ENRIQUE MANJAVACAS ARÉVALO

COMPUTATIONAL APPROACHES TO INTERTEXTUALITY

Universiteit Antwerpen

Faculty of Arts
Department of Literature

COMPUTATIONAL APPROACHES TO INTERTEXTUALITY: FROM RETRIEVAL ENGINES TO STATISTICAL ANALYSIS

Thesis submitted for the degree of DOCTOR OF LINGUISTICS AND LITERATURE at the University of Antwerp to be defended by

ENRIQUE MANJAVACAS ARÉVALO

SUPERVISOR: Prof. Mike Kestemont

Antwerp, 2021

Enrique Manjavacas Arévalo: *Computational Approaches to Intertextuality: From Retrieval Engines to Statistical Analysis* © 2021

LOCATION:
Salobreña

TIME:
March, 2021

Dedicado a mis padres:
Sonsoles Arévalo Gonzalez & Santos Manjavacas Aguilera

ABSTRACT

A common device exploited by literary writers consists in the reuse of texts originally authored by others. In doing this, literary writers—often unconsciously—establish references to other texts and, thus, situate their work in relation to others within a large network of texts. This device of literary referencing is not only responsible for a particular artistic experience—which is based on the recognition of the links—, but has also the potential effect of enriching the interpretation and the meaning of the work in a larger context.

A famous example from classic western literature is Ovid's beginning to the *Amores*.

> "arma gravi numero violentaque bella parabam"
>
> ("I was planning to write about arms and violent wars using a heavy meter")
>
> *Ov. Am. 1.1*

where Ovid reuses the beginning of Vergil's *Aeneid*:

> "arma virumque cano Troiae qui primus ab oris"
>
> ("I sing of arms and the man, who first [came] from the shores of Troy")
>
> *Verg. A. 1.1*

In this case, the recognition of the parallelisms at multiple levels—e. g. the reuse of the word "arma", the analogous syntactic construction using "-que" (en. and), the adoption of the theme of war ("violenta bella", en. violent wars) and the instruments of war ("arma", en. arms), as well as the resemblance in metrical pattern—not only constitutes a core artistic value to this passage, but also let us situate the *Amores*—a love elegy—in relation to the epic literature of Vergil, to which it relates from the point of view of literary genres.

The study of these referential processes in literary works has been systematized through the theory of "intertextuality", whose application can considerably profit from the identification of parallelisms at scale. In particular, the identification of biblical references in Medieval religious literature constitutes a task in which computational methods can make an impact, given the central role of the Bible in these writings.

This PhD thesis is concerned with the study of "intertextual" links from a computational point of view, and aims at fulfilling two goals.

The first one consists in improving the capacities of automated retrieval systems targeting cases of intertextual links. The second one emphasizes the usage of data-driven approaches to the study of intertextual processes. In order to accomplish these goals, this PhD thesis focuses on the use case of biblical references in the Latin Patrology—a large body of Medieval religious writings.

The first chapter introduces the theoretical context and the problems involved in studying intertextuality from a computational perspective.

The second chapter surveys current approaches to the retrieval intertextual references, providing discussions of the main algorithmic families and evaluation scenarios relevant to literary scholars. The goal is to establish reference values of performance and evaluation procedures for viable retrieval algorithms, providing an informative picture of the current state of such a technology. The core of the chapter is shaped by a set of experiments in which the discussed retrieval systems are compared across several datasets, exploring multiple evaluative conditions.

The application of retrieval systems to historical languages—e. g. Medieval Latin—poses specific challenges due to morphological complexity and a non-standard orthography of these languages. In order to tackle this challenge, the third chapter deals with the task known as lemmatization—i. e. the derivation of dictionary head-words from inflected words—in the context of historical languages, and presents a state-of-the-art approach that relies on modern Machine Learning architectures based on Neural Networks.

The fourth chapter narrows the scope of the automatic retrieval onto the case of literary allusions in the writings of Bernard of Clairvaux—an influential religious writer from the 12th century. The chapter emphasizes the application of distributional semantics in the form of contemporary vectorial representations of word meaning—known as word embeddings—in order to capture the semantic fields on which the allusions are based.

In the fifth and sixth chapters, the perspective shifts from that of the retrieval of intertextual references to that of the statistical analysis of intertextual processes. In chapter five, the results of an exploratory analysis of a large sample of biblical references in the Latin Patrology is presented. The analysis seeks to classify intertextual references with respect to the level of lexical overlap and thematic correspondence. Moreover, the analysis illustrates the role of several contextual factors—e. g. the role of authorial style—in the placement of intertextual references.

In the sixth chapter, the PhD thesis pursues the problem of the objectivity of intertextual references through an experiment in which three domain experts on the writing of Bernard of Clairvaux are shown retrieved candidate intertextual links and have to decide on their validity. Statistical techniques are, then, applied in order to estimate the extent to which experts agree in their judgments. The resulting statistical models are used in order to assess the difficulty of producing annotated resources of intertextuality, as well as the utility of different algorithms on the basis of how controversial the retrieved candidates are.

The PhD thesis concludes with the seventh chapter, in which the results of the previous chapters are synthesized and pointers to future research are provided.

SAMENVATTING

Een veelvoorkomende stijlfiguur bij literaire schrijvers bestaat uit het hergebruik van teksten die oorspronkelijk door andere auteurs werden geschreven. Hierdoor creëren auteurs—vaak onbewust—verbanden met andere teksten en plaatsen hun werk op die manier in een groter netwerk van literaire teksten. Dergelijke literaire citaten ressorteren niet alleen een specifieke artistieke ervaring bij de lezer, maar kunnen potentieel ook de interpretatie of betekenis van een werk in een bepaalde context verrijken.

Een bekend voorbeeld uit de klassieke westerse literatuur is Ovidius' begin van de *Amores*.

> "arma gravi numero violentaque bella parabam"
>
> ("Ik was van plan om over wapenfeiten te schrijven en gewelddadige oorlogen in een bombastisch metrum")
>
> *Ov. Am. 1.1*

waar de schrijver de start van Vergilius' *Aeneas* recupereert:

> "arma virumque cano Troiae qui primus ab oris"
>
> ("Ik bezing de wapenfeiten en de man die als eerste van de kusten van Troje [kwam]")
>
> *Verg. A. 1.1*

In dit geval, zullen de herkenning van de parallelismes op verschillende niveaus—bv. het hergebruik van het woord "arma", de gelijkaardige syntactische constructie met "-que" (nl. *en*), de introductie van het oorlogsthema ("violenta bella", nl. gewelddadige oorlogen) en wapens ("arma", nl. wapens), maar ook de metrische overeenkomsten—niet alleen bijdragen aan de kern van de artistieke waarde van deze passage, maar ons bovendien ook in staat stellen om de *Amores*—een liefdesklacht—te waarderen in het licht van Vergilius' epische literatuur, waaraan de tekst qua genre-conventies schatplichtig is.

De studie van deze verwijzingsprocessen in literaire teksten is gesystematiseerd geworden in de zogenaamde intertekstualiteitstheorie en de toepassing van deze theorie zou bijzonder gebaat zijn bij de automatische detectie van dergelijke verwijzingen op grote schaal. Computationele methodes zouden een grote impact kunnen hebben, in het bijzonder waar het om Bijbelse referenties gaat in middeleeuwse religieuze teksten, vanwege de centrale positie die de Schrift in deze literatuur inneemt.

Dit proefschrift gaat in op de studie van dergelijke intertekstuele verbanden vanuit een computationeel perspectief en heeft twee doelen voor oog. Het eerste doel is de verbetering van de performantie van geautomatiseerde systemen voor de ophaling (*retrieval*) van intertekstuele verwijzingen, maar dan specifiek in literaire contexten. Het tweede doel betreft het inzetten van data-gedreven benaderingen in de studie van intertekstuele processen. Om dit tweede doel te bereiken, richt dit proefschrift zich op de casus van bijbelse referenties in de Latijnse Patrologie—een groot corpus aan middeleeuwse religieuze teksten.

Het eerste hoofdstuk van dit proefschrift biedt een theoretische inleiding tot de problemen die gemoeid zijn met de studie van intertekstualiteit vanuit een computationeel perspectief.

Het tweede hoofdstuk biedt een overzicht van hedendaagse benaderingen inzake het ophalen van intertekstuele referenties, met een bespreking van de belangrijkste families van algoritmes en evaluatiescenario's die relevant zijn voor literatuurwetenschappers. Het doel is om zicht te krijgen op realistische waardes, wat betreft de performantie en evaluatieve procedures van de praktische toepassing van deze technologie. De kern van dit hoofdstuk wordt gevormd door een reeks experimenten waarin de voormelde ophalingssystemen worden vergeleken over verschillende datasets heen en zo een verkenning bieden van verschillende contexten waarin deze systemen kunnen worden geëvalueerd.

De toepassing van ophalingssystemen op historische talen, zoals middeleeuws Latijn, stelt ons voor specifieke uitdagingen vanwege de morfologische complexiteit en niet-gestandardiseerde orthografie van deze talen. Om deze uitdaging aan te pakken, gaat het derde hoofdstuk in op een taak die bekend staat als "lemmatisering", d.w.z. de automatische afleiding van het lemma of hoofdwoord dat bij een geïnflecteerde woordvorm hoort, in de context van historische talen. Een state-of-the-art benadering wordt voorgesteld die gebaseerd is op moderne systemen voor machinaal leren aan de hand van neurale netwerken.

In het vierde hoofdstuk wordt de toepassing van de automatische ophaling vernauwd tot de casus van literaire allusies in het oeuvre van Bernardus van Clairvaux—een invloedrijk religieus auteur uit de twaalfde eeuw. Dit hoofdstuk benadrukt het potentieel van de distributionele semantiek, in de vorm van hedendaagse vectoriële representaties van woordbetekenissen—bekend als "word embeddings"—om de betekenisvelden te vatten waarop allusies zijn gestoeld.

In het vijfde en zesde hoofdstuk verschuift onze aandacht van de *ophaling* van intertekstuele referenties naar de statistische *analyse* van intertekstuele processen. In hoofdstuk vijf worden de resultaten voorgesteld van een verkennende analyse naar een grote steekproef van bijbelse referenties in de *Patrologia Latina*. Deze analyse tracht om intertekstuele verwijzingen onder te brengen in categorieën op basis van de mate van lexicale overlap en thematische overeenkomst. Bovendien illustreert deze analyse de rol van uiteenlopende contextuele factoren—zoals bijvoorbeeld de rol van de individuele schrijfstijl van een auteur—op de plaatsing van intertekstuele referenties.

In het zesde hoofdstuk, behandelt dit proefschrift het probleem van de objectiviteit van intertekstuele verbanden via een experiment waarin drie domeinexperten op het gebied van Bernardus' oeuvre kandidaat-referenties kregen aangeboden en dienden te beslissen of deze valide waren. Vervolgens werden statistische methodes toegepast om te bepalen in welke mate de experten met elkaar in overeenstemming waren in hun oordelen. De resulterende statistische modellen worden gebruikt om in te schatten hoe moeilijk het is om geannoteerde datasets voor intertextualiteitsonderzoek aan te leggen, maar ook om in kaart te brengen hoe nuttig de verschillende ophalingstechnieken uiteindelijk zijn—op basis van hoe controversieel de aangereikte verbanden bleken.

Het proefschrift sluit af met een zevende hoofdstuk, waarin de resultaten uit de voorgaande hoofdstukken worden samengevat en suggesties worden geformuleerd voor toekomstig onderzoek.

RESUMEN

Un recurso literario del que muchos escritores se valen con frecuencia consiste en la reutilización de textos cuya autoría se debe a otros escritores. Por medio de esta práctica se establecen conexiones con otros textos—a menudo de manera inconsciente—y se sitúa la obra propia dentro de una gran red textual. El recurso de la referencia literaria no solo desencadena ciertas experiencias artísticas basadas en la identificación de dichas conexiones, sino que también tiene el potencial efecto de enriquecer la interpretación y el sentido de la obra en un contexto mas amplio.

Un conocido ejemplo dentro de la literatura clásica occidental es el comienzo de los *Amores* de Ovidio:

> "arma gravi numero violentaque bella parabam"
>
> ("Pensaba escribir sobre armas y violentas guerras usando una métrica pesada")
>
> *Ov. Am. 1.1*

donde Ovidio reutiliza el comienzo de la *Eneida* de Virgilio:

> "arma virumque cano Troiae qui primus ab oris"
>
> ("Canto a las armas y al hombre que [vino] primero de las costas de Troya")
>
> *Verg. A. 1.1*

En este caso, la identificación de los paralelismos en varios niveles como, por ejemplo, la reutilización de la palabra "arma", la construcción sintáctica análoga usando "-que" (es. y), la adopción del tema de la guerra ("violenta bella", es. guerras violentas) y de las armas ("arma", es. armas), así como el parecido en la métrica, no solo conforman un valor artístico fundamental, sino que también nos permiten situar los *Amores* de Ovidio—una elegía amorosa—con respecto a la literatura épica de Virgilio, con la cual guarda una íntima relación desde el punto de vista de los géneros literarios.

El estudio de este tipo de procesos referenciales en obras literarias ha sido sistematizado por la teoría de la intertextualidad, la aplicación de la cual en el campo de los estudios literarios puede beneficiarse considerablemente de la identificación de paralelismos a gran escala. Concretamente, la identificación de referencias bíblicas en la literatura religiosa medieval constituye un problema en el que las metodologías computacionales pueden marcar una gran diferencia, dado que la Biblia juega un papel fundamental en dicha literatura.

La presente tesis doctoral se ocupa del estudio de las conexiones "intertextuales" desde un punto de vista computacional, y pretende alcanzar dos metas. La primera consiste en mejorar la capacidades de los sistemas automatizados de búsqueda de nexos intertextuales en contextos literarios. La segunda enfatiza el uso de metodologías impulsadas por datos en el estudio de procesos de intertextualidad. Con el fin de lograr dichas metas esta tesis doctoral profundiza en el campo de las referencias bíblicas en la Patrología Latina—una amplia colección de escritos religiosos medievales.

El primer capítulo presenta el contexto teórico y los problemas resultantes del estudio del fenómeno de la intertextualidad desde un punto de vista computacional.

El segundo capítulo inspecciona las metodologías actuales mas relevantes para la búsqueda automatizada de referencias intertextuales aportando una descripción de las principales familias de algoritmos y de los escenarios de evaluación mas relevantes para los estudiosos de la literatura. La meta es establecer una serie de valores de referencia del rendimiento de los algoritmos, aportando un panorama informativo sobre el estado actual de dichas tecnologías. El núcleo del capítulo lo conforma un conjunto de experimentos en los que se comparan los sistemas de búsqueda anteriormente descritos en base a varios registros de datos y explorando múltiples contextos de evaluación.

La aplicación de estos sistemas de búsqueda a lenguas históricas como, por ejemplo, el latín medieval, supone un especial desafío debido a la complejidad morfológica de dichas lenguas y a una ortografía carente de estándar. Con el fin de abordar estos desafíos, el tercer capítulo se ocupa del problema conocido como "lematización"—es decir, la derivación automática del lema del diccionario a partir de la forma conjugada de la palabra—en el contexto de las lenguas históricas, y presenta un innovador sistema de lematización que usa arquitecturas contemporáneas de aprendizaje máquina (machine learning) basadas en redes neuronales.

El cuarto capítulo restringe el ámbito de aplicación a la búsqueda automática de alusiones literarias en los escritos de Bernardo de Claraval—un autor influyente del siglo XII. En este capítulo se enfatiza la aplicación de modelos de semántica distribucional basados en representaciones vectoriales del significado de las palabras—modelos conocidos como "word embeddings"—con el fin de registrar los campos semánticos sobre los cuales se establecen dichas alusiones literarias.

En los capítulos quinto y sexto la perspectiva se desvía de la búsqueda automática de referencias intertextuales y se centra en el análisis estadístico de los procesos intertextuales. En el capítulo quinto se presentan los resultados de un análisis exploratorio de una amplia muestra de referencias bíblicas tomadas de la Patrología Latina. El análisis ilustra el rol de varios factores contextuales—por ejemplo, el estilo del autor—en la localización de referencias intertextuales.

En el sexto capítulo se persigue el problema de la realidad objetiva de las referencias intertextuales a través de un experimento en el que tres expertos en Bernardo de Claraval han de juzgar la validez de una serie de nexos intertextuales extraídos computacionalmente. A través del uso de diversas técnicas estadísticas se consigue estimar el nivel de acuerdo obtenido por los expertos en sus juicios de valor. Los modelos estadísticos resultantes se usan a continuación para determinar la dificultad inherente a la producción de recursos lingüísticos en torno a la intertextualidad, así como para evaluar la utilidad de los diferentes algoritmos de búsqueda en base a la controversialidad de los nexos intertextuales que éstos extraen.

La presente tesis doctoral finaliza con el capítulo séptimo en el que se sintetizan los resultados de los capítulos anteriores y se sugieren campos de investigación para el futuro.

PUBLICATIONS

Substantial parts of the research presented in this thesis have appeared previously in the following publications:

Manjavacas Arévalo, **Enrique**, Ákos Kádár, and Mike Kestemont (June 2019). "Improving Lemmatization of Non-Standard Languages with Joint Learning." In: *Proceedings of the 2019 Conference of the North*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1493–1503. DOI: `10.18653/v1/N19-1153`. URL: `http://aclweb.org/anthology/N19-1153`.

Manjavacas Arévalo, **Enrique**, Folgert Karsdorp, and Mike Kestemont (2020). "A Statistical Foray into Contextual Aspects of Intertextuality." In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)* (Amsterdam, The Netherlands, Nov. 18, 2020–Nov. 20, 2020). CEUR Workshop Proceedings 2723. Aachen, pp. 77–96. URL: `http://ceur-ws.org/Vol-2723/long28.pdf`.

Manjavacas Arévalo, **Enrique**, Brian Long, and Mike Kestemont (June 2019). "On the Feasibility of Automated Detection of Allusive Text Reuse." In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 104–114. DOI: `10.18653/v1/W19-2514`. URL: `https://www.aclweb.org/anthology/W19-2514`.

OTHER PUBLICATIONS

The research presented in this thesis has been influenced by the following publications:

Emmery, Chris, **Enrique Manjavacas Arévalo**, and Grzegorz Chrupała (Aug. 2018). "Style Obfuscation by Invariance." In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 984–996.

Karsdorp, Folgert, Peter Kranenburg, and **Enrique Manjavacas Arévalo** (Nov. 2019). "Learning Similarity Metrics for Melody Retrieval." In: *Proceedings of the 20th International Society for Music Information Retrieval Conference* (Delft, The Netherlands). Delft, The Netherlands: ISMIR, pp. 478–485. DOI: `10.5281/zenodo.3527848`. URL: `https://doi.org/10.5281/zenodo.3527848`.

Karsdorp, Folgert, **Enrique Manjavacas Arévalo**, and Mike Kestemont (Oct. 2019). "Keepin' It Real: Linguistic Models of Authenticity Judgments for Artificially Generated Rap Lyrics." In: *PLOS ONE* 14.10, pp. 1–23. DOI: `10.1371/journal.pone.0224152`.

Manjavacas Arévalo, **Enrique**, Jeroen De Gussem, Walter Daelemans, and Mike Kestemont (2017a). "Assessing the Stylistic Properties of Neurally Generated Text in Authorship Attribution." In: *Proceedings of the Workshop on Stylistic Variation*. Association for Computational Linguistics, pp. 116–125. DOI: `10.18653/v1/W17-4914`. URL: `http://aclweb.org/anthology/W17-4914`.

Manjavacas Arévalo, **Enrique**, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont (Sept. 2017b). "Synthetic Literature: Writing Science Fiction in a Co-Creative Process." In: *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*. Santiago de Compostela, Spain: Association for Computational Linguistics, pp. 29–37. DOI: `10.18653/v1/W17-3904`.

Manjavacas Arévalo, **Enrique**, Mike Kestemont, and Folgert Karsdorp (2019). "Generation of Hip-Hop Lyrics with Hierarchical Modeling and Conditional Templates." In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 301–310. DOI: `10.18653/v1/W19-8638`.

Nature is like a genie
that answers a question truthfully
but only exactly as it is asked.

— Judea Pearl (Pearl and Mackenzie, 2018)

ACKNOWLEDGMENTS

A dissertation like the present one doesn't write itself. Besides, and not less obviously, the present dissertation isn't the sole work of the one appearing on the book cover, and that is why the following acknowledgments are due.

Thinking retrospectively about the people who have supported the present dissertation, I can hardly distinguish the bits contributed by others that are purely academic and those other bits that influenced me in a deeper manner, making me a richer and better person. Moreover, I cannot restrict myself to the period of 4 years that my PhD officially lasted, since many of the influences can be traced back to the beginnings of my work in Antwerp in 2014.

Chronologically speaking, I must start by thanking my previous supervisor, Peter Petré, who not only gave me the opportunity to start an academic career, but also ensured a smooth transition into a PhD before my contract was due. During my stay with the MBG team, I learned what it means to enjoy an engaging and hugely pleasant working environment, and was lucky to meet an inspiring group: Lynn, Sara, Odile, Emma-Louise, and, especially, Will, who so eloquently has helped me to endure the less beautiful bits of the PhD, and so effusively has joined me when the better ones called for celebration.

The most heartfelt acknowledgment goes to my supervisor, Mike, who has always been so keen on enriching my skills, enlarging the scope of my research, and ensuring that my projects can be carried out in the swiftest and smoothest manner—even before our relation as supervisor and supervisee became official. He never stopped giving me the confidence and freedom to focus on whatever research topics came closer to my heart. Besides, I must thank Mike for always being a joyful companion, and a honest and supportive friend when things went less good.

I am indebted to my friends and colleagues at CLiPS, who throughout all these years have shared not only space and time, but also countless discussions and fun conversations with me. I want to thank Walter for so generously welcoming me in Antwerp in the first place, and granting me the opportunity to join his research team with no hesitation. Without his generosity, I would have never met such a supportive group of researchers. In order of appearance these have been Ben Verhoeven, Guy and Janneke, who offered guidance from the very beginning; Robert and Michèle, who formed an extremely supportive society at a crucial time; Giovanni, who ensured we followed up on a discussion point until the last consequences; Chris and Stéphan, whose commitment to proper code and algorithmic fineness was a constant source of inspiration; Lisa, Simon, Pieter and Tim, who proved to be engaging office mates during the years; Madhumita, who has been an invaluable travel companion into the slippery grounds of Deep Learning; Pietro, Nicolay, Ben Burtenshaw and Wouter, the splinter group around Mike I was so lucky to join; and, finally, Maxime, Ehsan, Jeska and Ilia, the latest additions to the team, whose wit and knowledge I would have enjoyed longer, had it not been for the status quo brought in by the pandemic.

Besides my fellow researchers at CLiPS, the extended family of friends and colleagues spreads to the Netherlands and includes Florian, Chris van der Lee and Thiago, whose company I have always enjoyed during meetings at the most varied places around the world. An special mention goes to Ákos, whose input, eagerness and motivation was of paramount importance for many parts of this work. I must also thank Thibault and Jean-Baptiste—the French connection—from whom I received lots of useful resources and feedback on the application of Neural taggers to historical languages.

I am exceptionally thankful to Folgert, my unofficial second supervisor, for so many exciting and educating challenges throughout our now years-long collaboration. Folgert brought me into the world of Natural Language Generation, introduced me into Bayesian statistical modeling and exposed me to the intriguing research field of Cultural Evolution. Finally, I want to thank Lauren, my current supervisor, for placing confidence in me even before this thesis was drafted.

Many people around me and outside the strict margins of academia have been of substantial help in the completion of the present work. First of all, I must thank Iolanda, on whose selfless support I have relied during so many years: I hope our paths cross again at some point. I must thank my friends in Belgium, especially, Cintia and Jonathan, who showed genuine interest on my research and have so warmly accompanied me during this time. Further thanks go to my friends in Germany—Felix, Isa, Sonia and Axel—who have given me the pleasure of contemplating the passing of time during our now traditional yearly meetings. To Pjönghwa, whose effortless endurance continues to be an inspiration. Earnest thanks go to my friends in Spain—Violeta, Andy, Luis, Luismi, Serena and Jei—for being there all along. I am especially thankful to Rocío, for bringing me lots of joy and support during difficult times and for being a pivotal point in my life—as well as Pincho, whose faithful company I have enjoyed as much as he has enjoyed mine.

The largest part of this thesis was written in the Andalusian outpost of Salobreña, to where I came thanks to my dearest friends Jan and Sara. I must thank them for embracing me here and providing the best possible environment for the elaboration of the present work.

Finally, I want to thank my brother, for providing the best possible example of what research ought to be, as well as my parents, who truly made this research possible with their endless love and support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

TPR	True Positive Rate
FPR	False Positive Rate
P	Precision
R	Recall
P@K	Precision@K
AP	Average Precision
AUC	Area Under the Curve
ROC	Receiving Operator Characteristic
MAP	Mean Average Precision
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
LSH	Locality Sensitive Hashing
VSM	Vector Space Model
CV	Cross Validation
GVSM	Generalized Vector Space Model
ROPE	Region Of Practical Equivalence
LDA	Latent Dirichlet Allocation
LM	Language Modeling
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory

LSI	Latent Semantic Indexing
BOW	Bag Of Words
JSD	Jensen-Shannon Divergence
KLD	Kullback-Leiber Divergence
ELPD	Expected Log (Pointwise) Predictive Density
LOO	Leave One Out
MCMC	Markov Chain Monte Carlo
WAIC	Widely Applicable Information Criterion
NLP	Natural Language Processing
IR	Information Retrieval
CSLS	Cross-Domain Similarity Local Scaling

1 INTRODUCTION

This PhD thesis is concerned with the application of computational means to the extraction and study of textual references among texts. Literary authors do not conceive of their works in isolation. Instead, their writing constitutes a communicative act in which a set of past influences—including social and historical events as well as exposure to a specific cultural repertoire—is processed and addressed towards a future audience. In the quest to understand the resulting cultural artifacts, literary scholars often seek to trace the effects of these past influences and future audiences on the author's textual production. Certainly, the ways in which these external factors contribute to shaping the form and content of the literary work are varied. But, in particular, the realization that new readings of literary works emerge from the recognition of their connections to other literary works has proved an invaluable source of scholarship. Starting in the 1960s, a theory of the ways in which literary texts relate to each other and the implications of these links for literary scholarship crystallized around the term of "intertextuality" (Allen, 2000; Orr, 2003). While the development of intertextuality theory is a relatively recent event, a field in literary scholarship that has greatly profited from its application is the study of Medieval Western literary traditions, and, in particular, the case of religious writings (Moyise, 2002).

In these writings, the Bible has contributed extensively to the final form, both from the point of view of the influences on the authors—who take the Bible as a model that they seek to explain, imitate and disseminate—and the intended audience—which was assumed to possess a profound knowledge of the Bible which allowed them to recognize subtle allusions (Stahlberg, 2008, Chapter 2).

This PhD thesis focuses on the topic of biblical intertextuality in Medieval Western literature in order to contribute to the advancement of computational methods in intertextuality research. For this purpose, the presented studies exploit two sources of biblical intertextuality: the case of intra-biblical intertextuality and the case of biblical intertextuality in the Latin Patrology. The former comprises not only instances of textual echoes of the Old Testament within different books of the New Testament but also the philological landmark of the parallel accounts of events in the synoptical Gospels. The latter refers to the work of the so-called "Church Fathers", responsible for an extensive tradition of biblical exegetical Christian literature that spans over a millennium of Latin writing (from the 3rd to the 13th centuries).

