Latin and Greek are compared regarding their ability to identify semantic relationships among words. A smaller sample of the reuse datasets is analyzed manually, using further operations to understand the full richness of the reuse.

5.1.2 Datasets used

This analysis makes use of three datasets, i) the Bible reuse of Clement of Alexandria (see Sec. 3.1.1) together with Septuagint (Rahlfs, 1935) and the Greek New Testament (see

Sec. 3.1.1), ii) the Bible reuse of Bernard of Clairvaux (see Sec. 3.1.1) together with the Latin Bible (Sec. 3.1.1), and iii) Bernard of Clairvaux' bigger Bible reuse set (Sec. 3.1.1). Detailed information on the datasets can be fond in the relating chapter.

From the annotated reuse excerpts in the text of Clement of Alexandria, a total of 95 out of 128 mark-ups is selected, 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 a human expert[2] (skipping ten instances), (iii) the reference frame should be within five Bible verses (a verse is comparable with a sentence) to avoid too much noise and exclude strongly allusive references, which is beyond the investigation of this thesis (skipping nine instances), and (iv) reuse instances should not exceed a length of 40 tokens, again to cut the long tail, to avoid too much noise and keep the laborious work appropriately (skipping eight instances). Sometimes one reuse instance pointed to different Bible verses or one text passage contained more than one reuse instance. As a result, we come up with 199 verse-reuse pairs.

5.2 Detailed experiment description

5.2.1 Part-of-speech tagging

The automated and the manual approach also take POS information into account to understand the reuse transformation. The reuse instances originating from Clement and Bernard, as well as their source Bible verses are POS tagged using Perseus' tagging system (Bamman & Crane, 2011a). It maps POS and case information to single characters[3], which are shown in Tab. 5.1.

We introduce w for the POS gerund and F to denote a foreign word by ourselves, and POS tag the 199 reuse instances of Ancient Greek and the 162 of Latin, as well as the original Bible verses. b is furthermore introduced by ourselves to represent the Latin ablative case, which does not exist in Greek. Latin and Ancient Greek POS taggers often lack available trained models for certain epochs, or accuracy when exiting models are applied to a text that is different from the one it is trained on, or from the one for which grammar rules and vocabulary lists are defined for (Crane, 1991; vor der Brück *et al.*, 2015). For this reason, this step is performed manually to assure high accuracy. Further, we assign cases for the classes noun, article, adjective, and pronoun.

Figure 5.1 illustrates the whole procedure. Both, the text excerpt (reuse) and the relating Bible verse are manually assigned the POS and case information.[4]

[2] Andreas Wiederhold also assisted in the qualitative evaluation of this work. See Sec. 5.2.3.
[3] Documentation: `https://github.com/PerseusDL/treebank_data/tree/master/v1/greek`.
[4] Manual tagging is performed by Andreas Widerhold.

part of speech	tag	part of speech	tag
noun	n	pronoun	p
verb	v	numeral	m
participle	t	interjection	i
adjective	a	exclamation	e
adverb	d	punctuation	u
article	l	gerund	w
particle	g	foreign	F
conjunction	c	ablative (only Latin)	b
preposition	r		

Table 5.1: POS tagging following the tag system of the Perseus Digital Library. w, F and b are newly introduced.

5.2.2 Alignment supported by linguistic resources

Remember that the analysis of the reuse is inspired by the noisy-channel coding theorem (Shannon, 1948), which is about transferring data correctly through a noisy communication channel. One can understand the act of reusing as a similar problem, in which information between a primary author and the person who reuses a text unit is transmitted, and in which various kinds of noise affect and modify the data. As explained precisely in Ch. 4, the proposed approach is to model the transformation process in terms of parameterized operations applied to a word couple from the reuse and the source text to obtain the original words. These operations use lemma lists of classical Greek and Biblical Koine, and the Ancient Greek and Latin WordNet. For each reuse and its relating Bible verse a set of operations necessary to transform the reuse to its original is calculated.

Further linguistic resources

This small-scale experiment makes use of the following lemma lists to look up lemmatized forms of words—a prerequisite for looking up synsets:

- Biblindex' Lemma Lists contain entries for 65,537 Biblical Greek and 315,021 Latin words.
- Classical Language Tool Kit (CLTK) (Johnson *et al.*, 2014–2016) provides Ancient Greek and Latin lemma lists for 953,907 Greek words and 270,228 Latin words.

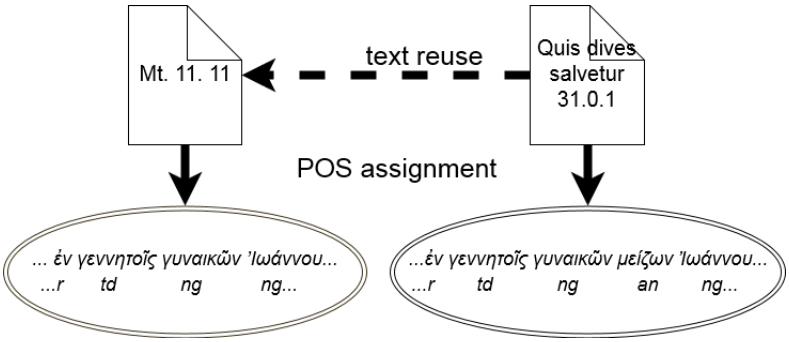


Figure 5.1: Illustration of our POS assignment. The original Bible verses and the reused text units are tokenized and each is assigned with the POS tag sequences respectively, which follows the same order as the text.

- SBLGNT&LXX refers to lemma lists extracted from the Greek New Testament of the Society of Biblical Literature (SBLGNT)[5] and the Septuaginta (LXX), a translation of the Old Testament (Rahlfs, 1935)[6] from the Center for Computer Analysis of Texts at UPenn. [7]

AGWN (Bizzoni *et al.*, 2014), which also contains Latin WordNet (LW) (Minozzi, 2009), is further used in this experiment to identify synsets, hypernyms, hyponyms, and co-hyponyms. In AGWN 33,910 synsets out of 98,950 contain Ancient Greek and 27,126 synsets contain Latin words. Words of the same meaning are aggregated in one synset. Hypernyms and hyponyms are associated via the tree-based structure that synset databases come with.

A second use case makes also use of BabelNet (BN) Navigli & Ponzetto (2012), which is a lexical resource made for modern languages, which also contains data in Latin. (see Sec. 4.5).

Coverage

For a first understanding, every token from the Bible verses and from its reuse is looked up in every single language resource independently. This means simply that the coverage of the vocabulary of each text sort by a resource is calculated. Table 5.2 shows how many

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

[6]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)

[7]We acknowledge code-page corrections by M. Munson. SBLGNT&LXX provide 59,510 word-lemma pairs.

of the words can be retrieved from each resource for the two datasets. Note that the total column simply shows the maximum that could be looked up on every single column. The resources used concern the lemma lists (a word is looked-up in the lemma list) and the synsets, hypersets, hyposets and co-hyposets. The latter return a success if the retrieved word exists in any of the sets from all synonym, hypernym, hyperonym or co-hyponym sets.

	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 5.2: Coverage of tokens by language resources. Note that every word with a hypernym (a mother) also has a co-hyponym (a sibling) and vice-verse, this is the reason for the identity of column hyper and co-hypo.

The table shows that CLTK covers the Bible data better than the Hellenistic Greek as used in Clement of Alexandria, the author from 2nd century AD, who writes in a rather archaic style, but uses Biblical vocabulary while also being influenced by Classical Greek. The coverage of lemmata stemming from the same source (Biblindex) as the reuse is checked for its ability to cover the reuse and Bible vocabulary too. However, Tab. 5.2 shows that the Greek vocabulary is covered best by the SBLGNT&LXX lemma lists. The Latin vocabulary indeed is covered best by the Bibleindex lemma lists. Finally, to not miss important information, all lemma resources are merged into one set of word-lemma pairs to be used by

the retrieval. In the lower part of Tab. 5.2 this turns out to ensure the best coverage. Every lemma of a word is then looked-up for semantic relations in the relating synsets, hypersets, hyposets and co-hyposets of AGWN.[8]

Experimenting with different ways of looking up lemmas showed that lower-casing all Latin tokens improved the success. For Greek (the Clement dataset), this step had the opposite effect, which indicates that the Greek text contains more entities that are not available in lowercase in the lemma lists, hence, these were not change in that case.[9]

Operations and grouping

For this setup, the replacement operations described in Sec. 4.1.1 are adapted to better address the use cases that are discussed in this chapter. Especially, since the use of lemma lists with associated normalized spelling to a given inflected form makes normalization (i.e., the *norm* operation) obsolete. Table 5.3 lists the operations for the computational approach. We distinguish case folding into the operations *upper* when a word was lower-cased during the reuse, and *lower* when it was written in lower cases in our Bible version—the reuse's source. Further, the operations *NOPmorph*, *repl_pos*, and *repl_case* are introduced for words having the same cognate, and *lemma_missing* is used when a word is not known to any of the lemma resources as well as *no_rel_found* when the relationship between a reuse word and each potential word from the original is not covered by AGWN.

Algorithm 1 shows the proposed approach to classify the reuse transformation by identifying the operations. Following this algorithm, the transformation from reuse instance to the Bible verse is the iterative concatenation (lines 30, 32) of the identified operations (lines 6, 8, 10, 13, 16, 18, 20, 22). For each reuse token (line 3), the algorithm identifies the first applicable operation matching the foremost word from the Bible verse (iterating over the verse—line 4) in the following order: exact word match (*NOP*, line 6), writing case changed from the Bible verse *lower* in the reuse to *upper* (line 7) or to *lower* (line 9). Thereafter, the algorithm looks up the lemma and returns *lem* if the lemma of the reused word matches the lemma of the original (line 13). For these four, the algorithm also checks the morphology (lines 6, 8, 10, 13), 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. This means that up to three operations can be returned per word. Finally, the algorithm checks for synonyms (*syn*), hyperonyms (*hyper*), hyponyms (*hypo*), and co-hyponyms (*co-hypo*), but does 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, i.e., it is black-listed (not in the pseudo code).

[8] The synonym, hypernym, hyponym and co-hyponym numbers depend on the lookup of the lemma lists.

[9] 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.

Algorithm 1: Reuse alignment 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   */
          $OP \leftarrow OP \cup (syn(t,b))$ **break**
        else if *id(s1)* = *parent_id(s2)* **then**
          /* t is hypernym of b   */
          $OP \leftarrow OP \cup (hypo(t,b))$ **break**
        else if *parent_id(s1)* = *id(s2)* **then**
          /* t is hyponym 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 */
          $OP \leftarrow OP \cup (co - hypo(t,b))$ **break**

        else
          $tmp\_op \leftarrow (no\_rel\_found(t,b))$

  end
  if *tmp_op* **then**
    $OP \leftarrow OP \cup tmp\_op$
  else
    $OP \leftarrow OP \cup (lem\_missing(t))$

end
return OP

operation	description	example
modified operation set of word modification, specific for the experiment		
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)*
semantic operation as common: syn, hyper, hypo, co-hypo		
operations taking morphological information as parameter		
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)* - genitive, dative
lem_missing(reuse_wrd,orig_wrd)	lemma unknown for reuse or original word	*lemma_missing(tentari,inlectus)*
no_rel_found(reuse_wrd,orig_wrd)	relation for reuse or original word not found in AGWN	*no_rel_found(gloria,arguitur)*

Table 5.3: Operation list for the automated approach

The algorithm is specially designed. Instead of figuring out the cheapest operation for each word pair after collecting any possible operation, it applies the soonest matching word operation of the counter verse to a given word from the input verse. This can lead to missing the cheapest operation in favor of a potential semantic one and, hence, gives an understanding of the possible presence of such semantic relationships for each word.

Latin WordNet versus BabelNet

In a second, slightly different setup targeting to address **RQ M2.2b**, the algorithm is modified so that it first finds all possible operations for a reuse word and a Bible word, and then applies the most literal operation using the counterpart Bible verse word, which fulfills this operation. This means, if no perfectly or lemmatized matching word (we summarize them as literal operation) is found, relationships of semantic closeness for a given word are retrieved, such as synonyms, etc. This algorithm is applied on the bigger dataset of Bernard's reuse (see Sec. 3.1.1), first using the relationships queried from LW and second, using BN. Afterwards, the operations identified are shown, and a support value for both processes is calculated.

5.2.3 Qualitative analysis

To obtain a deeper understanding of the limitations of linguistic resources for this study, two graduate students (one Classical Archaeologist for Greek and one Latinist for Latin)[10] manually analyze 100 from the Greek and 60 from the Latin reuses with their expert knowledge, using an extended set of operations. It has a richer set of replacement operations: those from the upper part of Tab. 5.3, but instead of only using *repl_case* when a word is inflected, the operation is refined and all changing morphological categories from Perseus' tagset are recorded for the respective modification between two words. For example, a resulting operation is *repl_case_a_g* when an accusative in the reuse is a genitive in the Bible version. Because there exist nine morphological categories with up to fourteen values, these are not listed in this thesis, instead the reader is forwarded to the Perseus project web page. For this case, *lem* simply remains to represent writing variants.

5.3 Results

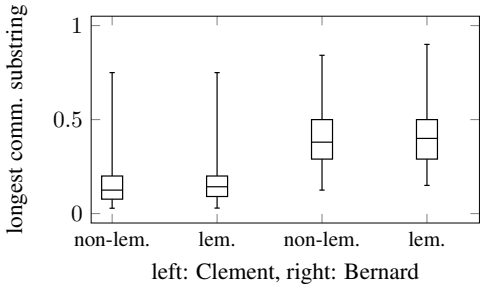
5.3.1 Literal share of reuse (RQ M1)

A first understanding of the reuse is obtained by looking at the percentage of overlapping words between reuse and Bible verse. There, we measure the longest common sequence of tokens. For example, the longest common sequence of tokens between "transfigurat se in angelum lucis" and "angeli lucis" is "lucis" = 1. The longest common sequence of lemmatized tokens however, is "angelus lucis" = 2. Figure 5.2a shows the distributions distinguishing between lemmatized and non-lemmatized word comparison.

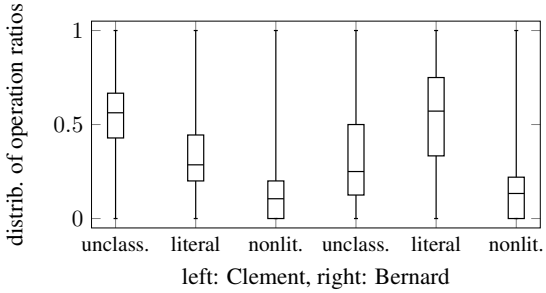
While lemmatizing words before comparison has only a small impact, there are differences between the data sets. In Bernard's reuse, the overlap is significantly higher than in Clement's reuse. 25% of Bernard's reuse instances have 50% or more tokens overlap with their original (see the upper quartile of the rightmost box of Fig. 5.2a). This is only the case for less than 25% in Clement's Greek data (see the upper part of the upper whisker of the left-hand boxes). Still, large overlaps of up to 75% (top whisker) in Clement and up to around 90% in Bernard exist—so, a small fraction of these reuses contain literal parts. This is an important outcome, which encourages to look deeper into the modifications that happen and to understand their type of modification as well.

For a more precise understanding of the literalness, the operations are grouped into *literal* (*NOP, upper, lower, lem*), *non-literal* (*syn, hyper, hypo, co-hypo*), and *unclassified* (*no_rel_found, lemma_missing*). Within each reuse, the relative occurrence of each of these groups is calculated using the results of the automated approach (see Alg. 1 and Sec. 5.2.2).

[10] namely, Barbara Pavlek and Andreas Wiederhold



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



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

Figure 5.2: Distribution of longest common substring and operation group rations in both data sets

Figure 5.2b shows the distribution of these relative occurrences for all reuse instances. It confirms Fig. 5.2a by showing a higher rate of literalness (see the right-hand box labeled *literal*) for the Latin dataset compared to the Greek dataset (see the left-hand box labeled *literal* in Fig. 5.2b). The figure further shows that the Latin reuse can be better classified by the approach, which takes the lemma lists and AGWN into account. Even if a significant part of words remain unclassified, that very fact points to two things: i) Bernard's Latin reuse is more often more literal than the Greek reuse of Clement and ii) Bernad's reuse can more often be classified by the synset database. This shows that the literal reuse identification can be well supported, but also that reuse identification on a non-literal level is more challenging simply because exiting resources lack the coverage of a diverse vocabulary.

5.3.2 Operations identified automatically (RQ M2.1)

Operation frequencies

Table 5.4 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 word couples and for 893 (67%) out of 1335 words in the Latin word couples, the algorithm was able to identify at least one operation, which already indicates to what extent the resources are helpful.

The first column group in Table 5.4 shows that about 25% to 30% of the tokens remain unmodified and that a high share of tokens experiences a morphological inflection. The second column group is especially interesting as it shows that some hypernym and hyponym relations are identified, but, more importantly, almost as many co-hyponyms are identified as are synonyms. The last column group especially raises attention to the ability of exiting resources to support the endeavor, which still is highly improvable. The lower part of the table shows operations that are applied next to operations from the upper part of the table for the exact same word couple.

Even though a big part of the operations is taken by unclassified or not found relationships, these figures clearly show that many word relations require require approaches that go beyond simple string matching, which is supported by the figures of *syn, hyper, hypo,* and *co-hypo*. Especially, the high ratio of identified co-hyponyms is important, because it is comparable with the synonym numbers identified and shows the impact of modifications beyond the rather tight relationship that synonym represents.

OP no.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**12**	**13**
		literal			**non-literal**				**unclassified**	
	NOP	**upper**	**lower**	**lem**	**syn**	**hyper**	**hypo**	**co-hypo**	**no_rel_found**	**lem_missing**
Clement	337	6	0	356	153	20	14	101	563	639
Bernard	587	0	44	102	60	14	28	68	347	85

OP no.	**9**	**10**	**11**
	NOPmorph	**repl_pos**	**repl_case**
Clement	420	49	258
Bernard	617	46	75

Table 5.4: Absolute numbers of occurring operations identified automatically for all reuse instances combined. Note that NOPmorph, repl_pos and repl_case operate on the PoS-tag, not on the word-level when a lemma relation was found. Punctuation is ignored. NOP figures are displayed for reasons of completeness.

Support measures

After having checked the overall coverage of the linguistic resources for all tokens, now the extent to what the resources support the identification of non-literal reuse transformation using the proposed approach is investigated specifically. Therefore, two specific measures are introduced sup_{lem} and sup_{AGWN} to calculate how often looking up a lemma or subsequently a synset element was successful. This is based on the operations from Tab. 5.4. Let $Occ(o)$ be the number of occurrences of an operation o obtained from Tab. 5.4. The operations that successfully looked up a lemma (before consulting AGWN) are described by the operation set *lem_success*={*lem, syn, hyper, hypo, co-hypo, no_rel_found*}. Recall that *lem_missing* represents the operation type for the case that a word from the reuse was not found in the lemma resources. Then, all occurrences of each operation from the operation set *lem_success* are totaled up and divided by a bigger operation set that also contains the *lem_missing operation*:

$$\mathrm{sup}_{\mathrm{lem}} = \frac{\sum_{\mathrm{Occ}(o)} o \in \mathrm{lem\_success}}{\sum_{\mathrm{Occ}(o)} o \in \mathrm{lem\_success} \cup \{\mathrm{lem\_missing}\}}. \tag{5.1}$$

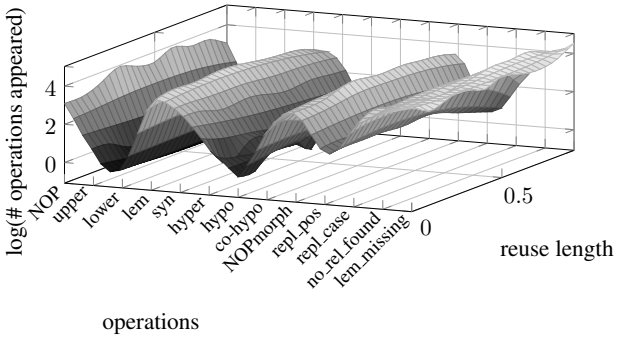
The summands of the numerator refer to the figures underneath the operations no. 4 to 8 and 12 from Tab. 5.4, because all of them use the lemma resources to look up a reused word upfront it is searched for in AGWN. The denominator refers to no. 4 to 8, 12 and 13, because no. 13 represents the case when a reuse word or its candidate was not found in the lemma resources. Finally, $\mathrm{sup}_{\mathrm{lem}}$ is 0.654 for the Greek reuse and 0.879 for the Latin reuse (0.848 without lower casing).

Similarly, the operations that successfully looked up from AGWN are *agwn_success*={*syn, hyper, hypo, co-hypo*}, with *no_rel_found* representing a failed lookup. Then:

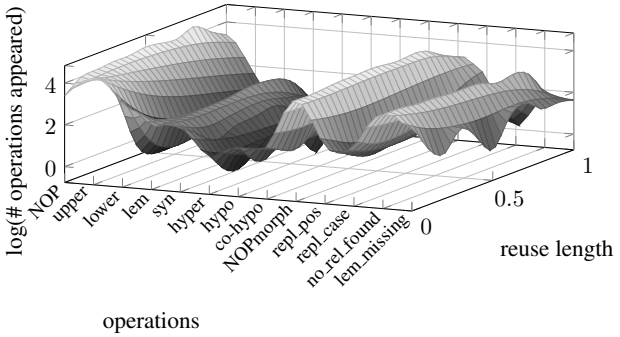
$$\mathrm{sup}_{\mathrm{AGWN}} = \frac{\sum_{\mathrm{Occ}(o)} o \in \mathrm{agwn\_success}}{\sum_{\mathrm{Occ}(o)} o \in \mathrm{agwn\_success} \cup \{\mathrm{no\_rel\_found}\}}. \tag{5.2}$$

Whereas the summands of the numerator refer to the figures underneath the operations number 4 to 8 from Tab. 5.4, because all of them use AGWN to look up a reused word. The denominator refers to number 4 to 8 and 12, because number 12 represents the case when a reuse word or its candidate was not found in AGWN. The result of the support value $\mathrm{sup}_{\mathrm{AGWN}}$ is 0.34 for Greek data and 0.33 for Latin data.

These values confirm that lemma resources for genre and time-specific text work comparably well for less-literal reuse. However, the resources to retrieve semantic relations (synset databases) show a lack of support and need further development and enhancement for specific domains.



(a) Clement of Alexandria



(b) Bernard of Clairvaux

Figure 5.3: 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. The Z-axis denotes the logarithm to compress space.