The focus of the PhD thesis lies on computational methods. The application of computational methods to aid the study of intertextuality has long appealed to literary scholars. Computational methods can offer insight in two ways. First, indirectly by helping researchers map cases of intertextual links in a more efficient and exhaustive manner than what they could possibly achieve through manual inspection. Second, in a direct way, by addressing a specific type of questions for which statistical models can deliver appropriate answers. In consonance with these two lines of quantitative research, we study two types of computational methods. The first one relates to algorithms commonly applied in the fields of Natural Language Processing (NLP) and Information Retrieval (IR) in order to extract textual references. In particular, we resort to retrieval algorithms based on the Vector Space Model (VSM) and seek to incorporate an often disregarded semantic component through the integration of distributional semantics in the form of word embeddings. The second one relates to recent developments in statistical modeling that have made it possible to fit complex models to the data through the mechanisms of Bayesian inference.

In the present study, our emphasis lies on the methodological challenges that computational approaches face. Specifically, we seek to stress the difficulties that intertextuality poses to computational approaches, underline the limitations and shortcomings of current approaches and, finally, stimulate progress by pointing out possibilities for future research. However, before delving into a discussion of these matters, we shall first introduce the theory of intertextuality more broadly.

1.1 WHAT IS INTERTEXTUALITY

A more precise definition of the theory of intertextuality emerges as the conceptualization of literary works as a dynamic network in which books acquire new interpretations and meanings through the recognition of "intertexts"—i. e. textual links—to other books.

The seminal conception of modern intertextuality theory, commonly attributed to J. Kristeva (Kristeva, 1967), rose to prominence in the late 1960s under the influence of post-structuralist readings of F. de Saussure's relational understanding of the linguistic sign, and Bakhtin's "dialogic" conceptualization of the linguistic utterance (Allen, 2000).

According to F. de Saussure (De Saussure, 2011), linguistic signs do not possess meaning on their own, but acquire them only through reference to other signs within the same linguistic system. They are, thus, relational. Translating this idea from the study of language to the study of literature, intertextuality posits that literary signs acquire meanings not just by referencing objects in the outside world but also through connections to other literary works on the basis of various kinds of resemblances. In the context of Western literature, an author conceiving of a story rooted on the motif of an eventful sea journey will hardly produce a narrative that forestalls the connections that readers, consciously or unconsciously, will draw to Homer's *Odyssey*.

The second major influence on the theory of intertextuality is Bakhtin's view of language as essentially "dialogic". Similarly to Saussure, Bakhtin also conceives of language as relational. In his view, the meaning of utterances depends both on their reception by others as well as on pre-existing utterances and discourses to which these utterances relate. But in contrast to Saussure's abstract view of language as a system that exists independently from its usage, Bakhtin's contends that language exists within a dynamic social process "as the product of the reciprocal relationship between speaker and listener, addresser and addressee" (Voloshinov and Bakhtin, 1986). The meaning and logic of utterances depends on the intention and context in which they have been used as well as on their future reception by others. Bakhtin's dialogism, thus, results in a dynamic view of literature that eschews the idea of books as complete end products. Instead, books must be understood through this dialogic—or, *avant la lettre*, intertextual—connection to past and future addressers and addressees, which is subject to change along with changes in the social context in which it lives.

From its conception, the idea of intertextuality greatly appealed to literary scholars. They saw in this concept a systematization of the scholarly practice of identifying and interpreting textual references that predates the emergence of intertextuality as a theory. Since intertextuality describes an exceptionally productive mechanism of meaning creation in the inception and reception of literary works, it also defines an ambitious research agenda. It is, thus, not a coincidence that, as philological research becomes more corpus-oriented, and the availability of digital editions spreads, more and more literary scholarship turns to computational methods as a means to conduct the systematic exploration of references in literary collections.

1.2 COMPUTATIONAL VIEW ON INTERTEXTUALITY

The advent of text digitization and the availability of computational methods have made it both possible and appealing to approach intertextuality from a computational perspective. If, as argued above, intertextuality posits an agenda for qualitative research, the same can be said about quantitative approaches, which thrive on large datasets and can be scaled up to address otherwise unfeasible research questions.

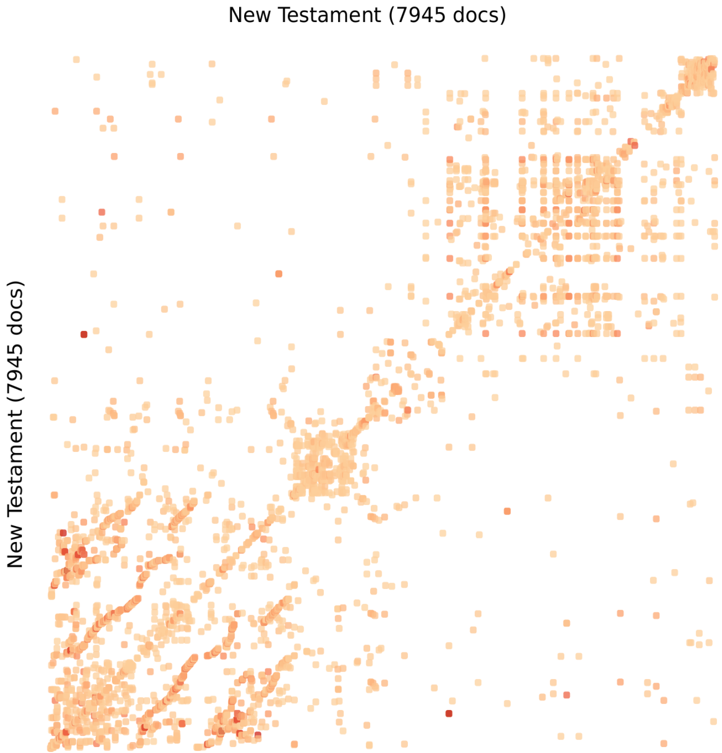


Figure 1.1: Dot-plot visualization of intra-biblical intertextuality within the New Testament. Color intensity corresponds to the degree of textual similarity between the related passages as assessed by a text alignment algorithm—discussed in detail in Section 2.5.2.3.

1.2.1 Finalities

In the present research landscape, three main use cases for large-scale computational applications can be identified in which particular methods are used in order to fulfill the needs of literary scholarship.

1.2.1.1 Visualization

The first one concerns the visualization of links across collections. This bird's-eye perspective on the specific relations identified across literary works is a frequent exploratory method—and an example of what is known as "distant reading" (Moretti, 2000)—that can facilitate the interpretation of collection-level intertextual patterns. A common approach is the "dot-plot" visualization, which charts detected intertexts in a two-dimensional scatter-plot, in which the respective axes represent the location of the relevant passages within either text (Jänicke et al., 2015; Yousef and Jänicke, 2021).

An example of the dot-plot visualization is shown in Figure 1.1, in which a text alignment algorithm is being deployed in order to identify structural parallelisms within the New Testament. In the resulting plot—which, in this case, is symmetric along the diagonal—each dot highlights a connection and the intensity of the color represents the textual similarity that underlies the connection. In the bottom-left corner, wiggly lines parallel to the diagonal correspond to parallelisms among the synoptic Gospels. Moreover, the checkerboard pattern in the upper-right comprises a set of recurrent expressions that are characteristic of the style of the Pauline epistles. Finally, the dense squares in the middle and in the top-right correspond, respectively, to John's Gospel and the Apocalypse, and indicate a writing style with a high degree of redundancy in the choice of expressions. In fact, these books have been traditionally attributed to a single author, although the hypothesis of single authorship has been discarded by modern scholarship—both qualitative (Ehrman, 1997) and quantitative (Erwin and Oakes, 2012).

1.2.1.2 *Study of Influence*

A second application can be found in the field of the study of influences (Bloom, 1973), where scholars have the goal of tracing the influence that an author has had on subsequent generations of writers. Singling out a recent example, it is worth referencing the HyperHamlet project,[1] which seeks to identify traces of Shakespeare on a large body of later literature and build an online database, in which these links can be searched for (Hohl Trillini, 2018). These resources can be used to find out, for instance, what Shakespearean plays and sonnets exerted the strongest influence at different times in history and on what genres or authors.

1.2.1.3 *Editorial Work*

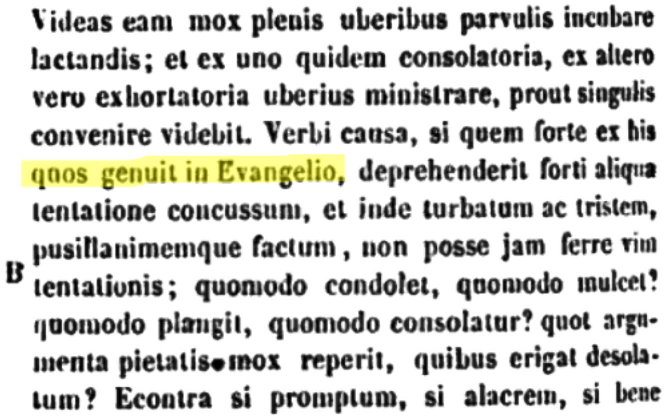
A third and final use case is to inform editorial work in which textual scholars seek to quasi-exhaustively identify the sources of borrowed passages in a given literary work for documentation and educational purposes. This is a common scenario in the preparation of scholarly editions of historical works, in which the text is supplemented by a "critical apparatus" documenting relevant meta-data (TEI Consortium, eds, 2020, Chapter 12).

An example of this practice from one of the digital editions used in the present PhD thesis is given in Listing 1.1, extracted from a modern born-digital re-edition of Bernard of Clairvaux's *Sermons on the Song of Songs*, carried out by the Sources Chrétiennes Institute. In the showed passage, XML markup is used—lines no. 9 to no. 22—to indicate an allusive reference to *1 Corinthians 4:15*. In this case, the modern edition has managed to identify a biblical link, where J.P. Migne's seminal edition from the 19th century—shown in Figure 1.2—had failed to recognize it (Migne, 1844-1855 (and 1862-1865)).

[1] Accessible via `http://www.hyperhamlet.unibas.ch/`

```
    <w xml:id="lat.w.sermo10.2.218">si</w>
    <w xml:id="lat.w.sermo10.2.219">quem</w>
    <w xml:id="lat.w.sermo10.2.220">forte</w>
    <w xml:id="lat.w.sermo10.2.221">ex</w>
 5  <w xml:id="lat.w.sermo10.2.222">his</w>
    <w xml:id="lat.w.sermo10.2.223">quos</w>
    <w xml:id="lat.w.sermo10.2.224">genuit</w>
    <w xml:id="lat.w.sermo10.2.225">in</w>
    <seg xml:id="lat.sQ.sermo10.2.a" type="scripturalQ">
10      <w xml:id="lat.w.sermo10.2.226">Evangelio</w>
        <span from="#lat.w.sermo10.2.226"
          to="#lat.w.sermo10.2.226"></span>
    </seg>
    <note xml:id="lat.sNote.sermo10.2.a"
15    n="a" place="foot" type="scripturalNote">
        <seg xml:id="lat.b.sermo10.2.1" type="bRef">
            <bibl type="biblical">
                <ref cRef="Vg:1_Co:4:15">1 Co 4, 15</ref>
            </bibl>
20      </seg>
        <link type="allusion"/>
    </note>
    <pc xml:id="lat.pc.sermo10.2.47">,</pc>
    <w xml:id="lat.w.sermo10.2.227">deprehenderit</w>
25  <w xml:id="lat.w.sermo10.2.228">forti</w>
    <w xml:id="lat.w.sermo10.2.229">aliqua</w>
    <w xml:id="lat.w.sermo10.2.230">tentatione</w>
```

Listing 1.1: Excerpt from a modern digital edition of Bernard of Clairvaux's 10th sermon from the book *Sermons on the Song of Songs,* highlighting the placement of a textual reference—lines 9 to 22—to *1 Corinthians 4:15* of allusive type.



Videas eam mox plenis uberibus parvulis incubare
lactandis; et ex uno quidem consolatoria, ex altero
vero exhortatoria uberius ministrare, prout singulis
convenire videbit. Verbi causa, si quem forte ex his
quos genuit in Evangelio, deprehenderit forti aliqua
tentatione concussum, et inde turbatum ac tristem,
pusillanimemque factum, non posse jam ferre vim
B tentationis; quomodo condolet, quomodo mulcet?
quomodo plangit, quomodo consolatur? quot argumenta
pietatis mox reperit, quibus erigat desolatum?
Econtra si promptum, si alacrem, si bene

Figure 1.2: Exerpt from J.P. Migne's 1859 edition of Bernard of Clairvaux's 10th sermon from the *Sermons on the Song of Songs*. Highlighted in yellow is the place of a missed biblical allusion.

1.2.2 Difficulties

The use cases mentioned above rely on fully operational intertextuality retrieval systems. However, as we shall see, the automatic detection of intertextuality faces a number of considerable challenges.

While the theory of intertextuality anticipates a large body of links between literary works, it does not necessarily pose any limitations on the material form that these links must take. In some cases, the links are supported by localized textual expressions that can be identified and isolated by readers. An illustrative example can be found in Herman Melville's *Moby Dick* (1851), where Shakespearean tragedies are frequently alluded to by borrowing short passages of text, or through the imitation of particular Shakespearean phrasings, in order to resort a dramatic effect with the reader (Matthiessen, 1968). Still, these intertexts can be subtle and require an enormous erudition on the part of the reader to be perceived (Olson, 1947). In other cases, intertexts involve a structural and organizatorial resemblance across entire book-length pieces—consider, for example, the role that Homer's *Odyssey* plays in the structuring of Joyce's *Ulysses* and Vergil's *Aenaid*, cases of literary adaptation that Genette termed "hyper-textuality" (Genette, 1982).

Due to sparsity and complexity reasons, computational approaches will struggle to automatically identify the latter type of intertexts. Those cases require a book-level understanding of narratives and offer but a few data points on which data-driven models can base the retrieval. In order to narrow down the scope of the computational treatment of intertexts, Forstall and Scheirer (2019) have recently introduced a helpful distinction between "large-scale effects"—comprising cases that can be hardly pin-pointed to specific passages—and "local effects" of intertextuality—consisting of localized phenomena such as motifs, quotations and allusions, in which links are established from and to specific passages. While large-scale effects remain outside of the scope of computational approaches to intertextuality, local effects—or "loci similes" in more traditional terms—are more amenable to computational modeling.

A second type of difficulty, relating especially to case studies dealing with historical languages, stems from the high degree of morphological variability and orthographical instability that these languages present. For languages with relatively simple morphology, computational models of intertextuality can still profit from the application of lemmatizers in order to abstract over morphological variants at a low cost. In fact, the lemmatization of standard modern languages—as well as other related tasks such as part-of-speech tagging or morphological analysis—is typically considered to be a "solved" task. The same, however, cannot be said of historical languages, which have so far often remained outside of the scope of traditional approaches in NLP, being mostly removed from common benchmark corpora—see (Piotrowski, 2012) for a specialized account. Still, given the emphasis on historical corpora of current intertextual studies—and certainly of the case studies in the present PhD thesis—, the problems posed by morphology transcend the status of a rarity, and deserve special attention. This situation highlights the fact that the relevance of certain topics is relative to the research community. While the NLP community has moved forward from lemmatization—focusing, instead, on the more general task of morphological re-inflection—, applications to literary studies find the current status of lemmatizers in some respects wanting.

A third difficulty relates to the general concern of the curation of linguistic resources for evaluation and model fitting purposes. Influenced by Bahktin's dialogistic view of literature as a dynamic social process in which both the addresser and addressee are responsible for the creation of meaning, more recent work on literary theory—in particular, reader-response criticism (Tompkins, 1980)—has shifted the focus from the author towards the reception of the work. The application of reader-response theory to intertextuality opens up a debate around the intentionality of the intertextual relationship, and challenges the reality of intertextual links beyond their perception by the reader. These matters pose challenges to computational approaches seeking to establish performance estimates of retrieval systems. In order for computational methods to make progress, evaluation resources are needed, which, in turn, depend on reliable manual annotation processes. However, the subjectivity implied in the process of identifying intertexts jeopardizes the required reliability. In fact, the curation of these resources is aggravated by reasons that go beyond ontological considerations on the existence of intertexts. As already mentioned, the recognition of an intertext requires a high degree of erudition, which for most case studies is hard to obtain. As a result, the cost of producing evaluation resources increases as the pool of potential annotators is dramatically reduced. At the same time, discussions about particular instances of intertextuality may turn into matters of scholarly debate on which agreement is hard to reach.

A final difficulty, that we shall stress, relates to the fact that in many cases the detection of intertexts is hard to achieve by computational means. In current applications of Machine Learning, researchers are often confronted with tasks like Object Classification that humans can do with high performance levels in few seconds—e. g. the time it takes for a human to identify a shown object as a cat. In those cases, computers can improve by leveraging large training datasets that were compiled at a low cost. In this vein, it has been hypothesized that if a task can be done by humans within seconds, it can be, thus, automated (Ng, 2016) by modern Deep Learning algorithms—known to efficiently leverage large quantities of labeled data (LeCun, Bengio, and Hinton, 2015). In other tasks, such as Authorship Attribution, computers succeed by exploiting abstract linguistic patterns which humans can hardly perceive. Still, in this case, the annotation is unproblematic, considering that the authorship of most texts is inherent to their publication, and, thus, learning algorithms can exploit these low-cost annotations to their own advantage. In order to tackle intertextual references, however, annotation is challenging for humans—as it has been mentioned above—and, as we shall see in this PhD thesis, computers need to model a set of varied set of syntactic and semantic phenomena as well as narrative structure and content analysis. Unfortunately in this case, the costs of compiling large-scale datasets means that computers cannot simply rely on data in order to improve.

1.3 TWO CULTURES

The use cases referred to above highlight certainly important applications of computational methods, which are, nevertheless, restricted to the retrieval of intertexts. In this PhD thesis, we argue that the scope for computational methods in intertextuality research transcends that of retrieval systems. A number of questions that revolve around intertextuality and that can receive proper quantitative treatment have already been hinted at, and many more could be envisioned.

First, a corpus-driven bottom-up derivation of intertextual typologies can contribute to theoretical debates in otherwise qualitative research contexts. Second, computational methods inspired from Authorship Attribution can help literary scholars to characterize distinct intertextual styles of authors, highlighting connections between the fields of intertextuality research and literary stylistics. Finally, a proper quantitative assessment of the agreement that can be expected from the expert judgments of literary scholars on the existence of particular links can shed light on the difficulties that preparing linguistic resources for intertextuality research actually entails. Interestingly, despite the appeal of these questions, current research has mostly focused on developing and deploying retrieval systems.

An interesting parallelism can be drawn to the situation in Statistical Modeling depicted by L. Breiman at the turn of the century (Breiman, 2001). In his seminal paper, Breiman described a rift inside statistics between, on the one hand, a data-modeling culture that is preoccupied with defining accurate models of the data generating process in order to explain the patterns observed in the data, and, on the other hand, an algorithmic culture that is concerned with producing predictive models beyond considerations for accurate statistical description. At the time, Breiman decried that statistical reservations on correctly modeling the data generating process were holding back research progress in a field meant to be solving real-world problems. Drawing a parallel—but inverting the terms—the current situation in quantitative intertextuality seems to be mostly preoccupied with solving the problem of intertext retrieval in a field where understanding the underlying data generating process can yield highly stimulating and equally valuable insights.

1.4 THESIS OVERVIEW

The present PhD thesis aims at advancing the current state of computational approaches to intertextuality, focusing on both aspects—the automatic retrieval of intertextual links, and the data-driven understanding of intertextual processes—as well as addressing some of the discussed difficulties that the field currently faces—most importantly, those relating to the curation of evaluation resources and the assessment and comparison of model performance.

1.4.1 Text Reuse Detection for Literary Texts

The evaluation of current retrieval algorithms is first approached in Chapter 2. We start by contextualizing the task of detecting intertextual connections within a set of related tasks in IR and NLP that go under the umbrella term of Text Reuse Detection—i. e. Plagiarism Detection, Paraphrase Identification, Semantic Textual Similarity and Historical Text Reuse Detection. Based on a discussion of the particularities and difficulties that intertextuality poses to computational retrieval approaches, we argue that future research would profit from a clearly delineated and differentiated treatment of this task—which we refer to by "Literary Text Reuse Detection"— avoiding the conflation with other related tasks.

A survey of current approaches in Literary Text Reuse Detection, highlighting the main strategies for evaluation, leads to the conclusion that the state of the art is difficult to assess due to the lack of benchmark corpora and established evaluation protocols. In the absence of benchmark corpora, most studies limit the scope of the evaluation to their own case study, often disregarding the definition of baselines and comparison to alternative approaches. However, researchers are often interested in the more general question of how the assessed performance of algorithms will generalize onto new datasets.

In order to alleviate the situation, we make the following contributions. In the theoretical part of the chapter, we target the goal of establishing a set of alternative approaches, identifying and discussing three main families of retrieval algorithms that underlie most current research. In order to facilitate the evaluation and interpretation of results, we consider three evaluation scenarios that relate to different use cases in the Humanities—one based on the classification paradigm, other on the retrieval paradigm and a hybrid one bridging between the two—, and define appropriate performance metrics. In the second part of the chapter, we focus on a systematic comparison of the discussed algorithms, targeting the question of the generalizability of results to unseen datasets. On the basis of three gold-standard datasets of intertextual references, spanning four corpora from different historical languages, we deploy modern Bayesian model comparison methods in order to compare the alternative methods across the evaluation scenarios.

Our experiments offer no evidence in support of a single best approach in absolute terms. Rather surprisingly, a bag-of-words VSM shows robust performance across different setups, and a semantically motivated extension to this method seems to prove beneficial, provided high-quality word embeddings can be obtained. Inspired by common Machine Learning evaluation practices, we conduct a set of transferability experiments, in which the performance costs incurred by fine-tuning a retrieval algorithm on a different but related dataset are assessed. The experiments show that algorithms differ in this aspect, not only depending on the number and type—e. g. categorical vs. numeric—of the hyper-parameters, but also with respect to the particularities of the target corpus. Finally, using Cross Validation (CV) we examine to what extent hyper-parameter fine-tuning on a subset leads to hyper-parameters that perform optimally on the entire dataset. Our experiments show that algorithms are likely to lag behind optimal performance in some evaluation contexts, and that in the absence of CV, evaluation procedures are likely to produce inflated estimates of performance.

1.4.2 Neural Lemmatization for Historical Languages

In Chapter 3, we approach the problem of lemmatizing historical languages. Lemmatization approaches range from methods that leverage lexical resources (i. e. lexicon-based approaches)—focusing on exhaustively listing possible morphological analyses congruent with the input token—to data-driven end-to-end approaches—leveraging annotated corpora to produce disambiguated predictions on the lemma underlying the input token in its sentential context. Due to its essentially non-disambiguating nature, the former approach can be considered less useful in the context of automatic identification of intertextual relations. Moreover, in the case of historical languages with non-standard orthography and high levels of morphological variation, common lexicon-based approaches struggle to provide robust morphological analyses since they typically rely on rule-based systems. A common solution in the field is to prepend a text normalization system (Baron and Rayson, 2008) to the pre-processing pipeline, and to run the lemmatizer on the normalized forms. The current PhD thesis, however, opts to explore data-driven lemmatization approaches that learn to lemmatize from annotated corpora while simultaneously abstracting over orthographical variation. This approach is conceptually simpler and is particularly promising when the patterns of orthographical variation in the target corpus resemble those in the training corpus.

Our study focuses on the comparison between current state-of-the-art lemmatizers based on Neural Encoder-Decoder architectures that learn to transduce the input token string into the target lemma, and a previous generation of lemmatizers that derive lemmata from input tokens through the application of binary edit-tree rules. These binary edit-tree rules are first induced from the training corpus and then become the target of a linear classifier that learns to predict which rule should be applied on a new input token. Our experiments show, first, that approaches based on binary edit-trees are very competitive on standard benchmark corpora, especially when Western European languages are most predominantly represented. The advantage of these approaches, however, becomes less apparent when languages of other morphological type are considered and, in the case of historical languages, are shown to lag behind modern neural lemmatizers.

Inspired by Multi-Task-Learning, we further enrich the neural lemmatizer with a joint Language Modeling (LM) loss. The LM objective—which has since become dominant in current NLP algorithms across tasks—consists in the task of predicting what word will appear next given the previous sequence of tokens. As current research in NLP shows, a LM objective guides the feature extractors implicitly learned by the neural network to model syntactic and semantic information that can help improving performance on the task of interest (Tenney et al., 2019). In our approach, we show that the joint loss helps the lemmatizer in retrieving the identity of the lemma underlying the morphologically and orthographically altered token. Furthermore, we conduct a series of so-called "probing" experiments, which confirm that the features extracted by the lemmatizer trained with the joint objective contain more morphologically relevant information than those extracted by the model learned with the simple lemmatization objective. The enriched neural lemmatizer sets a state-of-the-art result on lemmatization of historical languages and fulfills all lemmatization needs underlying the presented case studies.

1.4.3 Allusive Text Reuse Detection

Intertextual studies have been productive in terms of classifying intertexts and providing typologies according to factors such as the author's intentionality (Conte, 1988; Farrell, 2005; Knauer, 1965) or the function of the intertext in its context—e. g. parodic vs. satirical and non-satirical (Genette, 1982). In computational applications, the most frequent categorizations pertain to the axis of literality, which distinguishes intertexts with respect to how explicit the intertext is (Bamman and Crane, 2008; Büchler, 2013; Hohl Trillini and Quassdorf, 2010; Mellerin, 2014). While the distinction can be placed along a gradual continuum ranging from more to less literal quotations—based on the degree of re-phrasing and lexical overlap—, it is more common to divide intertexts into quotations, on the one hand, and allusions, on the other hand. In this case study, we focus on intertextual links that modern editors of the works of Bernard of Clairvaux have classified as allusions. These allusions are characterized by very low levels of lexical overlap—averaging around one token in the absence of lemmatization and just two after applying lemmatization. In view of such a fact, the question arises as to how semantic information can be integrated into retrieval systems and whether it can help improving retrieval performance.

On the basis of a dataset of over 600 allusions, we conduct a series of experiments, aiming at characterizing the difficulty of the task of Allusive Text Reuse Detection from two perspectives.

The first one consists of an assessment of the inter-annotator agreement on the task of identifying the span of the allusion. Annotators are asked to select the smallest span of Bernardine tokens that is maximally allusive to the target biblical verse. On the basis of a Fleiss's κ inspired inter-annotator agreement index for span annotations, we observe slightly compromising levels of agreement that question to what extent an automatic retrieval of such instances may be even feasible.

The second one focuses on the evaluation of retrieval models. We compare purely lexical models—based on bag-of-words representations and hand-crafted similarity functions targeting allusions—with purely semantic models—based on different applications of word embeddings like the bag-of-word embeddings and the Word Movers Distance—and, finally, hybrid retrieval models—which, making use of the Generalized Vector Space Model (GVSM), incorporate both lexical and semantic matching capabilities.

Our comparison shows, rather surprisingly, that purely semantic models strongly underperform their purely lexical counterparts. Furthermore, the hybrid model based on the GVSM is shown to provide a performance boost, especially when utilizing word embeddings in order to estimate word-level semantic similarity. A manual inspection of cases correctly retrieved by this model, but not by its purely lexical counterpart, helps highlighting the contexts in which the retrieval of allusions can be helped by lexical semantics—e. g. allusions relying on well represented semantic fields.

The results from both analyses, however, cast doubts over the general feasibility of retrieving this type of allusions from unrestricted corpora in a robust manner.

1.4.4 Contextual Aspects of Intertextuality

Retaking the discussion from Section 1.3 about the two cultures, Chapter 5 marks a transition from the algorithmic into the data-analysis "culture". We turn from an algorithmic perspective into intertextuality based around retrieval systems to a perspective based on the corpus-based statistical description of intertextual phenomena.

Our attempts at modeling allusive references in the works of Bernard of Clairvaux showed that this type of intertextual links is often supported by very scarce lexical evidence, and that the inclusion of word-level semantic information within the passages produces a beneficial but still mild performance boost. Motivated by this result, we question whether the placement of intertextual references is affected not only by lexical information—the degree of lexical overlap across the source and target passages—but also by the thematic context surrounding the reference.

We conduct a statistical analysis of a large dataset of annotated cases derived from the Latin Patrology and examine, first, whether intertextual links can be characterized along not one but two axes. The first axis of variation is represented by the already discussed continuum from literal to allusive referencing styles. A second axis, representing the extent to which the reference is thematically embedded within or disconnected from its context, is posited, and the question is examined as to whether this new axis can supplement the first axis. The hypothesis underlying this examination is that more or less allusive references may be more or less triggered by the thematic similarity between the contexts surrounding the source and target passages. In the case of more allusive references, a lower degree of thematic embedding represents a style of referencing in which ideas and motifs are transferred into a foreign context, whereas a high degree of thematic embedding characterizes a style in which the reader receives contextual cues and is prepared by the context in order to recognize the link. With the goal of providing a data-driven characterization of intertextual styles, we inspect whether the mentioned axes of lexical similarity and thematic embedding interact in particular ways.

Moreover, we look into additional factors of variation that could explain the placement of a particular link on the basis of the observed lexical similarity and thematic embedding. We first look at authorial signs, asking whether authors consistently prefer a more or less literal and thematically embedded style. Second, we look into the role of authority of the collection from which the sources originate—which in the case of the Latin Patrology corresponds to the Bible—, asking whether particular books trigger more or less literal references with more or less thematic preparation. Finally, we inspect whether particular topics can be observed to have an influence along the same terms.

Methodologically, we rely on two main computational tools. In order to capture thematic similarity, we resort to Latent Dirichlet Allocation (LDA) topic models. Secondly, in order to explore the variation in the data across all factors of interest, we deploy hierarchical multi-level statistical models—relying on modern Bayesian inference techniques in order to fit the models—, and study proportions of explained variance.

Our analysis shows that the thematic embedding axis presents high degrees of correlation with the lexical similarity axis—even after controlling for lexical overlap—and that the resulting plane of intertextual styles has, thus, reduced explanatory power for a bottom-up categorization of intertextual styles. However, the observed correlation varies across factors, being, for instance, less pronounced when characterizing the role of authority. An interesting finding in this regard is that references to the New Testament are more literal but not necessarily more thematically embedded. Inspection of the other two contextual factors shows that authorial style is best characterized in terms of the degree of lexical similarity, and that certain topics—like those relating to moral and philosophical questions—are likely to trigger more allusive intertextual styles.

1.4.5 Matters of Agreement

In the final chapter of this PhD thesis, we tackle the issue of inter-annotator agreement in intertextuality. The chapter is based on an inter-annotator agreement experiment in which three scholars involved in the modern edition of Bernard of Clairvaux's *Sermons on the Song of Songs* are tasked with annotating a set of promising biblical references. The criterion used for the annotation corresponds to whether they would consider incorporating the proposed reference into the critical apparatus of the edition—thus, enhancing the real-world significance of the study. In order to extract promising candidates, two competing algorithms—one capturing a more quotation-oriented intertextual style and a second one focusing more on semantic relations—are fine-tuned on the body of existing annotations and deployed on the remaining data.

We approach the matter of agreement with a double goal in mind. On the one hand, we aim at offering a complete picture of the reached agreement and the experimental factors that influence it. With this goal in mind, we, firstly, develop a statistical approach to estimating inter-annotator agreement indices. Based on a probabilistic formulation of Cohen's κ for multiple annotators, we compute the agreement indices using probability estimates derived from a multinomial multilevel statistical model, which we fit using modern Bayesian inference techniques. Basing the calculation on a statistical model allows us to control for a number of factors—including predictors as well as random effects—, thus increasing the accuracy of the agreement estimates. Moreover, the usage of Bayesian inference allows us to compute posterior densities of the estimates that directly incorporate uncertainty over the agreement indices. Finally, a post-experimental report offers help elucidating the main sources of disagreement.

As a second goal, we take advantage of the flexibility of the multilevel model in order to produce estimates of inter-annotator agreement, conditioned on the retrieval method underlying the proposed intertextual links. The differentiated estimates allow us to assess the relative utility of the deployed retrieval algorithms in terms of the controversiality of the intertexts that they propose.

Our study finds out that inter-annotator agreement can reach high levels under certain circumstances and with ample uncertainty, but that it is more likely to remain at lower levels under more realistic conditions. Moreover, the semantically inspired retrieval method is shown to produce slightly higher agreement scores. However, here again, the significance of the comparative result must not be over-estimated. Instead, it must be interpreted with respect to the uncertainty emanating from several contextual factors, including the level of lexical overlap between the borrowing and source passages and the biblical book from which the reference is taken.

2 TEXT REUSE DETECTION FOR LITERARY TEXTS

ABSTRACT In cultural studies, intertextual theory is concerned with the links that literary texts establish with each other through different types of textual referencing. Increasingly, computational methods are applied in this domain with an emphasis on the more narrowly defined task of Text Reuse Detection, which is occupied with the retrieval of reuse ranging from more or less exact quotations, to subtler allusions or paraphrases. While Text Reuse Detection in literary contexts bears obvious resemblances to a number of more established tasks in Computational Linguistics—such as Plagiarism Detection, Paraphrase Identification and Semantic Textual Similarity—it has a different finality and presents a number of challenges that we survey in this study. Specifically, we argue that progress in the field is currently hindered by the lack of representative benchmark corpora and clearly defined evaluation protocols. We report experiments with some of the main families of retrieval algorithms that can be discerned currently in the field—finger-printing approaches, text alignment approaches and VSMs. Additionally, we present a holistic set of measures to evaluate these, corresponding to different, real-word usage needs in the Humanities, and literary scholarship in particular. An important contribution of this study is that, in contrast to common practice, we evaluate the out-of-sample performance of the calibrated systems. The results are encouraging, but sobering in that they highlight the idiosyncratic nature of the available benchmarks, casting doubts over the feasibility of the task at a level that transcends that of an ad-hoc case study. We conclude the chapter with suggestions to stimulate future work in this domain and enumerate a number of open challenges that should be urgently addressed. The study is complemented by open-source code and datasets for replication purposes.

This chapter is based on Enrique Manjavacas Arévalo and Mike Kestemont (2021). *Evaluation in Text Reuse Detection for Literary Texts*. Forthcoming

2.1 INTRODUCTION

Text Reuse Detection refers to the task of automatically identifying passages in a text collection that have their origin in another text collection, and correctly mapping the identified instances of text reuse to the corresponding source passages. In literary studies, text reuse practices are commonly studied within the framework of "intertextuality", a theory which emphasizes the idea that literary works are permeated with references to other works, and acquire new readings and meanings through those links.

While the concept of intertextuality does not entail a strict definition of what counts as reuse, the type of link that it envisions goes certainly beyond mere textual similarity. However, in order for computational approaches to progress, a more concrete definition is needed, and, for this purpose, the distinction between large-scale effects and local effects of intertextuality introduced by Forstall and Scheirer (2019) offers help. While large-scale effects—such as the structural parallelism between Joyce's *Ulysses*, Vergil's *Aenaid* and Homer's *Odyssey* mentioned in Chapter 1—are difficult targets for computational methods, local effects—i. e. phenomena such as implicit quotations, motifs or allusions, in which a link is established from and to specific passages—are easier to operationalize.

In the context of local effects of intertextuality, the nature of the relationship between a reused passage and its source is varied and it can range from long (possibly re-phrased) quotations with high lexical overlap to single-word allusions and motifs with little or no lexical overlap, that primarily rely on semantic connections. These localized intertextual links have been categorized along different axes such as intentionality (Conte, 1988; Farrell, 2005; Knauer, 1965), function—parodic vs. satirical and non-satirical (Genette, 1982)—or the already mentioned one of "literality" (quotation vs. mention or allusion). This taxonomic activity has led to a considerable amount of intertext typologies, highlighting the complexity of the underlying phenomena. Moreover, as we shall discuss in Chapter 5, the variation in reuse styles can be shown to correlate with meaningful contextual variables of intertextuality such as the borrowing author or the source collection from which is being borrowed.

What is important for computational accounts of text reuse in literary texts is that intertextuality anticipates a large body of text reuse cases connecting literary works. Thus, intertextual theory implicitly lays out a research agenda in which computational approaches can play a substantial role: the automatic retrieval of intertextual links offers a promising research avenue in order to deepen our understanding of the co-dependencies between literary texts at scale.

In this chapter, we introduce and characterize Literary Text Reuse Detection as a task with distinct goals, applications and difficulties.

We argue that current research is hampered by a lack of systematic evaluation procedures and representative benchmark corpora, both of which are needed in order to stimulate scientific progress.

OUTLINE The remainder of this chapter is structured as follows. First, in Section 2.2, we situate the computational detection of intertextual links within the broader case of Text Reuse Detection, reviewing related tasks and common evaluation practices. Section 2.3 zooms in on evaluation practices in applications of Text Reuse Detection to literary cases in particular, and highlights the difficulties associated with it. In Section 2.4, we contribute a classification of three evaluation scenarios for computational approaches to text reuse that reflect realistic application cases in the Humanities. In Section 2.5, we detail Text Reuse Detection algorithms and categorize them into three broad families, covering most common approaches in current research. Next, in Section 2.6 we perform an set of experiments that provide a thorough comparison of the algorithmic families across a comprehensive set of corpora encompassing three different languages. In contrast to previous research, we deploy modern Bayesian evaluation techniques to assess the out-of-sample performance of the fine-tuned algorithms. Finally, Section 2.7 concludes the study, discussing the results and providing pointers to future work.

2.2 TEXT REUSE DETECTION

Outside the field of literary studies, Text Reuse Detection has been studied in a number of related tasks in Computational Linguistics—such as Plagiarism Detection, Paraphrase Identification or Semantic Textual Similarity. These tasks target phenomena which are certainly different from the local-effects of intertextuality that shape the interest of our study. Still, a discussion of these tasks will provide a point of reference for applications of text reuse to literary works both in terms of task definition and evaluation procedures.

2.2.1 Related Tasks

2.2.1.1 Plagiarism Detection

Plagiarism Detection is a well-established task in IR, and is supported by both the organization of long-running shared-tasks (Potthast et al., 2009) and consolidated evaluation procedures validated by the scientific community (Potthast et al., 2010). The task consists in identifying borrowed text in academic or, more generally, formal writing contexts, where due attribution is omitted. Plagiarism Detection has, thus, a largely uncontroversial task definition, which is implemented in practice as a binary classification problem—although fine-grained taxonomies also exist (Alzahrani, Salim, and Abraham, 2012). Plagiarism Detection corpora have been most often constructed following two procedures: artificially—by algorithmically emulating the act of plagiarism (Potthast et al., 2011)—or through simulation—by prompting human annotators to obfuscate the reuse of a given passage (Potthast et al., 2012, 2013b, 2010).

2.2.1.2 *Paraphrase Identification*

Paraphrase Identification is concerned with modeling the capabilities of languages to express the same facts with different words or phrases. The definition of a paraphrase can be formalized in terms of a bi-directional entailment with respect to the paraphrased text. This definition relates the task to other Natural Language Understanding and Textual Entailment Recognition tasks. Although even such a formal definition has been observed to lead to divergent implementations in practice (Bhagat and Hovy, 2013; Rus, Banjade, and Lintean, 2014), progress in Paraphrase Identification research has continued, not the least due to the proliferation of benchmark corpora—such as the prominent MSRP (or MRPC) corpus (Dolan and Brockett, 2005). Similarly to Plagiarism Detection, these corpora frame the task as a binary classification problem.

Corpus construction in Paraphrase Identification has followed various paths. For instance, for the MSRP corpus, a large database of related sentence pairs was first obtained by identifying phrases with shared lexical choices and similar positions within the respective documents. A preliminary paraphrase detector was then used to extract a more fine-grained subset, which was finally annotated considering "whether the two sentences mean the same thing". Another approach involves leveraging pre-existing resources. For instance, the Q&A Paraphrase Corpus exploits questions manually flagged as reformulations in the WikiAnswers platform and filters out irrelevant pairs in follow-up annotation processes (Bernhard and Gurevych, 2008). Regneri and Wang (2012) exploit parallel discourse structures from TV shows, and Ganitkevitch, Van Durme, and Callison-Burch (2013) and Creutz et al. (2018) exploit parallel translation data. Finally, other corpora have been built through elicitation, asking participants to produce paraphrases to a given input (McCarthy and McNamara, 2012).

2.2.1.3 *Semantic Textual Similarity*

Semantic Textual Similarity has recently flourished in parallel to developments in distributional meaning representations and renewed interest into semantic modeling tasks—as exemplified by the SemEval STS shared task (Agirre et al., 2012, 2013). Semantic Textual Similarity differs from Textual Entailment in that the entailment is considered to be bi-directional, as in Paraphrase Detection. In contrast to Paraphrase Detection, however, Semantic Textual Similarity entertains a gradual notion of similarity, and, thus, is traditionally casted as an ordinal classification or regression problem in which sentence pairs must be scored with respect to semantic similarity—typically in an ordinal scale from zero to five. The most prominent corpus, the STS Core, was constructed relying on sentence pairs already available from related resources—such as the MSRP corpus, machine translation evaluation data or parallel headline news—and re-annotated through crowd-sourcing, using the Amazon Mechanical Turk platform.

2.2.2 Literary Text Reuse Detection

As evident from the previous section, Plagiarism Detection, Paraphrase Identification and Semantic Textual Similarity can rely on specific task definitions to determine concrete annotation guidelines and produce evaluation protocols and resources. In contrast, the term "text reuse detection" seems to be used rather ambiguously to refer to a set of tasks, bearing a certain resemblance. In some cases, "text reuse" has been used as an umbrella term for plagiarism-related practices, such as quotation, translation, paraphrasing or summarization (Bär, Zesch, and Gurevych, 2012; Burrows, Potthast, and Stein, 2013; Metzler et al., 2005; Potthast et al., 2013a). Other cases have focused on the common journalistic practice of reusing text from newswires, considering either current (Gaizauskas et al., 2001) or historical journals (Salmi et al., 2020; Smith et al., 2014; Smith, Cordell, and Dillon, 2013). A third case involves text reuse within online communities, as exemplified in studies surrounding Wikipedia editors (Alshomary et al., 2019; Clough and Stevenson, 2011) or blogging (Seo and Croft, 2008). Finally, a last case, indeed, involves text reuse in literary works (Büchler et al., 2014a,b). Interestingly, the term "Historical Text Reuse" is commonly used in this context, likely due to the fact that text reuse is particularly frequent in Western Classical Literature traditions—e. g. Ancient Greek and Latin Literature (Orr, 2003).

In order to reduce this terminological ambiguity, at least in the present work, we shall refer to the application of Text Reuse Detection algorithms to literary contexts as "Literary Text Reuse Detection". Certainly, Literary Text Reuse Detection can be informed by algorithmic development in the discussed tasks of Plagiarism Detection, Paraphrase Identification and Semantic Textual Similarity, since these too need to model similarity at different levels ranging from literal borrowing (as in Plagiarism Detection) to sentential semantics (as in Semantic Textual Similarity). However, despite this correspondence at the level of methods, the focus of Literary Text Reuse Detection focuses on local effects of intertextuality presents strong differences with the targets of those related tasks.

For instance, many literary reuse cases are hardly identifiable as "plagiarism", a practice that implies an active and malicious (if not criminal) intention to silently borrow text from an undisclosed source. This intention is responsible for particular obfuscation strategies that are in stark contrast to the literary resources employed by writers in order to facilitate the reading of the intertextual link. Secondly, Plagiarism Detection aims to uncover omitted attribution where attribution was due. However, attribution is a misplaced concept in the context of literature where it can run counter to the very artistic effect text reuse—for instance, an allusive reference—aims to achieve. Moreover, the cultural or artistic status of text reuse has been subject to considerable shifts in history and plagiarized only acquired its negative connotations fairly recently. Next, paraphrasing represents a small subset of the linguistic repertoire by which writers, consciously or unconsciously, establish intertextual links. Finally, gradual notions of similarity entertained by Semantic Textual Similarity are foreign to applications in intertextuality, which are framed in a binary setting. Moreover, while being a crucial aspect of computational models of intertextuality, semantic similarity does not constitute a sufficient—nor perhaps even a necessary—condition for the identification of an intertextual link.[1]

As a consequence, subsuming computational approaches to the detection of intertextual links under the broadly defined task of Text Reuse Detection is problematic, and may be detrimental to the progress of Literary Text Reuse Detection as a task. In this regard, a further complication in the application of Text Reuse Detection algorithms to literary studies is that, in contrast to the reviewed tasks, consolidated evaluation practices are missing—as we shall see in the next section.

2.3 RELATED WORK

In terms of evaluation practices, Literary Text Reuse Detection studies are characterized by a relative scarcity of varied and representative benchmark corpora. When available, evaluation resources mostly consist of corpora constructed ad-hoc in order to report performance measures of proposed approaches on the target dataset. The reasons are twofold. First, representative corpora are comparatively more costly to obtain than in related Text Reuse Detection tasks. For instance, in contrast to Paraphrase Identification, eliciting reuse from participants for a given input text is not a promising strategy in the context of Literary Text Reuse. Crowd-sourcing is problematic, since the decision as to whether a given candidate pair represents a real

[1] It is not sufficient because not all semantically similar passages build intertextual links. And, as cases of reuse anchored on recurrent expressions like "to be or not to be" show, it is neither necessary.

case of reuse requires a skill set very scarce among annotators and, even when available, such question constitutes a highly interpretative matter.[2]

In some cases, manual post-hoc evaluation of retrieved instances has been explored (Coffee et al., 2012a). Moreover, as previously mentioned, literary text reuse comprises a large spectrum of reuse types—which are responsible for a variety of existing categorizations (Bamman and Crane, 2008), (Hohl Trillini and Quassdorf, 2010) or (Büchler, 2013, p. 77). Additionally, as we shall see in Chapter 5, borrowing styles tend to vary across authors, which, in view of the tendency for studies to focus on particular authors, further complicates the curation of varied benchmark corpora.

In view of such difficulties, research seeking to estimate the performance of reuse algorithms has most commonly relied on the work of specialized commentators and editors. For example, a series of studies exploited the work of four commentators to extract a set of parallels between the first book of Lucan's *Civil War* (695 verses) and Vergil's *Aeneid* (9,896 verses) (Coffee et al., 2012a,b; Forstall et al., 2015; Scheirer, Forstall, and Coffee, 2016). The resulting set comprises a few hundred validated pairs, and the authors have made available similar resources through their online platform.[3] Pioneering work on multilingual reuse by Bamman and Crane (2009) used a set of 151 allusions in Milton's *Paradise Lost* to Vergil's *Aeneid*, which were sourced from a specialist book on the matter (Verbart, 1995). Büchler et al. (2012) relied on the digitization of the "critical apparatus" of a historic edition of Athenaeus' *Deipnosophistai*, which records manually identified reuse cases from Homer's *Iliad* and *Odyssey*, amounting to 353 instances. Similarly, the on-going edition of the works of 12th century writer Bernard of Clairvaux has served other researchers interested in allusive text reuse (Moritz et al., 2016)—including the work presented in Chapter 4 of the present PhD thesis—and the digital edition of the *Index Tomisticus* was leveraged for a study targeting reuse in Thomas Aquinas' *Summa contra Gentiles* (Franzini et al., 2018), comprising a total of 24,416 sentences, of which 7,396 contain some type of reference. Ganascia, Glaudes, and Del Lungo (2014) used the output of a pre-existing study of the links between Balzac's *Human Comedy* and Theóphile Gautier. Although the exact number is not reported in the referenced paper, the attached visualization seems to indicate that the number may be a total 16 instances.

The emerging picture highlights a scarcity of varied and comprehensive benchmark corpora. Consequently, without such resources, it is difficult to understand how current performance estimates would generalize to new corpora. Moreover, the tendency to focus on sin-

[2] Though computational research on the issue of inter-annotator agreement of literary text reuse may contribute to clarify the importance of these considerations.

[3] Available through: `https://tesserae.caset.buffalo.edu/blog/benchmark-data/`

gle-author resources and the frequent omission of comparisons with baselines and alternative methods leads to a situation in which, first, progress cannot be reliably monitored, and, second, researchers seeking orientation about the efficiency of different methods will have their needs unfulfilled. Finally, evaluation protocols, which are commonly associated with existing benchmark corpora and defined on their basis, are, thus, scarce in the literature. However, evaluation protocols—which involve discussions on what are relevant performance metrics and what counts as true and false positives—are necessary in order to establish standard practices in a research community. Still, only a minority of studies in the field of Literary Text Reuse Detection have explored such issues—rarely abstracting over their own ad-hoc case study—and a systematic exploration of evaluation protocols is still missing.

2.4 EVALUATION IN LITERARY TEXT REUSE

In the present study, we assume that we are given a target collection (T) of n documents containing cases of reuse to be identified, and a source collection (S) with m documents, from which text is being borrowed. We use the term "document" rather loosely to refer to a passage—e. g. a sentence—that may hold an intertextual link to another document. For evaluation purposes, a gold standard specifies a set of links L, where each link $l = (d_i^T, d_j^S, r_{ij})$ represents a reference between documents of $d_i^T \in T$ and $d_j^S \in S$. Each link is assigned a relevance r_{ij}, which in most cases is just a fixed scalar indicating whether the link is relevant. However, for some setups—which will be discussed in Section 2.4.1.2—it is useful to consider degrees of relevance and r_{ij} then varies along a range. Further, each link $l \in L$ indicates that at least some part of document d_i^T has been interpreted by the annotators as a reference to at least some part of d_j^S.

The first consideration with respect to evaluation concerns the matter of identifying the exact location of the text reuse instance within the document, which can be done at the level of words or characters. In related fields such as Plagiarism Detection, the localization of the plagiarism is commonly done at the word-level, and it is part of performance comparisons between systems (Potthast et al., 2010). However, Literary Text Reuse is confronted with significant challenges in this regard. For example, as we will see in Chapter 4, inter-annotator agreement on the identification of the span of allusions is low. For these reasons, in the present study we exclude localization from the evaluation procedure.

Related to localization is the problem of segmenting the collections into individual documents. In some cases, a pre-existing segmentation into sentences may be given—for instance, studies on biblical references may use the segmentation into verses. If no clear boundaries are provided, it is common to resort to a disjoint or overlapping segmentation approach (Büchler et al., 2014a), although this situation introduces the complication that a single reference may spill over document boundaries. Again, evaluation procedures in Plagiarism Detection deal with this issue by requiring from successful algorithms that at least one of the overlapping documents is retrieved (Potthast et al., 2010). For the present discussion, however, we assume that segmentation is given, and that references are defined over single document pairs.

Given these conditions, a reuse algorithm is defined as a function $f(d_i^T, d_j^S) = s$, which is given a document pair and must produce a score s. Regardless the range of scores that f outputs, we will assume in the present work that the score can be interpreted as a similarity score. Even though some algorithms—like those based on edit distances—may produce dissimilarity scores, these can easily be transformed into similarity scores by taking their complement or inverting the sign.

2.4.1 Evaluation Scenarios

We now discuss three evaluation scenarios that correspond to different use cases.

2.4.1.1 Classification-based Evaluation

A first evaluation approach considers all n-by-m comparisons between documents in T and S and, for a given threshold t, classifies pairs as reuse if the score is equal or higher than the threshold. The corresponding use case is described by a researcher aiming at finding all possible links spanning the target and source collections. Well-known performance metrics such as Precision (P) and Recall (R) can be easily used in this scenario. As it is commonly known, P corresponds to the proportion of pairs in the gold standard L classified as reuse—i. e. the true positives (TPs)—over the total number of pairs classified as reuse—which is equal to the sum of true positives and false positives (FP).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.1}$$

R, also known as the True Positive Rate (TPR), corresponds to the total count of true positives over the total number of references in L—i. e. the sum of true positives and false negatives (FN).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.2}$$

In order to summarize P and R with a single number, one can use the F_β-Measure, defined as the harmonic mean of P and R, where β is a coefficient that weighs the importance of R—e. g. for $\beta = 1$, P and R are weighted equally.

Moreover, these measures require fixing a threshold value in order to classify pairs into positive and negative cases and, thus, it is often necessary to quantify how algorithms behave with respect to the underlying precision-recall trade-off that is obtained when varying the threshold: larger threshold values increase P at the cost of R, and viceversa, with the cost being dependent of the algorithm. The trade-off can be summarized by computing the average P weighted by the increase in R obtained by each increase in threshold. The resulting quantity, shown in Equation 2.3, is known as Average Precision (AP):

$$\text{AP} = \sum_{t_i} (R(t_i) - R(t_{i-1})) \cdot P(t_i) \tag{2.3}$$

where $R(t_i)$ and $P(t_i)$ refer to recall and precision evaluated at threshold t_i.

An alternative evaluation is given by the Area Under the Curve (AUC) measure. The P and R values obtained for a given set of thresholds draw a curve in a plane with P and R on the axes. The area under such curve can be approximated using interpolation techniques such as the linear trapezoidal interpolation, which results in the AUC-AP measure. Alternatively, it is common to consider the AUC of the Receiving Operator Characteristic (ROC) curve (AUC-ROC). The ROC curve is defined similarly to the AP curve, but considers the False Positive Rate (FPR)—i. e. the ratio of false positives over the total number of pairs from the negative class, shown in Equation 2.4—instead of P.

$$\text{FPR} = \frac{FP}{FP + TN} \tag{2.4}$$

However, both AUC measures are problematic in the context of Text Reuse. First, traditional interpolation methods for computing AUC-AP tend to over-estimate P (Davis and Goadrich, 2006), and, thus, problems with low positive rates like Literary Text Reuse Detection—i. e. problems known as "needle-in-a-haystack"—are better served with AP than AUC-AP. Second, since ROC is defined in terms of FPR, the amount of true negatives—ignored by P and R—is now being taking into account. However, in fields such as Text Reuse Detection, the positive class is typically very rare, and its importance will be downplayed if true negatives are included in the computation of performance measures.

2.4.1.2 *Ranking-based Evaluation*

An alternative evaluation scenario can be obtained from an IR point of view (Manning, Raghavan, and Schutze, 2008), by drawing an analogy between the Text Reuse Detection algorithm and a search engine. A search engine is tasked with processing information needs in form of user queries, and retrieving a set of candidate documents ranked according to their relevance to the query. The documents in the target collection ($d_i^T, \dots, d_n^T$) are, thus, interpreted as queries, and the relevant documents in the source collection must be ranked according to their relevance. As we can see, this evaluation is particularly fitting when documents in the target collection are linked to multiple source documents, and may have different levels of relevance.