Distribution of operations in the reuse length

Finally, Fig. 5.3 visualizes the distribution of the frequencies (z-axis) of each operation (x-axis) with regard to the operations' positions in a reuse (y-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. Figure 5.3 indicates that most operation types are quasi-equally distributed over the whole reuse length without a particular trend in both data sets. However, we encounter a more frequent use of *lower* in the beginning of verses from the Latin dataset, which means that Bernard often reuses Biblical verses starting from the beginning.

5.3.3 Operations identified qualitatively (RQ M2.2a)

As described before, transformation operations for 60 reuse pairs of the Ancient Greek data and for 100 of the Latin data are identified manually. Here operations are first distinguished into *NOPs*, *insertions* and *deletions*[11].

In the Greek data *NOPs* cover 9.3%, insertions 49.8 %, and deletions cover 30.5%. In the Latin data *NOPs* cover 26.1%, insertions 49.7%, and deletions 11.9%. This again shows that Bernard stays closer to the Bible then Clement does.

operation	Greek	Latin	operation	Greek	Latin
syn	78 (40.6%)	91 (40.4%)	repl_gender	6 (3.1%)	1 (0.4%)
ant	1 (0.5%)	0	repl_mood	11 (5.7%)	12 (5.3%)
hyper	3 (1.6%)	0	repl_number	17 (8.9%)	17 (7.6%)
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%)
co-hypo	0	1 (0.4%)	repl_tense	3 (1.6%)	9 (4.0%)
			repl_voice	0	8 (3.6%)
			repl_case	38 (19.8%)	36 (16%)

Table 5.5: Numbers of replacement operations identified for the manual reuse transformation.

Table 5.5 shows the precise ratios of the various replacement (modification) operations based on the remaining 10.4% and 12.2%. Similar to the automated approach, synonyms and other semantic-level operations are strongly used. Further, also a certain portion of switching morphological categories, which indicates paraphrastic reuse takes place (see the *lem* row in the left part of the table and the complete right part of the Tab. 5.5). 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 the Latin dataset, POS changes represent 15% of replacements: a pronoun changed to a noun (6 times) and a participle became a verb (12 times), and twice, a noun became a verb and a participle each. A verb also became a noun twice. Case changes are shown in Tab. 5.6. 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 paraphrastic expression changed. We can learn from these detailed modifications how diversely authors handle a text when they rephrase and copy it. Often, they add their personal signature and style, which can for example be measured by the morphological categories they change a word into, and the degree of modification in general.

[11] In the automatic approach *insertions* and *deletions* are partly represented in the form of the modification operations. However, many of them are skipped due to the paradigm to align the shortest sequence possible.

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 5.6: Numbers of case replacements

Following exceptions prevent applying the proposed technique in straight forward manner, because they are more complex and not covered by the operation set. In the Greek dataset, one word is replaced with its antonym[12], and once a synonym also changes its POS. Four times, more than one morphological category changes, twice an auxiliary is deleted, and five times inserted. One writing variance (yet called *lem*) is identified, and three times a synonym is replaced by a multi-word expression. In the Latin data, in 16 cases a synonym is replaced while 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 paraphrastic for any word to match semantic relationships (e.g., "judged calmly"—Bernard vs. "fake friend"—Sal 12 18). These special cases especially highlight the necessity of a manual complement to an automated approach when a detailed data analysis is required to improve tool to work well for the digital humanities.

5.3.4 Operations identified on the bigger Latin dataset using different lexical databases (RQ M2.2b)

In a last setup, the aim is to discover how a modern lexical resource—that also contains Latin vocabulary—and one made for the Classical Latin can support the task of measuring modification. For this purpose, Tab. 5.7 shows the identified operations.[13] Using the LW, one encounters a high ratio of synonyms (*syn*) and, again, almost as many co-hyponyms as synonyms. Which raises awareness to the necessity of a strong use of semanitcally looser operations when reuse needs to be recovered. Further, a significant number of hypernyms

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

[13]Table 5.7 shows that the values for NOP, lower and lem (matching words, and words with same lemma) slightly differ inbetween both databases. This is caused by a design decision of our algorithm, which pragmatically permits to reassign a word when it already was used in an association with an earlier word.

and hyponyms is identified. These figures are comparable with the figures from Tab. 5.4. Using BN these figures are about a tenth as high and prove that resources are not sufficient when they do not match with the domain of the data under investigation.

	literal				non-literal				unclassified		
	NOP	**upper**	**lower**	**lem**	**syn**	**hyper**	**hypo**	**co-hypo**	**no_rel_found**	**lem_missing**	**total**
LW	4521	1	396	770	397	125	124	316	2470	450	9570
BN	4526	1	397	771	25	22	35	27	3316	450	9570

Table 5.7: Absolute numbers of replacement operations identified by LW, which is included in AGWN, and BN.

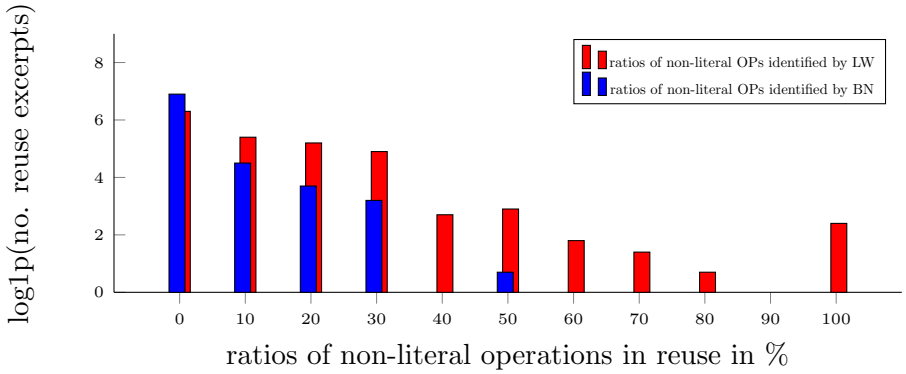


Figure 5.4: Ratio of non-literal (semantic) operations, aggregated in 10%-steps in relation to the whole reuse length. The reuse number is displayed logarithmically due to clarity reasons.

In Fig. 5.4 the reuse is ordered by its ratio of non-literal operations. The zero-bar represents reuse that does not contain any of the synonym, hypernym, hyponym and co-hyponym relations as modification. In contrast, the 100-bar represents reuse that only contains this class of operations. It shows that LW outperforms BN in identifying semantic relations, which represent non-literal text reuse, because these ratios are much lower for BN than for LW. One further encounters three significant descents: between 0% and 10%, 30% and 40%, and 50% and 60%.

Looking into samples (represented in Tab. 5.8) deeply, we find three patterns: i) the more semantic related words are replaced in a reuse, the more likely it is an allusion or analogy, and the less verbatim it is, ii) short allusions are better covered by the Latin synsets than paraphrases with a high ratio of semantic related words, and iii) paraphrases with a high literal ratio are covered best. These trends are displayed in Tab. 5.8. Summarizing, both word nets cover paraphrased reuse until a certain extend of replaced words, and LW better identifies allusions.

Furthermore, in Tab. 5.8 the samples are sorted according to their type of reuse. While the leftmost column (of the right part of the table) represents reuse that is often characterized by strong paraphrasing, the rightmost column concerns near literal copies of a text. The non-literal ratio of reuse decreases form the top to the bottom (left column). Ultimately, one can see a "hotter" diagonal—in the right part of the table—reaching from top-left to bottom-right. This shows that more allusive text reuse contains fewer stable words while literal text reuse contains many stable words.

Note that this one literal reuse in Tab. 5.8 that consists of 100% word replacements—being a contradiction—comes together as follows: "fragrantia" (fragrance, in the reuse) and "flagrantiam" (fragrance, in the Bible verse) are both inflections of the same word, the lemma lists however did not match their writing variants. Hence, the algorithm aligned "fragrantia" with another word from the Bible verse "odor", because the lexical database contains one synset where both words "fragrantia" and "odor" are synonyms.

non-literal ratio	analogy	allusion	paraphrased	near literal
100	6	1	1	1
60 - 80	5	1	1	0
40 - 50	1	1	1	2
10 - 30	0	0	0	5

Table 5.8: Sample classification for paraphrasticality. (Classification by L. Mellerin.)

Last, we also calculate the support value, which determines the ratio of non-literal operations (see Tab. 5.7) compared to them including unsuccessful resource look-ups (no_rel_found) in both, LW and BN. For LW this value is about 28%, for BN about 3%. Both values are to be understood as lower bounds, because often there is no reasonable relationship inbetween two words. Even if BN coverage is poor, its results tell which words of a dataset of Medieval Biblical Latin and Latin of the church fathers are prevailed in a current resource. Some examples are words such as "gloria" (glory) (contained in 17 synsets), "corona" (crown) (contained in 10 synsets), or "nemo" (nobody) (contained in 4 synsets).

5.3.5 Discussion

In the following, the research questions are answered **RQ M1:** The reuse is significantly non-literal and only lemmatizing words does not help discovering it. Results show that reuse in two medieval texts requires techniques beyond simple preprocessing (such as stemming or lemmatizing), which also explains why plagiarism-detection systems are challenged when paraphrases are used strongly (Alzahrani *et al.*, 2012). Further, 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 Tab. 5.5 and 5.6.

RQ M2.1: The results from the automated approach are encouraging, showing how reuse detection techniques can be supported with linguistic resources. Yet, it is not completely clear which precision and recall could be achieved and how existing techniques need to be adapted and calibrated in general. This investigation is beyond the scope of this study and subject to future work. However, first studies to investigate the issue in detail, is preformed by Franzini *et al.* (2018).

Next, linguistic resources support the automated 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.

RQ M2.2: The 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. This essentially extends current approaches that tend to focus on synonymy only. We further calculated the support of our approach by two lexical resources showing that language resources for Latin Biblical reuse are limited for certain time periods and that only a part of the required coverage is supported. This result raises awareness for the lack of resources for ancient data.

5.4 Threads to validity

5.4.1 External validity

External validity of this work is enhanced 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 lexical database for Ancient Greek and Latin.

5.4.2 Internal validity

A threat is that the ground truth has mistakes, as the POS tagging was done by one annotator only (Andreas Wiederhold) and relied on a manual post-correction. The selection criteria in Sec. 5.1.2 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 the sample of investigation 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, Fig. 5.5 displays the overlap of the whole 1128 instances of Bernard's extracted reuse. When compared to Fig. 5.2a (right) it supports the representativeness of the smaller sample.

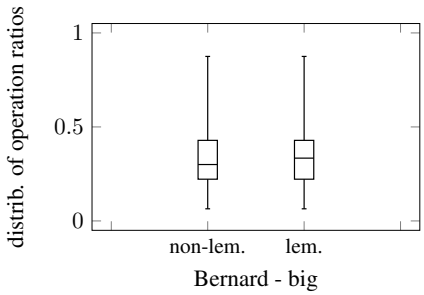


Figure 5.5: Ratios of literal overlaps in the whole Latin dataset

Last, the developed algorithm can only derive operation replacements when a word token was covered by the lemma sources that are 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.

5.5 Conclusion

This chapter presented a study of historical—and significantly non-literal—text reuse in two small Medieval Greek and Latin datasets. Reuse was automatically and manually characterized, and the extent to what existing linguistic resources are able to cover non-literal text reuse was identified. Further, the ratio of non-literal reuse in a bigger Latin dataset was identified and the support of two lexical resources was shown. The results show that language resources for Medieval Greek and Latin are limited and that only a small part of the required coverage is supported. This raises awareness for the lack of resources for ancient data, while language resources for modern languages are growing daily. The results show the potential as well as the necessity to develop robust techniques and to extend linguistic resources for analyzing and detecting such reuse, and proved the gap between current linguistic resources

and the characteristics that come with a degree of non-literal reusing. However, the results might also 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.

In future work, a smart 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. However, whether learning edit scripts on such intricate transformations is possible is an open question and valuable future research.

Chapter 6

A comprehensive analysis of paraphrastic text reuse

Until now the reader gained an understanding of how paraphrastic, historical text reuse—using the example of medieval Bible reuse—can be constituted (i.e., how which ratios are actually modified and in which regard they are modified), and how lexical databases are currently supporting the task of identifying these modifications. This chapter introduces a more comprehensive investigation of paraphasticality based on monolingual, parallel corpora of historical English and German Bible translations. Precisely, the chapter talks about the following main topics: i) the improvement of alignment accuracy in a parallel corpus of historical English, and ii) how reuse is modified in these corpora based on the figures that measure the introduced operations, and on figures that measure POS changes. This chapter contains an expansion of Moritz (2018)

6.1 Improving performance of writing variants (RQ B1)

This section is based on the publication:

- Maria Moritz. On the Impact of Time Proximity on the Alignment of Spelling Variants in old English Bibles: A Case Study. CRH 2018. Gerastree, Vienna.

6.1.1 Introduction

In this section, a prerequisite for the analysis of paraphrastic text reuse modifications is investigated, namely the word alignment of a subset of the used corpus that comes with a high ratio of historical writing variants. The question is to find out if time proximity can help to associate writing variants in Biblical text easier than between text that was published several hundred years apart from one another. Precisely, the procedure is to use temporally close Bible translations to, i) investigate the spelling modifications between them,

and ii) find out if the time proximity of these Bible translations can help to improve the alignment and, hence, the normalization of writing variants in old Bibles. In this study we are not interested in strongly paraphrased editions, hence editions such as the Bible in Basic English are omitted. Typical error cases that enable it to precisely improve the alignment method are also showcased and followed-up with an alternative alignment procedure. This is especially important when Bibles that are paraphrastic versions of each other need to be aligned. Considering the overall effort to investigate how paraphrased text is modified in detail, this section talks about a part that is a preparation step in the context of this thesis. Especially it is work on the improvement of word alignment of the parallel Bible corpus.

6.1.2 Complementing related work

In the field of specifically normalizing Early Modern English Yang & Eisenstein (2016) investigate application domain adaptation techniques to work with historical texts. Precisely, they apply POS tagging domain adaptation techniques to tag Early Modern English and Modern British English texts from the Penn Corpora of Historical English and find that embedding the entire lexical feature space (derived by means of co-occurrence statistics) outperforms simple word embeddings of individual words. This technique is also called structural correspondence learning or feature embedding. Combined with spelling normalization they yield an improvement of 5% (74% to 79%) in tagging accuracy on Early Modern English texts. Archer *et al.* (2003) report on (re)training the UCREL semantic and POS analysis system to cope with Early Modern English using news texts from 1653 and 1654 totaling in 613,000 words. They introduce a rule-based component for spelling normalization and template rules to identify morphologically modified words that are ambiguous in terms of POS. They achieve correct POS tags of about 94% when applying the system to a held-out dataset.

A more detailed overview of related work for the lemmatization of Historical English is discussed in Sec. 2.1.

In contrast to these related work, the herein presented approach makes use of a diachronic corpus to first improve word alignment in verse-aligned text, and, second, to attempts a way to normalizing Early Modern English using the same corpus.

6.1.3 Time proximity for variance alignment - overview

The goal is driven by diachronic data represented by temporally close Bibles. We investigate whether time proximity of Bible editions can help to map historical word variants to modern writing using only a simple character-distance measure. To this end, the following research questions that are tested on the **B**ibles are formulated: **RQ B1.1)** *Does the use of temporally close Bibles improve the alignment of historical writing variants?*, **RQ B1.2)** *Whether and*

how does time proximity in historical texts (i.e., text that are published within short period) help to normalize old variants of text to modern spelling?, and **RQ B1.3)** *What are specific problems to align a historical Bible corpus?*

Methodology

First, two time-proximate Bibles each are word-aligned by allowing relationships represented in terms of operations as displayed in Tab. 6.2, which is a subset of the overall operation set used in this thesis. This alignment is especially focussed on the explicit type of relationship (e.g., morphological modification, synonym replacement, etc.). Again, the texts are lemmatize using MorphAdorner by Burns (2013) to make sure to identify variants that the state-of-the-art can handle. Next, a simple character distance-based measure is defined that then is applied as the operation *editdist* (see Tab. 6.2) on top of the associations identified by lemmatizing performed using MorphAdorner. Last, ten verse pairs for each alignment (70 in total) are manually evaluated to give an overview of challenges that come with aligning historical text. Finally, statistical alignment is applied as an preprocessing step and the results are evaluated again.

Text data used

We use a subset of historical English Bible translations as described in Sec. 3.1.2. In the overall Bibleset we have twelve full English Bibles available from three different resources. However, only those Bibkes are selected that we think are suitable for the task. Hence, literal Bible translations such as Young's literal translation, Smith's literal translation and the English Septuagint by Brenton, are excluded, because these Bible editions have a very diverse vocabulary. The Darby Bible (1890) is also excluded, because a majority of its text was translated from other languages (c.f. Marlowe, 1867–1884). Table 6.1 lists the Bible subset used in this section next to the year of publication.

The text of the upper three Bibles (MATT, GREAT and GEN) is written in Early Modern English. This means that words appear different than today (e.g., "daye", "deuyde" instead of "day", "divide") or they are in old spelling (e.g., "heauen" instead of "heaven"). MorphAdorner (Burns, 2013) is able to cover such variants only when they follow certain rules. For example "catell" (GREAT) is correctly normalized to "cattle", "kynde" (GREAT) to "kind", and "likenes" (MATT) to "likeness". But "lycknesse" (MATT) and "licknesse" (GREAT) remain untouched. The lower five Bibles (RHE, DRC, KJV, WBT, ERV) do not contain a lot of historical writing. They contain a couple of words holding the typical archaic ending "eth", e.g., "creepeth", "yieldeth", etc.

Bible	date
Matthew Bible (MATT) (mys)	1537
Great Bible (GREAT) (mys)	1539
Geneva Bible (GEN) (mys)	1560
Douay-Rheims Catholic Bible (RHE) (bst)	1582-1609
Douay-Rheims, Challoner Revision (DRC) (mys)	1749-1752
King James (KJV) (ptp)	1611-1769
The Webster Bible (WBT) (bst)	1833
English Revised Version (ERV) (mys)	1881-1894

Table 6.1: Overview of used Bibles

no.	description	operation
1	perfect match	NOP(word1,word2)
2	lower-casing matches	lower(word1,word2)
3	lemmatizing matches	lem(word1,word2)
4	short levenshtein matches	editdist(word1,word2)
5	words are synonyms	syn(word1,word2)
6	word1 is hypernym of word2	hyper(word1,word2)
7	word1 is hyponym of word2	hypo(word1,word2)
8	fallback	-

Table 6.2: Transformation operations used for improving alignment accuracy. The lower part is shown for reasons of completion

6.1.4 Pairwise Bible alignment

We first align words of each verse in two Bibles following—as always—the order of operations in Tab. 6.2: *NOP, lower* (i.e., case-folding) and *lem*, as well as the newly introduced *editdist*. This allows to align words with an edit distance (Levenshtein, 1965) of 2/7 and it requires a minimum length of six characters for matching word candidates. This value was determined heuristically and showed to work best for our purpose. Only those resulting couples related by the operations *lem* and *editdist* are considered to measure variants in the parallel Bibles. Thereby, *lem* represents variants and modification that are already covered by MorphAdorner, while *editdist* represents newly found writing variants.[1] We use the word position stored with each operation to transitively link associated words across all Bibles together.

[1]Please see Sec. 6.2 for *fallback* operation counts in the resulting approach

Results - RQ B1.1

Now, the results of the first part of the work are presented. Recall that RQ B1.1 concerns the improvement of the alignment of historical writing variants in the Bible. In Tab. 6.3, the matching alignments that are already enabled by lemmatizing the words using MorphAdorner are displayed under "known lemmas". Table 6.3 distinguishes word types in the summed up operations from both, the source Bible and the target Bible. Further, tokens are considered—these are the same for the source and target Bible, because they simply represent the number of operations. Under "newly found edits" word types from the source and the target Bible, and tokens are listed. These numbers are determined based on the words that are aligned by allowing the introduced edit distance. Because of its strictness, this character distance measure works especially well for mapping proper names. About half as many types can be aligned with this measure compared to the types that can be aligned after lemmatization with MorphAdorner. Alignment between RHE and DRC, and KJV and WBT is particular similar (almost no differences between verses). This is, because in both cases the target Bible is a direct revision of its predecessor.

		known lemmas (*lem*)			**newly found edits (*editdist*)**		
source Bible	**target Bible**	**src types**	**target types**	**tokens**	**src types**	**target types**	**tokens**
MATT	GREAT	8,595	7,939	110,779	4,683	4,508	9,795
GREAT	GEN	7,531	6,105	147,671	3,178	2,753	9,359
GEN	RHE	5,300	4,534	115,027	1,471	1,424	6,296
RHE	DRC	392	406	777	349	359	1,212
DRC	KJV	2,713	2,747	24,206	1,235	1,199	4,316
KJV	WBT	706	717	7,242	594	592	2,233
WBT	ERV	1,734	1,816	11,908	974	958	2,772
sum		16,311	15,094	417,610	10,587	9,915	35,983
MATT	ERV	8,137	5,317	181,451	2,682	2,160	8,561

Table 6.3: Results of types and tokens identified between **source** and **target** Bibles each during alignment for the operations "lem" and "lev"

Comparing the overall alignments with the identified types and tokens[2] between MATT (the oldest Bible) and ERV (the most recent Bible), about four times as many types can be aligned with the *editdist* operation and about twice to three times as many word types with the *lem* operation (see last row of Tab. 6.3). Furthermore, the fact that much fewer types can be aligned between MATT and ERV indicates that aligning those hinders the alignment of rich-vocabulary texts. This shows that, indeed, more matches can be found using the advantages of temporally close Bibles. In general, coming back to RQ B1.1 we can learn that diachronic corpora, especially the herein used Bible corpus serves as a suitable dataset to align historical writing variants that do not exceed a certain edit distance.

[2]Types are collected as union set, i.e., ignoring duplicates.

For reasons of overview and comparison, a simple modification rate between two Bibles each is displayed in Tab. 6.4. This rate is determined by the number of all operations (no. 2 to 8 of Tab. 6.2) divided by the number of all operations displayed in Tab. 6.2. Table 6.4 shows that indeed most modification happens between MATT and GREAT. However, between GEN and GREAT we find slightly less of them than between GEN and MATT (basically skipping one Bible in the initial aliment chain). This correlates with the fact that Bibles that have a longer distance between their publications dates, can vary in terms of writing and grammar. In general the plot shows that more of the operations are found when older Bibles are compared (into any direction) rather than younger Bibles.