While less common, this situation arises in several cases. For instance, in situations where internal borrowing is present inside the source collection it may be unclear which of the parallel source documents should be included within the gold standard. More generally, research in Literary Text Reuse often has to deal with problematic cases in which interpretations differ as to the original source of the reference. From an annotation point of view, it may be advantageous to incorporate all links and assign relevances instead of binary decisions.

Finally, a ranking evaluation use case that can be particularly interesting is the crowd-sourcing of relevance judgments. Instead of tasking a small set of high-skilled annotators with fine-grained decisions on the relevance of candidate pairs, one can appeal to the intuitions of masses and use a voting scheme such that the relevance of a candidate pair amounts to the sum of all votes casted for this candidate pair. An example of such a dataset is described in Section 2.6.1.3.

For a set of target documents $d_i^T \in T$, let $Q_i = \{d_j^S : r_{ij} > 0\}$ be the set of relevant source documents that are relevant to d_i^T according the gold standard. Moreover, let Rank(Q_i) be the ranking of documents produced by the algorithm $f(d_i^T, d_j^S)$ for all $d_j^S \in S$, and Rank(Q_i, k) a subset of the ranking starting from the top and including documents until reaching at most k relevant documents. Then, a common ranking evaluation measure is given by Mean Average Precision (MAP), which takes the ranking for a given target document or query, computes a single precision score for each retrieved relevant document, averages these precision scores for the target document, and, finally, takes the mean of the average precision scores of all target documents:

$$\text{MAP} = \frac{1}{|T|} \sum_{i=1}^{|T|} \frac{1}{|Q|} \sum_{k=1}^{|Q|} P(\text{Rank}(Q_i, k)) \tag{2.5}$$

In cases where there is a single relevant document per query, or, when the goal is to evaluate the number of candidates a user needs to look at before satisfying the information need, a common metric is Mean Reciprocal Rank (MRR) (Voorhees, 1999):

$$\mathrm{MRR} = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{|\mathrm{Rank}(Q_i, 1)|} \tag{2.6}$$

where, as a reminder, $\mathrm{Rank}(Q_i, 1)$ refers to the set of top ranked documents retrieved for query d_i^T down to the first relevant document.

Both MAP and MRR compute arithmetic means of AP values over the entire set of queries. For queries with large numbers of relevant documents, the contribution of each of these relevant documents to the final score is diminished either because it is averaged into a single AP score (MAP) or because it is ignored (MRR). Moreover, MAP and MRR measure precision at all levels of recall, regardless the size of the ranked results list. More realistically, it may result unnecessary to evaluate relevant documents retrieved at a rank that no search engine user would consider inspecting. In this vein, it is common to only consider the first k ranked candidates and compute P for that shortened list, a measure known as Precision@K (P@K).

Since both MAP and MRR ignore non-binary relevance values, other relevance-sensitive measures like the Normalized Discounted Cumulative Gain (NDCG) have been developed. The main idea behind NDCG is to weigh the contributed relevance of retrieved documents by the rank at which they are retrieved—typically a logarithm scale of the ranks is used, which smooths the penalization of lower ranked relevant documents. For a given set of relevant documents Q_i associated with target document d_i^T, we first compute the DCG using the relevance scores r_{ij} from the gold standard:

$$\mathrm{DCG}(d_i^T) = \sum_{k=1}^{|Q|}\frac{r_{ik}}{\log_2(|\mathrm{Rank}(Q_i, k)| + 1)} \tag{2.7}$$

Finally, since the magnitude of DCG is determined by the range of relevance scores in the gold standard, it needs be normalized by the ideal DCG (iDCG) obtained by a perfect retrieval system—i. e. all relevant documents retrieved first in descending relevance order.

$$\mathrm{NDCG}(d_i^T) = \frac{\mathrm{DCG}(d_i^T)}{\mathrm{iDCG}(d_i^T)} \tag{2.8}$$

In order to obtain a single NDCG over all queries, a micro-averaged NDCG is computed by taking the arithmetic mean NDCG over target document or queries.

Ranking metrics have the advantage that they naturally incorporate estimates of precision at different recall values, and that they do not necessitate fine-tuned threshold values. However, they present two major disadvantages. The first one is that, in contrast to classification-based metrics such as P and R, ranking metrics are hard to interpret, and, while valuable in a context of baseline comparison, are of little use when considered in isolation. Secondly, ranking metrics rely on the assumption that all queries have relevant results. However, this assumption is problematic in Text Reuse, where rather the opposite is the case: most queries have no relevant documents. Thus, practitioners who wish to apply ranking metrics must restrict their target collections to documents appearing in the gold standard as true references. As a result, the obtained performance estimate must be interpreted as being conditioned on the existence of reuse, and the conclusions drawn from such experiments are somewhat limited.

In spite of these limitations, a ranking approach offers a fitting option when an evaluation procedure is desired that resembles the real-world use case of a researcher wishing to satisfy her information needs. Indeed, thanks to the researcher's intuitions, the queries actually run through such a search engine have a much higher likelihood of pointing towards a relevant document than what is expected from a random sample of a target collection.

2.4.1.3 *Hybrid Evaluation*

Finally, an alternative evaluation procedure arises when classification-based performance measures are considered but instead of isolated document pairs the evaluation handles ranked lists. The underlying setup is anchored in a frequent real-world use case, in which a literary scholar queries the source collection for a set of documents in the target collection that are suspected of containing text reuse. In contrast to the ranking-based scenario, these documents may still contain no text reuse. And in contrast to the pure classification-based scenario, the search engine user is not concerned about finding the entirety of text reuse cases. Instead, she focuses on a subset and is willing to inspect top-k ranked lists.

For each document d_i^T in T, let now $U_i(S, t)$ be the set of documents in S with a similarity score equal or higher than threshold t. Due to the thresholding, $U_i(S, t)$ can now be the empty set—i. e. no source document showed a similarity higher than t for the query document. Moreover, similar to P@K, we consider cut-off points in the retrieved ranked set of candidate documents at values aligned with what a search engine user would typically inspect—for instance, one can assume users would stop paginating through result lists after 20 negative instances. Thus, we extend the notation to $U_i(S, t, k)$ to denote the top-k documents in S that have at least a similarity of t.

We can now define True Positives, False Negatives and False Positives on the basis of such query result sets. True Positives correspond to query results that contain at least one true text reuse case. False

Negatives correspond to query results that contain not a single true text reuse case, although the gold standard specifies at least one such case. Finally, False Positives correspond to query results that contain at least one candidate even though the gold standard specifies no such text reuse case for the given query. Equations (2.9) to (2.11) formalize the former definitions:

$$\begin{aligned} TP(t,k) = \{&U_i(S,t,k) : U_i(S,t,k) \neq \varnothing \\ &\wedge \left(\bigcup_{d_i^T, d_j^S \in L} U_i(S,t,k) \cap \{d_j^S\} \right) \neq \varnothing\} \end{aligned} \tag{2.9}$$

$$\begin{aligned} FP(t,k) = \{&U_i(S,t,k) : U_i(S,t,k) \neq \varnothing \\ &\wedge \left(\bigcup_{d_i^T, d_j^S \in L} U_i(S,t,k) \cap \{d_j^S\} \right) = \varnothing\} \end{aligned} \tag{2.10}$$

$$FN(t,k) = \{(d_i^T, d_j^S) \in L : U_i(S,t,k) = \varnothing\} \tag{2.11}$$

On the basis of Equations (2.9) to (2.11) we can compute P and R at different threshold values and cut-off points, and AP over a range of threshold values. It is noteworthy that these measures receive a slightly different interpretation than those in Section 2.4.1.1. For instance, P now refers to the percentage of queries that are expected to be successful—i. e. to return positive results. These interpretations, however, are aligned with the use case on which the evaluation procedure is based. Moreover, in this approach, cases of multiple references per query are given the same treatment as in MRR. Any relevant documents ranked below the most highly ranked one are ignored. This limitation, however, is relatively uncompromising if the assumption is met that text reuse links in the gold standard involve a single source document per target document.

The merit of this evaluation procedure stems from the fact that results are highly interpretable—since it uses familiar classification-based measures that are meaningful to the users of the software—and, still, the evaluation is based on ranked results—which represent a widely spread use case for these retrieval algorithms.

2.4.2 Generalization of Evaluation Measures

A final issue regarding the evaluation of Text Reuse retrieval methods relates to the question of parameter and hyper-parameter fine-tuning. When evaluating a retrieval method, the goal may be a comparison with other methods on the target dataset only, or, more interestingly, we may seek to establish the out-of-sample performance of those methods—i. e. the performance that can be foreseen for future, unseen datasets.

In contrast to Machine Learning methods, where the expected prediction error needs to be considered against the background of the bias-variance trade-off (Hastie, Tibshirani, and Friedman, 2009), expected performance is rarely taken into consideration in the context of Text Reuse algorithms. A reason for this is that algorithms in Text Reuse Detection rarely rely on statistical learning procedures. Still, as we shall see, algorithms vary in the number of parameters, and may require exhaustive tuning in order to achieve competitive performance. Furthermore, the observed variety of reuse styles—not just across authors but also within given collections—casts doubts over the generalization of evaluation results to other subsets and datasets. Since, in real-world applications, manual fine-tuning can only be done on small subsets, practitioners assume that hyper-parameters selected on those subsets will perform similarly in other subsets. However, this assumption is rarely tested.

In order to approach these issues, we perform our evaluations using Cross Validation (CV). As it is commonly known, in CV the target collection is first split into k folds randomly, and, then, the algorithm is repeatedly fine-tuned on one fold and evaluated on the remaining $k-1$ folds. The resulting k estimates, provide a distribution of scores that represents the spread of uncertainty associated with the expected performance on new subsets.

The last aspect concerns fine-tuning procedures. One common approach is to perform a grid-search over pre-specified hyper-parameter ranges and select the best performing combination. However, since algorithms differ in the number of hyper-parameters, grid-search performs different number of fine-tuning rounds depending on the complexity of the algorithm. Thus, an evaluation based on grid-search can not only be unfair if the differences in complexity of algorithms is large, but it might also represent an artificial evaluation scenario, since researchers would often have a fixed amount of resources available for fine-tuning.

In order to offer a more realistic evaluation, we switch to an alternative approach—random-search—, which not only provides a more efficient parameter search procedure (Bergstra and Bengio, 2012), but also allows us to control for the total budget per algorithm in terms of fine-tuning rounds. In contrast to grid-search, random-search requires defining a distribution of values per parameter. For a fixed number of fine-tuning rounds, random-search samples a parameter value from the specified distributions. After the budget is exhausted, random-search selects the best model parameters based on the chosen evaluation measure. The total number of rounds can be kept constant across algorithms to ensure fairness of comparisons.

2.5 TEXT REUSE ALGORITHMS

The application of a Text Reuse algorithm in the literary domain implies a number of pre-processing steps that are relatively independent of the chosen algorithm, and that can have a crucial impact on the final results (Büchler et al., 2014a; Büchler, 2013). This situation is exacerbated in the presence of digital noise, such as OCR noise resulting from the digitization of the original sources, or when processing morphologically complex languages and historical languages with non-standard spelling (Piotrowski, 2012). These are common cases in literary text reuse studies, and, certainly, in the present one. Therefore, before delving into a discussion of retrieval algorithms in Section 2.5.2, we will consider several aspects of the preprocessing pipeline in Section 2.5.1.

2.5.1 Pipeline

Text Reuse Detection pipelines are concerned with the pre-processing of input texts to enhance and speed up the subsequent retrieval algorithms. Here, we highlight three main steps including normalization of the input text, enriching of the feature space through shingling and frequency-based feature selection.

2.5.1.1 Normalization

Once the input text has been tokenized, the next step is the normalization of the input. Due to morphological or spelling-related variation, the same input tokens may be realized differently. Algorithms can improve matching if the underlying forms are fed, through lemmatization, instead of the surface realizations. Two aspects of lemmatization are relevant to the preprocessing pipeline.

The first aspect relates to disambiguation. Lemmatizers that do not exploit sentential context—an example for Latin is `LemLat` (Passarotti et al., 2017)—cannot disambiguate cases in which the input form may correspond to multiple lemmata—for example, the token "living" can refer to lemma "live" or "living". In contrast, other lemmatizers (Chrupala, Dinu, and van Genabith, 2008; Cotterell, Fraser, and Schütze, 2015; Schmid, 2013) employ statistical learning techniques in order to make accurate predictions about the word form underlying each surface realization. It can be observed that, conditioned on lemmatization accuracy, a disambiguating lemmatizer potentially leads to higher precision in reuse detection than a non-disambiguating one, without necessarily compromising recall. This is because a Text Reuse algorithm loses discrimination ability when the normalization procedure conflates forms of different lemmata. In any case, the decision as to what lemmatizer type to use depends on the accuracy of the available disambiguating lemmatizers as well as the extent to which morphological ambiguity is an issue—as we shall see in Chapter 3.

A second aspect relates to lemmatization coverage. Even within the realm of statistical lemmatizers, some models are restricted to a closed-set of target lemma. In contrast, open-set lemmatizers can produce outputs for tokens for which no appropriate lemma is known to the model. For closed-set lemmatizers, two strategies are available for handling unknown tokens. If unknown tokens are assigned the input token as lemma (e. g. "aardvark" → "aardvark"), the resulting lemmatizer should lose overall reuse recall with respect to an accurate open-set lemmatizer. This is because the evidence for reuse is reduced when the same underlying lemma appears in both documents with different surface realizations but cannot be recovered by the closed-set lemmatizer because the lemma is out of vocabulary. However, if the unknown tokens are assigned a dummy token as lemma (e. g. "aardvark" → "unknown"), the situation resembles that of a non-disambiguating lemmatizer—i. e. different lemmata are conflated into the same form—and, thus, we should expect precision to take a hit. For the present study we resort to a neural open-set lemmatizer—described in Chapter 3—that outputs disambiguated lemmata.

After lemmatization, if the lemmatizer outputs cased lemmata or the pipeline skips lemmatization altogether, lowercasing the input can improve recall at the expense of precision. Finally, some of the algorithms also benefit from *punctuation removal* and *stop-word removal*. For the latter, it is usually necessary to carefully construct stop-words lists on a case by case basis, since the definition of a stop-word strongly depends on the domain of consideration.

2.5.1.2 *Shingling*

The next step in the pipeline involves the process of shingling, by which possibly overlapping sequences of contiguous words in the input are extracted and used to build a representation of the documents. The length of these sequences may vary from uni-grams to tetra-grams and higher orders, which can be used in order to introduce a bias towards more literal reuse styles. Shingling can be useful for algorithms that are based on representations that disregard word order—such as bag-of-words—since it can easily boost the contribution of longer borrowings to the final similarity score.

2.5.1.3 *Feature Selection*

Text reuse algorithms need to perform a high number of comparisons between documents—e. g. for collections sizes in the order of 10,000 documents, the number of comparisons is in the order of hundreds of millions. This sets the algorithms under strong space and speed constraints. Feature selection can offer some relief in terms of both memory and speed by reducing the feature set that is used to represent the documents.

An example reduction consists in eliminating "hapaxes", or more generally, words with a low *document frequency*. Since by definition hapaxes cannot help detection algorithms based on lexical matching, they are often removed straight away. Other algorithms that take into account semantics, however, may be able to include hapaxes and other low frequency words into the final similarity score if the semantics of those words imply a relation to words in the candidate match documents. For those algorithms, however, feature selection can be accomplished by dropping comparisons of words with a sufficiently low semantic similarity. Additionally, a threshold on raw word frequency is a common approach to reduce the size of the feature set. Due to the common Zipf-curve in language vocabularies, a seemingly low frequency threshold can already produce large gains in processing speed, without much performance harm.

2.5.2 Typology of Text Reuse Algorithms

In the present study we consider three broad families of text reuse algorithms. This categorization is not meant to be exhaustive but rather as a chart that helps the reader get oriented within the landscape of approaches in current research.

A conspicuous omission are Machine Learning based approaches. Machine Learning approaches to Text Reuse provide a promising route that would allow researchers to automatically extract text reuse patterns from the corpus of interest, jointly modeling semantics and the style of reuse of the authors. However, the application of Machine Learning algorithms to intertextuality faces two major hurdles. First, the lack of comprehensive training and evaluation corpora. Second, the costly running times that require efficient optimization for full collection-level extraction. Moreover, Machine Learning approaches in Literary Text Reuse Detection have been rather rare. Two exceptions are Forstall and Scheirer (2019, Chapter 7.2), who employ one-class Support Vector Machine to identify cases of self-plagiarism by Jonah Lehrer, and Liebl and Burghardt (2020), who incorporate a Siamese Neural Network into a system based on n-gram matching to retrieve cases of Shakespearean reuse in contemporary literature.

Machine Learning algorithms aside, Section 2.5.2.1 first presents methods that aim at efficiently computing set-based similarity functions over large collections—a subset of these methods comprises the well-known finger-printing methods. Next, Section 2.5.2.2 introduces VSMs models, which are based on vectorial representations of documents and allow for semantic modeling. Finally, Section 2.5.2.3, introduces text alignment algorithms, which compute similarity scores on the basis of word-level alignment between documents.

2.5.2.1 Set-based Approaches

Set-based approaches generate similarity scores on the basis of set-based similarity measures such as Jaccard similarity (Jaccard, 1901) or containment similarity (Broder, 1997). In order to apply set-based approaches, we need to represent documents as sets of features for a given vocabulary of features V. Formally, we let d_i^T and d_j^S represent—slightly overloading the notation—the respective sets of features $\{w \in V : w \in d_i^T\}$ and $\{w \in V : w \in d_j^S\}$.

SET-BASED SIMILARITY FUNCTIONS On the basis of such set representations, a number of similarity functions can be defined.

- **Jaccard**, also known as intersection over union, is defined as the ratio of the number of features in the intersection of the document sets over the number of items in the union of document sets:

$$\text{Jaccard}(d_i^T, d_j^S) = \frac{|d_i^T \cap d_j^S|}{|d_i^T \cup d_j^S|} \tag{2.12}$$

  Jaccard is, thus, defined in the $[0, 1]$ range and is symmetric, implying the following equality

$$\text{Jaccard}(d_i^T, d_j^S) = \text{Jaccard}(d_j^S, d_i^T) \tag{2.13}$$

  When the size of the source or the target documents is variable across the collection, Jaccard similarity shows a bias towards smaller documents due to the variance in the denominator. For such cases, a related similarity measure known as "containment" is preferred.

- **Containment** uses the length of the set representing target document in the denominator instead of the length of the union of sets and is therefore robust towards varying document lengths in the source collection. Similarly to Jaccard, containment is defined in the $[0, 1]$ range but it is not symmetric.

$$\text{Containment}(d_i^T, d_j^S) = \frac{|d_i^T \cap d_j^S|}{|d_i^T|} \tag{2.14}$$

A small modification to containment, however, can turn it into a symmetric measure:

$$\text{ContainmentMin}(d_i^T, d_j^S) = \frac{|d_i^T \cap d_j^S|}{\min(|d_i^T|, |d_j^S|)} \tag{2.15}$$

- **Cosine similarity** Finally, cosine similarity can be re-interpreted as a set-based similarity under certain conditions. Cosine similarity over vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$ is generally given by:

  $$\text{Cosine}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} \vec{x}_i \times \vec{y}_i}{\sqrt{\sum_{i=1}^{n} \vec{x}_i^2} \times \sqrt{\sum_{i=1}^{n} \vec{y}_i^2}} \tag{2.16}$$

  which corresponds to the dot product of the vectors $\vec{x}$ and $\vec{y}$ normalized to unit length. Cosine similarity expresses the similarity of two vectors by the cosine of the angle of the two vectors rooted at the origin. When considering binary input vectors—i. e. the equivalent vector representation of a set—the numerator is equivalent to the length of the intersection of the document sets, and the denominator equals the product of the square roots of the lengths of the document sets. Thus, we can define the cosine similarity for sets as:

  $$\text{Cosine}(d_i^T, d_j^S) = \frac{|d_i^T \cap d_j^S|}{\sqrt{|d_i^T|} \times \sqrt{|d_j^S|}} \tag{2.17}$$

OPTIMIZING THE RUNNING TIME When computing set similarities over large collections, collection-level statistics can be exploited in order to produce critical speed improvements at no or, at least, at a controllable cost in detection performance—see Leskovec, Rajaraman, and Ullman (2014, Chapter 3) for an introduction into such techniques. Here, we point at two major alternatives: signature-based and inverted-list approaches. Signature-based approaches are also referred to as finger-printing methods—see Lulu, Belkhouche, and Harous (2016) for an overview—and include the well-known Locality Sensitive Hashing (LSH).

1. **Signature-based approaches**, first introduced by Broder (1997) for the task of near-duplicate document detection, aim at substituting the feature sets representing the documents with a much smaller, fixed-size numeric representation—known as the signature—that allows efficient storing and fast similarity computations. An example of such a signature is the *minhash*, that is computed by considering a fixed number d of permutation functions operating over the feature set. Each permutation function produces an entry in a document signature, such that the probability that two documents receive the same value in that entry is equivalent to the Jaccard similarity of those two documents. This property is exploited systematically LSH to reduce the total number of comparisons between a target document and the source collection, when the goal is to retrieve source documents with a similarity higher than a given threshold. Interestingly, a general theory of LSH exists that allows for the development of signature functions that target other similarity scores than Jaccard.

2. **Inverted-list approaches** are especially effective for problems that entail full collection search such as the *set similarity join*—i.e. for two collections T and S and a given threshold t, find all pairs (d_i^T, d_j^S) with a similarity higher than t—or the *all-pairs search*—for a collection Q and a given threshold t, find all pairs (d_i^Q, d_j^Q) with similarity higher than t. In contrast to signature-based approaches, inverted-list approaches aim at computing the exact similarity of the document pairs, and exploit collection-level features statistics to discard comparisons between candidate pairs that do not share the required amount of features to meet the target similarity (Bayardo, Ma, and Srikant, 2007). For a given feature w in the total set of features V, an inverted-list I_w contains references to all documents in a collection that have feature w. In order to find candidates for a given document d_i^T under a strictly positive similarity threshold, we merge all documents found in the inverted-lists $\bigcup_{w \in d_i^T} I_w$. Any other document d_j^S not in the computed candidate set can be discarded since does not share any feature with d_i^T. While illustrative, such a heuristic does not provide sufficiently aggressive candidate filtering and can be easily improved by means of additional techniques. We shall briefly three such techniques: discuss size filtering, prefix filtering an positional filtering.

   a) First, **size filtering** (Arasu, Ganti, and Kaushik, 2006) reduces the set of candidates by taking into account the size of any candidate document d_j^S. For a given Jaccard threshold t, it can be observed that $\text{Jaccard}(d_i^T, d_j^S)$ will be below the threshold t if $|d_j^S| < |d_i^T| \times t$. Thus, when merging the inverted-lists for a given query document d_i^T, we can ignore any matching document that is too short to eventually meet the threshold.

b) Next, **prefix filtering** aims at reducing the candidate set of a given document by indexing only a subset of features of each document—known as the prefix (Chaudhuri, Ganti, and Kaushik, 2006). Prefix filtering assumes that the features of a document appear in a fixed order across documents in the collection—i. e. V follows a fixed predetermined order. Consider documents d_i^T and d_j^S and let now $\text{Prefix}(d_i^T)$ refer to the first $|d_i^T| \times \lceil t \times |d_i^T| \rceil + 1$ features of d_i^T in the pre-determined order of V. Then, if the Jaccard similarity of d_i^T and d_j^S is larger than t, it must be the case that $\text{Prefix}(d_i^T) \cap \text{Prefix}(d_j^S) \neq \varnothing$ (Doan, Halevy, and Ives, 2012, Chapter 4). Using this result, we can ignore any feature coming after the prefix during the construction of the inverted-lists and use $\text{Prefix}(d_i^T)$ to build the inverted-lists—i. e. $I_{w \in \text{Prefix}(d_i^T)}$ instead of $I_{w \in d_i^T}$. Moreover, further candidate set reductions can be obtained by sorting the features in increasing frequency, thus avoiding higher frequency words, such as stop-words, that typically produce larger inverted-lists.

c) Finally, **positional filtering** (Xiao et al., 2011) aims at boosting the reduction provided by prefix filtering by applying a size filter to the features following a specific position in the document. Thus, similarly to prefix filtering, positional filtering also assumes a fixed order of the feature set. Positional filtering is applied while constructing the candidate set through merging of the inverted-lists of features in the prefix of d_i^T. It relies on the idea of estimating an upper-bound of the similarity between d_i^T and d_j^S on the basis of the amount of remaining features at a given position. For a given feature w shared by documents d_i^T and d_j^S, and appearing respectively at positions p and q, the Jaccard similarity can only satisfy threshold t if the following inequality holds:

$$\frac{\min(|d_i^T| - p, |d_j^S| - q)}{\max(|d_i^T|, |d_j^S|)} >= t \tag{2.18}$$

After the application of all filtering techniques, the desired similarity of remaining set of candidates can be explicitly computed in an efficient way. Moreover, it must be noted that while the present discussion employs Jaccard similarity, minor variations of the discussed filtering strategies exist that target other set-based similarity measures.

One of the most representative toolboxes based on finger-printing used in current Literary Text Reuse Detection is TRACER (Büchler et al., 2014b), which has served as the basis for numerous studies. Another toolbox using finger-printing techniques is InterText (YaleDHLab, 2017), which uses min-hashing to identify potential cases of reuse.

2.5.2.2 Vector Space Models

A VSM[4] is based on vectorial representations of documents in a multidimensional feature space and relies on proximity metrics in order to retrieve similar documents. A VSM, thus, represents a collection of documents C as a document-term matrix in which each vector $\overrightarrow{d}$ corresponds to a document and the i^{th} entry of each vector to the weight of the i^{th} term of the vocabulary V in document d. There are several strategies to implement the way term weights are computed. The bag-of-words representation, for instance, takes the raw term frequency as the weight.

Term frequency, however, results in representations that are dominated by most frequent words in the the vocabulary. A generally more efficient representation is given by applying a term frequency–inverse document frequency (Tf-Idf) transformation on the word frequencies, taking into account a notion of document-specific importance of words. The Tf-Idf score for the i_{th} word is computed as the product of the term frequency (Tf) in d—denoted $\text{Tf}(w, d)$—and its inverse document frequency—$\text{Idf}(w, d)$—defined by Equation 2.19:

$$\text{Idf}(w, d) = \log\left(\frac{|C|}{1 + |\{d \in C : w \in d\}|}\right) \tag{2.19}$$

where $C = T \cup S$ refers to the entire collection of documents.

An extension of VSMs that is able to incorporate semantic relations between words is Latent Semantic Indexing (LSI) (Deerwester et al., 1990). LSI exploits correlations between term to induce a latent semantic space in which document proximity indicates thematic resemblance. The key is a dimensionality reduction that generates a low-rank approximation $M_{LSI} \in \mathbb{R}^{kx|C|}$ to the document-term matrix $M \in \mathbb{R}^{|V|x|C|}$. In M_{LSI} the vocabulary is replaced by k factors that maximally explain the variance in the original document-term matrix, and the weights now correspond to the importance of those factors to the document. These latent factors have been shown to represent multiple semantic aspects of the underlying corpora.

While document matching on the basis of bag-of-words representations relies on lexical similarity, LSI solely relies on the induced semantic factors. A third approach is a hybrid in which both lexical correspondence and word-level semantic relations are exploited. This approach can be implemented on the basis of a generalization of the

[4] As it has been noted by Dubin (2004), the paper by Gerard Salton *A Vector Space Model for Information Retrieval*, allegedly published in 1975 and often credited with the introduction of VSMs, does not actually exist.

VSM. If we assume that the document vectors $\vec{a}, \vec{b} \in \mathbb{R}^{|V|}$ are already normalized, cosine similarity reduces to the vector product: $\vec{a} \cdot \vec{b}$ which is computed by:

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^{|V|} \vec{a}_i \times \vec{b}_i \tag{2.20}$$

Equation 2.20 indicates that document vectors are compared feature $\vec{a}_i$ by feature $\vec{b}_j$. Thus, how similar two documents are with respect to different, but semantically related terms is not taken into account—geometrically, this means that the basis dimensions underlying different terms are considered to be pairwise orthogonal. For illustration, if document a consists of terms "car", "bike" and "dolphin", and document b consists of terms "car", "bicycle" and "tree", only the respective weights associated with the term "car" will contribute to the resulting cosine similarity, even though "bike" and "bicycle" are technically synonym.

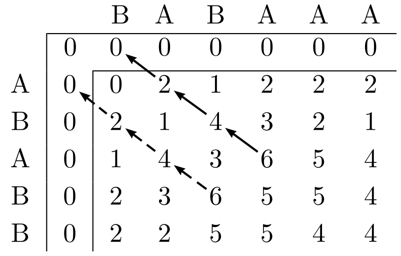
We can, instead, generalize the cosine similarity in order to additionally incorporate the similarity of two documents with respect to distinct features $\vec{x}_i$ and $\vec{y}_j$, as shown in Equation 2.21:

$$\text{SoftCosine}(\vec{a}, \vec{b}) = \frac{\sum_{i,j}^{|V|} W_{i,j} \vec{a}_i \vec{b}_j}{\sqrt{\sum_{i,j}^{|V|} W_{i,j} \vec{a}_i \vec{b}_j} \sqrt{\sum_{i,j}^{|V|} W_{i,j} \vec{a}_i \vec{b}_j}} \tag{2.21}$$

where $W_{i,j}$ is a scalar capturing the relatedness between the i^{th} and j^{th} terms. The resulting measure, known as "soft cosine" (Sidorov et al., 2014), can be interpreted as the cosine similarity computed in a GVSM—introduced already in the late 80s by Wong, Ziarko, and Wong (1985) and Wong et al. (1987)—in which the dimensions corresponding to individual terms are now expressed as linear combinations of $2^{|V|}$ latent vectors.[5]

In order to estimate the values of matrix W, co-occurrence statistics can be used or semantic resources such as WordNet (Fellbaum, 2012) can been exploited. A more recent approach involves using distributional word representations—word embeddings—. In particular, we focus on prediction-based word embeddings (Mikolov et al., 2013), which have been shown to excel over their count-based counterparts on semantic tasks such as word analogies (Baroni, Dinu, and Kruszewski, 2014). Word embeddings represent words as points in a vector space in such a way that semantically related words appear in close proximity of each other, and, thus, pairwise word similarities can be computed using standard VSM similarities, such as cosine similarity.

[5] The parallelism between soft cosine and GVSM seems to have gone unnoticed in the literature, which has "rediscovered" the soft cosine measure as an independent development.



$$\begin{array}{c|c|cccccc}
 & & B & A & B & A & A & A \\
\hline
 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
A & 0 & 0 & 2 & 1 & 2 & 2 & 2 \\
B & 0 & 2 & 1 & 4 & 3 & 2 & 1 \\
A & 0 & 1 & 4 & 3 & 6 & 5 & 4 \\
B & 0 & 2 & 3 & 6 & 5 & 5 & 4 \\
B & 0 & 2 & 2 & 5 & 5 & 4 & 4
\end{array}$$

Figure 2.1: Score matrix illustrating Smith-Waterman on two sequences of As and Bs. Solid and dashed arrows represent the two possible alignments found, which correspond respectively to sequences A-B-A and B-A-B. The scores correspond to a match of 2 and a mismatch and gap of -1.

VSM-based approaches have been employed to tackle semantically motivated cases of reuse, such as allusions. For example, Scheirer, Forstall, and Coffee (2016) use LSI to detect allusions in Latin literature. Lund et al. (2019) use Anchor-word Topic Models (Arora et al., 2013) to extract intra-biblical references.

2.5.2.3 *Text Alignment Algorithms*

The last family of algorithms that we consider is text alignment algorithms. Text alignment algorithms try to find the set of word-by-word correspondences between two texts that produces the highest score, as determined by the number of matches, mismatches and gaps. The most popular alignment algorithm is perhaps the Levensthein distance (Levenshtein, 1966), which is a variant of the more general Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). Needleman-Wunsch is a *global* alignment algorithm since it tries to align all words in the input texts, and does not allow gaps in the final alignment. Text reuse, however, is commonly localized in possibly discontinued subsequences, and thus, it is more suitable to consider a local alignment algorithm such as Smith-Waterman (Smith and Waterman, 1981), on which the discussion below shall focus.

As a dynamic programming algorithm, Smith-Waterman is best illustrated through the score matrix of size n-by-m that encodes all possible alignments between sequences a (of size n) and b (of size m) and their score—see Figure 2.1.

The score matrix encodes the maximum score that can be obtained at each cell when processing the sequences from left to right. In order to construct the score matrix, each cell is reached either from the top-left (representing a matching or mismatching of the sequences), top (representing a gap in the a—vertical gap) or left (representing

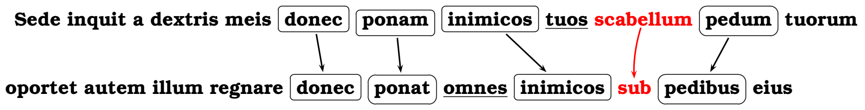

Sede inquit a dextris meis donec ponam inimicos tuos scabellum pedum tuorum
oportet autem illum regnare donec ponat omnes inimicos sub pedibus eius

Figure 2.2: Example of local alignment computed by the Smith-Waterman algorithm, matching a passage from Bernard's 6th Sermon—document on top—and the biblical verse *1 Corinthians, 15:25*—document on the bottom. Boxes mark tokens that matched based on their lemmata. Underlined tokens represent gaps and tokens highlighted in red represent mismatches.

a gap in b—horizontal gap) cell. The score of any given cell $c_{i,j}$ is accumulated with the recurrence shown in Equation 2.22:

$$c_{i,j} = \max \begin{cases} 0 & \text{end of local alignment} \\ c_{i-1,j-1} + W_{a_i,b_j} & \text{diagonal match/mismatch} \\ c_{i-1,j} - \text{gap} & \text{vertical gap} \\ c_{i,j-1} - \text{gap} & \text{horizontal gap} \end{cases} \tag{2.22}$$

where W_{a_i,b_i} represents a matching score or a negative penalty between words a_i and b_j; and gap sets a penalty for a gap in one of the sequences—i. e. skipping over items of one of the sequences.

Once the score matrix has been constructed, the alignment can be retrieved by identifying the highest score in the matrix and backtracking until we reach a cell at the margin, where the cumulative score is zero, indicating the end of the alignment. Note that multiple "best" alignments are possible. For instance in Table 2.1 two possible alignments of the input sequences are shown. As in the case of soft cosine, W_{a_i,b_j} can be constructed such that word embedding based similarities are exploited to boost the score of text reuse cases where words are replaced by their synonyms.[6] Finally, a common extension to Smith-Waterman is to distinguish between penalizations for opening a gap and extending a gap. By penalizing less the continuation of a gap than its opening, Smith-Waterman can often achieve more intuitive alignments.

In order to illustrate the Smith-Waterman algorithm, Figure 2.2 visualizes an alignment between two documents from one of the datasets used in this study. Using lemmatization Smith-Waterman is able to draw an alignment spanning 6 tokens, including two gaps and a mismatch. Using a gap penalty of 1, a matching score of 3 and a mismatch penalty of 2, the resulting alignment score amounts to 8.

Text alignment algorithms are notoriously costly to compute, with quadratic complexity in the length of the documents. For reasonably

[6] In preliminary experiments, this approach consistently underperformed the variant that does not use word embeddings to estimate the matching score, and was, thus, left out of consideration.

sized collections, computing all n-by-m comparisons is unfeasible. An optimization strategy involves a pre-filtering step using a cheaper algorithm that can be optimized for recall. On the basis of the high-recall full-collection comparison, the text alignment algorithm can be run only on of the k most promising candidate pairs for a given budget of k comparisons. Such an approach underlies BLAST (Altschul et al., 1990)—a high-performant software package used for sequence similarity search in Bioinformatics and has been applied to historical newspaper collections successfully (Vesanto et al., 2017). Another approach using text alignment algorithms is `passim`[7], which was also employed to extract reuse in historical newspaper collections (Smith et al., 2014).

2.6 EXPERIMENTS

In order to estimate the performance of the discussed families of algorithms, we ran a set of fine-tuning experiments across a number of datasets, allowing us to explore different evaluation scenarios. More concretely, we focus on 4 algorithms. First, `Set-based` is a set-based algorithm implementing an inverted-list approach with size, prefix and position filtering optimizations, using a 0.1 threshold for feature filtering. Second, `Tf-Idf` represents a VSM-based approach implementing a Tf-Idf weighting scheme and computing document similarities using cosine. Third, `Soft-Cosine`, implements a VSM-based variant that also uses the Tf-Idf weighting scheme but that relies on the soft cosine similarity function using cosine similarity between word embeddings to estimate word similarity. Finally, `Smith-Waterman` corresponds to the text alignment algorithm using `Soft-Cosine` as pre-filtering step with a budget of two million comparisons.

2.6.1 Datasets

2.6.1.1 Bernard of Clairvaux

Bernard of Clairvaux (1090–1153) was a French abbot who became one of the most important religious leaders during the High Middle Ages. He was influential not only within the monastic order of the Cistercians, to which he belonged, but also within the Christian Church, counseling as many as five popes and playing an active role during the development of the Second Crusade (1147–1150). Moreover, through his extensive writing, he is considered one of the most prolific Fathers of the Church. His style of writing is rich in biblical references and known for his recurrent use of allusions (Mcguire, 2007). These circumstances have led to careful work on the manual identification of biblical

[7] Available through the following URL: `https://github.com/dasmiq/passim`

intertext, which can be utilized for benchmarking. In the present study, we made used of a digitized version of Bernard's *Sermons on the Song of Songs* (Bernard of Clairvaux, 1998, 2000, 2003, 2006, 2007)—made available at the Sources Chrétiennes Institute as part of the BiblIndex project (Mellerin, 2013, 2014)—which was kindly provided to us for computational analysis.[8] In total, the dataset amounts to 6,689 references across 85 sermons comprising 199,508 words. The references are classified according to "exact quotations" (2092), "inexact quotations (or mentions)" (3580) and "allusions" (1017). The source collection is Jerome's Vulgate, which was acquired from Perseus (Crane, 1996), and amounts to 36,664 verses and 624,249 words. While the Vulgate offers a natural segmentation into verses, Bernard's sermons must be segmented. For the present experiments, documents were extracted from the sermons by shingling with a length of 15 words and an overlap of 10 words. Both collections were further lemmatized using a neural lemmatizer trained on Medieval Latin charters that will be introduced in Chapter 3. We refer to this dataset by the name of `Sermons`.

2.6.1.2 Intra-biblical References

The website `blueletterbible.org` offers an online interface to different Bible versions aligned at the verse level. As additional resources it offers a dataset of intra-biblical references. The phenomenon of intra-biblical intertextuality comprises several study cases, such as parallelism across synoptic Gospels or reuse from the *Old Testament* in the *New Testament*. The dataset in `blueletterbible.com` focuses on the latter, and consists of 918 references classified into "allusion" (483), "plain reference" (300) and "unclassified" (135). The basis of the references is Bagster's Bible edition as well as the works of American evangelist George F. Pentecost.

In order to expand the linguistic coverage of the experiments, we acquired the Ancient Greek version of the Vulgate (known as the *Septuagint* or LXX), as well as a version in Early Modern English—King James Version of the Bible (KJV). The LXX version was lemmatized using a model of Ancient Greek trained on a large collection of text, comprising more than 50,000 sentences.[9] For the English version, we employed the lemmatizer provided by Stanford CoreNLP (Manning et al., 2014). We refer to this dataset by the name of `blueletterbible`.

[8] We are indebted to Elysabeth Hue-Gay, Lou Cecile and Charles Bourdot for providing access, and Jean Figuet, Marie-Imelda Huille and Laurence Mellerin for the biblical analysis underlying the corpus. BiblIndex is available online via `http://www.biblindex.mom.fr/`, and provides access to various biblical editions, including English translations.

[9] The data comes from the sources curated by means of the Perseids Platform: (Gorman, 2019; Harrington et al., 2021; Keersmaekers et al., 2019; Mambrini, Francesco, 2020); and from the Perseus Treebanks (Bamman and Crane, 2011b). The model achieves 97% accuracy on a random test split.

2.6.1.3 *Crowd-sourced Intra-biblical References*

The last dataset we consider consists of a set of crowd-sourced relevance judgments curated and made freely available by the website `openbible.info`. The cross-references were first seeded from the *Treasury of Scripted Knowledge* and other various public domain datasets, and displayed through an online interface to users, who casted their votes on the relevance of the shown reference. The resulting dataset consists of a total of 344,441 cross-references and has been used in previous computational studies on cross-referencing (Lund et al., 2019). For our case study, we filtered references from the *New Testament* to the *Old Testament*. The filtered dataset amounts to a total of 5,020 *New Testament* verses cross-referenced to 21,359 *Old Testament* verses and a total of 99,542 casted votes. We refer to this dataset by the name of `openbible`.

2.6.2 Word Embeddings

In order to deploy the `Soft-Cosine` algorithm, word embedding matrices are required, which serve the purpose of estimating word-level similarity. For the current experiments, we trained word embedding matrices for Latin, English and Ancient Greek. A set of fine-tuning experiments for Latin, leveraging the word similarity benchmark dataset of Sprugnoli, Passarotti, and Moretti (2019), can be seen in Appendix A.1.

For Latin, we used the *Corpus Corporum* (Roelli, 2014), which comprises about 162 million tokens of a diachronically representative sample of Latin. The corpus was lemmatized with the same neural lemmatizer described in Section 2.6.1. For Ancient Greek, we trained word embeddings on the `First1KGreek` collection, using the lemmatization model described in Section 2.6.1.2.[10] Finally, for English we trained word embeddings on the non-tokenized version of the BookCorpus (Zhu et al., 2015). We note that this entails a shift in domain with respect to the target data, since the language in the BookCorpus corresponds to contemporary English novels, while King James Version of the Bible is written in Early Modern English.

All embedding matrices were obtained using the `FastText` algorithm (Bojanowski et al., 2017). Since the situation of Ancient Greek in terms of available resources is problematic, and no word similarity benchmark corpora are available, we, transferred the hyper-parameters selected on the Latin word similarity benchmark dataset—see Appendix A.1—to Ancient Greek. For the BookCorpus, we used used default hyper-parameters.

[10] The dataset represents one of the largest databases of Ancient Greek—comprising more than 23 million tokens—and is available through the following URL: `https://opengreekandlatin.github.io/First1KGreek/`.

Parameter	Stop-word	Lemmatization	Shingling	Min-Freq
Distribution	U(Yes, No)	U(Yes, No)	1-grams 1+2-grams 1+3-grams 2+3-grams 2+4-grams	U(1, . . . , 10)

Table 2.1: Distributions employed for sampling during random-search for hyper-parameters common to `Set-based`, `Tf-Idf` and `Soft-Cosine`. "U" refers to uniform sampling from the given values. "Stop-word" refers to stop-word filtering.

2.6.3 Model Comparison

2.6.3.1 Methodology

RANDOM SEARCH We first estimate the relative model capacity on our benchmark corpora. Even though the tested algorithms do not involve statistical learning, they include hyper-parameters that need tweaking in order to achieve optimal performance. As argued in Section 2.4.2, the process of hyper-parameter tuning introduces a question about performance generalization to unseen datasets, which we set to address employing a 10-fold CV procedure with 500 iterations of random search. Pre-processing hyper-parameters were sampled equally for `Set-based`, `Tf-Idf` and `Soft-Cosine`, following the distributions specified in Table 2.1. For `Smith-Waterman`, neither a minimum frequency threshold nor stop-word filtering were applied. Algorithm-specific hyper-parameters and their distributions can be seen in Table 2.2.

BAYESIAN ANALYSIS FOR MODEL COMPARISON The traditional practice in model comparison involves conducting a paired t-test—or a non-parametric alternative—over the results of a k-fold CV procedure. The goal is to establish whether the mean difference in performance is significantly different from zero—more specifically, whether the null-hypothesis that the difference in performance is zero can be rejected. However, this procedure has long been known to suffer from an under-estimation of the Type-I error—i. e. wrongly rejecting the null-hypothesis. The problem lies with the unmet assumption that the differences in performance between any two systems be mutually independent, which cannot be guaranteed since the CV folds overlap (Dietterich, 1998; Nadeau and Bengio, 2003).

A recent resurgence of Bayesian testing methods provides a promising alternative. The increased availability of Bayesian inference software like Stan (Carpenter et al., 2017) allows researchers to reap the benefits of the new paradigm, without having to implement efficient samplers themselves. In contrast to frequentist analyses based on p-values, Bayesian comparison produces estimates that can be inter-

Set-based	
Parameter	**Distribution**
Similarity	U(Jaccard, Containment, ContainmentMin, Cosine)

Soft-Cosine	
Parameter	**Distribution**
β	TruncNorm(1, 10, 3, 2)
Cut-off	TruncNorm(Min($W_{i,j}$), Max($W_{i,j}$), Percentile($W_{i,j}$, 75%), 2)

Smith-Waterman	
Parameter	**Distribution**
Match	TruncNorm(1, 10, 5, 2)
Mismatch	TruncNorm(-10, -1, -3, 2)
Gap-Penalty	TruncNorm(-10, -1, -3, 2)

Table 2.2: Algorithm-specific distributions employed for the random-search. "U" refers to uniform sampling from the given values. "TruncNorm" corresponds to a truncated normal distribution parameterized respectively by the minimum and maximum value, mean and standard deviation. $W_{i,j}$ refers to the embedding-based word similarity as measured by the cosine similarity. Parameter β refers to the power to which $W_{i,j}$ is raised in order to boost or reduce the relative difference in similarity between the upper and lower quantiles of the similarity distribution (Charlet and Damnati, 2017). "Cut-off" refers to a minimum threshold similarity below which the similarity of two words is ignored.

preted as the probability that a method is superior to an alternative one on a given evaluation measure. Moreover, Bayesian methods naturally provide a measure of uncertainty in the estimates of performance difference—while the frequentist counterpart is restricted to a binary decision on rejecting the null-hypothesis.

Following Corani et al. (2017), we employ a hierarchical model of the performance differences of two retrieval systems on multiple datasets. Let $\vec{x}_i = \{x_{i1}, \dots, x_{ik}\}$ be the vector of differences over k folds on the i^{th} dataset, then Equation 2.23 defines the target likelihood for q datasets:

$$\begin{aligned} \vec{x}_i &\sim \text{MVN}(\mathbf{1}\delta_i, \Sigma_i) \\ \delta_1, \dots, \delta_q &\sim \text{Student-T}(\delta_0, \sigma_0, \nu) \\ \sigma_1, \dots, \sigma_q &\sim \text{Uniform}(0, \bar{\sigma}) \end{aligned} \tag{2.23}$$

Here, the entries of vector $\vec{x}_i$ are modeled as coming from a multivariate normal with single mean δ_i—transformed into a k-dimensional vector by operator $\mathbf{1}$—, and a covariance matrix Σ_i with variances in the diagonal equal to σ_i^2 and off-diagonal co-variances equal to $\rho\sigma_i^2$.

The ρ term accounts for the mentioned fact that, due to overlapping folds, the observed differences in performance are correlated and, thus, not independent. Note that ρ is not directly modeled, but instead is kept fixed at $\rho = \frac{1}{k}$, following the heuristic from Nadeau and Bengio (2003). Finally, the mean differences δ_i are modeled with a Student-T distribution[11] with mean δ_0, scale factor σ_0 and degrees of freedom ν, and the standard deviations σ_i are assumed to be drawn from a Uniform distribution in the range 0 to $\bar{\sigma}$.[12]

Using the inferred posterior distribution over d_0,[13] we can draw samples of the expected difference in performance on unseen datasets. Finally, the arising distribution can be used to ask questions about the relative performance of the competing systems. As recommended by Benavoli et al. (2017), we based the comparisons on the notion of a Region Of Practical Equivalence (ROPE). The ROPE designates a range of performance differences within which the competing algorithms can be deemed to be equivalent. The selection of the ROPE depends on both the domain and the evaluation metric. For illustration, a difference of 1% or less in retrieval measures such as accuracy is often considered to be indicative of performance equivalence (Szymański and Gorman, 2020).

Finally, while the sketched comparison procedure is defined over system pairs, in our case we wish to conduct comparisons of multiple systems. In traditional frequentist procedures, multiple comparisons incur an inflated Type-I error rate and the significance threshold needs to be adjusted. In the case of Bayesian analysis, however, does not suffer from inflated false positive rates, as long as ROPE ranges are carefully chosen—as shown by Kruschke (2013).

2.6.3.2 Results

HYBRID EVALUATION We first focus on the evaluation of the candidate algorithms on the `blueletterbible` and `Sermons` datasets. We chose the hybrid evaluation procedure from Section 2.4.1.3, assuming that the relevant range of retrieved results extends to the first 20 items of the ranked list—or fewer if less are retrieved. As evaluation measure, we compute AP, which, as discussed in Section 2.4.1.1 allows us to describe the overall precision-recall trade-off without the need for fine-tuning the similarity threshold.

An alternative approach, that we do not pursue in the present experiments, is to include the similarity threshold as an additional hyper-parameter and compare systems on P, R and F-measure. Such an approach, however, would be counter-intuitive, since thresholds are commonly applied post-hoc and, in contrast to true hyper-parameters,

[11] Instead of the more common Normal distribution, a Student-T is used since it provides for more robust estimation in the presence of outliers.

[12] We refer to Corani et al. (2017) for a full description of the model and the priors.

[13] We use the `baycomp` package in order to fit the Bayesian models.

do not require the recomputing of the entire matrix of similarities. From the point of view of deployment, thresholds can be left to the user, who decides whether to focus on precision or recall.

For each pairwise comparison between algorithms, the Bayesian model outputs a posterior distribution of performance differences—which, as discussed above, can be interpreted as the expected performance difference on an unseen datasets.[14] We use this posterior distribution in order to compute the probability of each of the following three hypotheses: one in support of the practical equivalence of both systems (≈), and two more in support of the relative improvement of one of the systems over the other (left over right: >, and right over left: <). For each pairwise comparison we report the probabilities of the hypotheses considering three different ROPEs: 0.01, 0.02 and 0.05, which allows us to nuance the results against different values.

The left-column of Table 2.3 summarizes the results of the Bayesian model comparison on the hybrid evaluation setting using the AP scores obtained through 10-fold CV. As we can see, `Set-based` appears to be the weakest approach, being outperformed by the remaining three algorithms with a probability of at least 0.73 for all ROPE values. Next is `Smith-Waterman`, which is outperformed by `Tf-Idf` and `Soft-Cosine` with probabilities 0.64 and 0.63 at 0.01 ROPE—although at ROPE 0.05, the probability has diminished considerably, and the hypothesis of practical equivalence emerges as most plausible. Thus, the winning methods are `Tf-Idf` and `Soft-Cosine`, at least at ROPE 0.01. Between these two contenders, mild evidence of the superiority of `Soft-Cosine` is available (0.39 probability at ROPE 0.01). Nonetheless, practical equivalence emerges as the most plausible conclusion at higher ROPEs.

It is noteworthy that all runs produced high standard-deviations[15], which highlights problems with the practice of manual fine-tuning on small subsets. Furthermore, we observed that AP scores are markedly lower for the `blueletterbible` dataset than for the `Sermons` dataset. This may be a result of the smaller size of the gold standard, which implies that less data is available per fold for fine-tuning. But it may also indicate differences in the subtlety and difficulty of reuse. Interestingly, the variance that is found within the `blueletterbible` dataset differs across languages—being highest for the Latin corpus with an average 4.7 points vs. 3.3 and 3.7 for Ancient Greek and English—, which points at language-specific issues. Section 2.6.4 will zoom into this matter.

Lastly, a relevant aspect of hyper-parameter fine-tuning regards how capable the algorithms are of achieving optimal performance. For this purpose, we compared CV results obtained with an oracle that

[14] A full visualization of the distributions can be found in Appendix A.2.2.

[15] Figures A.2 and A.3 in the Appendix display the distribution of scores over the 10 CV folds for the `blueletterbible` and the `Sermons` datasets.

		Average Precision			NDCG		
		>	≈	<	>	≈	<
Set-based	Smith-Waterman	.08	.00	**.92**	.02	.00	**.98**
		.08	.00	**.92**	.01	.12	**.86**
		.07	.21	**.73**	.00	**.99**	.01
	Tf-Idf	.02	.00	**.98**	.06	.00	**.94**
		.02	.00	**.98**	.07	.02	**.91**
		.01	.08	**.90**	.03	**.82**	.15
	Soft-Cosine	.04	.00	**.96**	.01	.00	**.99**
		.03	.00	**.97**	.00	.00	**.99**
		.03	.10	**.88**	.01	.15	**.85**
Smith-Waterman	Tf-Idf	.31	.05	**.64**	.20	.07	**.73**
		.21	.33	**.47**	.12	**.45**	.43
		.06	**.83**	.10	.04	**.91**	.05
	Soft-Cosine	.15	.22	**.63**	.02	.00	**.97**
		.09	**.61**	.30	.02	.07	**.91**
		.02	**.95**	.03	.01	**.98**	.01
Tf-Idf	Soft-Cosine	.33	.27	**.39**	.12	.09	**.79**
		.14	**.68**	.18	.08	**.57**	.35
		.01	**.97**	.01	.02	**.96**	.02

Table 2.3: Summary of the Bayesian comparison for the hybrid evaluation using AP on the Sermons and blueletterbible datasets (left column), and the ranking-based evaluation using NDCG on the openbible dataset. For each pairwise system comparison, we show the probability that the left-hand side system performs better (>), both are practically equivalent (≈) and the right-hand side system performs better (<). Results are reported for three different ROPEs: 0.01, 0.02 and 0.05, which are shown respectively in the first, second and third lines of each block.

has access to the best hyper-parameters for each fold. In practice, the oracle is represented by the distribution of scores obtained by selecting the best random-search parameterization for each test fold, and, thus, corresponds to the maximum attainable performance per fold on the basis of the sampled parameterizations.

Overall, we observed only mild drops in average performance with respect to the oracle setting, which emphasizes the effectiveness of the fine-tuning process.[16] Still, since the magnitude of this drop seemed to vary across algorithms, a closer examination was carried. For this

[16] The full distribution of oracle scores can be seen in Figures A.2 and A.3 in Appendix A.2.1.