	KJV	GEN	RHE	GREAT	ERV	WBT	MATT
DRC	30.0	46.0	1.0	58.0	31.0	30.0	59.0
KJV	-	35.0	33.0	54.0	9.0	5.0	54.0
GEN	-	-	47.0	45.0	38.0	37.0	47.0
RHE	-		-	60.0	32.0	30.0	60.0
GREAT	-	-	-	-	55.0	55.0	28.0
ERV	-	-	-	-	-	13.0	56.0
WBT	-	-	-	-	-	-	56.0

Table 6.4: Modification rate based on non-*NOP*-operations

Results - RQ B1.2

A product of the time proximate alignment is a dictionary with 5,803 entries that contains types of the aligned words where the key entry is chosen to be the first appearance of a word that finished an alignment chain, i.e., the word from the youngest Bible. The other variants that are stored next to the key entry are all other types of words that appear in one or more alignment chains that result in the same finishing word (that one from the youngest Bible). This means that a dictionary entry set is build from more than one such alignment chains. An example of one single such alignment chain with words according to their Bibles is shown in Tab. 6.5. This dictionary was generated only based on verses that appear in every Bible, i.e., at least seven alignments per chain must exist. Because POS is not distinguished in this process, this dictionary is not aware of mixed POS information in one dictionary entry. Here are two examples:

- offering
  - offreth (.5, fail), offeryng (.1, pass), offring (.1, pass), offereth (.4, fail), offeringe (.1, pass), offer (.2, pass), offered (.27, pass), offred (.4, fail), offerynge (.2, pass), offrynges (.5, fail), offryng (.27, pass), offerings (.1, pass), offrynge (.4, fail)
- require
  - requier (.28, pass), requyre (.1, pass), requyreth (.5, fail), requireth (.25 pass), requere (.1, pass)

MATT	GREAT	GEN	RHE	DRC	KJV	WBT	ERV
requyre	requyre	require	require	require	require	require	require

Table 6.5: Example of one alignment chain over all eight Bible versions (neighboring words fulfill the 2/7 threshold)

The reader further finds a corresponding digit next to each of the word variants of a given dictionary entry.[3] These are thresholds that denote the distance between the respective word and the dictionary entry. Saying so, the *pass/fail* label next to that threshold denotes whether or not the 2/7 (including) mark had allowed or disallowed the alignment with the latest (youngest) writing form, hence, not making use of the temporally closeness of other available Bibles. These randomly chosen examples already show (40% and 20% fail) how many words had remained un-aligned, had the approach been ignored. This is an important result that shows the usefulness of diachronic corpora. Coming back to RQ B1.2, we can learn that these corpora and the here presented technique are simple mechanisms that can support normalization in the field essentially and open possibilities for many tasks in the digital humanities. Especially since normalization of historical writing variants is an essential step in almost all of the text-based research sub-fields and tasks.

Results of the error classification - RQ B1.3

Next, 70 verses are manually evaluated (ten verses from each Bible alignment pair). Table 6.6 shows how precisely the proposed edit distance works, how well its recall is, and how often lemmatizing enables a correct alignment. It also lists which other operations are identified, and it shows a first classification of errors found during the evaluation of alignments.

[3] As shown in the example, it cannot be ensured that the leading entry in the dictionary is actually a lemma. Still, a lot of variants are found and stored together in one set, and the key word indeed is a modern word form coming form the ERV.

Bible		lem alignments		editdist alignments			other operations			error types		
source	target	correct	wrong	true pos	false pos	false neg	syn	hyper	hypo	WN	PP	AUX
MATT	GREAT	32	0	2	0	3	2	1	0	3	2	0
GREAT	GEN	56	1	0	0	4	2	2	0	1	2	2
GEN	RHE	33	0	1	0	0	9	0	3	0	0	2
RHE	DRC	2	0	0	0	0	0	0	0	0	0	0
DRC	KJV	5	0	0	0	0	6	2	0	1	0	2
KJV	WBT	1	0	0	0	0	0	0	0	0	0	0
WBT	ERV	7	0	1	0	0	1	1	0	0	0	0

Table 6.6: Detailed list of error classes, manually evaluated between the alignment

In general, we can see that alignment by lemmatization works well with one exception of a false positive. Alignment by *editdist* has a high precision, but due to the strict conditions, a comparably bad recall (see false negatives of 3 and 4 in the first two rows of Tab. 6.6).

In Tab. 6.6 further, three error classes are distinguished: i) WN (word net) errors, ii) PP (preprocessing) errors such as wrongly tokenized words, and iii) AUX (auxiliary) errors. The first class represents the case of two words that can not be aligned with each other, simply because the synset database used does not store these words in the respective relation, or does not contain all of the words. The latter is the most frequent error. It appears when two auxiliary verbs are aligned, because their lemmas are identical. In many cases, however, these associations represent false couples. Examples of each error class are listed in Tab. 6.7. Relating to RQ B1.3, we can learn from this that the alignment—even though improved—is still challenging. An implication of this evaluation is an extra step to reduce the error during the alignment (note that the co-hyponym relation was ignored up until now to avoid even more wrong alignments). Consequently, the next section reports on the alignment accuracy of another experiment in which a statistical alignment is inserted at the end of the preprocessing step.

source	swalowe	my	Selah	for	faythfulnes
target	eate	me	Sela.	forth	treuth
error class	WN error	recall error	PP error	recall error	WN error

source	shall	wold	eate	vp	shall
target	will	would	swallowe	-	wil
error class	AUX error	recall error	WN error	-	AUX error

Table 6.7: Error class examples. In the example above, it appeared that the algorithm aligned "wold" and "will", which is wrong, and further could not align "shall/will" and "shall/wil"

6.1.5 Statistical alignment for preprocessing as an implication

Now, the Bibles are prealigned on the token level. To this end, Berkeley Word Aligner (DeNero & Klein, 2007) is used. The Berkeley Aligner is a statistical, unsupervised word aligner that was originally designed for machine translation. It combines two asymmetric alignment models based on HMM that are trained jointly to maximize their agreement in a combined symmetric alignment model. This mechanism especially makes the prioritized order of applying an operation as a relation between to words obsolete.

Bible		**lem alignments**		**editdist alignments**			**other operations**				**error types**		
source	target	correct	wrong	true pos	false pos	false neg	syn	hyper	hypo	co-hypo	WN	PP	AUX
MATT	GREAT	31	0	2	0	2	2	0	0	4	0	2	0
GREAT	GEN	55	0	0	0	3	2	0	0	2	0	2	0
GEN	RHE	30	0	1	0	0	8	0	2	2	0	0	0
RHE	DRC	2	0	0	0	0	0	0	0	0	0	0	0
DRC	KJV	4	0	0	0	0	6	2	0	2	0	0	0
KJV	WBT	1	0	0	0	0	0	0	0	0	0	0	0
WBT	ERV	4	0	1	0	0	0	0	0	0	0	0	0

Table 6.8: Detailed list of error classes, manually evaluated between the alignment with statistical prealignment

Table 6.8 shows—compared to Tab. 6.6—that the alignment errors are reduced drastically (see fewer "false neg" in the column *editdist alignments* and no wrongly aligned words in *lem alignments* anymore). In fact, only preprocessing errors remain. Minor differences in the number of relations (see column *other operations*) are displayed. This is attributed to the following reasons.

First, co-hyponyms are enabled that were disabled in the former experiment to reduce false positives in the comparably simple alignments approach. This enabling allows words that are placed on a similar position within two aligned sentences to be rather related as a co-hyponym than via the *editdist* operation. E.g., "my-me" is now aligned via the co-hyponym relation whereas it was a false negative alignment of *editdist* before (due to the minimum word length). It also compensates the WN error from the former experiment. However, lemma, hyponym, and hypernym relations are slightly decreased now. This is a problem of using a statistical prealignment. Word couples such as "13:13 syn(performeth,done);" could not be aligned because their statistical probabilities differ too much from each other. Depending on a sentence's available alignment candidates (i.e., if the words among two sentences remain the same to a high degree) a word couple such as "12:9 lem(him,he);" is aligned or not. In the sample, both happens once. Further, word couples with distant positions in the sentences as "8:11 lem(he,him);", "0:3 lem(Exalt,exalted);" are likewise not aligned. However, this also contributes to an accuracy increase of the local alignment. Specifically, in the former alignment experiment, often function words are aligned with each

other, even when they have sentence positions highly distant from each other. This often causes false positives, but can be prevented by statistical prealignment. Note that in the first experiment an alignment was considered correct if the words can be considered relatives of each other, even though an alignment partner was not necessarily the correct one if multiple candidates existed.

In summary, using the Berkeley Aligner as a preprocessing step does not yield many disadvantages, but assures precision with the disadvantage of some words not being aligned anymore. Hence, for the following experiments, only those words that are not aligned by Berkeley Word Aligner are fed into the earlier proposed routine to address the recall problem, and ensure accuracy at the same time. For the current set-up, that did not compensate all lost alignments, but returned two missing lemma alignments on top of Tab. 6.8.

6.1.6 Conclusion

In this section, experiments to optimizing the alignment of historical writing variants were presented and discussed. Such alignments are relevant for analyzing the characteristics of modification in text reuse. The experiments showed that not only alignment was improved by adding additional modifications operations and preprocessing steps, the proposed alignment furthermore paves the way for normalization techniques that make use of diachronic data. An additional outcome of the investigation is to combine statistical alignment as a preprocessing step, and a postprocessing step to associate words that were not aligned by the statistical alignment. The *editdist* operation furthermore is added from now on to the operation set furthermore. For future experiments, the use of derivation dictionaries and categorial databases also helps to align words with different appearances across the corpus, such as nouns and verbs of the same family that have a different POS, hence a *deriv* operation is added to the overall operation set as well.

6.2 Empirical analysis of reuse modification in German and English Bible translations

This section gives an overview of the raw figures that represent the modification among paraphrastic monolingual text reuse that we analyze—independently—in one corpus of parallel English Bible translations, and in one corpus of parallel German Bible translations. First, modification is measured—by means of the operations introduced earlier—in the English corpus, afterwards in the German corpus. For each corpus, we first look at the operations and their ratios, afterwards, we show how POS can change among the Bibles, and discuss reasons that cause these changes.

6.2.1 Research questions asked

In this section the following research questions shall be answered (see also Sec. 1.3.2). As mentioned above, the experiments designed to answer theses questions make use of English and German parallel Bible corpora.

- **RQ B2.1** *How are the different types of modification distributed in paraphrastic text reuse and how does the use of different lexical resources affect theses distributions?* This questions is addressed by running the introduced overall methodology (the operation-based alignment) on the data.
- **RQ B2.2** *What does the number of POS changes tell when measured in the parallel Bible corpora?* The main approach applied here, is to measure changes between the POS tags of the aligned tokens form the former step.

The following sections answer both questions—first, by applying the techniques to the English Bible corpus (Sec. 6.2.3 and Sec. 6.2.4), afterwards, by applying the techniques to the German Bible corpus (Sec. 6.2.5 and Sec. 6.2.6).

6.2.2 Part-of-speech tagset selection and unification

MorphAdorner's POS classes are very granular and sum up to over 230. To reduce the length of the tagset and to find an easier generalization between the English and the German POS tagset, MorphAdorner's POS classes are first grouped into twelve coarse-grained classes. Because MorphAdorner's POS taxonomy follows regular compound rules, it is not too difficult to extract the main POS tag from a POS tag string of a length of up to seven characters. The resulting general tags match with both, the tagging system of MorphAdorner (NUPOS)[4] and TreeTagger's German models (STTS)[5]. As a further result, this tagset matches a subset of the tagset from the Perseus Digital Library already used in Ch. 5. We use the same abbreviations to be consistent with them. These are—for English as well as for German:

- the open class POS: noun (n), verb (n), adjective (a) and adverb (d)
- the closed classes: preposition (r), pronoun (p), determiner/article (l), conjunction (c), as well as:
- exclamation (e), cardinal/numeral (m), particle (g), and foreign material (F).

[4] Available under: `http://panini.northwestern.edu/mmueller/nupos.pdf`

[5] `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html`

6.2.3 Empirical analysis of operations in English Bibles (RQ B2.1)

Running the alignment strategy first by considering the lexical synset database BabelNet only, second, considering ConceptNet alone, and third, considering BabelNet and ConceptNet at the same time, returns the operations as listed in Tab. 6.9.

operation	BabelNet	ConceptNet	BabelNet & ConceptNet
NOP	25,298,690	25,298,689	25,298,690
lower	844,971	844,971	844,971
norm	2,949,020	2,949,020	2,949,020
lem	1,876,553	1,876,553	1,876,553
deriv	84,501	75,416	84,473
editdist	255,372	256,586	255,360
syn	697,234	732,021	1,090,107
hyper	219,846	0	173,025
hypo	218,726	0	167,551
co-hypo	337,217	39,600	325,956
fallback	6,125,443	6,826,359	5,845,033
total	**38,907,573**	**38,899,215**	**38,910,739**

Table 6.9: Overview of operations identified. Semantic relations considering BabelNet, ConceptNet and BabelNet plus ConceptNet for the alignment of the whole historical Bibles corpus consisting in eleven Bibles. *deriv*—containing POS change—and *editdist* are treated equally, hence they can be chosen randomly.

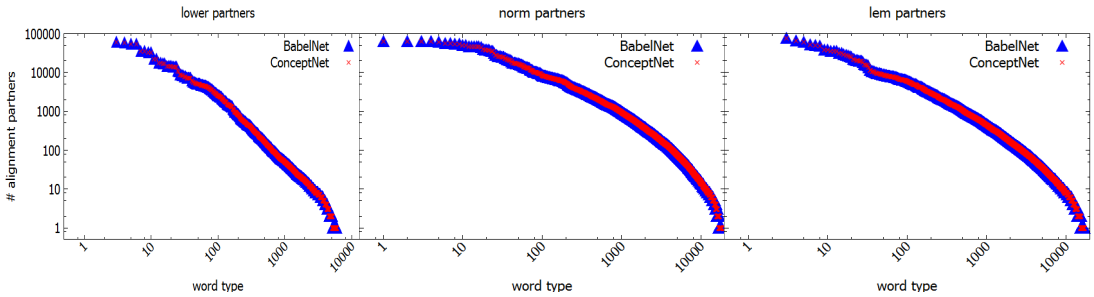
The figures presented are absolute. *NOPs* are presented for reasons of completion. The upper part—containing operations that represent morphological modification—does not differ when using BabelNet only, or both lexical databases as source to look-up semantic relations. Morphological modification—including case-folding, *deriv* and the *editdist*—cover about 15% of all operations. The figures for *deriv* and *editdist* differ slightly among the three columns. This is caused by the algorithm that chooses only one word couple if Berkeley Aligner used one word (identified by its sentence position) to align it to more than one word in the counter verse. That choice, however, can change depending on how many operations are covered by semantic relations, because they also influence which potential alignment (token) is preferred over, say, a fallback.[6]

The operations of semantic relations (see lower part of Tab. 6.9) differs strongly in all three columns. As already indicated in Sec. 4.5, BabelNet rather identifies relationships as co-hyponyms where ConceptNet supposes them to be synonyms. Hence, when both resources—or only ConceptNet respectively—is enabled, the synonym operation is applied before a

[6]This also affects the number of total operations.

potential co-hyponym relation can be applied. That is because the synonym operation represents more similarity and co-hyponymy represents the loosest type of similarity that is possible in the context of this work. This again, is due to the way a synset database is designed and shows the lack of standards for building them.

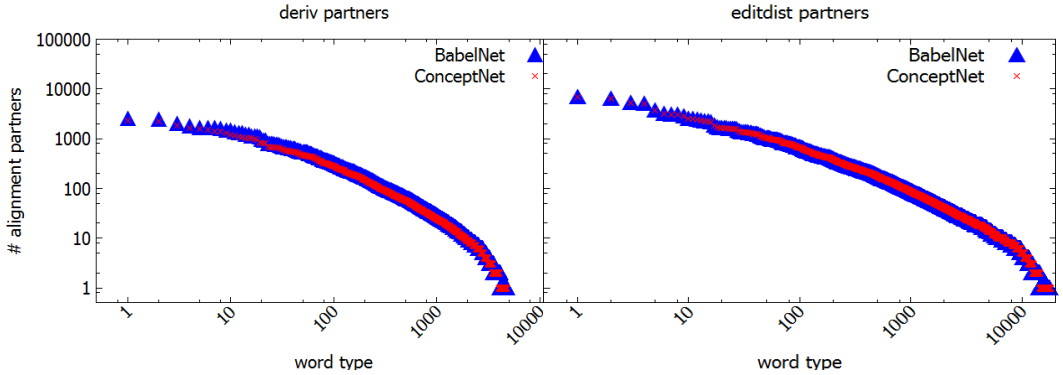
Further, for all columns, we again can see a substantial use of co-hyponyms. Hence, operations of semantic relation cover about 4% of the total when BabelNet is enabled. When ConceptNet is enabled as the only lexical data source, the appearances of the operations of semantic relations changes (see also Fig. 6.2a and Fig. 6.2d). That again is caused by many-to-many alignments and the decision, which word aligns as a potential couple first. Again, one word is only counted in once, even if Berkeley Aligner aligns a token multiple times. When ConceptNet is enabled alone, hypernym and hyponym relations that BabelNet supports are not supported.



(a) Distribution of alignment partners for the *lower* operation

(b) Distribution of alignment partners for the *norm* operation

(c) Distribution of alignment partners for the *lem* operation



(d) Distribution of alignment partners for the *deriv* operation

(e) Distribution of alignment partners for the *editdist* operation

Figure 6.1: Distribution of alignment partners for the operations that represent morphological modification

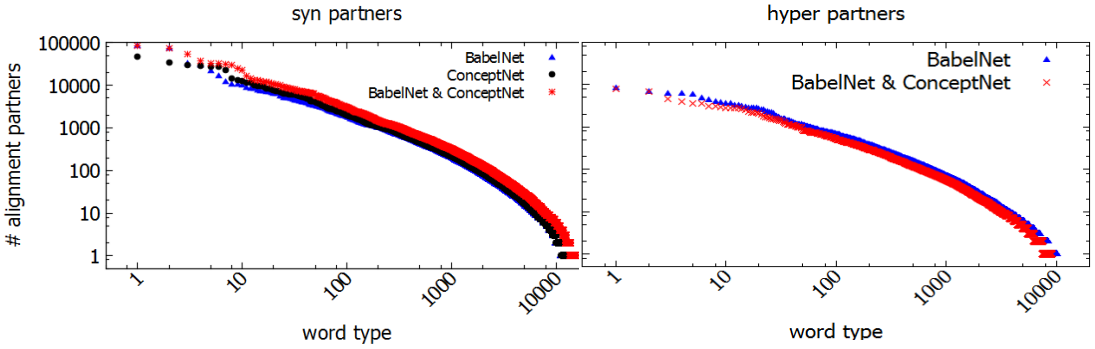
Figure 6.1 shows the distribution of replacements for operations that do not differ significantly when a different lexical database is used.[7] The figure shows the number of different alignment partners (word forms) for each operation (x-axis) and each word. The distribution of these numbers of alignment partners are presented as the y-value. The scales of the plots are logarithmic to avoid long tails. As the reader can see, every operation's plot follows a power law distribution. Note that even though the distributions follow a power law, they cannot necessarily be assumed to follow Zipf's law, because this only applies to the words of a natural language text. The alignment partners of the *deriv* and *editdist* operations are less frequent, hence, the according plots do not reach as high numbers as the plots of Fig. 6.1a to Fig. 6.1c. To avoid skew, and because Bibles are aligned only once, without a certain direction, these figures are counted for each operation into both directions.

Figure 6.2 is based on operations calculated when using BabelNet alone, ConceptNet alone as well as both at the same time as lexical databases. It shows that most hypernyms and hyponyms alignments are enabled when BabelNet is enabled alone. Synonym distributions are highest for ConceptNet alone, because ConceptNet prefers these relations above hyper and hyponyms. Co-hyponym distributions are higher when BabelNet is enabled. Hypernyms and hyponyms are not covered when ConceptNet is enabled alone. Looking at sameples it is obvious that the taxonomy of ConceptNet is much too detailed and too modern to operate on the Biblical vocabulary. For example, hyper and hyponym relations supported are accordionist-/musician and destroyer-/weapon. Compared to BabelNet ConceptNet does not enable a lot of co-hyponym relations. Again, this is because synonyms are aligned instead and ConceptNet's coverage is not much higher than BabelNet's even though, the enabling of both resources reduces about 5% of the fallbacks.

6.2.4 Empirical analysis of part-of-speech changes in English Bible translations (RQ B2.2)

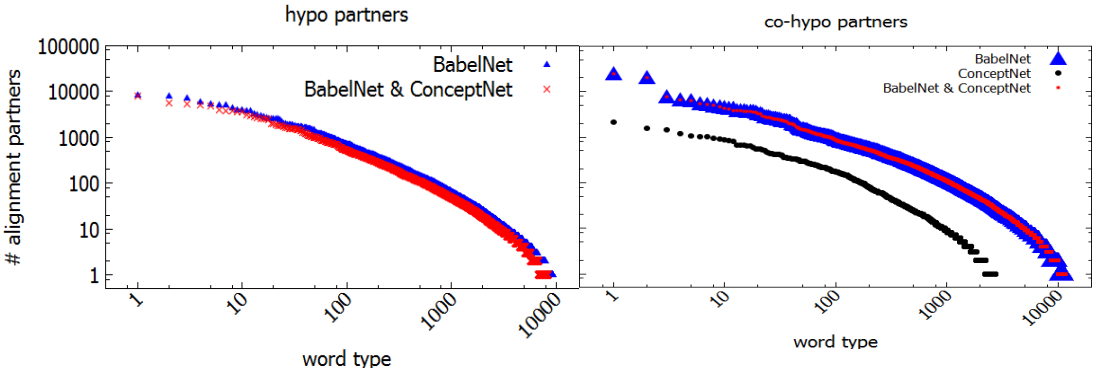
Modification is not only considered on the lexical level—for example by replacing words—it is also interesting to learn in detail how morphology changes. For this purpose, the POS tag of two aligned words are investigated. Hence, the POS-tagged version of two Bibles are linked with the word-aligned version in which two Bibles are compared/aligned. Based on this information in place, changes in the POS tag can be counted, as well as stable remaining POS, and unsupported drop-outs. The latter are mainly tokens that could not be assigned, because those are unfiltered punctuation marks or other tokens that are not interpreteable. This share is also very low for each Bible coupling (below 0.2%). Table 6.10 shows a detailed

[7] Following outliers are cut: Two at around 200,000 in the *lower* curve, and two around 200,000 in the *lem* curve.



(a) Distribution of alignment partners for the *syn* operation

(b) Distribution of alignment partners for the *hyper* operation



(c) Distribution of alignment partners for the *hypo* operation

(d) Distribution of alignment partners for the *co-hypo* operation

Figure 6.2: Distribution of alignment partners for the operations that represent lexical modification

overview of the staple, changing, and unsupported POS and their shares.[8] POS changes cover 9.4% of all aligned tokens with a minimum of .4% (between DRC and RHW—both revisions of each other) and a maximum of 15.6% (between MATT—the oldest Bible in the corpus and YLT—one of the literal translations) showing again the wide range of degrees in the alignments of the diverse types of Bible translations. To give the reader a better understanding of the actual (un-)change, note that one Bible has between 700K and 1 mio. tokens. Numbers are displayed absolutely and relatively.

[8] In total, 39,698,801 token pairs in 55 Bible pair alignments were calculated. The total POS alignment count differs from the total number of operations in Tab. 6.9, because, during alignment only one alignment operation is considered for a multi-word alignment.

Bible1	Bible2	changing POS	stable POS	drop-outs	total
DBY	DRC	65,214 (9.28%)	636,371 (90.54%)	1,279 (0.18%)	702,864
DBY	ERV	38,116 (5.07%)	713,052 (94.75%)	1,360 (0.18%)	752,528
DBY	GEN	61,763 (8.43%)	669,611 (91.4%)	1,229 (0.17%)	732,603
DBY	GREAT	81,080 (11.48%)	623,451 (88.28%)	1,689 (0.24%)	706,250
DBY	KJV	39,260 (5.24%)	708,510 (94.58%)	1,310 (0.17%)	749,080
DBY	MATT	83,555 (11.94%)	614,904 (87.85%)	1,509 (0.22%)	699,968
DBY	RHE	66,077 (9.39%)	636,558 (90.42%)	1,356 (0.19%)	703,991
DBY	SLT	69,715 (9.62%)	654,206 (90.24%)	1,052 (0.15%)	724,973
DBY	WBT	41,269 (5.51%)	705,992 (94.31%)	1,292 (0.17%)	748,553
DBY	YLT	67,544 (9.18%)	666,985 (90.64%)	1,336 (0.18%)	735,865
DRC	ERV	60,540 (8.47%)	653,848 (91.49%)	249 (0.03%)	714,637
DRC	GEN	68,468 (9.67%)	639,339 (90.29%)	290 (0.04%)	708,097
DRC	GREAT	81,897 (11.84%)	608,735 (88.03%)	883 (0.13%)	691,540
DRC	KJV	57,653 (8.07%)	656,375 (91.89%)	260 (0.04%)	714,288
DRC	MATT	83,769 (12.18%)	603,333 (87.72%)	684 (0.1%)	687,786
DRC	RHE	**2,879 (0.37%)**	772,452 (99.6%)	221 (0.03%)	775,552
DRC	SLT	83,763 (12.12%)	607,150 (87.84%)	311 (0.04%)	691,224
DRC	WBT	61,784 (8.65%)	651,820 (91.31%)	276 (0.04%)	713,880
DRC	YLT	88,208 (12.63%)	609,909 (87.32%)	342 (0.05%)	698,459
ERV	GEN	46,965 (6.29%)	699,823 (93.69%)	160 (0.02%)	746,948
ERV	GREAT	70,496 (9.78%)	649,705 (90.11%)	771 (0.11%)	720,999
ERV	KJV	17,779 (2.32%)	749,917 (97.67%)	124 (0.02%)	767,820
ERV	MATT	74,402 (10.42%)	639,248 (89.5%)	606 (0.08%)	714,256
ERV	RHE	61,405 (8.58%)	653,917 (91.38%)	299 (0.04%)	715,621
ERV	SLT	75,721 (10.4%)	652,150 (89.58%)	172 (0.02%)	728,043
ERV	WBT	28,499 (3.71%)	738,921 (96.27%)	144 (0.02%)	767,564
ERV	YLT	75,342 (10.19%)	663,923 (89.78%)	223 (0.03%)	739,488
GEN	GREAT	62,270 (8.58%)	662,948 (91.31%)	820 (0.11%)	726,072
GEN	KJV	37,590 (4.99%)	715,027 (94.98%)	173 (0.02%)	752,790
GEN	MATT	66,572 (9.27%)	651,023 (90.64%)	635 (0.09%)	718,230
GEN	RHE	69,706 (9.83%)	639,119 (90.12%)	336 (0.05%)	709,161
GEN	SLT	89,191 (12.45%)	626,884 (87.52%)	225 (0.03%)	716,300
GEN	WBT	46,432 (6.17%)	705,402 (93.8%)	192 (0.03%)	752,026
GEN	YLT	91,281 (12.61%)	632,557 (87.36%)	257 (0.04%)	724,095
GREAT	KJV	61,651 (8.49%)	664,056 (91.4%)	789 (0.11%)	726,523
GREAT	MATT	33,721 (4.5%)	714,250 (95.36%)	1,040 (0.14%)	749,036
GREAT	RHE	83,216 (12.02%)	607,953 (87.84%)	906 (0.13%)	692,097
GREAT	SLT	102,604 (14.88%)	585,955 (84.99%)	818 (0.12%)	689,406
GREAT	WBT	67,954 (9.37%)	656,678 (90.52%)	808 (0.11%)	725,466
GREAT	YLT	106,200 (15.22%)	590,571 (84.65%)	823 (0.12%)	697,624
KJV	MATT	67,066 (9.34%)	650,513 (90.58%)	605 (0.08%)	718,184
KJV	RHE	58,653 (8.2%)	656,270 (91.76%)	247 (0.03%)	715,170
KJV	SLT	75,716 (10.42%)	650,534 (89.55%)	180 (0.02%)	726,430
KJV	WBT	**13,270 (1.7%)**	768,917 (98.28%)	150 (0.02%)	782,337
KJV	YLT	76,149 (10.34%)	660,396 (89.63%)	231 (0.03%)	736,776
MATT	RHE	85,056 (12.35%)	603,117 (87.55%)	714 (0.1%)	688,887
MATT	SLT	104,721 (15.3%)	579,234 (84.61%)	631 (0.09%)	684,586
MATT	WBT	72,577 (10.12%)	643,785 (89.79%)	629 (0.09%)	716,991
MATT	YLT	**107,622 (15.57%)**	583,086 (84.34%)	648 (0.09%)	691,356
RHE	SLT	85,113 (12.29%)	607,027 (87.66%)	377 (0.05%)	692,517
RHE	WBT	62,658 (8.77%)	651,875 (91.19%)	317 (0.04%)	714,850
RHE	YLT	89,520 (12.79%)	609,886 (87.15%)	393 (0.06%)	699,799
SLT	WBT	75,898 (10.46%)	649,519 (89.51%)	199 (0.03%)	725,616
SLT	YLT	68,797 (9.32%)	668,745 (90.64%)	258 (0.03%)	737,800
WBT	YLT	73,820 (10.03%)	661,733 (89.93%)	246 (0.03%)	735,799

Table 6.10: Overview of changing and stable POS in the English Bible corpus

Table 6.10 sows that the majority of POS stays stable in all Bible couplings. Couples such as ERV-KJV, ERV-WBT, KJV-WBT, MATT-GREAT, as well as RHE-DRC show exceptionally high values in column "stable POS". This can be explained, because WBT and ERV are revisions of KJV, the DRC is a revision of the catholic RHE Bible, and MATT and GREAT are both followers of the Tyndale Bible. All these groups have a high ratio of material in common.

Concerning the column of changing POS in Tab. 6.10, it is still impressive how high the degree of modification happens to be between MATT and YLT. This is especially an important outcome to consider during the work with lexical resources such as synset databases. A change in the POS tag inevitably affects the recall of a synset database on a certain lexical domain, because relations such as synonyms etc. are only stored word class-wise. Endeavors to store different aspects of a word such as what a concept consists of or what an action entails (Speer & Havasi, 2012) do not necessarily solve this issue, because a thing still consist of things (nouns) and an action follows or requires another action (all verbs).

	YLT	SLT	GREAT	DBY	KJV	MATT	DRC	RHE	GEN	WBT
ERV	10.19	10.4	9.78	5.07	2.32	10.42	8.47	8.58	6.29	3.71
YLT	-	9.32	15.22	9.18	10.34	15.57	12.63	12.79	12.61	10.03
SLT	-	-	14.88	9.62	10.42	15.3	12.12	12.29	12.45	10.46
GREAT	-	-	-	11.48	8.49	4.5	11.84	12.02	8.58	9.37
DBY	-	-	-	-	5.24	11.94	9.28	9.39	8.43	5.51
KJV	-	-	-	-	-	9.34	8.07	8.2	4.99	1.7
MATT	-	-	-	-	-	-	12.18	12.35	9.27	10.12
DRC	-	-	-	-	-	-	-	0.37	9.67	8.65
RHE	-	-	-	-	-	-	-	-	9.83	8.77
GEN	-	-	-	-	-	-	-	-	-	6.17

low frequent — high frequent

Table 6.11: Frequencies of changing POS in the English Bible corpus in %

Table 6.11 displays the changing POS between two Bibles each. The figures are displayed percentage-wise. One obvious cell, the brightest cell of the heatmap, is KJV-WBT. This is, because WBT is a direct successor of the KJV in the history of its revisions. Similar effects cause the bright color of cells ERV-KJV and ERV-WBT (see also Tab. 6.10), which all belong to the same revision path. Further, cell GREAT-MATT is only slightly colored. Again, both are revisions from another, the GREAT Bible is a revision of Matthew's Bible.

On the other hand, especially the two rows next to YLT and SLT are deeply colored, which means that POS changes are especially high when Bibles are coupled with these two

literal translations. Remember that YLT and SLT are English Bible translations that were meant to be translated literally from the primary languages (Hebrew, Greek, Latin). Hence, their choice of vocabulary often leads to changes in POS such as nominalization (more details in short)[9]. Their choice of grammar and syntax can force the aligner to couple obviously different word classes (details in short)[10], which further distinguishes these translations from the more modern, standard English translations.

	v	a	d	r	p	l	c	e	m	g	F
n	487,324	343,965	108,482	33,269	69,857	25,823	17,553	12,492	12,074	5,167	19,438
v	-	204,329	180,012	86,439	142,395	40,859	91,347	23,108	1,913	50,156	3,701
a	-	-	69,443	11,223	4,259	27,146	4,304	1,562	9,776	1,741	1,669
d	-	-	-	142,082	52,870	59,872	241,475	8,469	14,522	36,433	313
r	-	-	-	-	12,993	56,022	258,502	2,080	540	77,389	110
p	-	-	-	-	-	163,893	58,794	9,426	14,259	39,431	284
l	-	-	-	-	-	-	75,206	3,070	16,874	48,085	166
c	-	-	-	-	-	-	-	27,280	800	154,398	88
e	-	-	-	-	-	-	-	-	103	19,952	965
m	-	-	-	-	-	-	-	-	-	230	380
g	-	-	-	-	-	-	-	-	-	-	5

low frequent — high frequent

Table 6.12: Numbers of POS changes in the English Bible corpus according to POS class

Finally, Tab. 6.12 shows the POS changes according to the POS classes. For this purpose, every POS class is listed in form of a matrix in the x/y dimension. Because the direction is not considered, and the number of POS changes shall be treated symmetrical, we fold together modification from one Bible a to one Bible b and vice-versa. Very high frequent changes from nouns to verbs and vice-versa (see columns n-v and n-v) and among the open class in general. This happens for example when “to shine” (in DRC) becomes “bring lights” (in DBY).

Further, highly frequent changes are shown from (c) to (r), and from (c) to (d). A (c)-(r) change is often accompanied by an alignment error when two quite literal Bibles are aligned such as “I will not take from a thread even to(r) a [...]” (KJV) and “I will not take a thread nor(c) a [...]” (ERV). In two Bibles that are not revisions from each other, and hence,

[9] Remember, these also affect the lexical operations in the lower part of Tab. 4.1.
[10] These also affect the morphological changes in the upper part of Tab. 4.1.

more paraphrastic to each other, a change from (c) to (r) rather indicates that alignment is challenged as the following example shows: "But as(c) touchynge the tre of knowledge [...]" (GREAT) and "but of(r) the tree of the knowledge [...]" (ERV). A (c)-(d) change happens for example when "soever(d)" and "that(c)" are aligned in the following texts: "in what day soever(d) thou shalt eat" (DRC) aligned with "in the day that(c) thou eatest" (DBY). The two words normally are considered to be incorrectly aligned. However, for a statistical aligner, both words serve the same purpose, being frequency, positioning in the sentence, and also a sort of binding of two clauses. In most cases, however, it happens that the clauses such as "And(c) God said" (DBY, ERV, etc.) are aligned with the clause "God also(d) said" (DRC). We can learn from the aligner's choice to align "and" and "also", and the fact that the changes of these POS are less intuitive—than for example a nominalization from a verb to a noun—that these alignments indeed happen when paraphrastic reuse is analyzed. (More insights on how inflection can be a marker for un-similar text will follow in Sec. 7.1.)

To further investigate changes, we also calculate the significance values from the chi-squared test. This is important to find out if the appearance of a given POS change (e.g., n-v) is significant according to the overall probability of verbs and nouns in the alignment couples in general. Table 6.13 shows these chi-squared values. Considering a degree of freedom of 1.0 and a significance degree (p-value) of 5%, all values above 3.84 indicate that this POS change is significant, but it is not when the chi value is below 3.84.

Next to changes that are already discussed above, especially changes containing a nu(m)eral, a particle (g) and foreign Material (F) are significant. For example, numerals often happen to become nouns when MorphAdorner does not make nominalization explicit (and vice-versa) such as "one" in "[...] the one(m) to his place to present him the cup [...]" (DCR) and "cup-bearers" in "[...] the cup-bearers(n) to [...]" (DBY). (m)-(v) changes on the other hand often indicate strong paraphrasing from passive voice to active voice—and with it a questionable choice of the aligner, even though no better choice exists. An example is the alignment of "was" in "And it was(v) told [...]" (DRC) with "one" in "And one(m) told [...]" (DBY). Significant changes in the newer ERV and the older MATT relate to numeral-preposition changes. For example, "one" in "[...] coupled five curtains one(m) to another" (ERV) and "by" "[...] coupled .v. curtaynes by(r) them selues" (MATT). These alignment errors, however, mainly affect short sentences. They are significant compared to the little appearance of the POS tags (numeral and preposition) in the corpus, but seldom in sheer appearance (540 out of 40 mio. POS changes). A particle is aligned with a noun for example between "thing" in "[...] the thing(n) entrusted guard thou [...]" (YLT) and "which" in "[...] that which(g) is committed [...]". Finally, changes such as "Pondre" in "Pondre(F) the path" and "straight" in "Make straight(d) the path" is caused by weaknesses of both, the normalizer and the aligner. First, MoprhAdorner does not recognize "pondre" as a verb,

second, straight is aligned with "ponder", because "make"—being a much more frequent word—does not suffice well as a candidate.

	v	a	d	r	p	l	c	e	m	g	F
n	1.35	2.44	2.41	2.73	2.57	2.77	3.43	2.83	39.93	7.59	2.8
v	-	1.82	2.6	4.03	2.0	5.78	3.0	2.4	45.83	2.3	2.48
a	-	-	4.19	4.76	5.79	4.59	3.52	30.54	4.78	4.86	4.86
d	-	-	-	3.59	5.13	3.23	2.31	3.55	3.51	3.37	122.9
r	-	-	-	-	4.73	4.3	2.24	4.85	46.76	4.11	4.87
p	-	-	-	-	-	3.97	3.18	5.71	38.84	5.3	5.85
l	-	-	-	-	-	-	3.09	6.39	6.15	5.67	6.44
c	-	-	-	-	-	-	-	24.24	3.54	5.01	123.93
e	-	-	-	-	-	-	-	-	30.98	7.25	119.98
m	-	-	-	-	-	-	-	-	-	46.97	46.87
g		-	-	-	-	-	-	-	-	-	7.71

not significant significant

Table 6.13: Chi-squared numbers of POS changes in the English Bible corpus according to POS class. Statistical significance of a POS change is measured towards the overall probability of the given POS in the overall alignments.

6.2.5 Empirical analysis of operations in German Bibles (RQ B2.1)

Now, we apply the alignment strategy to the German Bible corpus. Running it first by considering the lexical synset database BabelNet only, second, considering ConceptNet alone, and third, considering BabelNet and ConceptNet at the same time, returns the operations listed in Tab. 6.14.

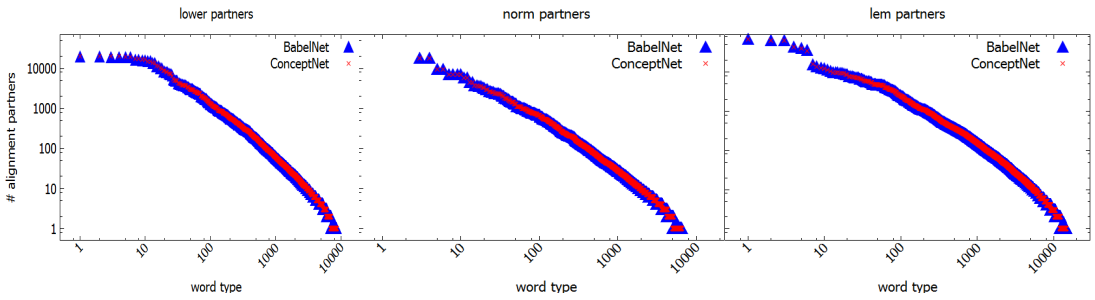
operation	BabelNet	ConceptNet	BabelNet & ConceptNet
NOP	6,569,336	6,638,065	6,569,336
lower	411,463	424,960	411,463
norm	328,495	330,614	328,495
lem	737,307	742,295	737,307
deriv	17,132	17,394	17,124
editdist	213,945	216,805	213,937
syn	136,017	365,835	425,632
hyper	32,534	0	11,251
hypo	13,610	0	6,540
co-hypo	36,674	8,736	39,802
fallback	4,006,286	3,801,626	3,738,513
total	**12,502,799**	**12,546,330**	**12,499,400**

Table 6.14: Operations identified during the alignment of two Bibles each from seven German Bible translations. *deriv*—containing POS change—and *editdist* are treated equally, hence they can be chosen randomly.

Again, figures are presented in absolute numbers. *NOP*s are presented for reasons of completion. Similar to the English corpus, operations that represent morphological modification cover about 14%—including derivation and *editdist*. However, in the German corpus the *lem* operation is much more often present compared to *norm*. This has two reasons. The first reason concerns the age of the English corpus. In average, it is much older, and, hence, it uses more writing variants (covered by *norm*) in the older Bibles. Second, German is a language with a richer inflection in both, historical and modern texts than a compared text in English.

Next, derivations are identified by a certain extend, even though not a big one. Although, the German derivation dictionary contains many more family entries than the English one (235,000 vs. 62,000), only 17,000 of them are non-singleton families, which means, only 17,000 contain more than one entry. This reduces the hit rate drastically. Furthermore, the accuracy of the normalizer (Norma) used is lower compared to the one used for English (MorphAdorner). This results in a procrastination of the word couples that can not be identified in the derivation dictionary (this certainly also applies for *lem* and the operations of semantic relations).

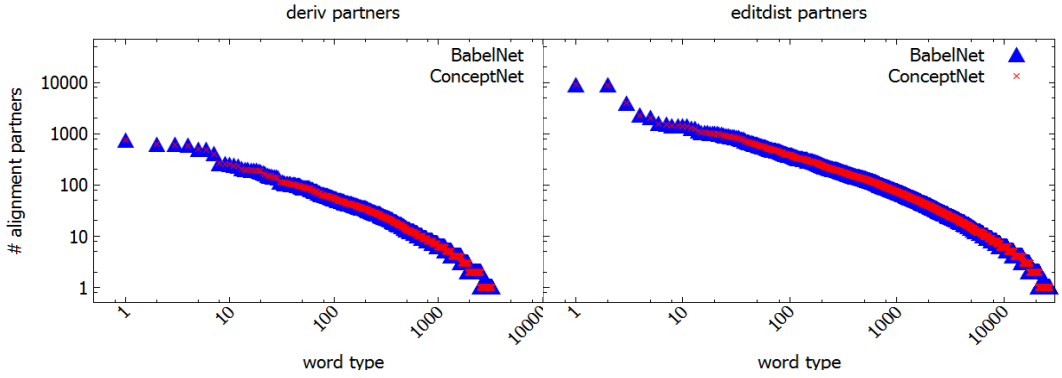
Operations of semantic relations in the German corpus cover roughly 2% when BabelNet is enabled, almost 3% when only ConceptNet is enabled, and, about 8% when ConceptNet is enabled on top of BabelNet. Again, ConceptNet rather assign synonym relationships than co-hyponyms, which is a matter of design.