Method	AP			NDCG		
`Set-based`	.00	.10	**.94**	**.99**	**1.00**	**1.00**
`Smith-Waterman`	.26	**.84**	**1.00**	**.96**	**.99**	**1.00**
`Tf-Idf`	.38	**.82**	**.99**	**1.00**	**1.00**	**1.00**
`Soft-Cosine`	.01	.20	**.93**	**.98**	**.99**	**1.00**
ROPE	0.01	0.02	0.05	0.01	0.02	0.05

Table 2.4: Probability that the competing algorithms reach maximal performance in a CV evaluation at different ROPEs. Estimates are derived from the hierarchical model outlined in Section 2.6.3.1 and are inferred across all benchmark corpora considered in this study. Scores are highlighted in bold when the probability is higher than 0.5.

purpose, we conduct a comparison of the results of each algorithm in the CV and oracle setting. We utilize the Bayesian comparison method outlined in Section 2.6.3.1, but with the modification that the hyperprior for δ_0 is now lower-bounded by zero, since it is impossible for the CV results to improve on the oracle. The results of this comparison for the hybrid evaluation are shown in the left hand-side of Table 2.4.

As we can see, both `Set-based` and `Soft-Cosine` are least likely to reach maximum performance with negligible probabilities for ROPE 0.01. In contrast, `Smith-Waterman` and `Tf-Idf` are very likely to do so with probabilities higher than 0.8 for ROPE 0.02. Interestingly, this cannot be explained in terms of the number of hyper-parameters only, since `Smith-Waterman` is the algorithm with the largest number of hyper-parameters.

RANKING-BASED EVALUATION For the `openbible` dataset, the nature of the dataset points already at a ranking-based evaluation. We focus on the NDCG score to compare the models across the three Bible versions. The results of the Bayesian model comparison on the basis of NDCG are summarized in the right column of Table 2.3.

Overall, the results resemble the situation in the hybrid evaluation setting. `Set-based` is outperformed by the remaining candidates, although now only `Soft-Cosine` is plausibly better at ROPE 0.05 with 0.85 probability, while `Smith-Waterman` and `Tf-Idf` are likely to be practically equivalent to `Set-based` with probabilities of 0.99 and 0.82, respectively. In this setup, `Smith-Waterman` is more clearly outperformed by `Tf-Idf` and `Soft-Cosine` at ROPE 0.01—with probabilities of 0.73 and 0.97, respectively—but practically equivalent at ROPE 0.05. Finally, `Soft-Cosine` emerges more clearly as the "winner" from this evaluation perspective, outperforming `Tf-Idf` with 0.79 probability—while in the hybrid evaluation the probability was just 0.39. Still, at higher ROPEs, the hypothesis of practical equivalence becomes most plausible, reaching 0.96 at ROPE 0.05.

In contrast to the hybrid evaluation scenario, we now observe that the distribution of cross-validated scores stays close to their optimal values as per the oracle.[17] The right hand-side of Table 2.4 illustrates this matter, showing that all algorithms are very likely to reach maximum performance for all considered ROPEs.

2.6.4 Transferability

Besides a model comparison targeting the generalization capabilities of hyper-parameter choices to unseen datasets, we also provide a direct quantification of the effect of fine-tuning on a dataset that is different from the target dataset. In contrast to the experimental setting from Section 2.6.3, which assumes that fine-tuning data is available for the new dataset, we now consider the frequent case where researchers resort to their knowledge of hyper-parameter combinations that were efficient on previous datasets in order to parameterize the algorithm for a new dataset. This situation arises commonly in Literary Text Reuse Detection, where—as discussed in Section 2.3—annotated resources are scarce.

For this scenario, we run a set of experiments in which hyper-parameters are selected based on a dataset and applied to a different dataset. For these experiments, we dispense with the CV scenario, since we need to identify a single hyper-parameter combination per algorithm and dataset, and CV produces one for each fold. As a result of this choice, the performance drop must be interpreted as the expected drop in an ideal oracle setting in which the best hyper-parameters are known.

For the hybrid evaluation setting, 4 datasets are available—three belonging to the `blueletterbible` and the `Sermons` dataset—which amounts to 12 transferability experiments. For NDCG, only the three `openbible` datasets are available, which results in 6 transferability experiments. We report the average performance drop (and standard deviation) across the two major corpora in the present study: the three Bible versions vs. Bernard's Sermons. For NDCG we can only transfer within the Bible corpora, since no relevance labels are available for the `Sermons` dataset.

The results are shown in Table 2.5. For AP, transferring to the `Sermons` corpus is generally less costly than transferring to the Bible corpora. `Smith-Waterman` is the most sensitive algorithm in this respect, with an average drop of 14.94% when transferring from `Sermons` and 9.51% from other Bible versions. `Smith-Waterman` together with

[17] Similarly to the hybrid evaluation scenario, Figure A.4 in the Appendix A.2.1 shows the distribution of cross-validated NDCGs on the `openbible` dataset for the three corpora, including as well the distribution of scores for the oracle setting.

			Drop (%) – μ (σ^2)			
Source	Target	Method	AP		NDCG	
Sermons	Bible	`Set-based`	4.79	(4.87)		
		`Smith-Waterman`	14.94	(11.17)		
		`Tf-Idf`	1.91	(1.97)		
		`Soft-Cosine`	2.80	(2.57)		
Bible	Bible	`Set-based`	.0	(.0)	.48	(.6)
		`Smith-Waterman`	9.51	(8.92)	1.69	(1.25)
		`Tf-Idf`	2.77	(1.65)	.34	(.69)
		`Soft-Cosine`	5.17	(4.83)	7.28	(6.84)
	Sermons	`Set-based`	.83	(.0)		
		`Smith-Waterman`	.53	(.33)		
		`Tf-Idf`	1.39	(1.13)		
		`Soft-Cosine`	.77	(.91)		

Table 2.5: Summary of the hyper-parameter transfer experiments for the hybrid evaluation using AP (left hand-side) and the ranking-based evaluation using NDCG (right hand-side). Results correspond to average percentage drop (and standard deviation), and are grouped with respect to the underlying corpora.

`Soft-Cosine` are the algorithms with the largest number of hyper-parameters, and most are in a continuous scale. In this respect, the drop in performance observed for `Smith-Waterman` is understandable since a larger number of hyper-parameters should imply more flexibility to fit the reuse patterns of a corpus. In order to inspect this hypothesis, Figure 2.3 visualizes the AP and NDCG scores for 500 random parameterizations of `Smith-Waterman` and `Soft-Cosine`, highlighting the top performance runs in each measure. As we can see, in the case of `Soft-Cosine` a different parameterization is responsible for the highest AP than for the highest NDCG score. For `Smith-Waterman`, however, the same degree of flexibility cannot be observed. The region occupied by all parameterizations is comparatively smaller, and the winning parameter combination for one evaluation is very close in performance to the best combination in the other evaluation setting.

In contrast, the algorithms suffering the least in the transfer experiment are `Set-based` and `Tf-Idf`, both of which have the fewest number of hyper-parameters to fine-tune, and no hyper-parameters in a continuous scale. Finally, for NDCG, `Soft-Cosine` suffers a remarkable drop, which may be explained by the fact that two of the hyper-parameters relate to embeddings and are, thus, language-specific. However, such an effect is not observed when transferring from the Bibles into Sermons, where the drop for `Soft-Cosine` is less than 1%.

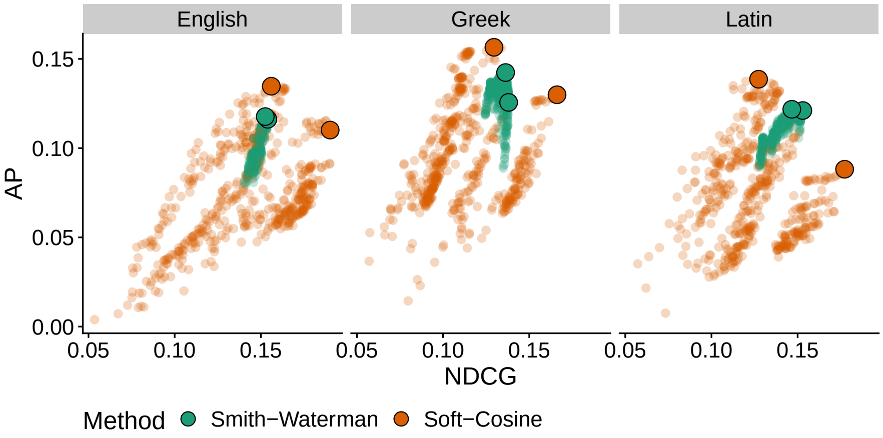


Figure 2.3: Scatter-plot of AP (on the y-axis) and NDCG scores (on the x-axis) for 500 randomly sampled parameterizations of the competing algorithms across three Bible versions. Highlighted are the top performance runs for each algorithm in each evaluation measure.

2.7 CONCLUSION & FUTURE WORK

The present study has highlighted a number of methodological issues surrounding the task of Literary Text Reuse Detection. First, we have clarified an ambiguity in the usage of the term "text reuse" within applications in the field of IR and NLP, where it is used as an umbrella term to refer to a set of related tasks—such as Plagiarism Detection, Paraphrase Identification or Semantic Textual Similarity—rather than to a well-defined task on its own. Second, we underlined how, despite the common subsumption of literary text reuse studies under the rubric of Text Reuse Detection, Literary Text Reuse Detection differs from the previously discussed specific Text Reuse Detection tasks not only in definition and goals but also in the difficulty to compile evaluation resources. Finally, we argued that these difficulties—most importantly, the costs associated with manual curation of benchmark corpora—may be responsible for the lack of established evaluation protocols in Literary Text Reuse Detection, where, as a result, it is not straight-forward to observe progress.

In order to address these shortcomings, we have identified three evaluation scenarios—a classification-based, a ranking-based and a hybrid one—that correspond to specific application cases of Text Reuse Detection algorithms, and have carried out a set of evaluation experiments involving three major algorithm families—one based on set-similarity and finger-printing approaches, one based on text alignment and two based on the VSM—and two benchmark datasets, spanning three corpora in three different languages.

To our knowledge, the experiments reported in Section 2.6 represent the first systematic comparison of a broad range of Text Reuse Detection algorithms in literary studies, taking into account generalization of hyper-parameter fine-tuning to unseen corpora, across different evaluation scenarios.

Using a hierarchical model of differences in cross-validated scores across multiple corpora, our experiments show that approaches based on set similarity measures are plausibly outperformed by approaches based on text alignment algorithms and VSMs. While set-based approaches are appealing due to their efficient run-times for large-scale datasets—e. g. web document de-duplication—, many use cases in the Humanities are located in a middle-size data regime and can certainly profit from more costly algorithms.

From those algorithms with stronger performance, it would be misleading to select a "winner" on the basis of the present evidence. Generally, statistical estimates of generalization favor VSMs. In particular, `Soft-Cosine`—a Generalized Vector Space variant that computes word-level similarities using word embeddings—obtains a positive probability estimate of superiority over the basic VSM variant, when considering a region of practical equivalence of 0.01 AP or NDCG. However, the most plausible hypothesis at larger regions is that of practical equivalence between the two.

Since the main difference between these two models involves the application of word embeddings, and the considered benchmark corpora vary in language, the question arises as to whether the quality of the word embeddings has an effect. In our case, the involved languages—Latin and Ancient Greek—belong to the category of low resource by current standards, considering that, in the best case, just over 100 million words are available for training, and word similarity benchmarks were only available for Latin.[18] For the third considered language, Early Modern English, the employed dataset—i. e. the BookCorpus—is commensurable to modern standards in size, but shifts in domain and semantics with respect to the target language may compromise the quality of the word representations. However, besides the quality of the word representations, the fact that the Generalized Vector Space Model does not outperform its simple counterpart as decisively as could be expected may be explained by the fact that not all corpora require semantic modeling to the same extent.

Moreover, our experiments highlight the importance of accounting for generalization, an aspect that is commonly omitted in Literary Text Reuse Detection studies. Especially in classification-based approaches using measures like AP, we observed that algorithms are overall likely to underperform their optimal performance, as estimated by an oracle that has access to the best hyper-parameter combinations on each CV fold. However, in Literary Text Reuse studies results are every so often reported without considering CV folds or standard splits—see, for instance, (Büchler, 2013, Chapter 5.3)—and may be, thus, inflated.

[18] See Appendix A.1 for experiments evaluating the quality of word embeddings for Latin.

Finally, we conducted a set of cross-lingual hyper-parameter transfer experiments, where the best performing hyper-parameters on a given dataset were applied to a different dataset, measuring the drop in performance incurred by different algorithms. This experiment replicates a common setting where a researcher applies previously acquired knowledge on new datasets. We showed that the drop in performance depends not only on the type—continuous vs. boolean or categorical—and number of hyper-parameters, but also on the target corpus—which indicates that some types of reuse may require more exhaustive fine-tuning than others.

We hope that our study provides a reference of evaluation approaches for further research to come. At the same time, we wish to emphasize the need for more substantial work before Text Reuse Detection can be established as a task, which should incentivize progress in the design of algorithms. Literary writers (un-)consciously resort to different strategies to establish links to other works, and as a results a variety of patterns emerge that require modeling different linguistic components—syntactic patterns, semantic relationships, etc. Manually constructing algorithms that target such patterns to a satisfactory degree is cumbersome, and future work should explore Machine Learning approaches that model the target phenomena in an end-to-end fashion. In particular, Siamese Networks (Bromley et al., 1994; Chopra, Hadsell, and LeCun, 2005), focused on modeling local interactions at lower levels and hierarchical interaction patterns at higher levels (Guo et al., 2016), constitute a promising venue of research. In contrast to representational-based Siamese Networks, which compress the documents into single vectors and match them on the basis of abstract semantic representations, so-called interaction-based approaches are able to capture word-level matches at lower levels, while still modeling more complex patterns at higher levels. As we argue, an important reason why statistical learning approaches, as the ones described, have not been as present as the contemporary research in Computational Linguistics would suggest, is the lack of benchmark corpora.

3 NEURAL LEMMATIZATION FOR HISTORICAL LANGUAGES

ABSTRACT Lemmatization of standard languages is concerned with two main problems. The first one consists in abstracting over morphological differences. The second one relates to resolving token-lemma ambiguities of inflected words in order to map them to a dictionary headword. In the present chapter we aim to improve lemmatization performance on a set of non-standard historical languages in which the difficulty is increased by a third aspect: spelling variation due to lacking orthographic standards. We approach lemmatization as a string transduction task with a neural encoder-decoder architecture, which we enrich with sentence context information using a hierarchical sentence encoder. We show significant improvements over the state of the art when training the sentence encoder jointly for lemmatization and language modeling. Crucially, our architecture does not require part-of-speech tags or morphological annotations, which are not always available and are particularly costly to obtain for historical corpora. Additionally, we test the proposed model on a set of typologically diverse standard languages showing results on par with or better than a model without enhanced sentence representations and previous state-of-the-art systems. Finally, our training procedure is shown to produce improved sentence-level representations in a set of "probing" experiments, in which these representations are used to predict available morphological tags.

This chapter is based on Enrique Manjavacas Arévalo, Ákos Kádár, and Mike Kestemont (June 2019). "Improving Lemmatization of Non-Standard Languages with Joint Learning." In: *Proceedings of the 2019 Conference of the North*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1493–1503. DOI: `10.18653/v1/N19-1153`. URL: `https://www.aclweb.org/anthology/N19-1153`

3.1 INTRODUCTION

Lemmatization can be defined as the task of mapping a token to its corresponding dictionary head-form (Knowles and Mohd Don, 2004). Thanks to lemmatizers, downstream applications can abstract away orthographic and inflectional variation. Lemmatization is considered an unproblematic—and, under circumstances, even a solved—task for resource-rich languages such as English, German or Spanish, which, coincidentally, present morphological systems of the analytic type. In contrast, lemmatization of languages with more involved morphological systems—e. g. Estonian or Latvian—still remains an open challenge. Additionally, historical languages—which, in comparison to modern languages, can be often considered low-resource languages—offer additional challenges due not only to complex morphological systems but also to unstable orthography. This latter case has come to the forefront in recent years with the emergence of computational applications in the Humanities. Common downstream tasks in this area involve Topic Modeling, Stylometry or Text Reuse Detection, in which lemmatization plays a crucial role as a preprocessing step.

In the case of standard languages, lemmatization complexity arises primarily from two sources. First, morphological complexity affects the number of inflectional patterns a lemmatizer has to model. Second, token-lemma ambiguities—e. g. the fact that a surface form "living" can refer to multiple lemmata: "living" or "live"—must be solved based on contextual information available in the sentence. In the case of historical languages, however, the aforementioned spelling variation introduces further complications. For instance, the regularity of the morphological system is drastically reduced since the evidence supporting token-lemma mappings becomes more sparse. As an example, while the modern Dutch lemma "jaar" (en. year) can be inflected in 2 different ways ("jaar", "jaren"), in a Middle Dutch corpus used in this study it is found in combination with 70 different forms ("iare", "ior", "jaer", etc.) Moreover, spelling variation increases token-lemma ambiguities by conflating surface realizations of otherwise unambiguous tokens—e. g. Middle Low German "bath" can refer to lemmata "bat" (en. bad) and "bidden" (en. bet) due to different spellings of the dental occlusive in final position.

Spelling variation is not exclusive of historical languages and it can be found in contemporary forms of communication like micro-blogs with loose orthographic conventions (Crystal, 2001). An important difference, however, is that while for modern languages normalization is feasible (Schulz et al., 2016), for many historic languages such is not possible, because one is dealing with an amalgam of regional dialects that lacked any sort of supra-regional variant functioning as target domain (Kestemont et al., 2016).

In the present study, we apply representation learning (LeCun, Bengio, and Hinton, 2015) to lemmatization of historical languages. In contrast to traditional Machine Learning approaches, which rely on hand-crafted features, Deep Learning approaches incorporate the feature extraction phase into the statistical learning. In our study, we supplement the lemmatization objective of the statistical learner with an additional Language Modeling (LM) objective in order to incentivize the extraction of more informative features. Our method shows improvements over a plain encoder-decoder framework, which reportedly achieves state-of-the-art performance on lemmatization and morphological analysis (Bergmanis and Goldwater, 2018).

CONTRIBUTIONS In particular, this study makes the following contributions:

1. We introduce a simple joint learning approach using on an ancillary bi-directional LM loss and achieve relative improvements in overall accuracy of 7.9% over an encoder-decoder trained without the joint LM loss and 30.72% over alternative edit-tree based approaches.

2. We provide a detailed analysis of the linguistic characteristics and corpus-based particularities of the target languages that help explain the amount of improvement can be expected from the proposed joint LM training.

3. We probe the hidden representations learned with the joint loss and find them significantly better predictors of part-of-speech tags and other morphological categories than the representations of the simple model, confirming the efficiency of the joint loss for feature extraction.

Additionally, we test our approach on a typologically varied set of modern standard languages and find that the joint LM loss significantly improves lemmatization accuracy of ambiguous tokens over the encoder-decoder baseline (with a relative increase of 15.1%), but that, in contrast to previous literature (Bergmanis and Goldwater, 2018; Chakrabarty, Pandit, and Garain, 2017), the overall performance of encoder-decoder models is not significantly higher than that of edit-tree based approaches. Taking into account the type of inflectional morphology dominating in a particular language, we show that the benefit of encoder-decoder approaches is highly dependent on the type of the morphological system. Finally, to assure reproducibility, we release all corpus preprocessing pipelines and train-dev-test splits. With this release, we hope to encourage future work on processing of lesser studied non-standard varieties.[1]

[1] Datasets and training splits are available at `https://www.github.com/emanjavacas/pie-data`, code can be obtained through the `pie` repository `https://www.github.com/emanjavacas/pie`.

OUTLINE The present chapter is structured as follows. First, in Section 3.2 we review recent relevant work on lemmatization. Next, in Section 3.3, we introduce the proposed architecture, detailing the basic encoder-decoder module (Section 3.3.1), the integration of information from the sentential context in order to allow for token-lemma disambiguation (Section 3.3.2), and our extension to improve the quality of the learned sentential representations (Section 3.3.3). Third, in Section 3.4, we describe the set of experiments underlying the present study, including a description of the datasets in Section 3.4.1, lemmatization baselines in Section 3.4.2, and the results in Section 3.5. Next, in Section 3.6, we present a discussion of the implications and proposed explanations of the results, and offer a fine-grained error analysis, illustrating difficulties of processing historical languages, as well as highlighting the relative advantages of the considered methods. Lastly, Section 3.7 ends the chapter with final conclusions.

3.2 RELATED WORK

Modern data-driven approaches typically treat lemmatization as a classification task where classes are represented by binary edit-trees induced from the training data. Figure 3.1 visualizes the binary edit-tree induced from the German token-lemma pair "vorgelegt" (en. presented) → "vorlegen" (en. to present). Given a token-lemma pair, its binary edit-tree is induced by first computing the prefix and suffix around the longest common subsequence, and recursively building a tree until no common character can be found. In the present case, at the first iteration the longest common subsequence corresponds to "leg", which splits the token into a 5-character long prefix ("vorge") and 1-character long suffix ("t"). At the second level, we first compare the suffixes "vorge" from the token and "vor" from the lemma and find the longest common subsequence "vor". The third level of this sub-branch already produces modification rules at the leaves, which, in this case, indicate that "vor" should be keep and "ge" deleted. On the other branch, the first-level suffix "t" results in a replace operation ("t" → "en"), since no common characters can be found. Importantly, these binary edit-trees are de-lexicalized since they only retain information about modifications on prefixes and suffixes, but are agnostic with respect to the stems. Thus, this same edit-tree correctly analyzes similarly structured token-lemma pairs—e. g. "vorgeplant" (en. preplanned) → "vorplanen" (to preplan) and "mitgesagt" (en. agreed) → "mitsagen" (en. to agree)—even though the prefixes are different and the stems vary in length. Edit-trees manage to capture a large proportion of the morphological regularity, especially for languages that rely on prefixation or suffixation for morphological inflection (e. g. Western European languages), for which the method was primarily designed.

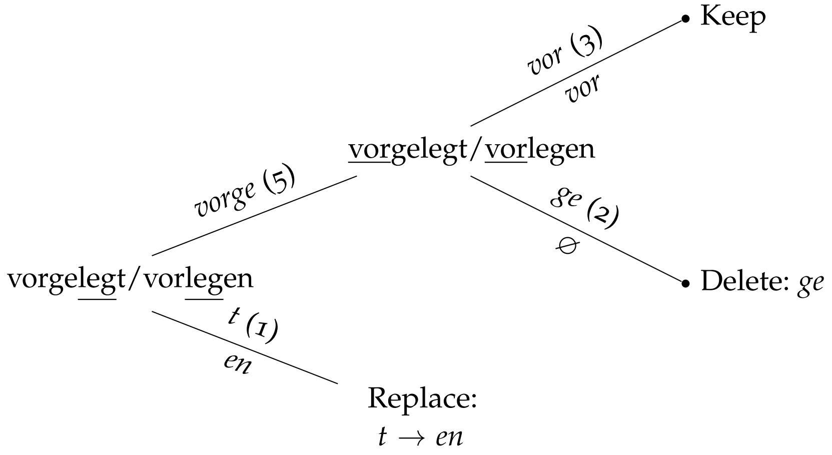


Figure 3.1: Example of an induced binary edit-tree for inflected German verb "vorgelegt" (en. presented) from lemma "vorlegen" (en. to present). Underlined characters correspond to the longest common subsequences at each node. In red are shown characters that have already been processed and are not being considered at that node. Leaves correspond to replace, delete or keep operations.

Based on edit-tree induction, different lemmatizers have been proposed. For example, Chrupala, Dinu, and van Genabith (2008) use a log-linear model for classification on top of a set of hand-crafted features to decode a sequence of edit-trees together with the sequence of part-of-speech tags using a beam-search strategy. A related approach is presented by Gesmundo and Samardži (2012), where edit-trees are extracted using a non-recursive version of the binary edit-tree induction approach. More recently, Cotterell, Fraser, and Schütze (2015) have used an extended set of features and a second-order Conditional Random Field to jointly predict part-of-speech tags and binary edit-trees with state-of-the-art performance. Finally, Chakrabarty, Pandit, and Garain (2017) employed a softmax classifier to predict edit-trees based on sentence-level features implicitly learned with a neural encoder over the input sentence.

With the advent of contemporary encoder-decoder architectures, lemmatization as a string transduction task has gained interest, partly due to the success of these architectures in Neural Machine Translation. For instance, Bergmanis and Goldwater (2018) apply a state-of-the-art Neural Machine Translation system with the lemma as target and as source the focus token, using a fixed window over neighboring tokens. Most similar to our work is the approach by Kondratyuk, Gavenčiak, and Straka (2018), which conditions the decoder on sentence-level distributional features extracted from a sentence-level bi-directional Recurrent Neural Network (RNN) and additionally predicted morphological tags. Finally, a string transduction approach that predates neural lemmatizers is presented by Juršic et al. (2010), who induce suffix replacement rules, or "ripple-down-rules", in order to transduce an inflected form into its lemma.

With respect to lemmatization of non-standard historical varieties, recent work has focused on spelling normalization using rule-based, statistical and neural string transduction models (Bollmann and Søgaard, 2016; Pettersson, Megyesi, and Nivre, 2014; Tang et al., 2018). Previous studies on lemmatization of historical variants focused on evaluating off-the-shelf systems. A particularly relevant example is given by Eger, Gleim, and Mehler (2016), who evaluate different preexisting models on a dataset of German and Medieval Latin. Dereza (2018) focuses, instead, on Early Irish. In this area, the most similar study to the present one is work by Kestemont et al. (2016), which tackled lemmatization of Middle Dutch with a neural encoder that extracts character and word-level features from a fixed-length token window and predicts the target lemma from a closed set of lemmata.

Using an LM loss in a Transfer Learning setup—i. e. jointly optimizing a neural architecture on an ancillary task in order to improve the quality of extracted features—has gained momentum in the last few years, partly due to overall performance improvements obtained across multiple tasks such as Named Entity Recognition, part-of-speech tagging, Question Answering, etc. Different models have been proposed around the same idea varying in implementation, optimization and ancillary task. For instance, Howard and Ruder (2018) present a method to fine-tune a pre-trained Language Model for text classification. Peters et al. (2018) learn task-specific weighting schemes over different layer features extracted by a pre-trained bi-directional Language Model. Recently, Akbik, Blythe, and Vollgraf (2018) used context-sensitive word embeddings extracted from a bi-directional character-level Language Model to improve Named Entity Recognition, part-of-speech tagging and chunking.

3.3 ARCHITECTURE

In this section, we describe the proposed encoder-decoder architecture for lemmatization. In Section 3.3.1 we start by describing the basic formulation of the encoder-decoder as it is known from the literature on Neural Machine Translation. Next, Section 3.3.2 shows how sentential context can be integrated into the decoding process as an extra source of information. Finally, Section 3.3.3 describes how to learn richer representations for the encoder through the addition of the ancillary LM task.

3.3.1 Encoder-Decoder

We employ a character-level encoder-decoder architecture that processes input tokens x_t character-by-character. The goal of this architecture is to decode the target lemma l_t character by character, conditioned on an intermediate representation of x_t.