(a) Distribution of alignment partners for the *lower* operation

(b) Distribution of alignment partners for the *norm* operation

(c) Distribution of alignment partners for the *lem* operation



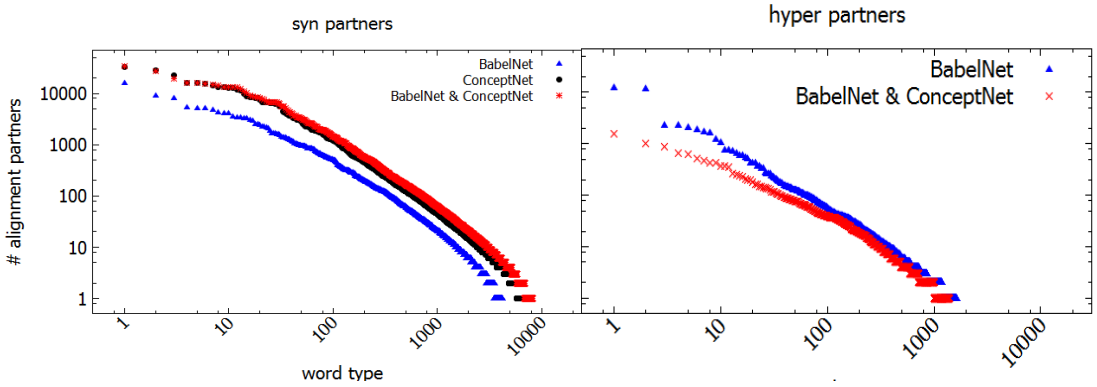
(d) Distribution of alignment partners for the *deriv* operation

(e) Distribution of alignment partners for the *editdist* operation

Figure 6.3: Distribution of alignment partners for the operations that represent morphological modification in the German Bible corpus

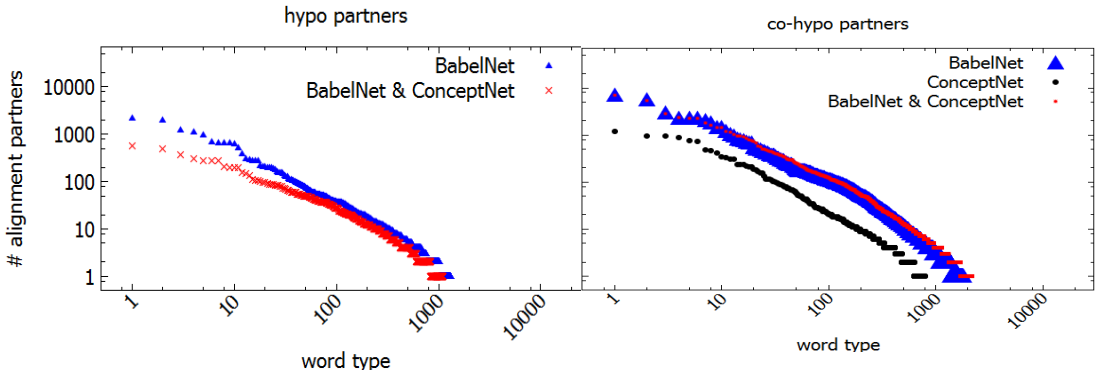
Figure 6.3 shows the distribution of alignment partners for each word type for the operations *lower*, *norm*, *lem*, *deriv* and *editdist*. Again, the plots show the numbers of different alignment partners (word forms) for a given word type (x-axis). The number of alignment partners are presented as the y-value. The scales of the plots, again, are logarithmic. The distributions of all operations follow the power law. Also, similar as in the English data, the plot of the synonym operation reaches a higher frequency on the y-axis in the "BabelNet & ConceptNet" plot. The alignment partners of the *deriv* and *editdist* operations are much less frequent than those of the former three operations. Especially eye-catching is the over-

averaged increase of the very first word types in the plot of Fig. 6.3c and Fig. 6.3e. These show an especially high raise in the beginning of the power law distribution and prove the concept that only a few tokens cover the main part of the data points.[11]



(a) Distribution of alignment partners for the *syn* operation

(b) Distribution of alignment partners for the *hyper* operation



(c) Distribution of alignment partners for the *hypo* operation

(d) Distribution of alignment partners for the *co-hypo* operation

Figure 6.4: Distribution of alignment partners for the operations that represent lexical modification in the German Bible corpus

Figure 6.4 presents the distribution of operation partners for the modeled operations based on the German Bibles, all processed in one run. Again, hypernym and hyponym relations are empty for ConceptNet alignments. Replacement figures of *syn* and *co-hypo* are most frequent when BabelNet and ConceptNet are enabled both at the same time as shown in Fig. 6.4. This is, again, owed to the preference of synonym relations over co-

[11] To avoid skew, and because Bibles are aligned only once without a certain direction, these figures are counted for each operation into both directions.

Bible1	Bible2	changing POS	stable POS	drop-outs	total
ELB1	ELB2	**1,784 (0.25%)**	711,104 (99.72%)	203 (0.03%)	713,091
ELB1	GB	100,014 (16.54%)	504,222 (83.38%)	492 (0.08%)	604,728
ELB1	LB1	167,820 (26.5%)	465,238 (73.45%)	332 (0.05%)	633,390
ELB1	LB2	82,459 (12.85%)	559,212 (87.12%)	220 (0.03%)	641,891
ELB1	NeÜ	110,221 (19.01%)	469,182 (80.93%)	313 (0.05%)	579,716
ELB1	TB	85,555 (13.35%)	554,952 (86.59%)	364 (0.06%)	640,871
ELB2	GB	100,283 (16.62%)	502,691 (83.3%)	509 (0.08%)	603,483
ELB2	LB1	167,945 (26.56%)	464,013 (73.38%)	355 (0.06%)	632,313
ELB2	LB2	82,019 (12.73%)	561,807 (87.23%)	248 (0.04%)	644,074
ELB2	NeÜ	110,677 (19.13%)	467,550 (80.81%)	344 (0.06%)	578,571
ELB2	TB	86,185 (13.47%)	553,291 (86.47%)	390 (0.06%)	639,866
GB	LB1	172,266 (28.92%)	422,830 (70.99%)	519 (0.09%)	595,615
GB	LB2	111,802 (19.15%)	471,674 (80.78%)	414 (0.07%)	583,890
GB	NeÜ	111,952 (20.02%)	446,828 (79.89%)	510 (0.09%)	559,290
GB	TB	92,700 (15.29%)	512,863 (84.62%)	534 (0.09%)	606,097
LB1	LB2	125,551 (18.72%)	544,778 (81.24%)	244 (0.04%)	670,573
LB1	NeÜ	175,717 (30.71%)	396,009 (69.22%)	364 (0.06%)	572,090
LB1	TB	**172,666 (27.75%)**	449,066 (72.18%)	405 (0.07%)	622,137
LB2	NeÜ	122,807 (21.9%)	437,608 (78.05%)	273 (0.05%)	560,688
LB2	TB	102,503 (16.77%)	508,331 (83.18%)	297 (0.05%)	611,131
NeÜ	TB	108,778 (18.76%)	470,686 (81.17%)	386 (0.07%)	579,850

Table 6.15: Overview of changing and stable POS in the German Bible corpus

hyponymy that comes with ConceptNet's design. On the other hand, ConceptNet does not support hypernym and hyponym relations. Hence, operation relations are increased slightly in favor of co-hyponymy pairings (from 37,000 to 40,000). ConceptNet supports co-hyponym relations of about one fourth of the number that BabelNet does.

6.2.6 Empirical analysis of part-of-speech changes in German Bible translations (RQ B2.2)

In the current experiment, POS changes as well as not changing POS are determined for the German Bible corpus.[12] The exact same coarse-grained POS classes as used in Sec. 6.2.4 are used again. Table. 6.15 gives an overview of changing, stable and not classifiable POS tags between two German Bible translations each including their percentages. POS changes for about 18.8% in average. With a minimum value of .3% (between the two Elberfelder versions, which are revisions of each other, and are published 50 years apart from each other), and a maximum change of 30.7% (between LB1—the oldest Bible—and NeÜ—the youngest Bible; between these two Bibles' publication dates almost 500 years passed). The

[12] Altogether, 12,873,355 token pairs are considered.

first example, again, shows that Bibles remain more similar to each other when they i) follow the same revision line, ii) their ages/publication years are not far apart from each other, and iii) both versions are relatively young, hence, they are both not strongly affected by historical writing variants. The second example shows the opposite. Bibles can strongly differ from each other, not only in terms of a different vocabulary, also tense, mood, person, or POS can change strongly. All that results in a different POS tag. Characteristics of the publication situation are then, for example, the following: First, Bibles do not share the same revision tradition, and are translated under different conditions with different intentions[13], second, the publication dates of the Bibles are very time distant from each other, and third, at least one of the two Bibles compared is published a couple of centuries earlier that the other. Hence, writing variants and a different syntax[14] in the older Bible, and spelling normalization in the younger Bibles affect the change of POS.

	ELB2	TB	LB2	GB	NeÜ	LB1
ELB1	0.25	13.35	12.85	16.54	19.01	26.5
ELB2	-	13.47	12.73	16.62	19.13	26.56
TB	-	-	16.77	15.29	18.76	27.75
LB2	-	-	-	19.15	21.9	18.72
GB	-	-	-	-	20.02	28.92
NeÜ	-	-	-	-	-	30.71

low frequent  high frequent

Table 6.16: Frequencies of changing POS in the German Bible corpus in %

Table 6.16 shows the numbers of changing POS tags between two Bibles each represented as a heat map. The reader can discover four regions of different degrees of modification. The lowest degree, again, is between the two Elberfelder versions. A slightly darker square is formed by the comparison of the two ELB, the young Luther, and the Text Bible. In fact, these four Bibles do not necessarily share the same revision tradition. However, they are all published around the same time (end of the 19th and beginning of the 20th century). Actually, at most 50 years passed between the publications of any two of them. As an opposite, it is obvious that when compared to the older Luther Bible, all Bible versions

[13] Consider also that different primary versions are consulted by the translators or that the primary text used for the Bible versions that lead a translation/revision tradition differ.

[14] A historical word ordering together with very exotic spelling variation can confuse even a specified tagger and lead to mixing up for example numerals and nouns.

show a high frequency of changes (last column). Again, this can be explained by the old age of the text of LB1 and the accompanying difference in word spelling and syntax order.

	v	a	d	r	p	l	c	e	m	g	F
n	424,847	243,012	63,300	31,119	94,361	26,242	60,332	459	9,720	15,187	292
v	-	178,788	65,018	30,144	125,363	10,716	30,934	288	2,192	25,719	273
a	-	-	36,409	15,375	50,033	9,196	7,764	210	11,068	8,117	57
d	-	-	-	23,305	62,479	8,310	147,290	245	1,214	26,907	16
r	-	-	-	-	52,442	71,529	56,935	4	197	50,238	21
p	-	-	-	-	-	160,960	78,073	79	2,423	23,881	79
l	-	-	-	-	-	-	26,170	1	786	4,938	1
c	-	-	-	-	-	-	-	23	22	16,493	4
e	-	-	-	-	-	-	-	-	1	38	-
m	-	-	-	-	-	-	-	-	-	62	-
g	-	-	-	-	-	-	-	-	-	-	3

low frequent high frequent

Table 6.17: Numbers of changes in the German Bible corpus according to POS class

Figure 6.17 shows changes according to the POS classes in the German Bible corpus. Again, changes among the open class (noun, verb, adjective and adverb) are often due to concepts such as nominalization. A p-n change can indicate error. In "Jehova Gott ließ aus dem Erdboden allerlei(p) Bäume wachsen" (ELB1) and "Allerlei(n) Bäume, lieblich zur Schau [...]" (GB), the upper-cased "Allerlei" is wrongly POS tagged as a noun. Another erroneous instance of p-n happens in: "von allem Lebendigen von allem Fleische zwei von jeglichem(p) sollst du [...]" (ELB1) and "Von allem Lebenden von jedem Fleischeswesen(n) sollst du [...]" (GB). However, this happens when one word is aligned to more than one word. Hence, another alignment partner of "Fleischeswesen" (n) is "Fleische" (n), which is the correct association in this example.[15]

Surprisingly, we find the cause of the POS change (p)-(v) in "[...] vnd heiliget jn(p) darumb das er [...]" (LB1) and "[...] und heiligte ihn(v) darum daß er [...]" (LB2) being a problem that the POS tagger has. While after normaliazation "jn" is correctly normalized into "ihn" and tagged as (p), in the newer Luther Bible (LB1), TreeTagger is confused by

[15] Modeling operations, only the cheapest operation couple is considered when multiple (unequal fallback) are identified.

the pronoun “ihn” and tags it as verb instead. Similar as in the English Bible corpus, it happens very often that the typical leading conjunction “and”, e.g., in ELB1 is aligned to words that are structurally correctly aligned, but accompany the POS (d), e.g., “da”, “so”, etc. We find more examples of paraphrasing when we look at changes from (l) to (p) and vice verse. For example, “alles” in “Herdenvieh und wilde Tiere und alles(p) was kriecht” and “das” in “Vieh, Gewürm und das(l) Wild der Erde” is aligned. On first sight, this seems to be an alignment error. However, both snippets represent an enumeration, and the two aligned words are both referring to the last item in the list.

	v	a	d	r	p	l	c	e	m	g	F
n	0.82	1.13	1.68	1.8	2.29	5.52	1.69	1.93	53.15	1.87	1.93
v	-	1.4	1.81	5.24	2.11	5.89	1.96	2.11	2.1	1.98	2.11
a	-	-	3.96	5.56	2.57	5.93	4.48	3.45	51.22	10.96	3.45
d	-	-	-	4.14	2.49	5.95	2.84	4.49	69.53	4.09	4.49
r	-	-	-	-	4.8	4.61	4.71	5.94	72.16	8.42	5.93
p	-	-	-	-	-	3.34	3.52	1,411.17	66.64	9.88	2.95
l	-	-	-	-	-	-	5.52	6.17	6.15	11.21	2,692.73
c	-	-	-	-	-	-	-	4.6	4.6	10.37	4.6
e	-	-	-	-	-	-	-	-	1,492.79	11.6	-
m	-	-	-	-	-	-	-	-	-	11.6	-
g	-	-	-	-	-	-	-	-	-	-	11.6

not significant significant

Table 6.18: Chi-squared numbers of changes in the German Bible corpus according to the POS class. Statistical significance of a POS change is measured towards the probability of the given POS in the overall alignments.

Finally, in Tab. 6.18 the significance values for POS changes from Tab. 6.17 in the German corpus are shown. Very significant changes are shown for exclamations (e), numerals (m) and particles (g). For example, an exclamation change happens when words such as “Ach(e)” as in “Und er sprach Ach(e) Herr” (ELB1) are aligned to pronouns as in “Mose sprach aber Mein(p) HERR” (LB2). A pronoun-numeral change happens in cases such as “zwey(m)” as in “Aber die zwey [...]” (LB1) and “beiden(p)” as in “und die beiden(p) [...]”. A change from a particle (g) to a conjunction (c) happens for example when “zu(g)” as in “[...] sprach Gott den Himmel zu(g) So ward Abend” (GB) and “Und(c)” from “Gott nannte die Ausdehnung Himmel Und(c) es ward Abend”. The number of the words between “Himmel” and “Abend” is equal in both texts. Hence, the aligner chooses to align—incorrectly—“zu” and “Und”.

6.2.7 Summarizing discussion

Following, the answers of the research questions of this section are summarized.

RQ B2.1: Modification in historical, paraphrastic text are distributed following a power-law. However, the frequency of the unique measuring points of the distribution also depends on how well a linguistic resource supports the vocabulary of the historical text. Further, the empirical figures of the modification also depend on the tools that are available for preprocessing, and their performance and flexibility to process out-of-domain or out-of-time data. The lack of resources was addressed by considering multiple synset databases. However, especially the semantic operations can only be identified if a lexical database comes with a certain coverage. Hence, the findings are only considered a lower bound of what actually can be found.

RQ B2.2: We further learned that POS changes appear vastly, which is shown by the percentage of almost 16%[16] in the English Bible corpus, and up to 30% in the German Bible corpus based on a very coarse-grained tagset. These changes mainly indicate strong paraphrasing—e.g., caused by the alignment of an adverb and a conjunction ("God also said" and "And God said" both taking a similar role introducing a new sentence), or the typical nominalization ("to shine" vs. "bring light"). These examples can not be considered to be recognized by the modification model, because they are neighter inflection of each other nor do they hold the same POS which excludes the application of a synonym relation or similar. That is, because most current synset (see e.g., Fellbaum (1998); Miller & Fellbaum (2007)) databases mainly store words POS-wise—i.e., words with the same meaning are stored together, but they also have the same POS. But the presented results show that in paraphrastic text it is very usual to change the POS when one or more words are replaced and a sentence is repeated following the same meaning, but different vocabulary. Hence, it is important to find ways to store semantically similar words not only when sorted by POS, possibly also as so called "semantic word families" that include words of the same meaning that also come with different POS.

[16] An evaluation of the alignment performance of a sample of English Bible translations is presented in Sec. 6.1

Chapter 7

Measuring paraphrasticality

This chapter is an expansion of the following papers:[1]

- Maria Moritz, Johannes Hellrich, & Sven Büchel. Towards a Metric for Paraphrastic Modification. DH 2018. ADHO.
- Maria Moritz, Johannes Hellrich, & Sven Büchel. A Human-Interpretable Method to Predict Paraphrasticality. LaTeCH-CLfL 2018.

The following chapter talks about the evaluation of the newly introduced, human-interpretable feature-based method against existing methods to measure and predict modification. First, paraphrasticality is measured by designing and validating a score that is used to measure distance among Bibles in a DH use case. Thereafter, the modification analysis is conducted within a task of determining semantic similarity in three parallel text datasets. Running the technique on three corpora, comparable accuracy with current similarity scores can be achieved, significantly beating them in one of the three corpora, which indicates the potential of the method. The similarity scores used for comparison were initially designed to evaluate machine translation output.

7.1 Towards a metric for paraphrastic modification

7.1.1 Overview

The previous chapter discussed how to improve the alignment recall and accuracy in historical English Bibles, and it presented details on the modification that happens when text is reused paraphrastically. However, the proposed operation-driven technique can also be used to design a metric that measures modification in historical text reuse. This is important,

[1] J. Hellrich and S. Buechel showed how to use a library of regression functions and together we discussed some experimental details. I run the experiments myself. Regression is used in the paper to compare the prediction of equivalency on the test data. The data frames used by the regression function come from my own code base. I wrote the paper myself.

because, to a human reader, the introduction of, say, spelling variations is a minor modification compared to substituting entire words. Yet, how can the different degrees of alterations, which are intuitively clear to scholars, be captured in an algorithmic way? The hereby proposed technique thereby is outstanding in this regard that it is human-interpretable, because it explicitly measures the type of modifications and the ratio of each of them.

Therefore, this section presents a first approach for designing a metric for paraphrastic modification in historical text. Based on an English Bible corpus (consisting in three Bible editions literally translated from Hebrew and Greek and three standard translations) the frequency of different classes of textual variations between each pair of Bibles is measured. We then use the probability of these variations in a binary classification experiment based on regression to determine important features for these classes of modifications. Ultimately, this allows for defining a metric for paraphrasticality which we validated with promising results.

7.1.2 Complementing related work

Measuring the similarity or distance between two spans of text is relevant to many areas in and related to NLP (see e.g., Levenshtein (1965); Xu *et al.* (2015); Papineni *et al.* (2002)). In stylometry, different kinds of delta metrics are used to compute the difference between the writing style of authors or texts (Jannidis *et al.* (2015)). These are typically based on the frequency distribution of the most frequent words. Many of them have in common that they rely on features at the token and character-level alone and do not incorporate semantic proximity. In contrast to that, computing the semantic similarity between two sentences is a popular task within NLP as shown in (Xu *et al.*, 2015). However, approaches in this field are typically not indented for manual inspection and are thus not suited for the use in the humanities. Instead, they usually focus on measuring if and how frequent a text has been modified, rarely determining the degree and explicit character of paraphrastic modification. In contrast to these contributions, this work aims to develop a measure, which is both, semantically informed as well as human interpretable by identifying the degree based on different modification types. Doing so, it also makes the degree of modification transparent and interpretatable to the humanist.

7.1.3 Methods

Research question

The following research question is formulated: *How can the proposed method be used to measure distance between two Bibles with regard to both, the translation background, and the time distance between their ages, and which operations are important for this task?*

Text data used

For this experiment, we use a subset of the parallel corpus described in Sec. 3.1.2 that consists of Bibles from the 19th century, half of them being literal translations that closely follow the primary source texts' language and syntax while the other half are standard translations following the tradition of the Anglican Church (see Tab. 7.1 for precise Bible information next to publication data).

Bible	published	type
The Webster Bible (WBT)	1833	standard
English Septuagint (LXXE)	1851	literal
Young's Literal Translation (YLT)	1862	literal
Smith's Literal Translation (SLT)	1876	literal
Darby Bible (DBY)	1867-1890	standard
English Revised Version (ERV)	1881-1894	standard

Table 7.1: Overview of English Bible translations used

Preprocessing and alignment

As always, punctuation and verse identifiers are removed before pairing up the six Bibles in every possible combination (15 in total) and aligning them at the token level using the Berkeley Word Aligner (DeNero & Klein, 2007) (see Sec. 4.1.3).

Counting modification operations

Building on these word-aligned pairs of Bibles, we can describe the divergence between a pair of verses in terms of the modification operations which would be necessary to convert one version into another. As usual, the modification operations introduced before are automatically applied and counted for each verse and Bible pair (see Tab. 7.2).

Weight identification

By counting modification operations, we gain a fine-grained description of the exact differences between two spans of text. However, to construct a metric, it is necessary to find a way to condense these modification frequencies down to a single number. For that the fact that we deal with two classes of Bible translations is exploited, literal and standard ones. Thus, to estimate a human judgment of deviation, one can assume that standard translations are more homogeneous to each other than literal translations (since the latter demand for more creative language use). Hence, we train a classifier to distinguish whether a pair of Bible verses is from the same class (both Bibles being standard or literal translations, respectively)

operation	description	estimated coeff. θ_{rel}	weights based on gini impurity
lower	case-folding matches	.070	.123
lem	lemmatizing matches	.226	.215
editdist	writing variant	.079	.043
syn	synonyms match	.221	.108
hyper &	source w1 is hypernym of target w2		
hypo	source w1 is hyponym of target w2	.170	.086
co-hypo	co-hyponyms match	.142	.089
fallback	other	.091	.336

Table 7.2: Operations used for distance measuring next to weighted features

or from different classes. For this task, a maximum entropy classifier[2] is used and the relative frequencies of the modification operations serve as features.[3] Now, the key part of the contribution is that the coefficients of the fitted model can be exploited as the empirical estimate of the relative importance of these modification operations for paraphrasticality.

7.1.4 Results

Metric

After applying the weighted features to the whole dataset, the maximum entropy classifier decides which features (operations) are more/less important to predict the label (correct class) in the test dataset best. Table 7.2 lists the final, normalized (summing up to 1) feature weights of the fitted model. Lemmatization, hyponym, hypernym[4] and synonym relations turn out to be especially important for the classification task.

Based on these coefficients, we define the paraphrasticality metric *par* between two word-aligned text spans a and b as:

$$par(a,b) = \sum_{i=0}^{n} \theta_i x_i^{a,b} \tag{7.1}$$

where n is the total number of features (or classes of operations), θ_i is the absolute weight for feature i determined via the classification experiment and $x_i^{a,b}$ is the relative frequency of the respective operation. In order to gain face validity for this newly defined metric, the

[2] Using the `scikit-learn.org` implementation. Training for this binary classification task was done using 10-fold cross validation achieving an accuracy of .68.
[3] Relative operations are use to normalize the impact of each feature on the training examples.
[4] Hyperonyms and hyponyms are folded to receive symmetric relations.

paraphrasticality score can be computed for each one of the 15 Bible pairs in the corpus (as average of their verse paraphrasticality). The results are presented in Tab. 7.3.

		b				
		ERV	WBT	*LXXE*	*YLT*	*SLT*
a	DBY	0.16	0.16	0.35	0.41	0.4
	ERV	-	0.12	0.35	0.43	0.42
	WBT	-	-	0.34	0.44	0.40
	LXXE	-	-	-	0.52	0.47
	YLT	-	-	-	-	0.40

similar — distant

Table 7.3: Deviation between each pair of Bibles in terms of the newly developed paraphrasticality metric; higher values indicate higher distance

Qualitative validation

Three regions can be identified in the plot of Tab. 7.3. The upper left triangle shows that the standard translations do not differ much from each other (as expected), especially since WBT and ERV are revisions of the same Bible. The 3x3 rectangle in the upper right corner represents pairs of one literal and one standard translation, respectively. One can see that the distance between those is about 0.3 thus displaying increasing paraphrasticality compared to pairs of only standard translations. The highest deviation however is between the literal translations by Smith (SLT) and Young (YLT) compared to the English Septuagint (LXXE). This can be explained by the choice of vocabulary by each translator and by the purpose they follow in their translations. For example, SLT and LXXE use “firmament” when YLT uses “expanse”, SLT and YLT use “rule” when LXXE uses “regulating”. Coming back to the research question, it can thus be concluded that the proposed metric yields valid and—perhaps even more important for applications in the humanities—interpretable results to measure, visualize and validate distance in a parallel monolingual corpus (see Fig. 7.1 for a simplified alignment example between LXXE and YLT).