3.3.1.1 Encoder

For a given token x_t, a sequence of character embeddings $c_1^x, \dots, c_n^x$ is extracted from embedding matrix $W_{enc} \in \mathbb{R}^{|C| \times d}$—where $|C|$ and d represent, respectively, the size of the character vocabulary and the dimensionality of the character embeddings. These character embeddings are passed to a bi-directional RNN encoder that computes a forward and a backward sequence of hidden states: $\overrightarrow{h_1^{enc}}, \dots, \overrightarrow{h_n^{enc}}$ and $\overleftarrow{h_1^{enc}}, \dots, \overleftarrow{h_n^{enc}}$. The final representation of each character is the concatenation of the forward and the backward states: $h_i^{enc} = [\overrightarrow{h_i^{enc}}; \overleftarrow{h_i^{enc}}]$. Each h_i^{enc} represents a token-level abstract representation of the corresponding character.

3.3.1.2 Decoder

Similarly to the encoder, at each decoding step j, the character-level RNN decoder takes embeddings of the previously decoded lemma character c_{j-1}^l from embedding matrix $W_{dec} \in \mathbb{R}^{|L| \times d}$ and processes it in combination with the previous hidden state h_{j-1}^{dec} in order to generate a new hidden state h_j^{dec}. Each hidden state h_j^{dec} can be transformed into a richer vector representation r_j in order to incorporate additional token-level contextual information. In the most traditional form, r_j is a summary vector obtained via an attentional mechanism (Bahdanau, Cho, and Bengio, 2015), which distills any relevant information for the prediction of the next lemma character from the encoder activations $h_1^{enc}, \dots, h_n^{enc}$ and the previous decoder state h_{j-1}^{dec}. Attention mechanisms have been proven to generate implicit alignments between the input and the currently decoded sequence that facilitate the decoding of further items.[2]

Finally, the output logits for the j^{th} character are computed by a linear projection of the current summary vector r_j and a learned projection matrix $O \in \mathbb{R}^{H \times |L|}$—where H corresponds to the dimensionality of the hidden layer and $|L|$ to the length of the output vocabulary of lemma characters. In order to produced an output probability distribution over the lemma characters, the logits are normalized using the $\mathtt{softmax}$ function. Formally, Equation 3.1 shows the computation for

[2] We refer to Bahdanau, Cho, and Bengio (2015) for the full description and analysis of the attentional mechanism employed in this study.

the probability of the k^{th} character in the vocabulary L during the decoding of the j^{th} lemma output character.

$$P(c^l_{j,k}|c^l_{1...j-1}, x_t) = \frac{\exp(O \cdot r_j)_k}{\sum_{k' \in L} \exp(O \cdot r_j)_{k'}} \tag{3.1}$$

The model is trained to maximize the probability of the target character sequence expressed in Equation 3.2, using a teacher-forcing training regime. In teacher-forcing, the true lemma character is fed to the decoder during training, independently of whether the logits generated in the previous step by the decoder are consistent with the gold output character.

$$P(l_t|x_t) = \prod_{j=1}^{m} P(c^l_j|c^l_{<j}, r_j; \theta_{enc}, \theta_{dec}) \tag{3.2}$$

where m is the number of characters in the target lemma.

3.3.2 Incorporating Sentential Context

Lemmatization of ambiguous tokens can be improved by incorporating sentence-level information. Our architecture is similar to the one used by Kondratyuk, Gavenčiak, and Straka (2018) in that it incorporates sentence-level representations of the input tokens using an additional bi-directional RNN deployed—as we shall see—in a hierarchical manner. More concretely, we re-use the last hidden state of the character-level bi-directional RNN encoder from Section 3.3.1 as word-level token representations: $w_t = [\overrightarrow{h^{enc}_t}; \overleftarrow{h^{enc}_t}]$. Optionally, these word-level representations can be enriched with extra word embeddings coming from an additional matrix $W_{word} \in \mathbb{R}^{|V| \times e}$—where V and e denote respectively the vocabulary size in words and the word embedding dimensionality. During development experiments, however, word embeddings did not contribute significant improvements on historical languages and we therefore exclude them from the rest of the experiments. It must be noted, however, that word embeddings might still be helpful for lemmatization of standard languages where the type-token ratio is smaller as well as when pre-trained embeddings are available. The additional sentence-level bi-directional RNN uses these word-level features w_t and computes sentence-level token representations s_t as the concatenation of forward and backward activations $s_t = [\overrightarrow{s_t}; \overleftarrow{s_t}]$ over the input sequence of w_t.

In order to perform sentence-aware lemmatization for token x_t, we condition the decoder on the sentence-level encoding s_t and optimize the probability given by Equation 3.3.

$$P(l_t|x_t) = \prod_{j=1}^{m} P(c^l_j|c^l_{<j}, r_j, s_t; \theta_{enc}, \theta_{dec}) \tag{3.3}$$

In practice, the conditioning is implemented by concatenating s_t to each summary vector r_j before computing the output logits.

This hierarchical architecture ensures that both the word-level and the character-level features of each input token can contribute to the extraction of sentence-level features at any given step, and, due to the conditioning of the decoder on these sentence-level features, to the lemmatization of any other token in the sentence. Figure 3.2 visualizes the proposed hierarchical architecture. We hypothesize that the inclusion of sentence-level features in this manner enables the decoder to resolve ambiguities in the token-lemma pairs. From this perspective, our architecture is more general than those presented in Kestemont et al. (2016) and Bergmanis and Goldwater (2018), where sentence information is included by running the encoder over a predetermined fixed-length window of neighboring characters. Moreover, we re-use the character-level embedding matrix of the encoder in order to compute character embeddings for the decoder—i. e. $W_{enc} = W_{dec}$—in a process that resembles weight sharing (Inan, Khosravi, and Socher, 2017; Press and Wolf, 2017).

3.3.3 Improved Sentence-level Features

We hypothesize that the training signal from lemmatization alone might not be enough to extract sentence-level features with sufficiently strong disambiguating power. For this reason, we include an additional bi-directional word-level LM loss over the input sentence, which is defined as follows.

Given the forward $\overrightarrow{s^t}$ and backward $\overleftarrow{s^t}$ sub-vectors of the sentence encoding, we train two additional softmax classifiers to predict token x^{t+1} given $\overrightarrow{s_t}$ and x^{t-1} given $\overleftarrow{s_t}$ with parameters O_{LMfwd} and $O_{LMbwd} \in \mathbb{R}^{S \times |V|}$. Equation 3.4 shows formally the computation of the probability of the k^{th} output token in the vocabulary V for the forward and backward LM:

$$\begin{aligned} P_{LMfwd}(x_{t+1,k}|x_{1...t}) &= \frac{\exp(O_{LMfwd} \cdot \overrightarrow{s_t})_k}{\sum_{k' \in V} \exp(O_{LM} \cdot \overrightarrow{s_t})_{k'}} \\ P_{LMbwd}(x_{t-1,k}|x_{n...t}) &= \frac{\exp(O_{LMfwd} \cdot \overleftarrow{s_t})_k}{\sum_{k' \in V} \exp(O_{LM} \cdot \overleftarrow{s_t})_{k'}} \end{aligned} \qquad (3.4)$$

where n corresponds to the length of the sequence of tokens.

During development experiments, we have found that the joint loss is most effective when both forward and backward classifiers shared parameters, which reduces the risk of overfitting.

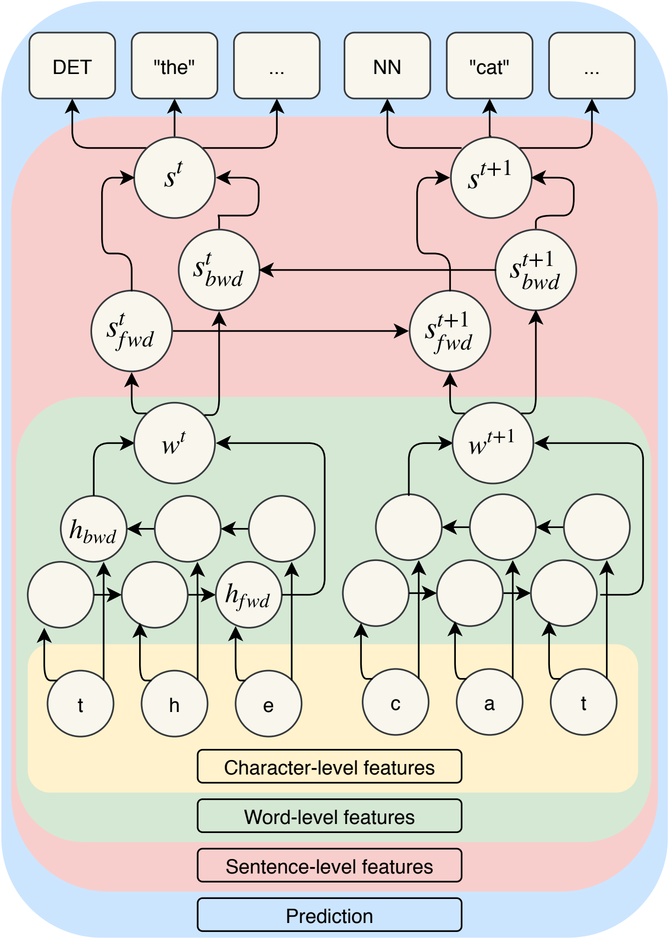


Figure 3.2: Diagram of the hierarchical sentence encoding architecture, highlighting feature extraction at character, word and sentence-level. The outermost layer corresponds to any given number of morphological classifiers trained on top of the hierarchically extracted features.

We train our model to jointly minimize the negative log-likelihood of the probability defined by Equation 3.3 and the LM probability defined by Equation 3.5.

$$\begin{aligned} P_{LM}(\mathbf{x}) = 1/2 \prod_{t=2}^{n} P(x_t|x_1, \ldots, x_{t-1}) \\ + 1/2 \prod_{t=1}^{n-1} P(x_t|x_n, \ldots, x_{t+1}) \end{aligned} \tag{3.5}$$

Following ideas from Multi-Task Learning introduced by Caruana (1997), we set a weight on the LM loss, which controls the contribution of this objective in the overall training loss. Importantly, this weight is decreased during training based on lemmatization accuracy on development data, which allows for reducing the influence of the

Language	Dataset	Code
Middle Dutch	Gys (Admin)	cga
	Gys (Literary)	cgl
	Religious	cgr
	Adelheid	crm
Medieval French	Geste	fro
Medieval Latin	Capitularia	cap
	LLCT1	llat
Middle Low German	ReN	gml
Slovenian	goo300k	goo

Table 3.1: Corpus identifier and description in the historical languages dataset. "Gys" refers to the Gysseling corpus, which consists of several subsets.

ancillary tasks on the overall training regime, once the ancillary tasks have reached convergence.

3.4 EXPERIMENTS

We now report on the conducted experiments. Section 3.4.1 first discusses the employed datasets, both the newly introduced dataset of historical languages, and the dataset of modern standard languages sampled from Universal Dependencies (v2.2) corpus (Nivre et al., 2016). Finally, Section 3.4.2 describes model training and settings in detail.

3.4.1 Datasets

3.4.1.1 Historical Languages

In recent years, a number of historical corpora have appeared thanks to an increasing number of digitization initiatives (Piotrowski, 2012). For the present study, we chose a representative collection of medieval and early modern datasets. In order to improve reproducibility, we favored corpora that were publicly available. Moreover, we tried to utilize corpora that had already been used in similar research, in order to improve the informativeness of our experiments. Finally, we tried to cover as many genres and historic periods as it was possible. The resulting dataset includes a total of 8 corpora covering Middle Dutch, Middle Low German, Medieval French, Historical Slovene and Medieval Latin. Table 3.1 shows the dataset sources and codes from our Historical Languages.

The following paragraph contains a discussion of the sources of the used corpora as well as a short description of the literature they cover.

For **Medieval Dutch**, both `cga` and `cgl` contain medieval Dutch material from the Gysseling corpus curated by the Institute for Dutch Lexicology[3] `cga` is a collection of charters (administrative documents), whereas `cgl` concerns a variety of literary texts that greatly differ in length. `crm` is another Middle Dutch charter collection from the 14th century with wide geographic coverage (Van Reenen and Mulder, 1993; van Halteren and Rem, 2013). Finally, `cgr` is a smaller collection of samples from Middle Dutch religious writings that include later medieval texts (Kestemont et al., 2016).

For **Medieval French**, `fro` offers a corpus of Old French heroic epics, known as *chansons de geste* (Camps et al., 2019). **Medieval Latin** is represented by `llat`, a dataset taken from the Late Latin Charter Treebank consisting of early Medieval Latin documentary texts (Korkiakangas and Lassila, 2013), and `cap`, which is a corpus of early Medieval Latin ordinances decreed by Carolingian rulers. The `cap` corpus has served as the basis to a number of Latin lemmatization studies (Eger, Gleim, and Mehler, 2016; Kestemont and Gussem, 2017; vor der Brück, Eger, and Mehler, 2015).

Middle Low German is represented by `gml`, which corresponds to the reference corpus of Middle Low German and Low Rhenish texts, consisting of manuscripts, prints and inscriptions (Barteld et al., 2017). Finally, `goo` comes from the reference corpus of **historical Slovene**, sampled from 89 texts from the period 1584–1899 (Erjavec, 2015).

3.4.1.2 Standard Languages

For a more thorough comparison between systems across domains and a better examination of the effect of the LM loss, we evaluate our systems on a set of 20 standard languages sampled from the UD corpus, trying to guarantee typological diversity while selecting datasets with at least 20k words. We use the pre-defined splits from the original UD corpus (v2.2). Table 3.2 shows the languages from the UD corpus that were sampled for the study. We have used ISO 639-1 codes (instead of the more general ISO 639-2) in order to avoid clutter in the presentation of results.

In the cases where train-dev-test splits were not pre-defined, we randomly split sentences using 10% and 5% for test and dev respectively.[4] Figure 3.3 visualizes the test set sizes in terms of total, ambiguous and unknown tokens for both historical and standard languages.

[3] `https://ivdnt.org/taalmaterialen`.

[4] The splits can be reproduced using the code and data release via `https://www.github.com/emanjavacas/pie-data`.

Language	Dataset	Code
Arabic	Arabic-PDAT	ar
Bulgarian	Bulgarian-BTB	bg
Czech	Czech-CAC	cs
German	German-GSD	de
English	English-EWT	en
Spanish	Spanish-AnCora	es
Estonian	Estonian-EDT	et
Basque	Basque-BDT	eu
Persian	Persian-Seraji	fa
Finnish	Finnish-TDT	fi
French	French-GSD	fr
Hebrew	Hebrew-HTB	he
Hungarian	Hungarian-Szeged	hu
Italian	Italian-ISDT	it
Latvian	Latvian-LVTB	lv
Norwegian (Bokmaal)	Norwegian-Bokmaal	nb
Russian	Russian-SynTagRus	ru
Slovenian	Slovenian-SSJ	sl
Turkish	Turkish-IMST	tr
Urdu	Urdu-UDTB	ur

Table 3.2: Standard language corpora from the Universal Dependencies (v2.2) dataset.

3.4.2 Models

We now present the models used for the experiments. We refer to the full model trained with joint LM loss by `Sent-LM`. In order to test the effectiveness of sentence information and the importance of enhancing the quality of the sentence-level feature extraction, we compare this model against a simple encoder-decoder model without sentence-level information (`Plain`) and a model trained without the joint LM loss (`Sent`). Moreover, we compare to previous state-of-the-art lemmatizers based on binary edit-tree induction: `Morfette` (Chrupala, Dinu, and van Genabith, 2008) and `Lemming` (Cotterell, Fraser, and Schütze, 2015), which we run with default hyper-parameters.

For all our models, we use the following hyper-parameter values. All recurrent layers have 150 cells per layer using the Gated Recurrent Unit (GRU) (Cho et al., 2014) as recurrent cell, which is known to have comparable performance to Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) with smaller number of parame-

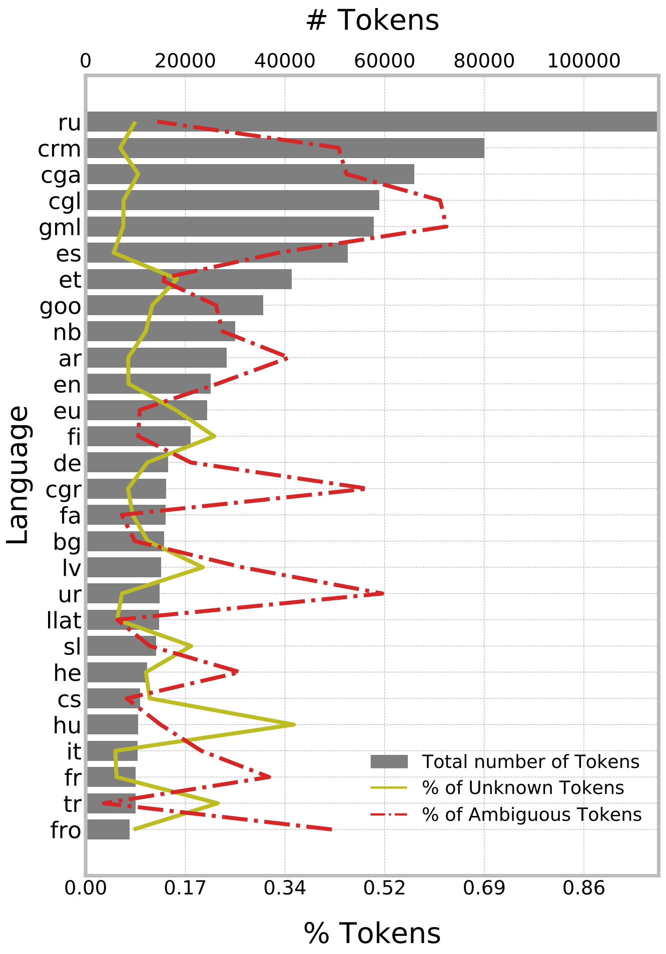


Figure 3.3: Statistics of total number of tokens, ambiguous and unknown tokens in the test sets. The full list of languages for both historical and standard corpora as well as the corresponding ISO 639-1 codes used in the present study can be found in Tables 3.1 and 3.2. Statistics for unknown and ambiguous tokens are shown as percentages.

ters. Encoder and decoder have two layers but the sentence encoder has only one. We apply 0.25 dropout (Srivastava et al., 2014) after the embedding layer and before the output layer and 0.25 variational dropout (Gal and Ghahramani, 2016) in between recurrent layers. Models are optimized using Adam (Kingma and Ba, 2015) with an initial learning rate of $1e-3$ which is reduced by 25% after each epoch without improvement on development accuracy. Models are trained until failing to achieve an improvement for three consecutive epochs. Initial LM loss weight is set to 0.2 and it is halved each epoch after two consecutive epochs without achieving any improvements on development perplexity.

In order to segment the corpora into sentences, we use sentence boundaries when given and otherwise use part-of-speech tags corresponding to full stops as clues. In any case, sentences are split into

	Full	Ambiguous	Unknown
`Edit-Tree`	91.11	91.79	35.48
`Plain`	91.61	87.39	65.69
`Sent`	93.4	91.14	**66.98**
`Sent-LM`	**94.0**	**92.81**	65.39

Table 3.3: Average accuracy across historical languages for the compared models. The combined results for `Lemming` and `Morfette` are shown aggregated by taking the best performing model per dataset.

chunks of maximum 35 words to accommodate to limited working memory. Target lemmata during both training and testing are lowercased in agreement with the implementation of `Lemming` and `Morfette`, which also do so. For models with joint loss, we truncate the output vocabulary to the top 50k most frequent words for similar reasons. We run a maximum of 100 optimization epochs in randomized batches containing 25 sentences each. The learning rate is decreased by a factor of 0.75 after every 2 epochs without accuracy increase on held-out data and learning stops after failing to improve for 5 epochs. Decoding is done with beam search with a beam size of 10, which resulted in relatively small gains ranging from 0.1% to 0.5% in overall accuracy for all languages.

3.5 RESULTS

As is customary, we report exact-match accuracy on target lemmata. Besides overall accuracy, we also compute accuracy of ambiguous tokens—i. e. tokens that map to more than one lemma in the training data—and unknown tokens—i. e. tokens that do not appear in the training data). All numbers correspond to micro-averaged accuracies.

3.5.1 Historical Languages

Table 3.3 shows the aggregated results over all datasets in our historical language corpus. We aggregate both edit-tree based approaches by selecting the best performing model for each corpus. We note that when `Lemming` managed to converge, the results were overall better than those obtained by `Morfette`. The latter approach, however, managed to converge on a larger set of corpora. Specifically, `Lemming` failed to converge in 4 cases—`cga`, `cgl`, `crm` and `gml`—due to memory requirements exceeding the available 250G RAM. We hypothesized that this is due to the large amount of edit-trees caused by orthographic and

	Full	Ambiguous	Unknown
K-2016	91.88		51.64
Edit-Tree	89.01	91.18	20.46
Plain	90.21	87.4	61.93
Sent	92.55	91.6	**64.61**
Sent-LM	**93.25**	**93.31**	62.1

Table 3.4: Average accuracy for the Gysseling subcorpora comparing all considered models as well as results from Kestemont et al. (2016).

morphological variation. Following Søgaard et al. (2014), we compute p-values with a Wilcoxon's signed rank test.

As we can see, Sent-LM is the best performing model with a relative improvement of 7.9% ($p < .01$) over Sent and 30.72% ($p < .01$) over the edit-tree approach when considering entire datasets. When considering ambiguous tokens only, the improvement amounts to a 10.27% ($p < .1$) over Sent and 18.66% ($p < .01$) over the edit-tree approach. Moreover, the edit-tree approach outperforms encoder-decoder models Plain and Sent on ambiguous tokens, and it is only due to the joint loss that the encoder-decoder paradigm gains an advantage. Finally, for tokens unseen during training, the best performing model is Sent with a relative error reduction of 47% ($p < .01$) over the edit-tree approach and 4.77% ($p < .1$) over Sent-LM.

Table 3.4 compares scores for a subset of the corpora coming from the Gysseling corpus, which have been used in previous work on lemmatization of historical languages. The model described by Kestemont et al. (2016) is included as K-2016 for comparison. Unfortunately, scores on ambiguous tokens were not reported on that study and, therefore, the model cannot be compared in this regard.

It is apparent that both Sent and Sent-LM outperform K-2016 on full and unknown tokens. It is worth noting that K-2016, a model that uses distributed contextual features but no edit-tree induction, performs better than Plain—which highlights the importance of context for the lemmatization of historical languages—, and also better than the edit-tree approaches—which highlights the difficulty of tree induction on this dataset. We find that Sent-LM has a significant advantage over Sent on full and ambiguous tokens, but a disadvantage with respect to both Sent and Plain on unknown tokens.

	Full		
	Type 1	Type 2	Type 3
`Edit-Tree`	96.34	93.02	**98.37**
`Plain`	96.21	94.6	97.5
`Sent`	96.42	94.41	97.84
`Sent-LM`	**96.52**	**94.62**	97.86
	Ambiguous		
`Edit-Tree`	92.47	**92.47**	**97.5**
`Plain`	88.87	90.51	95.34
`Sent`	91.12	92.01	96.65
`Sent-LM`	**93.01**	92.07	97.48
	Unknown		
`Edit-Tree`	84.99	74.66	**91.39**
`Plain`	85.43	**83.78**	86.37
`Sent`	**85.44**	84.02	86.67
`Sent-LM`	85.15	83.32	85.36

Table 3.5: Average accuracy per language group for standard languages.

3.5.2 Standard Languages

Table 3.6 shows overall accuracy scores aggregated across languages.[5] We observe that on average `Sent-LM` is the best model on full datasets. However, in contrast to previous results, the edit-tree approach has an advantage over all encoder-decoder models for both ambiguous and unknown tokens.

Since the differences in performance are not statistically significant at $p > 0.05$, we seek to shed light on the advantages and disadvantages of the encoder-decoder and edit-tree paradigms by conducting a more fine-grained analysis with respect to the morphological typology of the considered languages. To this end, we group languages into morphological types based on the dominant morphological processes of each language and compute aggregate performance scores over languages in each type. We note that this is not meant as a strict systematic categorization of languages in morphological terms but rather as an informative classification aimed at facilitating a more nuanced interpretation of the results.

We used the following three morphological groups, which we briefly characterize in the next paragraph.

[5] Similarly to results on historical languages, we aggregate `Morfette` and `Lemming` due to the later failing to converge on `et`.

	Full	Ambiguous	Unknown
Edit-Tree	96.1	**94.35**	**83.26**
Plain	95.93	91.44	83.02
Sent	96.19	93.25	82.61
Sent-LM	**96.28**	94.08	82.58

Table 3.6: Average accuracy across all 20 standard languages. Lemming and Morfette are shown aggregated by taking the best performing model per dataset.

Type 1. Balto-Slavic languages, which are known for their strongly suffixing morphology and complex case system. This group is represented in our corpora by Bulgarian (bg), Czech (cs), Latvian (lv), Russian (ru) and Slovenian (sl).

Type 2. Uralic and Altaic languages, which are characterized by agglutinative morphology and a tendency towards mono-exponential case and vowel harmony. In our corpora, this group comprises Estonian (et), Finnish (fi), Hungarian (hu) and Turkish (tr).

Type 3. Western European languages with a tendency towards synthetic morphology and partially lacking nominal case. Type 3 encompasses German (de), English (en), Spanish (es), French (fr), Italian (it) and Norwegian Bokmål (nb).

Table 3.5 shows accuracy scores per morphological group for each model type. It is apparent that the Edit-Tree approach is very effective for Type 3 languages both in ambiguous and unknown tokens. In both Type 1 and Type 2 languages, the best overall performing model is Sent-LM. In the case of ambiguous tokens, Sent-LM achieves highest accuracy for Type 1 languages, but it is surpassed by the Edit-Tree approach on Type 2 languages. Finally, in the case of unknown tokens, we observe a similar pattern to the historical languages where Plain and Sent have an advantage over Sent-LM.

3.6 DISCUSSION

For clarity, we group the discussion of the main findings according to four major discussion points.

3.6.1 How does the Joint Language Model Loss Help?

As Section 3.5 shows, Sent-LM is the overall best model, and its advantage is biggest on ambiguous datasets, always outperforming the second-best encoder-decoder model on ambiguous tokens. For a more

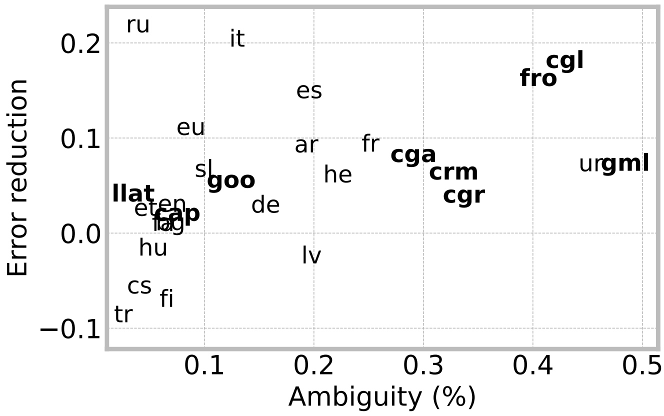


Figure 3.4: Error reduction of `Sent-LM` vs. `Sent` by percentage of ambiguous tokens (Spearman's $R = 0.53$; $p < .01$). Labels of historical languages are shown in bold.

detailed comparison of the two models, we tested the following two hypotheses.

Hyp. 1 The joint LM loss helps by providing sentence representations with stronger disambiguation capacities.

Hyp. 2 The joint LM loss helps in cases where the evidence of a token-lemma relationship is sparse—e. g. in languages with highly synthetic morphological systems and in the presence of spelling variation.