The approach also enables to judge distance on a fine-grained level based on pure operation counts. In Tab. 7.4 the top 3 operations for each Bible pair are shown. As one can see, most of the top 3 operations per Bible pair relate to semantic relations between the

aligned word pairs (matching lemma, synonymy, or co-hyponomy). Even though lemmatizing is the most frequent operation—one might assumes that this simply indicates weak paraphrasing—together with an essential ratio of synonymy and co-hyponymy it represents very strong paraphrasing. That is, because restructuring a sentence—while retaining the same meaning—comes together with changing the tense, mood, number of words etc. This furthermore underscores the advantage (the interpretability)[5] that our metric has as opposed to approaches that only work based on token and character ngrams (to textual similarity) such as Levenshtein distance or delta measures.

Bible pair	**operation 1**	**operation 2**	**operation 3**	**classes**
DBY-ERV	lem (1.6%)	syn (1.1%)	cohypo (.9%)	standard
DBY-WBT	lem (1.6%)	syn (1.1%)	cohypo (.9%)	standard
ERV-WBT	lem (1.6%)	syn (.7%)	cohypo (.6%)	standard
DBY-LXXE	lem (3.1%)	syn (2%)	cohypo (1.9%)	standard/literal
DBY-YLT	lem (6.6%)	low (4.7%)	syn (2.6%)	standard/literal
DBY-SLT	lem (5.9%)	syn (2.6%)	cohypo (2.2%)	standard/literal
ERV-LXXE	lem (3.5%)	low (2.1%)	syn (1.9%)	standard/literal
ERV-YLT	lem (6.6%)	low (4.7%)	syn (2.5%)	standard/literal
ERV-SLT	lem (5.9%)	syn (2.6%)	cohypo (2.2%)	standard/literal
WBT-LXXE	lem (3.4%)	low (2.2%)	syn (1.9%)	standard/literal
WBT-YLT	lem (6.8%)	low (4.8%)	syn (2.7%)	standard/literal
WBT-SLT	lem (5.8%)	syn (2.6%)	cohypo (2.2%)	standard/literal
LXXE-YLT	lem (7.%)	low (4.4%)	syn (2.6%)	literal
LXXE-SLT	lem (5.8%)	cohypo (2.6%)	syn (2.6%)	literal
YLT-SLT	lem (5.4%)	low (4.8%)	syn (2.5%)	literal

Table 7.4: Top 3 most frequent operations (without fallback) per Bible pair

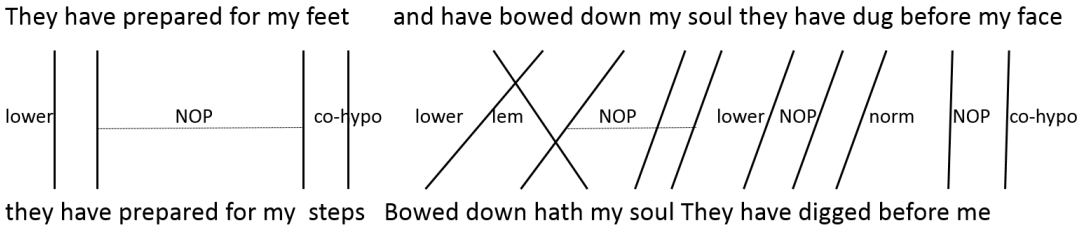


Figure 7.1: Example for alignment with associated operations - the program output is not ordered and uses the word position for identifying a token.

[5]This also can leave the judgment of the similarity of two sentences and verses to the expert.

An alternative weighting strategy

In a further experiment, the weights of features are estimated with a different strategy. In this experiment, we use a meta tree that fits several randomized decision trees[6] on different samples of the whole dataset. This way the so called "gini impurity" (first introduced by Breiman *et al.*, 1984, 1993) of each feature can be determined by the total decrease of the impurity (i.e., the incorrectly labeling) happening at each node. A node represents a discriminating feature. The impurity is then weighted by the probability—of a sample—to reach certain node. In general, the gini impurity measures how often a randomly chosen example from a dataset would be incorrectly labeled under the condition that it was randomly labeled. It is defined as:

$$I_G(p) = 1 - \sum_{i=0}^{J} p_i^2 \tag{7.2}$$

where J is the number of labels (features) and p_i is the fraction of examples labeled with this label J. The gini impurity, hence, is similar to the information gain.

We also modify the earlier design in that regard that instead of using two labels (0–Bibles form the same class aligned; 1–Bibles form different classes aligned), the new experiment now makes use of three labels distinguishing for Bibles aligned that come from the same class, which class that is. Hence, introducing labels for two Bibles aligned being literally translated (label 2) and two Bibles aligned being both a standard translation (label 0). Again, two Bibles form different classes aligned receive the label 1.

Table 7.2 displays the resulting weights next to those of the first experiment. The weights are distributed slightly different. Especially, the operations of semantic relations (i.e., *syn, hyper(o) and co-hypo*) are weighted even lower than before. *fallback*, however, which—from experience—strongly correlates with the morphological and lexical modification degree between two verses, experiences a much higher weighting.

To validate these weights as well, the distance scores are calculated for the gini weights and shown in Tab. 7.5. Principally, these scores do not differ in a meaningful way from the visualization of the first experiment. Again, the three regions of aligned Bibles from the different classes are visible. As one can see, the scores differ just slightly. Especially, the column of LXXE is a bit darker than before which comes with the increase of the weights of the fallback operation while synonyms, hyponyms and co-hyponyms (operations that appear much fewer) are downgraded.

[6] The ExtraTreesClassifier implementation that comes with the sklearn library is used.

a \ b	ERV	WBT	*LXXE*	*YLT*	*SLT*
DBY	0.16	0.17	0.39	0.36	0.35
ERV	-	0.11	0.38	0.38	0.38
WBT	-	-	0.36	0.39	0.36
LXXE	-	-	-	0.53	0.47
YLT	-	-	-	-	0.36

similar — distant

Table 7.5: Deviation between each pair of Bibles in terms of applying the gini impurity. Higher values indicate higher distance. Resulting scores are scaled up by multiplying them by 10 to better compare both weighting approaches

Adaption to German Bibles

The experiment requires a specifically-selected dataset to perform the regression classification. This means that is can not simply be applied to any other parallel corpus. For example, when the coefficients estimated via the experiment that was based on the English Bibles are applied to a parallel corpus of German Bibles, results would be skewed because the operations do not necessarily come in the same order (many to a few) as do their weights—considering they come from a different dataset (see Sec. 7.1.5).

However, since the approach also returns the explicit set of operations, we can investigate the top three operations of the alignment of the German Bibles. These are displayed in Tab. 7.6. Again, lemmatization is first ranked, followed by case-folding (lower) and the edit distance similarity measure, which does not confirm with the top operations from the experiment of the English Bibles. The reason for this is, again, the influence of lemmatization to the experiment.

7.1.5 Restrictions

One important aspect of the proposed method is that the weighting of the features is purely based on the aligned Bible data according to the use case that distinguishes literally and standardly translated Bibles. This means that i) it may not be adaptable to texts not being editions of each other, and ii) that distance is measured and weighted based on how

Bible pair	operation 1	operation 2	operation 3
ELB1-ELB2	lower (.2%)	editdist (.07%)	lem (.07%)
ELB1-GRU	lem (6.3%)	lower (3.6%)	editdist (1.4%)
ELB1-LU1	lem (15.8%)	lower (4.7%)	editdist (3.5%)
ELB1-LU2	lem (6.5%)	lower (2.1%)	editdist (1.3%)
ELB1-NEU	lem (6.9%)	lower (3.1%)	syn (1.3%)
ELB1-TXT	lem (5.4%)	lower (2.1%)	editdist (1.3%)
ELB2-GRU	lem (6.2%)	lower (3.6%)	editdist (1.3%)
ELB2-LU1	lem (15.7%)	lower (4.7%)	editdist (3.5%)
ELB2-LU2	lem (6.4%)	lower (2.1%)	editdist (1.3%)
ELB2-NEU	lem (6.8%)	lower (3.1%)	syn (1.3%)
ELB2-TXT	lem (5.4%)	lower (2.1%)	editdist (1.3%)
GRU-LU1	lem (14.4%)	lower (5.9%)	editdist (3.0%)
GRU-LU2	lem (7.%)	lower (4.2%)	syn (1.4%)
GRU-NEU	lem (7.%)	lower (3.7%)	syn (1.5%)
GRU-TXT	lem (6.1%)	lower (3.7%)	editdist (1.5%)
LU1-LU2	lem (15.9%)	lower (6.0%)	editdist (5.0%)
LU1-NEU	lem (13.3%)	lower (4.2%)	editdist (2.7%)
LU1-TXT	lem (14.5%)	lower (4.6%)	editdist (3.2%)
LU2-NEU	lem (7.2%)	lower (2.9%)	syn (1.6%)
LU2-TXT	lem (6.5%)	lower (2.3%)	syn (1.2%)
NEU-TXT	lem (6.9%)	lower (3.1%)	syn (1.4%)

Table 7.6: Top 3 most frequent operations (without fallback) per German Bible pair

strongly features are considered, and which features are considered most relevant for the distinction by the regression model. Further prerequisites of the method clearly are the existence of appropriately working preprocessing tools and the lexical databases that allow for querying semantic relations such as synonymy, etc. Established similarity measures such as the cosine similarity and the tf-idf measure are more robust in this regards even though they do not come with a detailed identification of the different modification types as the proposed approach does.

7.1.6 Conclusion

Summary

This section presented the first study on designing a metric for paraphrasticality. Different from existing approaches on measuring distance or similarity between texts, here, paraphrasticality is described as frequency of specific modification operations for which empirically adequate weights were found via a machine learning experiment. As demonstrated, the ap-

proach is specifically useful for applications in the humanities as operation frequencies and feature weights, as well as paraphrasticality scores are open to manual inspection.

Revisiting the role of semantic relations

After finding a way to weigh the operations from the proposed method, and forming a paraphrasticality score from it, it is interesting to see the results of the examples from Sec. 4.5 when that procedure is applied to them. Recall, this are the distance scores based on Meteor:

1. the *dog* eats the bone & the *hound* eats the bone → 5% distance
2. the *poodle* eats the bone & the *dachshund* eats the bone → 65% distance
3. the elephant eats the *orange* & the elephant eats the *pear* → 60% distance
4. the elephant eats the *peanut* & the elephant eats the *nut* → 60% distance
5. the elephant eats the *peanut* & the elephant eats the *groundnut* → 5% distance

With the new method, the following distance scores returned are:[7]

1. the *dog* eats the bone & the *hound* eats the bone → 44% distance
2. the *poodle* eats the bone & the *dachshund* eats the bone → 29% distance
3. the elephant eats the *orange* & the elephant eats the *pear* → 29% distance
4. the elephant eats the *peanut* & the elephant eats the *nut* → 34% distance
5. the elephant eats the *peanut* & the elephant eats the *groundnut* → 44% distance

Which shows room for adapting scores of semantic textual similarity more adequately.

7.2 Comparison against existing techniques of semantic equivalency prediction

7.2.1 Overview

As we already learned in Sec. 2.4.2, a lot of effort is put into constantly improving plagiarism detecting methods, c.f. Potthast *et al.* (2011); Ferrero *et al.* (2017). However, algorithmic support that addresses both, high recall and precision for the detection of paraphrastic reuse in historical text is much more limited. As such, current techniques—such as embedding-based methods—, which are preferably applied to NLP tasks in modern texts, are often able to tell if and how frequent a text has been modified. However, it is especially important to determine the degree and specific type of modification such as the morphological

[7] The coefficients are trained on the Bibles, which have a comparably small ratio of modification per verse. Hence, distance figures are high. Further, to distinguish aligned Bible classes, the classifier choose synonyms to be weighted over co-hyponyms, which influences the weight of synonymy in these examples as well.

and lexical change. What precisely constitutes these changes (i.e., which "features" represent these modifications) is further an important prerequisite for enhancing reuse detection techniques. That being said, remember that a strong paraphrastic reuse, for example, does not only come with morphological change, but also with a certain degree of derivation and lexical substitution. The proposed method—while being both human-interpretable and semantically informed—can also be used to determine paraphrastic modification. In contrast to recent techniques that can identify semantic similarity in sentences (Wieting *et al.*, 2015; Brlek *et al.*, 2016), the presented technique exhibits detailed feature information such as the ratio of word substitution and the semantic relationships among them.

7.2.2 Complementing related work

Generally, computing the semantic similarity between two sentences is a popular task in NLP. Examples for techniques from the field of paraphrase detection are those of semantic similarity between sentences, and entailment. These are undertaken for example by Wieting *et al.* (2015) who use embedding models to identify paraphrastic sentences in a mixed NLP task based on the Paraphrase Database (Ganitkevitch *et al.*, 2013), a huge corpus of short phrases associated with paraphrastic relatives. Their simplest model represents a sentence embedding based on the averaged vectors of its tokens, the most complex model is a long short-term memory recurrent neural network. Their results show that the simple, word averaging model performs best on similar sentences and entailment.

7.2.3 Research questions and approach

Overarchingly, the following question is asked: **RQ P** *Does the degree of modification measured based on the operations applied between the words of two sentences serve as a good feature for paraphrase prediction?* The degree is thereby determined to be the frequency of the operations of each type of operation. Hence, some operations represent stronger modification (e.g., *hyperonymy*) and others weaker modification (e.g., *lower* casing). These relationships between two words can reach from exact copy (*NOP*) to co-hyponymy, see Tab. 4.1. Compared to scores such as Meteor that make use of synonymy, but do not model other relationships, the proposed score also integrates information on hypernymy, hyponymy, and co-hyponymy. This is especially useful in historical text, since meaning and, therefor, relationships change over time. For example, Meteor would rate two sentences (one containing the word "husky", one the word "poodle") with a much lower similarity (ca. 40%) than two sentences that contain the word "dog" and "hound" (ca. 95%).

Remember, the order of applying the scores follows typical prepossessing steps that one would perform to reduce variance in a text corpus before running a retrieval task. These are based on the token-level, such as normalization and lemmatization, and finally addressing

words that are semantically related, but do not share the same root or cognate (see Ch. 4 for details). Each operation is distinguished into the modification represented by the operation alone, and the case when the operation is accompanied by a change in POS. If the POS changes, the operation name is assigned with the suffix *POSch*. The relative numbers of the operations serve as features in a classification task.

To answer the research question stated above, the approach described is applied to the folling three datasets: **RQ P1** *a modern English parapahrase corpus*, **RQ P2** *a parallel Bible corpus*, and **RQ P3** *a Medieval Latin reuse dataset*.

7.2.4 Material used

Tools and lexical resources used

Just like in the previous experiments, BabelNet by Navigli & Ponzetto (2012) is used to retrieve relationships among two words of two verses. Given the lemma of a word BabelNet provides related words for that given lemma. For the Latin dataset Minozzi's Latin WordNet (Minozzi, 2009) is consulted. MorphAdorner by Paetzold (2015) is used to normalize, lemmatize, and POS tag the English text. Finally, sentences from the given parallel corpus (see next section) are aligned on the token level utilizing Berkeley Word Aligner. On top, the relation operations defined in Tab. 4.1 are modeled.

Contemporary parallel text data

As already introduced in Sec. 2.5, Madnani *et al.* (2012) conduct a comprehensive study on the usefulness of automated MT evaluation metrics, such as BLEU, NIST and Meteor, for the task of paraphrase identification. As a side product, they release a monolingual corpus of semantic equivalence, which is extracted from the PAN 2010 plagiarism detection challenge corpus. As a Gold dataset for paraphrase prediction, we use Madnani *et al.* (2012)'s corpus (see Sec. 3.2 for details).

Madnani *et al.* (2012) created negative pairings to the Gold dataset by sampling non-aligned sentences with an overlap of four words. The training and test set comprise 10,000 and 3,000 sentence pairs, respectively. Both datasets are balanced regarding positive and negative labels.

Historical text data

Again, the experiment makes use of a subset of Bibles from the parallel Bible corpus as described in Sec. 3.1.2 Tab. 3.3 last column. Again, they come from two classes: literal translations—those being literally translated from the primary languages Hebrew and Ancient Greek coming with rich linguistic diversity—and translations that mainly follow the

tradition of the Anglican Church. Table 7.7 lists the detailed edition names accompanied by its publishing date.

Bible	published	type
Douay-Rheims Challoner Rev. (DRC)	1749-1752	standard
King James Version (KJV)	1769	standard
The Webster Bible (WBT)	1833	standard
English Septuagint (LXXE)	1851	literal
Young's Literal Translation (YLT)	1862	literal
Smith's Literal Translation (SLT)	1876	literal
Darby Bible (DBY)	1867-1890	standard
English Revised Version (ERV)	1881-1894	standard

Table 7.7: Overview of English Bible translations used

For the current experiment, we extract parallel verses from two different editions and try to predict if they come from the same or different translation classes (literal vs. standard). For the experiment we conduct on this dataset, we do not need negative training data.

Latin reuse

As the oldest dataset, and to cover a wider range of reuse in terms of language and age, Bernard's Latin reuse dataset from Sec. 3.1.1 is considered. As its parallel version, the relating Bible verses of the Biblia Sacra Juxta Vulgatam Versionem is used next to Bernard's reuse (see 3.1.1). Negative training data of equal size were obtained by randomly shuffling the initial dataset.

7.2.5 Experiment method and metrics

Abstractly spoken, this section describes how the experiments are conducted. It demonstrates the performance of the proposed approach in a task to predict semantic similarity of verses in parallel text. It shows a use case that carves out the strengths of the approach in historical data, and it shows the performance of state-of-the-art metrics in the same task.

Proposed approach

The method relies on the relative frequencies of modification operations (see Tab. 4.1) in an aligned sentence pair which later serve as features for a classifier:

$$x_i = \frac{\#o_i}{\sum_{j=0}^{m} \#o_j} \tag{7.3}$$

where x_i is the relative frequency of a modification operation i in an aligned sentence or Bible verse pair, m is the number of features, and o_i is the absolute frequency of operation i.[8] Remember that words are aligned using Berkeley Aligner, and operations are modeled on top by the main approach of this thesis. The relative frequency of an operation is its rate in an aligned sentence/verse pair. This method, hence, can be understood as a collection of features that are represented as relative frequencies of edits obtained from empirical values. These features are used as input to a maximum entropy classifier to predict if two sentences are paraphrases of each other. MaxEnt was chosen due to its simplicity, relying on a linear combination of features. Thus, feature weights can be roughly interpreted as importance of the respective modification operation after fitting the model. See the alignment example presented in Tab. 7.8, which illustrates the high interpretability of the proposed approach, because it comes with precise operation output.

OP	NOP	NOP	cohypo	NOP	syn	NOP	fallback	NOP	NOP	NOP	NOP	NOP	syn	fallback
sent. 1	It	is	unlawful	he	contends	to	co-operate	with	any	one	who	is	doing	wrong
sent. 2	It	is	law	he	argues	to	-	with	any	one	who	is	performing	-

Table 7.8: Example of operation (feature) based alignment

The method is evaluated by comparing it to several reference methods based on MT evaluation metrics.[9] To adapt these to the different paraphrase detection tasks, the source Bible provides the reference sentence (ref) and the target Bible (and Bernard's reuse respectively) provides the system output (sys). From the Gold corpus, also the source text (numbered in the repository with 1, see Madnani *et al.* (2012) for the data) serves as reference, and the paraphrastic reuse of it (numbered with 2) provides the system output.

Other metrics to compare the proposed method to

Often, MT metrics are based on simple edit distance measures such as the Position-independent Error Rate (**PER**) (Tillmann *et al.*, 1997), which uses a bag-of-words approach. Popović & Ney (2007) define PER based on counts of independent words that system output and reference sentence have in common. For the purpose of this study their document-wide score is adapted to the sentence level:

[8] $m = 15$. Table 4.1 shows in total 11 operations. All of them—except fallback–are distinguished into two sub operations: with and without changing POS. Because we dropped three features after development experiments, i.e., *NOP, lem and hyper*—six in tiota—we are encountered with 15 features.

[9] Some of the metrics that capture distance (instead of a similarity) needed to be modified by using their complement.

$$PER = \frac{1}{2 \cdot N_{ref}}(|N_{ref} - N_{sys}| + \sum_{e} |n(e, ref) - n(e, sys)|), \tag{7.4}$$

where N is the size of system output and reference sentence respectively, $n(e, ref)$ is the number of a given word e in the reference, and $n(e, sys)$ is the number of a given word e in the system output. The PER score is used as it is since it defines a distance.

The translation edit rate (**TER**) (Snover *et al.*, 2006)) is the number of edits that a system output needs to experience so that it matches a reference sentence. TER is normalized by the length of the reference input: $TER = \frac{\#edits}{\#w_{ref}}$ The experiment makes use of the implementation of the TER score by Snover *et al.* (2008).
Following Papineni *et al.* (2002), a sentence-based, hence, slightly modified **BLEU** score is define as:[10]

$$BLEU = BP \times \exp(\sum_{n=1}^{N} \frac{1}{N} \log p_n) \tag{7.5}$$

where N is the maximum $n$gram size, which is set to 2[11]. p_n is a precision score that is calculated based on ngrams in both, source and target texts (Papineni *et al.*, 2002). BLEU's brevity penalty (BP) is omitted, which would otherwise dominate the sentence level analysis.

The last measure considered is **Meteor** 1.5 (Denkowski & Lavie, 2014). Meteor especially differs from other scores by considering not only precision, but also recall. It further takes synonymy and paraphrases into account. Meteor introduces so called matchers that are represented by exact match, stem match, synonym match or paraphrase match. The hypothesis (system) and reference texts h and r are split into content words h_c and r_c, and function words h_f and r_f. Precision and recall measures are then used to determine the harmonic mean F_{mean}. Together with a fragmentation penalty that measures the degree of chunks, the Meteor score is calculated by $Meteor = (1 - penalty) \times F_{mean}$.

Similar to Madnani *et al.* (2012), the MT scores are used separately in a classification task to predict paraphrasticality with a maximum entropy classifier on the three datasets.

7.2.6 Results

Determining Paraphrases (RQ P1): Recall that using the relative operation count from the alignment operations as features in a classification task, the classification accuracy of the proposed approach is determined on the Gold corpus. A maximum entropy classifier is

[10] The following equation appears different than Eg. 2.4, because here, it is used on sentences.
[11] This means that unigrams and bigrams are considered; Setting this value to 1 seemed to drastically first. However, experimenting with N=1 resulted in an accuracy increase of 2.0% at the Gold dataset and 5.3% at Bernard hitting place three after our technique and Meteor

run on the one feature of the introduced metrics (as a threshold) and the operation features of the hereby proposed operation ratio-based approach. The results in Tab. 7.9 show that Meteor performs best on that task, followed by the operation ratio-based approach proposed in this thesis.

Determining Translation Classes (RQ P2): The goal is to find out whether it is possible to distinguish which types of Bible translations are aligned by using the features measured between them as classification features. The operations equip us with a fine-grained description of the degree of modification of two text excerpts. The Bible corpus is a suitable source for measuring the degree of modification, since it holds a wide range of paraphrastic reuse. Thus, to estimate a human judgment of deviation, it can be assumed that standard translations are more homogeneous to each other (based in their evolution history) than literal translations that demand for more creative language use. A 10-fold cross validation is applied. The results in Tab. 7.9 show that this task is achieved by all measures comparably well, but accuracy drops slightly when features of semantic relations are dropped (see upper part of Tab. 7.9). When WordNet is used solely for identifying relations performance increases slightly, which can be attributed to noise that comes with using BabelNet.

Determining Latin Reuse (RQ P3): Finally, reuse is predicted in the Medieval Latin dataset of Biblical reuse. The conjecture is that the proposed method is especially suited for this task, since especially co-hyponymy is a common means of substitution in historical text reuse. This came especially clear when looking at samples from the first task (Determining Paraphrases, RQ P1) revealed false positives that were enabled by allowing the rather loose co-hyponym relations in the modern plagiarism dataset.[12] Again, 10-fold cross validation is applied. Table 7.9 shows that dropping features such as co-hyponyms indeed worsens the accuracy of the method.

name	**Gold dataset**	**Bibles**	**Bernard**
par only WordNet	87.6	67.2	-
par synonyms only	87.7	67.1	88.9
par w/o cohyponyms	87.9	67.3	89.8
par	87.6	67.3	**90.7**
TER	85.8	67.0	61.9
PER	85.4	67.4	87.6
BLEU	83.9	**68.1**	83.6
Meteor	**89.5**	67.8	88.9

Table 7.9: Accuracy of semantic equivalency determination in %

[12]Note also that Meteor only contains synonym data in English. This can influence its accuracy.

7.2.7 Threats to validity

The proposed metric is especially depending on the quality of the input data. Though a comprehensive accuracy test of the results from the preprocessing steps was not run yet, one can expect the tools we use to work reasonably well for this purpose. Especially the POS tagger and normalizer's performance MorphAdorner is denoted with an error rate of 1.9% for historical data (c.f. Basu (2014);Wilkens (2008)).

Further, only operation relations based on the richness of BabelNet can be derived, and we must trust that these relations are correct. Again, a certain security comes with the fact that candidates for relationships only stem from a certain window—a verse length. Finally, note that the alignment works symmetrically, meaning that it only looks into one direction between two Bibles. This means that hyperonymy and hyponymy relations can slightly differ depending on the chosen pairing.

7.2.8 Conclusion

In this section, a method for evaluating paraphrastic similarity in parallel text was presented. The method describes reuse based on the frequency of specific modification operations and is thus easily interpretable for humans, because it returns the precise ratio of, for example, lemma-aligned words or synonymy within two text excerpts. It was shown that modeling reuse in historical text using semantic relations beyond synonyms achieves results comparable or better to using features derived from machine translation metrics. Moreover, the proposed method is especially useful for applications in the humanities as operation frequencies—and if necessary their respective feature weights—and individual model decisions—are open to manual inspection.

That being said, it is one of the first works that considers looser semantic relations—beyond synonymy as Meteor simply does—to model reuse and predict semantic equivalency. The operation-based approach to measure and predict reuse in historical text can thus be understood as a step into the right direction.

Chapter 8

Conclusion

This chapter summarizes the thesis, lists the main contribution to the field of reuse detection in historical texts, and presents possible links to ongoing research, for example in other domains or for other languages.

8.1 Contributions

This thesis presented an automated method to model modification in historical text reuse based on different parallel datasets and corpora. Modification was measured by applying operations in a prioritized order from no modification (words are stable) up until co-hyponym substitution[1] between two sides of mono-lingual bi-text. These operations represent morphological change—such as inflection and derivation—and lexical change—such as synonymy, hyper or co-hyponym substitution. The modification was collected for two datasets of Medieval Greek and Latin, a parallel English Bible corpus and a German Bible corpus—both spanning about 400 years. This thesis' main contribution is to show-case explicitly that modification beyond synonymy needs to be taken into account for future automated reuse detection techniques that are supposed to work on historical text.

8.1.1 Method to measure modification in historical text reuse

The first contribution is the design of an operation-based method to model and measure modification in parallel corpora of historical text (see Ch. 4). This method is based on the preprocessing steps that are performed when an information retrieval task shall be performed on a text base covering morphological and lexical change. Morphological operations are NOP (word re-appears in the reuse), case-folding, normalization, lemmatization, derivation, and an edit-distance-based character distance. Beyond word similarity, also semantic relations—driven by lexical databases—were included. These are synonymy, hypernymy, hyponymy, and co-hyponymy. The main advantage of this approach is that modification is given an

[1] Many-to-many word substitution was analyzed on smaller dataset manually only.

explicit name and ratio. Opposed to recent developments that measure modification and similarity without lexicon knowledge, this approach and its results are much clearer and easier to interpret by humans, because the returned operation names—holding the aligned words as parameters—offer more precise information on what changed and how it changed between to sides of a paraphrastic parallel text or reuse. The source code of this work is freely available.[2]

8.1.2 Application of the method in a small-scale use case of Medieval Greek and Latin

The developed method was applied to one dataset of Medieval Greek and one dataset of Medieval Latin (see Ch. 5). Both consist of couples of sentences where one is a Bible verse, and one is the modified (rephrased or slightly modified) reuse of that verse. Empirical figures of these modifications were presented and discussed. The results show that, especially, co-hyponym substitution occurs about as often as synonym substitution, and that substitution represented by semantic relations (i.e., synonymy, hypernymy, etc.) cover about 10% of all operations (stable and changing), and about 20% of the identified modification operations. The results further show that resources such as lexical databases for ancient languages can support the task of reuse modification analysis to some extent, but still lack coverage of vocabulary.

8.1.3 Application of the method in a bigger use case of historical Bible translations

In a bigger use case, the proposed method was applied to a parallel corpus of Bible translations in English (1500–1900) and German (1545–2010, see Ch. 6). First, a new operation was defined and assessed to cover a gap between morphological changes of words and lexical replacement. This operation (an edit-distance-based string similarity) proved to be especially useful to align named entities as these do often have a certain length and a high diversity of writing variants.

Next, modification among these Bible corpora was measured and analyzed based on changes in part-of-speech. The results showed that a high percentage of changing part-of-speech correlates with a longer distance—in terms of writing variance and time of publishing. Applying the operation-based method showed that normalization is the most frequent operation in the English corpus, and lemmatization is the most frequent operation in the German corpus. It was also shown that—especially in the English corpus—co-hyponym relations are

[2] `https://bitbucket.org/mariamoritz/reuse_modification_analysis/src/master/`

a common way to replace words when a text is paraphrased. Co-hyponym replacement is in fact used about half as often as synonym replacement.

Additionally, two lexical databases were compared regarding the recall of operations based on semantic relations in the experiments processed on the Bibles. The results from this experiment make clear that the definition of what is considered a synonym or a co-hyponym is not unified yet, and could benefit from better standards.

8.1.4 A measure for textual distance and paraphrase prediction

In Sec. 7.1 a score was derived based on the operations designed to measure modification. A special setup of comparing English Bible translations from two classes enabled it to determine a distance score by weighting features (coming form the operations) according to their importance to discriminate whether two Bibles from the same or different classes are aligned. The classes are *standard* translations, and Bibles translated *literally* from the primary texts' languages (i.e., Ancient Greek, Latin and Hebrew). After learning more important and less important features, the weighted features were applied to compute a distance value between two Bibles each. A qualitative validation showed that the score represents distance caused by writing variance, style and time passed between the publishing of any two Bibles.

The feature-based distance score, derived from the operations, was also compared against exiting techniques in sentence similarity prediction (see Ch. 7.2). Compared to machine translation evaluation scores, it was shown that in a paraphrase prediction test, the proposed method performs best on a Latin reuse dataset—due to the characteristics of historical text and the fact that it takes co-hyponym relations into account. In modern English plagiarism detection, the proposed approach works comparably well as existing techniques, always considering that the new approach needs more preparation time and data to be applied to a dataset.

8.2 Future work

8.2.1 Application from and to other domains and languages

Further development in reuse analysis and the topic of deriving one text version from another is to learn and apply so-called edit scripts (Kehrer, 2014; Chawathe *et al.*, 1996). Edit scripts come from the domain of software engineering and are used to track modifications that software developers perform on a codebase. The edits can be insertions, deletions or modifications of classes or functions, and provide deeper insights into changes to some source code, as opposed to textual diffs. Whether learning edit scripts on such intricate operations is possible is an open question.

In the opposite direction, certainly an interesting topic is to develop techniques that enable modification analyzes in domains other than prose text. Hence, the vast field of software engineering sees itself encountered with the problem to record, evaluate, and manage modification in codebases and platforms on a daily base. For instance, Cortés-Coy *et al.* (2014) propose a method to automatically generate commit messages using a commit's change set.[3] These modifications are categorized as *structural, behavioral, creational, and collaborational*[4], and help to add extra information to the commit message. Such work might benefit from techniques that consider external language sources to trace, record, and measure modification. Similar as work performed by Lu *et al.* (2015) who use WordNet to expand search queries, and, thus, improve the efficiency of code search.

8.2.2 Reuse detection using transfer and cross-lingual learning

With the advent of transfer and cross-language learning (Shi *et al.*, 2015; Sasaki *et al.*, 2018), an interesting direction can be to make use of higher resource languages, such as English, to find further important features automatically using machine or deep-learning models. These can be applied in historical text reuse detection in less-resourced language text such as historical corpora that are often still under construction or require laborious manual cleaning. The task is also known as Dynamic NLP, i.e., the generation of tools that automatically select the best parameter settings to choose cross-lingual and domain adaptation techniques given a specific text genre to be processed. The portability of NLP tools across the diverse textual typologies is still an ongoing question in the NLP and in the DH community.

One way to address this is the focus on language independent features. First work exists already, for example, van der Goot *et al.* (2018) "bleach" their texts from lexical information and instead use different substitutions of the words, such as POS tags or word length, as additional features to improve prediction accuracy in standard NLP tasks that otherwise solely make use of lexical information-based features. Other examples for potential features are those of cognates or lexical concepts (the placeholder that summarizes all words from a whole family) for example.

Further, a recent work of this thesis' author (Moritz & Steding, 2018), predicts paraphrastic text reuse in Medieval Latin by cross-applying classifiers—that were trained for paraphrase prediction on modern English text corpora—and applied to historical text reuse. The authors analyze the impact of different language-independent features on the result-

[3] A change set contains all unique modifications applied to a codebase. Changes can be additions, deletions, modifications, renamed files, but also the stereotype of changed methods, which describe the effect a method has in a class.

[4] For example, creational methods create objects; structural methods get and set attributes; collaborational methods communicate between objects. (Cortés-Coy *et al.*, 2014)

ing reuse-detection accuracy. This can be solved comparably easy in reuse prediction by the number or ratio of surface features (e.g., number of words or similar words that two text excerpts have in common) that the source and the retrieved text have in common. This frequency is mainly language independent. Moritz & Steding (2018) find out that the angle—calculated based on two sentence's embedding—can help to drastically improve accuracy if features (such as the overlap of similar words) fail to achieve a satisfying precision and accuracy. But more important, the experiments showed that a smart choice of training corpora (which shares some characteristics with the data of the target language, even though the data is not obviously similar) can essentially help these cross-language learning tasks. Such a fortunate choice of resources is enhanced by Barbara Plank (Plank *et al.*, 2016), who proposes the use of what she names *fortuitous data*. These fortuitous data are extra data sources that can help to improve NLP tasks, but are not necessarily known already.

8.2.3 Future work in resource creation for historical languages

Attempts to create and enlarge language resources for modern languages are vastly growing. In addition to FrameNet, one trend is to expand existing resources to the multilingual level: Multilingual FrameNet (Boas, 2005) and Open Multilingual Wordnet (Team, 2018) are such examples. One important trend is not only to create these resources, but to design them in a way that makes information and knowledge flow between these resources easy. This way, it might be possible to link different sources of lexica. For example, one lexicon storing semantic equivalence, and one lexicon that stores inflected variances of a word family. Hence, both types of relationship can be made use of at the same time.

In the field of old languages, an initiative was recently established by Marco Passarotti to create, enrich, and combine a comprehensible resource for Latin. The motivation of the project is to unify linguistic resources and tools for automatically processing Latin, making them compatible. As such, the initiative addresses the gap between raw/low language resource data, NLP and knowledge descriptions, and contributes to a linked knowledge base for Latin resources (Passarotti, 2018). Previous attempts to combine language resources are, for example, CLARIN (Váradi *et al.*, 2008), which collects material for the humanities and social sciences, and the German Text Archive (Jurish *et al.*, 2014), which collects tokenized and POS-tagged versions of historical German literature.

Finally, once real gold corpora of historical text reuse exist, it will be worth pursuing the analysis on these data as well. Franzini *et al.* (2018) recently performed first work on the evaluation of reuse detection based on a gold standard of Medieval Latin reuse of Thomas Aquinas, which is now also available online. Such works provide the basis to learn and measure, even more precisely, different types of modification in the real.

Bibliography

Kurt Aland & Barbara Aland (eds.) (1966): *The Greek New Testament.* Deutsche Bibelgesellschaft-United Bible Societies, 27 edition.

Salha Alzahrani & Naomie Salim (2010): Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection. In: *Braschler and Harman*, 1176: 1–8.

Salha M. Alzahrani, Naomie Salim & Ajith Abraham (2012): Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. In: *Trans. Sys. Man Cyber Part C*, 42(2): 133–149.

Dawn Archer, Tony McEnery, Paul Rayson & Andrew Hardie (2003): Developing an automated semantic analysis system for Early Modern English. In: *Corpus Linguistics 2003 conference*, pp. 22–31. UCREL.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto *et al.* (1999): *Modern Information Retrieval*, volume 463. ACM, New York.

Collin F. Baker, Charles J. Fillmore & John B. Lowe (1998): The Berkeley FrameNet Project. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pp. 86–90. ACL.

David Bamman & Gregory Crane (2008): The Logic and Discovery of Textual Allusion. In: *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data.* ELRA.

David Bamman & Gregory Crane (2011a): The Ancient Greek and Latin Dependency Treebanks. In: *Language technology for cultural heritage: Selected papers from the LaTeCH Workshop Series*, pp. 79–98. Springer.

David Bamman & Gregory Crane (2011b): Measuring Historical Word Sense Variation. In: *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 1–10. ACM.

Alistair Baron & Paul Rayson (2008): VARD2: A tool for dealing with spelling variation in historical corpora. In: *Postgraduate Conference in Corpus Linguistics*.

Alberto Barrón-Cedeño, Marta Vila, M.Antònia Martí & Paolo Rosso (2013): Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. In: *Computational Linguistics*, 39(4): 917–947.

Anupam Basu (2014): MorphAdorner v2.0: From Access to Analysis. Spenser Review 44.1.8. `http://www.english.cam.ac.uk/spenseronline/review/volume-44/441/digital-projects/morphadorner-v20-from-access-to-analysis`. Accessed Nov. 2017.

Imene Bensalem, Imene Boukhalfa, Paolo Rosso, Lahsen Abouenour, Kareem Darwish & Salim Chikhi (2015): In: *FIRE Workshop*, pp. 111–122. Gandhinagar, India.

Monica Berti, Bridget Almas & Gregory R Crane (2016): The Leipzig Open Fragmentary Texts Series (LOFTS). In: *DHQ: Digital Humanities Quarterly*, 10(2).

Douglas Biber & Victoria Clark (2002): Historical shifts in modification patterns with complex noun phrase structures. In: *Teresa Fanego, Maria Lépez—Couso and Javier Perez—Guerra (eds.). English Historical Morphology. Selected Papers from*, 11: 43–66.

Bible-Study-Tools (2018): Bible Study Tools. `http://www.biblestudytools.com/`. Accessed: July 2018.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini & Gregory Crane (2014): The Making of Ancient Greek WordNet. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. ELRA.

Hans C. Boas (2005): Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. In: , 18(4): 445–478.

Marcel Bollmann (2012): (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool. In: *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*.

Marcel Bollmann, Florian Petran & Stefanie Dipper (2011): Rule-Based Normalization of Historical Texts. In: *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pp. 34–42.

Robert S Boyer & J Strother Moore (1977): A Fast String Searching Algorithm. In: *Communications of the ACM*, 20(10): 762–772.

Leo Breiman, J H Friedman, Richard A Olshen & Charles J Stone (1984, 1993): *Classification and regression trees*. Cole Advanced Books & Software, Google Scholar.

Eric Brill (1995): Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. In: *Computational Linguistics*, 21(4): 543–565.

Eric Brill & Robert C Moore (2000): An Improved Error Model for Noisy Channel Spelling Correction. In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 286–293. ACL.

Arijana Brlek, Petra Franjic & Nino Uzelac (2016): Plagiarism Detection Using Word2vec Model. In: *Text Analysis and Retrieval 2016 Course Project Reports*, pp. 4–7. University of Zagreb, Faculty of Electrical Engineering and Computing.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra & Robert L Mercer (1993): The Mathematics of Statistical Machine Translation: Parameter Estimation. In: *Computational Linguistics*, 19(2): 263–311.

Frederick Fyvie Bruce (1978): *History of the Bible in English*. 3rd ed., Oxford University Press.

Tim vor der Brück, Steffen Eger & 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)*, pp. 105–113. ACL.

John Bryant (2002): *The Fluid Text: A Theory of Revision and Editing for Book and Screen*. University of Michigan Press.

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

Marco Büchler, Gregory Crane, Maria Moritz & 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*, pp. 95–100. Springer.

Philip R Burns (2013): Morphadorner v2: A Java Library for the Morphological Adornment of English Language Texts. `http://douglasduhaime.com/blog/cross-lingual-plagiarism-detection-with-scikit-learn`.

Sudarshan S Chawathe, Anand Rajaraman, Hector Garcia-Molina & Jennifer Widom (1996): Change Detection in Hierarchically Structured Information. In: *ACM SIGMOD Record*, volume 25, pp. 493–504. ACM.

Titus Flavius Clemens (1905-1909): Werke (in Greek). In: Otto Stählin (ed.) *Die Griechischen Christlichen Schrifteller, Berlin, v. 12, 15, 27*. Leipzig.

Clément d'Alexandrie (2011): *Quel riche peut-être sauvé*. éditions du Cerf, Paris, sources chrétiennes 537 edition.

L. F. Cortés-Coy, M. Linares-Vásquez, J. Aponte & D. Poshyvanyk (2014): On Automatically Generating Commit Messages via Summarization of Source Code Changes. In: *2014 IEEE 14th International Working Conference on Source Code Analysis and Manipulation*, pp. 275–284. IEEE.

Carl P. Cosaert (2008): *The Text of the Gospels in Clement of Alexandria*. New Testament in the Greek Fathers, Society of Biblical Literature.

Gregory Crane (1985): Perseus Digital Library. `http://www.perseus.tufts.edu/hopper/`. Accessed: Dec. 2017.

Gregory Crane (1991): Generating and Parsing Classical Greek. In: *Literary and Linguistic Computing*, 6(4): 243–245.

Maxime Crochemore & Wojciech Rytter (2003): *Jewels Of Stringology: Text Algorithms*. World Scientific.

Scott Crossley, Tom Salsbury & Danielle McNamara (2010): The Development of Polysemy and Frequency Use in English Second Language Speakers. In: *Language Learning*, 60(3): 573–605.

David Daniell (2003): *The Bible in English: Its History and Influence*. New Haven, Conn: Yale University Press.

John Nelson Darby, Carl Brockhaus & Julius Anton von Poseck (1905): Die Bibel (Elberfelder Übersetzung 1905). `https://www.offenesbuch.com/g135719`.

Epistle Dedicatorie (1611): The Authorized King James Version of the Holy Bible. Wikisource link to Epistle Dedicatorie `https://en.wikisource.org/wiki/Bible_(King_James_Version,_1611)/Epistle_Dedicatorie`.

John DeNero & Dan Klein (2007): Tailoring Word Alignments to Syntactic Machine Translation. In: *Proceedings of the Annual Meeting on Association for Computational Linguistics*, volume 45, pp. 17–24. ACL.

Michael Denkowski & Alon Lavie (2011): Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 85–91. ACL.

Michael Denkowski & Alon Lavie (2014): Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380. ACL.

George Doddington (2002): Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In: *Proceedings of the Second International Conference on Human Language Technology Research*, pp. 138–145. Morgan Kaufmann Publishers Inc.

Bill Dolan & Chris Brockett (2005): Automatically Constructing a Corpus of Sentential Paraphrases. In: *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Douglas Duhaime (2015): Cross-lingual plagiarism detection with scikit-learn. `http://douglasduhaime.com/blog/cross-lingual-plagiarism-detection-with-scikit-learn`. Accessed: Jan. 2019.

Editorial Team of Elberfelder Bibel (1985): *Die Elberfelder Bibel. Vorwort.* R. Brockhaus Verlag.

Gertrud Faaß & Kerstin Eckart (2013): *SdeWaC–A Corpus of Parsable Sentences from the Web*, pp. 61–68. Springer Berlin Heidelberg.

Christiane Fellbaum (1998): WordNet: An Electronic Lexical Database: Bradford Book.

Pablo Malvar Fernández (2008): *Improving Word-to-word Alignments Using Morphological Information.* Ph.D. thesis, San Diego State University.

Samuel Fernando & Mark Stevenson (2008): A Semantic Similarity Aapproach to Paraphrase Detection. In: *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, pp. 45–52.

Jérémy Ferrero, Frédéric Agnès, Laurent Besacier & Didier Schwab (2017): Using Word Embedding for Cross-Language Plagiarism Detection. In: *CoRR*, abs/1702.03082.

Andrew Finch, Young-Sook Hwang & Eiichiro Sumita (2005): Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence. In: *Proceedings of the Third International Workshop on Paraphrasing*, pp. 17–24. ACL.

Greta Franzini, Marco Passarotti, Maria Moritz & Marco Büchler (2018): Using and evaluating TRACER for an Index fontium computatus of the Summa contra Gentiles of Thomas Aquinas. In: *Proceedings of the Fifth Italian Conference on Computational Linguistics*. Italian Association of Computational Linguistics.

Charles-Émile Freppel (1865): *Clement d'Alexandrie*. Bray et Retaux Libraires–Éditeurs.

Juri Ganitkevitch, Benjamin Van Durme & Chris Callison-Burch (2013): PPDB: The Paraphrase Database. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 758–764. ACL.

Alexander Geyken & 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 Tübingen.

Bela Gipp & Jöran Beel (2010): Citation Based Plagiarism Detection: A New Approach to Identify Plagiarized Work Language Independently. In: *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pp. 273–274. HT '10, ACM.

Bela Gipp & Norman Meuschke (2011): Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence. In: *Proceedings of the 11th ACM Symposium on Document Engineering*, pp. 249–258. DocEng '11, ACM.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim & Barbara Plank (2018): Bleaching Text: Abstract Features for Cross-lingual Gender Prediction. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Vishal Gupta & Gurpreet S. Lehal (2009): A Survey of Text Mining Techniques and Applications. In: *Journal of Emerging Technologies in Web Intelligence (JETWI)*, 1(1): 60–76.

Dan Gusfield (1997): *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Nizar Habash & Bonnie Dorr (2003): A Categorial Variation Database for English. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 17–23. ACL.

Andreas Hauser, Markus Heller, Elisabeth Leiss, Klaus U Schulz & Christiane Wanzeck (2007): Information Access to Historical Documents from the Early New High German Period. In: *Dagstuhl Seminar Proceedings: Digital Historical Corpora–Architecture, Annotation, and Retrieval*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Johannes Hellrich, Sven Buechel & Udo Hahn (2018): JeSemE: Interleaving Semantics and Emotions in a Web Service for the Exploration of Language Change Phenomena. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pp. 10–14.

Johannes Hellrich & Udo Hahn (2017): Exploring Diachronic Lexical Semantics with JeSemE. In: *Proceedings of ACL 2017, System Demonstrations*, pp. 31–36. ACL.

Arthur Sumner Herbert & T H Darlow (1968): *Historical Catalogue of Printed Editions of the English Bible, 1525–1961.* London: British and Foreign Bible Society, New York: American Bible Society.

Gerhard Heyer & Marco Büchler (2010): Some Challenges Posed to Computer Science by the eHumanities. In: *44. Jahrestag der Gesellschaft für Informatik e.V., Service Science - Neue Perspektiven für die Informatik*. Springer-Verlag, Leipzig.

Gerhard Heyer, Uwe Quasthoff & Thomas Wittig (2006): Text Mining: Wissensrohstoff Text - Konzepte, Algorithmen, Ergebnisse.

Hugh A G Houghton (2013a): Patristic Evidence in the New Edition of the Vetus Latina Iohannes. In: L. Mellerin & H.A.G. Houghton (eds.) *Biblical Quotations in Patristic Texts (Studia Patristica 54)*, pp. 69—85. Peeters, Leuven.

Hugh A G Houghton (2013b): The Use of the Latin Fathers for New Testament Textual Criticism. In: B.D. Ehrman & M.W. Holmes (eds.) *The Text of the New Testament in Contemporary Research. Essays on the Status Quaestionis second edition. NTTSD*, pp. 375—-405. Brill, Leiden.

James Wayne Hunt & M Douglas MacIlroy (1976): *An Algorithm for Differential File Comparison.* Bell Laboratories Murray Hill.

Fotis Jannidis, Steffen Pielström, Christof Schöch & Thorsten Vitt (2015): Improving Burrows' Delta–An empirical evaluation of text distance measures. In: *Abstract Proceedings of Digital Humanities 2015: Annual Conference of the Alliance of Digital Humanities Organizations*, p. no page numbers. Alliance of Digital Humanities Organizations.

Jay J Jiang & David W Conrath (1997): Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: *Proceedings of the 10th Research on Computational Linguistics International Conference*, pp. 19–33. ACL.

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*, pp. 128–134. ACL.

Bernadette Elaine Johnson (2009): *Using the Levenshtein algorithm for automatic lemmatization in Old English.* Ph.D. thesis, University of Georgia.

Kyle P Johnson, Patrick J Burns, Luke Hollis, Martín Pozzi, Amit Shilo, Stephen Margheim, Gitter Badger & Eamonn Bell (2014–2016): CLTK: The Classical Language Toolkit. `https://github.com/cltk/cltk`. DOI 10.5281/zenodo.44555 v0.1.32.

Bart Jongejan & Hercules Dalianis (2009): Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike. 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 1-Volume 1*, pp. 145–153. ACL.

Bryan Jurish (2008): Finding canonical forms for historical German text. In: *KONVENS*, pp. 27–38.

Bryan Jurish (2010): More Than Words: Using Token Context to Improve Canonicalization of Historical German. In: *Journal for Language Technology and Computational Linguistics*, 25(1): 23–39.

Bryan Jurish, Christian Thomas & Frank Wiegand (2014): Querying the Deutsches Textarchiv. In: *MindTheGap@iConference*.

Emil Kautzsch (1894): Textbibel. `http://vorwort.textbibel.de/`. Accessed: Jan. 2019.

Timo Kehrer (2014): Generierung konsistenzerhaltender Editierskripte im Kontext der Modellversionierung. In: Wilhelm Hasselbring & Nils Christian Ehmke (eds.) *Software Engineering 2014, Fachtagung des GI-Fachbereichs Softwaretechnik*, volume 227 of *LNI*, pp. 57–58. GI.

Mike Kestemont, Vincent Christlein & Dominique Stutzmann (2017): Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts. In: *Speculum*, 92(S1): S86–S109.

Erik Ketzan & Christof Schöch (2017): What Changed When Andy Weir's The Martian Got Edited? In: *Abstract Proceedings of Digital Humanities 2017: Annual Conference of the Alliance of Digital Humanities Organizations*, p. 4. Alliance of Digital Humanities Organizations.

Andrei N Kolmogorov (1963): On Tables of Random Rumbers. In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 369–376.

John Lafferty, Andrew McCallum & Fernando Pereira (2001): Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289. Morgan Kaufmann Publishers Inc.

Kirsopp Lake (1911): *Codex Sinaiticus Petropolitanus*. Oxford.

Alon Lavie & Michael J Denkowski (2009): The METEOR Metric for Automatic Evaluation of Machine Translation. In: *Machine Translation*, 23(2): 105–115.

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*, pp. 472–479. ACL.

Vladimir I Levenshtein (1965): Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. In: *Doklady Akademii Nauk SSSR*, 163(4): 845–848. (1966) Russisch, Englische Übersetzung. In: Soviet Physics Doklady Vol. 10, No. 8: 707–710.

Ming Li & Paul Vitáni (2008): *An Introduction to Kolmogorov Complexity and Its Applications 3. Auflage*. Springer New York.

Percy Liang, Ben Taskar & Dan Klein (2006): Alignment by Agreement. In: *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 104–111. ACL.

Dekang Lin (1998): An Information-Theoretic Definition of Similarity. In: *Proceedings of the 15th International Conference of Machine Learning (ICML)*, pp. 296–304. Citeseer.

Meili Lu, X. Sun, S. Wang, D. Lo & Yucong Duan (2015): Query Expansion via Wordnet for Effective Code Search. In: *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, pp. 545–549. IEEE.

Christoph Mackert (2014): Luthers Handexemplar der hebräischen Bibelausgabe von 1494. Objektbezogene und besitzgeschichtliche Aspekte. In: Irene Dingel und Henning

P. Jürgens (ed.) *Meilensteine der Reformation. Schlüsseldokumente der frühen Wirksamkeit Martin Luthers*, pp. 70–78, 255–257. Gütersloher Verlagshaus.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett & Ruth Reeves (1998): Nomlex: A lexicon of nominalizations. In: *Proceedings of EURALEX*, volume 98, pp. 187–193.

Nitin Madnani & Bonnie J Dorr (2010): Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. In: *Computational Linguistics*, 36(3): 341–387.

Nitin Madnani, Joel Tetreault & Martin Chodorow (2012): Re-examining Machine Translation Metrics for Paraphrase Identification. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 182–190. ACL.

David Malone (2010): Julia Smith bible translation (1876). `https://recollections.wheaton.edu/2010/12/julia-smith-bible-translation-1876/`. Accessed: Jan. 2019.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard & David McClosky (2014): The Stanford CoreNLP Natural Language Processing Toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60. ACL.

Christopher D Manning & Hinrich Schütze (1999): *Foundations of Statistical Natural Language Processing*. MIT press.

Mitchell P Marcus, Mary Ann Marcinkiewicz & Beatrice Santorini (1993): Building a Large Annotated Corpus of English: The Penn Treebank. In: *Computational Linguistics*, 19(2): 313–330.

Michael Marlowe (1867–1884): John Nelson Darby's Version. `http://www.bible-researcher.com/darby.html`. Accessed: Nov. 2017.

Thomas Mayer & Michael Cysouw (2014): Creating a Massively Parallel Bible Corpus. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 3158—-3163. ELRA.

Laurence Mellerin (2014): New Ways of Searching with Biblindex, the online Index of Biblical Quotations in Early Christian Literature. In: Claire Clivaz, Andrew Gregory & David Hamidovic (eds.) *Digital Humanities in Biblical, Early Jewish and Early Christian Studies*, chapter 11, pp. 175–192. Brill.

Laurence Mellerin (2016): Biblindex. `http://www.biblindex.mom.fr/`.

Bruce Metzger (1960): The Geneva Bible of 1560. In: , 17(3): 339.

Tomas Mikolov, Kai Chen, Greg Corrado & Jeffrey Dean (2013): Efficient Estimation of Word Representations in Vector Space. In: *Proceedings of the International Conference on Learning Representations*.

George Miller & Christiane Fellbaum (2007): WordNet. `http://wordnet.princeton.edu/`.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller (1990): Introduction to WordNet: An On-line Lexical Database. In: *International Journal of Lexicography (special issue)*, 3(4): 235–312.

Stefano Minozzi (2009): The Latin WordNet Project. In: Peter Anreiter & Manfred Kienpointner (eds.) *Latin Linguistics Today–Proceedings of the 15th International Colloquium on Latin Linguistics*, volume 137, pp. 707—716. Innsbrucker Beiträge zur Sprachwissenschaft.

Krisztián Monostori, Arkdy Zaslavsky & Heinz Schmidt (2000): Document Overlap Detection System for Distributed Digital Libraries. In: *Proceedings of the fifth ACM conference on Digital libraries*, pp. 226–227. ACM.

Maria Moritz (2018): On the Impact of Time Proximity on the Alignment of Spelling Variants in Old English Bibles: A Case Study. In: *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities*, pp. 1–9. Gerastree, Wien.

Maria Moritz & Marco Büchler (2017): Ambiguity in Semantically Related Word Substitutions: an investigation in historical Bible translations. In: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pp. 18–23. 133, Linköping University Electronic Press.

Maria Moritz & David Steding (2018): Lexical and Semantic Features for Cross-lingual Text Reuse Classification: an Experiment in English and Latin Paraphrases. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pp. 1976–1980. ELRA.

MySword (2011–2018): MySword. `www.mysword.info/`. Accessed: July 2018.

Daniel Naber (2005): OpenThesaurus: ein offenes deutsches Wortnetz. In: *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung, Bonn, Germany*, pp. 422–433.

Roberto Navigli & Simone P Ponzetto (2012): BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. In: *Artificial Intelligence*, 193: 217–250.

Saul B Needleman & Christian D Wunsch (1970): A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. In: *Journal of Molecular Biology*, 48(3): 443–453.

Cardinal John Henry Newman (1859): The Text of the Rheims and Douay Version of Holy Scripture. In: *The Rambler*, 1(New Series, Part II): 145–169.

Mark EJ Newman (2005): Power Laws, Pareto Distributions and Zipf's Law. In: *Contemporary Physics*, 46(5): 323–351.

Franz Josef Och & Hermann Ney (2000): Improved Statistical Alignment Models. In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 440–447. ACL.

Ahmed Hamza Osman, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb & Albaraa Abuobieda (2012): An Improved Plagiarism Detection Scheme Based on Semantic Role Labeling. In: *Applied Soft Computing*, 12(5): 1493—-1502.

Hans Otte (2014): *Pietismus und Neuzeit*, volume 40, chapter Halle, Stuttgart und anderswo. Zur Bedeutung der Bibelgesellschaften im 19. Jahrhundert. Vandenhoeck & Ruprecht.

Gustavo Paetzold (2015): Morph adorner toolkit: Morph adorner made simple. `http://morphadorner.northwestern.edu`. Accessed: Jan. 2019.

Kishore Papineni, Salim Roukos, Todd Ward & Wei-Jing Zhu (2002): Bleu: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. ACL.

Marco Passarotti (2018): LiLa: Linking Latin: Building a Knowledge Base of Linguistic Resources for Latin. `https://lila-erc.eu/`. Accessed: Dec. 2018.

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

Barbara Plank, Anders Johannsen & Željko Agić (2016): Improving language technology with fortuitous data: Course held at the ESSLLI summer school, August 15-19, Bozen-Bolzano. `https://fortuitousdata.github.io/`. Accessed: Dec. 2018.

Alfred W. Pollard (2003): *Biographical Introduction. The Holy Bible: 1611 Edition, King James Version*. Hendrickson.

Hugh Pope (1910): The Origin of the Douay Bible. In: *The Dublin Review*, 147(294-295): 97–118.

Maja Popović & Hermann Ney (2007): Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. In: *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 48–55. ACL.

Martin F Porter (1980): An algorithm for suffix stripping. In: *Program*, 14(3): 130–137.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein & Paolo Rosso (2011): Cross-language plagiarism detection. In: *Language Resources and Evaluation*, 45(1): 45–62.

Alfred Rahlfs (ed.) (1935): *Septuaginta, id est Vetus Testamentum Graece juxta LXX interpretes*. Württembergische Bibelanstalt. 2 vol., 1950.

Editorial Team of Revised Version (1989): *The Holy Bible. Revised Version*. London: Cambridge University Press. Synopsis.

Paul Rießler & Rupert Storr (1934): Mainzer Bibel Grünewald Bibel. `https://www.bibelpedia.com/index.php?title=Rie%C3%9Fler,_P_%26_Storr,_R.` Accessed: July 2018.

Leonardo Rigutini, Marco Maggini & Bing Liu (2005): An EM Based Training Algorithm for Cross-Language Text Categorization. In: *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 529–535. IEEE Computer Society.

Nicole Roger (1958, 1959): New Testament Use of the Old Testament. p. 137–51. Baker, Grand Rapids. The Tyndale Press, London.

G. Salton, A. Wong & C. S. Yang (1975): A Vector Space Model for Automatic Indexing. In: *Communications of the ACM*, volume 18, pp. 613—-620. ACM.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh & Kentaro Inui (2018): Cross-Lingual Learning-to-Rank with Shared Representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 458–463. ACL.

Philip Schaff (1858–1890): Luther's Translation of the Bible. In: *History of the Christian Church*.

Silke Scheible, Richard J Whitt, Martin Durrell & Paul Bennett (2011): A Gold Standard Corpus of Early Modern German. In: *Proceedings of the 5th Linguistic Annotation Workshop*, pp. 124–128. ACL.

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

Thomas Schoenemann (2010): Computing Optimal Alignments for the IBM-3 Translation Model. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp. 98–106. ACL.

Karin Kipper Schuler (2005): *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, USA.

Claude E Shannon (1948): A Mathematical Theory of Communication. In: *The Bell System Technical Journal*, 27(3): 379–423.

Helen Shenton (2009): Virtual Reunification, Virtual Preservation and Enhanced Conservation. In: *Alexandria*, 21(2): 33–45.

Tianze Shi, Zhiyuan Liu, Yang Liu & Maosong Sun (2015): Learning Cross-lingual Word Embeddings via Matrix Co-factorization. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 567–572. ACL.

David A Smith, Ryan Cordell, Elizabeth Maddock Dillon, Nick Stramp & John Wilkerson (2014): Detecting and Modeling Local Text Reuse. In: *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 183–192. IEEE Press.

Temple F Smith & Michael S Waterman (1981): Comparison of biosequences. In: *Advances in Applied Mathematics*, 2(4): 482–489.

William Smith (2010): *Catholic Church Milestones: People and Events That Shaped the Institutional Church*. Indianapolis: Left Coast.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul (2006): A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*. The Association for Machine Translation in the Americas.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul (2008): TRANSLATION ERROR RATE. `http://www.cs.umd.edu/%7Esnover/tercom/`.

Richard Socher, Cliff C Lin, Chris Manning & Andrew Y Ng (2011): Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 129–136.

Yangqiu Song & Dan Roth (2015): Unsupervised Sparse Vector Densification for Short Text Similarity. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1275–1280. Association for Computational Linguistics.

Robert Speer & Catherine Havasi (2012): Representing General Relational Knowledge in ConceptNet 5. In: *LREC*, pp. 3679–3686.

Efstathios Stamatatos (2009): Intrinsic Plagiarism Detection Using Character n-gram Profiles. In: *Proceedings of the 3rd International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, PAN'09*, pp. 38—-46.

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi & Thomas Griffiths (2004): Probabilistic Author-topic Models for Information Discovery. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 306–315. KDD '04, ACM.

Christina Stockmann-Hovekamp (1991): *Untersuchungen zur Strassburger Druckersprache in den Flugschriften Martin Bucers: graphematische, morphologische und lexikologische Aspekte*. Carl Winter, Universitätsverlag, Heidelberg.

Open Multilingual Wordnet Team (2018): Open Multilingual Wordnet. `http://compling.hss.ntu.edu.sg/omw/`. Accessed: Dec. 2018.

Ken Thompson (1968): Programming techniques: Regular expression search algorithm. In: *Communications of the ACM*, 11(6): 419–422.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga & Hassan Sawaf (1997): Accelerated DP Based Search for Statistical Translation. In: *Proceedings of Eurospeech-97*, pp. 2667–2670.

William Tyndale (1989): *Tyndale's New Testament*. Yale University Press.

Karl-Heinz Vanheiden (2018): NeÜ. Vorwort. bibel.heute. `https://neue.derbibelvertrauen.de/#vorwort`. Accessed: July 2018.

Tamás Váradi, Peter Wittenburg, Steven Krauwer, Martin Wynne & Kimmo Koskenniemi (2008): CLARIN: Common language resources and technology infrastructure. In: *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

C. Vercellonis & J. Cozza (1868): *Bibliorum Sacrorum Graecus Codex Vaticanus*. Roma.

Aleksi Vesanto, Asko Nivala, Heli Rantala, Tapio Salakoski, Hannu Salmi & Filip Ginter (2017): Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771-1910. In: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pp. 54–58. 133, Linköping University Electronic Press.

M. Vinzent, Laurence Mellerin & Hugh A G Houghton (eds.) (2013): *Biblical Quotations in Patristic Texts (Studia Patristica 54)*. Theory and Applications of Natural Language Processing, Peeters, Leuven.

Stephan Vogel, Hermann Ney & Christoph Tillmann (1996): HMM-Based Word Alignment in Statistical Translation. In: *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pp. 836–841. ACL.

Piek Vossen (1998): Introduction to Eurowordnet. In: *EuroWordNet: A multilingual database with lexical semantic networks*, pp. 1–17. Springer.

Denny Vrandečić & Markus Krötzsch (2014): Wikidata: A Free Collaborative Knowledgebase. In: *Communications of the ACM*, 57(10): 78–85.

Ivan Vulić (2010): Term Alignment. State of the Art Overview. In: .

Gribomont J. Weber R., Fischer B. (ed.) (1969, 1994, 2007): *Biblia Sacra Juxta Vulgatam Versionem*. Deutsche Bibelgesellschaft.

Noah Webster (1833): Preface of the Noah Webster Bible. `https://www.hbu.edu/museums/dunham-bible-museum/reprints-from-the-collection/prefaces-to-major-bible-editions/noah-webster-bible-1833/`. Accessed July 2018.

John Wieting, Mohit Bansal, Kevin Gimpel & Karen Livescu (2015): Towards Universal Paraphrastic Sentence Embeddings. In: *CoRR*, abs/1511.08198.

Matthew Wilkens (2008): Evaluating POS Taggers: Basic MorphAdorner Accuracy. `https://mattwilkens.com/2008/11/22/evaluating-pos-taggers-accuracy-1/`. [Acc. Nov. 2017].

Wei Xu, Chris Callison-Burch & Bill Dolan (2015): SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT). In: *Proceedings of the 9th International Workshop on Semantic Evaluation*, pp. 1–11. ACL.

Yi Yang & Jacob Eisenstein (2016): Part-of-Speech Tagging for Historical English. In: *CoRR*, abs/1603.03144.

Robert Young (1898a): Young's Literal Translation. `http://www.bible-researcher.com/young.html`. Accessed Jan. 2019.

Robert Young (1898b): Young's Translation: Publisher's Note and Preface. `http://www.ccel.org/bible/ylt/ylt.htm`. Accessed Jan. 2019.

Britta Zeller, Jan Snajder & Sebastion Padó (2013): DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1201–1211.

Chi Zhang, Shagan Sah, Thang Nguyen, Dheeraj Peri, Alexander Loui, Carl Salvaggio & Raymond Ptucha (2017): Semantic sentence embeddings for paraphrasing and text summarization. In: *Signal and Information Processing (GlobalSIP), 2017 IEEE Global Conference on*, pp. 705–709. IEEE.

Shiqi Zhao, Xiang Lan, Ting Liu & 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*, pp. 834–842. ACL.

George Kingsley Zipf (1949): *Human behavior and the principle of least effort: An introduction to human ecology.* Addison-Wesley Press. Re-published 2016 at Ravenio Books.

Imed Zitouni (2014): Natural Language Processing of Semitic Languages.

Sven Meyer Zu Eissen & Benno Stein (2006): Intrinsic plagiarism detection. In: *European Conference on Information Retrieval*, pp. 565–569. Springer.

Maria Moritz
Birkenweg 7
OT Panitzsch
04451 Borsdorf

Ehrenwörtliche Erklärung zu meiner Dissertation mit dem Titel:
"Modification Analysis in Historical Paraphrastical Parallel Text - An Empirical Work on Stable and Changing Elements in Historical Text Reuse"

Sehr geehrte Damen und Herren,

hiermit erkläre ich, dass ich die beigefügte Dissertation selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel genutzt habe. Alle wörtlich oder inhaltlich übernommenen Stellen habe ich als solche gekennzeichnet.

Ich versichere außerdem, dass ich die beigefügte Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und, dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.

Göttingen, 28. Februar 2019