In order to approach Hyp. 1, we compute the Spearman correlation between the improvement of `Sent-LM` over `Sent` and the percentage of token-lemma ambiguity per corpus—computed by the percentage of ambiguous tokens in the corpus. Figure 3.4 shows a scatter-plot of performance improvement and percentage of token-lemma ambiguity underlying the resulting Spearman correlation of 0.53 ($p < .01$). The existing correlation provides evidence in favor of Hyp. 1.

In order to tackle Hyp. 2, Figure 3.5 visualizes the relationship between improvement of `Sent-LM` over `Sent` and the token-lemma ratio. Similarly, we obtain a positive correlation of 0.47 ($p < .05$), suggesting that, in line with Hyp. 2, the learned representations help in cases where the input token is harder to identify.

These two aspects help explain the efficiency of the joint learning approach on non-standard languages, where high levels of spelling variation provide increased ambiguity by conflating unrelated forms and also lower evidence for token-lemma mappings. Another factor certainly related to the efficiency of the proposed joint LM loss is the size of the training dataset. Still, while sufficient dataset size is a necessary condition for the viability and efficiency of the joint LM loss, it does not, however, offer a sufficient reason and has therefore weak

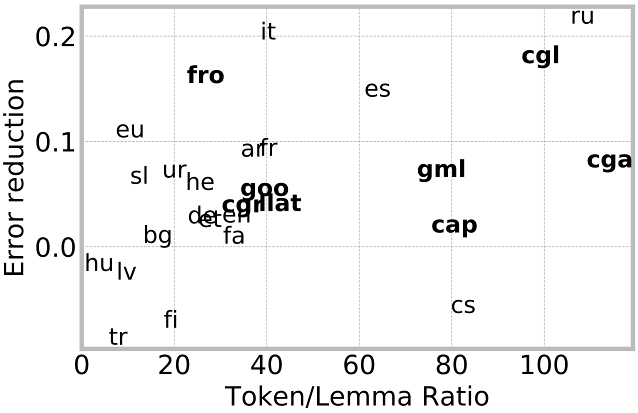


Figure 3.5: Error reduction of `Sent-LM` vs. `Sent` by Token-Lemma ratio on 50k tokens of the training set (Spearman's $R = 0.47$; $p < .05$). Labels of historical languages are shown in bold.

explanation power for the observed improvements of the proposed approach.

3.6.2 Better Representations from Joint Learning

In order to analyze the representations learned with the joint loss, we turn to "representation probing" experiments following recent approaches on interpretability (Adi et al., 2017; Linzen, Dupoux, and Goldberg, 2016). Using the same train-dev-test splits from the current study, we exploit additional part-of-speech tags, and other morphological tags such as number, linguistic gender, case as well as syntactic function annotations provided in the UD corpora, and compare the ability of the representations extracted by `Sent` and `Sent-LM` to predict these tags. Since not all tasks are available for all languages—due to some corpora not providing all annotations and some categories not being relevant for particular languages—the number of corpora per task varies.

For these probing experiments, we freeze all model parameters and train an additional linear softmax classifier $Q \in \mathbb{R}^{H \times V}$ per task using a cross-entropy loss function. We train these classifiers for 50 epochs using the Adam optimizer with default learning rate and early-stop training after 2 epochs without an increase in accuracy on the development set. The difference in accuracy obtained by classifiers trained on the features extracted by the different encoders can be interpreted as the effect of the quality of the underlying sentence-level feature extractors.

The results of this experiment are reported in Table 3.7. As we can see, the classifier trained with `Sent-LM` outperforms the one with `Sent` on all considered labeling tasks. Moreover, the accuracy obtained with representations coming from the `Sent` model is frequently lower than

	Pos	Dep	Gender	Case	Num
Majority	82.61	62.93	85.83	80.76	83.92
Sent	79.27	64.14	83.39	83.01	81.01
Sent-LM	**83.62**	**68.36**	**86.55**	**84.44**	**84.37**
Support	29	20	15	19	20

Table 3.7: Overall accuracy of Sent and Sent-LM models and a Majority baseline on 5 probing tasks and actual number of languages per morphological category. All differences over Sent except for Case are significant at $p < 0.05$. Support is shown in terms of the number of languages exhibiting such grammatical distinctions.

the one obtained by a majority baseline. This experiment, thus, confirms the efficiency of the LM loss at extracting more morphologically relevant representations.

3.6.3 Edit-tree vs. Encoder-Decoder

The fine-grained analysis on standard languages presented in Table 3.5 suggests that the performance of the edit-tree and encoder-decoder approaches depends on the underlying morphological typology of the studied languages. Neural approaches seem to be stronger for languages with complex case systems and agglutinative morphology. In contrast, edit-tree approaches excel on more synthetic languages (e. g. Type 3) and languages with lower ambiguity (e. g. Type 2), where edit-tree models outperformed Sent-LM on ambiguous tokens. .

Figure 3.6 illustrates how as the number of edit-trees increases the encoder-decoder models start to excel. This is most likely due to the fact that, from an edit-tree approach perspective, a large number of trees creates a large number of classes, which leads to higher class imbalance and more skewness and sparsity. However, edit-tree based approaches do outperform representation learning methods for languages with lower number of trees, which leads to the intuition that the edit-tree formalism does provide a useful inductive bias to the task of lemmatization and it should not be discarded in future work. Our results, in fact, point to a future direction which applies the edit-tree formalism, but alleviates the edit-tree explosion by exploiting the relationships between the edit-tree classes potentially using representation learning methods.

3.6.4 Accuracy on Unknown Tokens

We observe that, while the addition of the joint LM loss to the encoder-decoder results in an overall stronger lemmatizer, it seems, however,

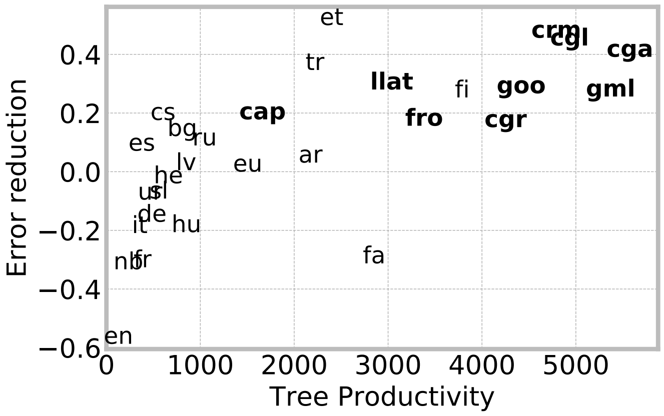


Figure 3.6: Error reduction obtained by the best encoder-decoder model with respect to the best tree-edit model over tree productivity computed as number of unique binary edit-trees in the first 50k tokens of the training corpora (Spearman's $R = 0.79$; $p < .001$). Labels of historical languages are shown in bold.

detrimental to the accuracy on unknown tokens. This discrepancy is probably due to two facts. First, unknown tokens are likely unambiguous and therefore less likely to profit from improved context representations. Second, our design choice used word-level predictions for the LM objective, and, thus, the model was forced to predict UNK for unknown words. As `Sent-LM` is the overall best model, in future work we shall explore a character-level LM objective in order to harness the full potential of the joint-training approach, even on unknown tokens.

3.6.5 Error Analysis

In order to illustrate the advantages and disadvantages of the proposed joint training procedure with respect to the simple encoder-decoder architectures and the binary edit-tree models, we have selected a number of examples where the `Sent-LM` model produces a different lemmatization than either of the baselines. Using these examples as a departure point, we can observe the specific behavior of the different models and highlight specific situations in which some models perform better than others.

In order to facilitate the interpretation of the examples, we have added the sentential context that was accessible to the lemmatizers, following the corpus segmentation. Moreover, we provide token-level and sentence-level translations, as well as morphological annotations of tokens, in cases where these can help to interpret the analyses. Cases where sentence segmentation has rendered a token difficult to interpret due to missing context, appear as "[...]" in the translation. On the left-hand side, we show the example with translation and analysis,

	Sent-LM	Edit-Tree
(1) *ubi* ***uocauolum*** *est* *Centu* where term is hundred *porche,* *et* *[...]* pig.PL.FEM, and "where the term is "hundred female pigs", and"	**vocabulum**	vocabolum
(2) *rignante* *domnu* *nostru* *uirum* ruling lord our man ***exselentissimum*** *Radchisi* *rige* eminent Radchisi king "under the rule of our lord king Radchisi, oh eminent man	**excello**	exselentus
(3) ***Escripsi*** *ego* *Appo* *rogatus* wrote.1P.SG.PAST I Appus asked *a* *Donatum* by Donatus.ABL "I, Appo, wrote it asked by Donatus"	**scribo**	escribo

Table 3.8: Examples from the LLCT (`llat`) corpus highlighting situations in which the string transduction approach has an advantage over a binary edit-tree lemmatizer due to robustness against orthographic variation.

highlighting the focus token in **bold**. On the right-hand side, we show the proposed lemmatizations, highlighting the correct one in **bold**.

The first set of examples, shown in Table 3.8, highlights the robustness of encoder-decoder architectures with respect to orthographic variation. In these cases, orthographic variation challenges the binary edit-tree approach in different ways. In Example (1), we can see that the induced edit-tree covers orthographic idiosyncrasies like the "u"-"v" alternation in the onset of "uocauolum" and even the "v"-spelling of "b". Moreover, the lemmatizer has correctly identified the matching rule. However, a further "u"-"o" alternation (i. e. "uocau**u**lum" → "uocau**o**lum")—reflecting an opening of the vowel in the scribe's dialect—is not captured by the rule, and as a result the lemmatizer produces an incorrect analysis.

Example (2) represents a more aggravating type of mistake. Here, the "-xs-" spelling of the consonant cluster represented by "-xc-" (i. e. "ex**c**elentissimum" → "ex**s**elentissimum") (mis-)leads the edit-tree lemmatizer to interpret the input token as a derivation of the non-existent adjective "exselentus" instead of the non-finite form of "excello" (en. to excell).

Finally, in Example (3) the predicted edit-tree rule identifies the prothetic "e-" as a prefix and it is, thus, wrongly kept in the predicted

		Sent-LM	Edit-Tree
(1)	*et quarto, mense die* ***calende*** and fourth-(year), month day calends *martia, indictione prima.* Mars, declaration first. "during the fourth (year), in the first day of the month of Mars, first declaration"	**kalendae**	calens
(2)	*Et accepit* ***ad*** *te pretium* and received.3P.SG.PAST from you money *pro suprascripta casa* for mentioned-before house "And he received from you the value for money for the house mentioned before"	**ab**	ad

Table 3.9: Examples from the LLCT (`llat`) corpus highlighting situations in which `Sent-LM` model has an advantage over a binary edit-tree lemmatizer due to relying on contextual information to solve ambiguity arising from orthographic variation.

lemma. In contrast, the string transduction approach can generalize from the total number cases of prothetic "e-"-s in the corpus and produce the correct analysis.

Table 3.9 shows two examples in which orthographic variation complicates the identification of the underlying lemma but the sentential context offers cues that add to the evidence of the correct lemma. In Example (1), the originally greek word "kalendae" (i. e. the first day of every month in the Roman calendar) is spelled with a Latin "c". This rare spelling leads the `Edit-Tree` lemmatizer to interpret the token as a non-existent adjective "calendus", despite abundant cues in the context linking to a time reference—"mense", "die", "martia".

In Example (2), spelling variation is conflating the forms of the prepositions "ad" (en. towards) and "ab" (en. from). This is due to the phonological process by which mono-syllabic prepositions suffer the lenition of the consonant in the coda—and eventually the distinction between the two prepositions disappears. At this stage, scribes hesitate about the correct form—readers, in turn, can rely on context to interpret the correct meaning. In this case, the intransitive verb "accipio" (en. receive) requires "ab", a fact that is exploited by the neural lemmatizer trained with the joint loss, but neither by the neural variant trained without joint loss nor the edit-tree approach.

The next set of examples, shown in Table 3.10, highlights two common artifacts produced by the string transduction approach. The first one relates to the appearance of unexpected characters. In Example (1) the capitalization of "Limite" (en. border, but, in this case likely used as a toponym) appears to derail the subsequent string transduction

		`Sent-LM`	`Edit-Tree`
(1)	*in loco qui uocitator* ***Limite,*** *ubi* in place that name "Limite", where *uocauolum est Centu* term is "Centu" ”in the place named "Limite", where the term is "Centu"”	limitte	**limes**
(2)	*ut dixi, nulla iuidem* as say.1P.SG.PAST, no there ***molestationem,*** *uel deuisionem* violence, or-even dispute *facientem.* doing ”as I said, without exerting there any violence or even provoking a dispute”	molestia	**molestatio**
(3)	*qui supra ad signa eorum* who before towards signs they *contrascripsi* oppose ”I, who opposed their signs”	constribo	**contrascribo**

Table 3.10: Examples from the LLCT (`llat`) corpus highlighting situations in which `Sent-LM` model has a disadvantage with respect to the binary edit-tree lemmatizer, due to the former being misled by letter case or the length of the input token.

causing the neural lemmatizer to output a morphologically invalid lemma: ”limitte”. Capitalized forms are rare but can offer information relevant to lemmatization when proper names should receive special treatment. In this case, the neural lemmatizer—unlike the edit-tree lemmatizer—has failed to recognize that the capitalized form behaves in a morphologically similar manner to the non-capitalized form.

Example (2) in Table 3.10 relates to the commonly known issue that, under auto-regressive models of language—like the neural string transduction lemmatizers discussed in this chapter—, the addition of further items to the sequence causes the overall probability assigned to the sequence to decrease. As a result, input tokens with particularly long lemmata are likely to be lemmatized incorrectly, since shorter and superficially similar lemma candidates are outputted with higher probability, even if these shorter lemmata have not appeared during training or no apparent relation to the token can be drawn. In particular, Example (2) illustrates the case where a shorter derivation (”molestia”) of the same stem (”molest-”) is outputted instead of the correct one (”molestatio”). In contrast, Example (3) shows a case in which a Latin-like form ”constribo” is ”hallucinated”.

		Sent-LM	Sent
(1)	*an ende ghinc ligghen bi* ***aren,*** *man ende* and went lay by her, husband and *nam al slapende* took while sleeping "[...] she lied down with her husband and took while sleeping"	**haar**	arend
(2)	*dat de heleghe gheest in v wont ende v al* that the holy spirit in you lives and you all *leert* teach "that the holy spirit lives inside you and teaches you everything"	uw	**gij**

Table 3.11: Examples from the cgl corpus highlighting situations in which Sent-LM model shows stronger disambiguating capabilities leading to correct and incorrect lemma guesses.

Table 3.11 shows two examples from the Middle Dutch corpus, in which, alle gedly, the stronger disambiguating capabilities of the Sent-LM lemmatizer result, respectively, in correct and incorrect predictions. In Example (1), the problem arises by a compounding of two factors: the inclusion of a "," between the tokens "aren" and "man" and the drop of initial "h" in the possessive pronoun "haar". First, the inclusion of a comma obscures the correct interpretation of "bi aren man" (en. next to her husband) as a prepositional phrase. Second, the drop of "h" likens the token to the form "arend" (en. eagle). However, the interpretation suggested by the lemmatization—i. e. "and she lied with eagle and man, took while sleeping"—is much less likely.

In Example (2), we encounter a similar situation in which alternative lemmatizations are backed by two different interpretations of the prepositional phrase "in v wont". Analyzing "v" as a possessive pronoun—the path taken by Sent-LM—results in an interpretation of Medieval Dutch "wont" as English "wound", while analyzing it as a personal pronoun—the path taken by Sent–results in an interpretation "wont" as the 3rd person singular of the verb "wonen" (en. to live). In this case, one can argue that Sent-LM is overfitting on a local syntactic construction—i. e. "in" followed by a noun phrase—which is arguably highly frequent in the corpus. Still, from the point of view of sentential semantics, this analysis implies a highly unplausible interpretation of the sentence—which translates as "that the holy spirit [is] in your wound and teaches you everything".

Finally, Table 3.12 displays two examples in which the strong disambiguating capabilities of the Sent-LM mislead it towards incorrect—yet plausible—lemmatizations. In Example (1), the Sent-LM lemmatizer

		Sent-LM	Sent
(1)	*ad te semper defensus et* ***protector*** to you always defended and protector *esse diueas ad prauis* be owe.2P.SG.PRES.SBJ from crooked *hominibus* man “and you ought to defend and protect [them] from crooked men”	protectus	**protector**
(2)	*land voerdise van* land transported-him-they from *caspia. dar* ***se*** *alexander* caspian-location. there them Alexander *sloech derna.* defeated afterwards. “[...] they transported him from the Caspian. There Alexander defeated them afterwards.”	zee	**zij**

Table 3.12: Examples from the LLCT1 (llat)—example (1)—and cgl—example (2)—corpora, highlighting situations in which Sent-LM is misled by context.

outputs “protectus” following the syntactic trend of copular constructions to pair nouns with similar morphological endings. In the Medieval Dutch Example (2), the problem arises from the morphologically ambiguous “se” form that can refer to both the personal pronoun “zij” (en. they) and the noun “zee” (en. sea). In this case, one can argue that semantic triggers in the clause—“land” (en. land), “voerd” (en. to transport, to transport by sea) and, even, “caspian” (which can refer to both a political region and a sea)—lead Sent-LM to prefer the geographical interpretation, which, with a slightly contrived syntax, can be rendered by “[...] they transported him from the Caspian, in that sea Alexander defeated them afterwards”.

3.7 CONCLUSION

We have presented a method to improve lemmatization with encoder-decoder models by enriching the information modeled by contextual word representations through the incorporation of a joint bi-directional LM loss. Our method sets a new state of the art for lemmatization of historical languages across a varied set of benchmark corpora, and is still competitive on standard languages.

A set of representation probing experiments indicate that the representations learned with the proposed joint LM loss manage to capture more morphologically relevant information than representations learned with the lemmatization objective, and, thus, present stronger disambiguation capabilities.

In view of a typologically informed comparison of approaches based on the encoder-decoder and the binary edit-tree paradigms, we have shown that the latter can be very effective for highly synthetic languages and provide an inductive bias with respect to which current end-to-end neural models are at a disadvantage. This situation may have been overlooked in previous studies due to considering only a reduced number of languages (Chakrabarty, Pandit, and Garain, 2017) or because of undifferentiated pooling of results across typologically different languages (Bergmanis and Goldwater, 2018). With respect to languages with higher ambiguity and token-lemma ratio, the encoder-decoder approach appears as the preferable option and the joint LM loss generally provides a substantial improvement.

Finally, while other models use morphological information to improve the representation of context (e. g. edit-tree approaches), our joint LM loss does not rely on any additional annotation, an aspect that can be crucial in low resource and non-standard situations where annotation is costly and often not trivial.

4 ALLUSIVE TEXT REUSE DETECTION

ABSTRACT The detection of allusive text reuse is a particularly challenging task due to the sparse evidence on which allusive references rely—commonly based on none or very few shared words. Arguably, lexical semantics can be resorted to since uncovering semantic relations between words has the potential to increase the support underlying the allusion and alleviate the lexical sparsity. A further obstacle is the lack of evaluation benchmark corpora, largely due to the highly interpretative character of the annotation process. In the present chapter, we aim to elucidate the feasibility of automated allusion detection. We approach the matter from an Information Retrieval (IR) perspective in which referencing texts act as queries and referenced texts as relevant documents to be retrieved, and estimate the difficulty of benchmark corpus compilation by a novel inter-annotator agreement study on query segmentation. Furthermore, we investigate to what extent the integration of lexical semantic information derived from distributional models and ontologies can aid retrieving cases of allusive reuse. The results show that despite low agreement scores, using manual queries considerably improves retrieval performance with respect to a windowing approach, and that retrieval performance can be moderately boosted with distributional semantics.

This chapter is based on Enrique Manjavacas Arévalo, Brian Long, and Mike Kestemont (June 2019). “On the Feasibility of Automated Detection of Allusive Text Reuse.” In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 104–114. DOI: `10.18653/v1/W19-2514`. URL: `https://www.aclweb.org/anthology/W19-2514`

4.1 INTRODUCTION

In the 20th century, intertextuality emerged as an influential concept in literary criticism. Originally developed by French deconstructionist theorists, such as J. Kristeva and R. Barthes, the term broadly refers to the phenomenon where texts integrate (fragments of) other texts or allude to them (Orr, 2003). In the minds of both authors and readers, intertexts can establish meaningful connections between works, evoking particular stylistic effects and interpretations of a text. Existing categorizations emphasize the broad spectrum of intertexts, which can range from direct quotations, over paraphrased passages to highly subtle allusions (Bamman and Crane, 2008; Büchler, 2013; Hohl Trillini and Quassdorf, 2010; Mellerin, 2014).

With the emergence of computational methods in literary studies over the past decades, intertextuality has often been presented as a promising application, helping scholars identifying potential intertextual links that had previously gone unnoticed. Much progress has been made in this area and a number of highly useful tools are now available—pieces of representative software are TRACER (Büchler et al., 2014b) or Tesserae (Coffee et al., 2012a), to name a few. This chapter, however, aims to contribute to a number of open issues that still present significant challenges to the further development of the field.

> **Source** *Ephesians 3:19* "scire etiam supereminentem scientiae caritatem Christi ut impleamini in omnem plenitudinem Dei"
>
> "and to know the love (caritas) of Christ that is beyond knowledge, such that you'd be filled with all fullness of God"
>
> **Target** *S. 8, 7* "Osculum plane dilectionis et pacis, *sed dilectio illa supereminet omni* ***scientiae***, et pax illa omnem sensum exsuperat"
>
> "It is a kiss of love and peace, but of that kind of love (dilectio) that is beyond any knowledge, and of that kind of peace that surpasses all senses."

Quote 1

Most scholarship continues to focus on the detection of relatively literal instances of so-called "text reuse", as intertextuality is commonly, and somewhat restrictively, referred to in the field. Such instances are relatively unambiguous and unproblematic to detect using n-gram matching, fingerprinting and string alignment algorithms. Much less research has been devoted to the detection of fuzzier instances of text reuse holding between passages that lack a significant lexical correspondence. This situation is aggravated by the severe lack of openly available benchmark datasets. An additional hindrance is that the establishment of intertextual links is to a high degree subjective—both regarding the existence of particular intertextual links and the exact scope of the correspondence in both fragments. Studies of inter-annotator agreement are surprisingly rare in the field, which might be partially due to to the fact that existing agreement metrics are hard to port to this problem.

CONTRIBUTIONS In this chapter, we report on an empirical feasibility study, focusing on the annotation and automated detection of allusive text reuse. We focus on biblical intertext in the works of Bernard of Clairvaux (1090–1153), an author known for his pervasive references to the Bible. The chapter has two main parts. In the first part, we formulate an adaptation of Fleiss's κ that allows us to quantitatively estimate and discuss the level of inter-annotator agreement concerning the span of the intertexts. While annotators show considerably low levels of agreement, we show that manual segmentation has nevertheless a big impact on the automatic retrieval of allusive reuse. In the second part, we offer an evaluation of common IR techniques for allusive text reuse detection. We confirm that semantic retrieval models based on word and sentence embeddings do not present advantages over hand-crafted scoring functions from previous studies, and that both are outperformed by conventional retrieval models based on `TfIdf`. Finally, we show how the soft cosine similarity allows us to combine lexical and semantic information to obtain significant improvements over any other considered model.

OUTLINE The remainder of this chapter is structured as follows. First, in Section 4.1.1 we review previous related work on allusive text reuse detection. In Section 4.2, we discuss our experiments on the annotation of allusions, introducing the dataset, as well as the formulation of the inter-annotator agreement index for span annotations. Next, in Section 4.3, we discuss the experiments on the retrieval of allusive text reuse, introducing the different models involved and analyzing the results. Finally, Section 4.4 offers concluding remarks and pointers for future work.

4.1.1 Related Work

Previous research on text reuse detection in literary texts has extensively explored methods such as n-gram matching (Büchler et al., 2014b) and sequence alignment algorithms (Lee, 2007; Smith et al., 2014). In such approaches, fuzzier forms of intertextual links are accounted for through the use of edit distance comparisons or the inclusion of abstract linguistic information such as word lemmata or part-of-speech tags, and lexical semantic relationships extracted from

WordNet (Fellbaum, 2012). More recently, researchers have started to explore techniques from the field of distributional semantics in order to capture allusive text reuse. Scheirer, Forstall, and Coffee (2016), for instance, have applied LSI to find semantic connections and evaluated such method on a set of 35 allusive references to Vergil's *Aeneis* in the first book of Lucan's *Civil War*.

Previous research in the field of text reuse has also focused on the more specific problem of finding allusive references. One of the first studies (Bamman and Crane, 2008) looked at allusion detection in literary text and exploited features at a variety of linguistic levels (including morphology and syntax) but collected only qualitative evidence on the efficiency of such approach. More ambitiously, Bamman and Crane (2009) approached the task of finding allusive references across texts in different languages using string alignment algorithms from machine translation. Besides the afore-mentioned work by Scheirer, Forstall, and Coffee (2016), the work by Moritz et al. (2016) is highly related to the present study, since the authors also worked on allusive reuse from the Bible in the works of Bernard. In their work, the authors focused on modeling text reuse patterns based on a set of transformation rules defined over string case, lemmata, part-of-speech tags and synset relationships such as synonymy, hyponymy or co-hyponymy. More recently, Moritz, Hellrich, and Büchel (2018) conducted a quantitative comparison of such transformation rules with paraphrase detection methods on the task of predicting a paraphrase relation between text pairs.

4.2 ANNOTATION EXPERIMENT

The basis for the present study stems from the BiblIndex project (Mellerin, 2014), which aims to index biblical references found in Christian literature.[1] More specifically, we use the dataset with manually identified biblical references from Bernard of Clairvaux to Jerome's Vulgate, which was introduced in Chapter 2. Importantly, BiblIndex distinguishes three types of references: "quotation", "mention" and "allusion". While the links in the first two types are in their vast majority exact or near-exact lexical matches, the latter type comprises mostly references that fall into what is commonly known as allusive text reuse. Although our focus lies on the allusive category, Table 4.1 displays statistics about all these types in order to appreciate the characteristics of the task. As shown in the last column of Table 4.1, allusions are characterized by low Jaccard coefficients—i. e. , in set-theoretical terms, the ratio of the intersection over the union of the sets of words of both passages. On average, annotated allusions share 6% of the word forms with their source passages and 13% of the lemmata. In comparison,

[1] http://www.biblindex.mom.fr/

	Quotation	Mention	Allusion	Allusion (Post)
Token	0.37 (± 0.23)	0.26 (± 0.18)	0.02 (± 0.04)	0.06 (± 0.07)
Lemma	0.37 (± 0.22)	0.31 (± 0.18)	0.04 (± 0.05)	0.13 (± 0.1)
Source	15.12 (± 5.99)	16.24 (± 6.20)	17.22 (± 6.58)	
Target	6.69 (± 4.55)	7.47 (± 5.52)	1.10 (± 0.85)	6.86 (± 4.83)
Count	1768	3150	876	729

Table 4.1: Full dataset statistics for link types originally provided by the editors. Last column shows statistics for allusive references in Bernard after annotation. Jaccard coefficients are shown for both tokenized (first row) and lemmatized (second row) sentences, together with text lengths of (source and target) sentences and total instance counts.

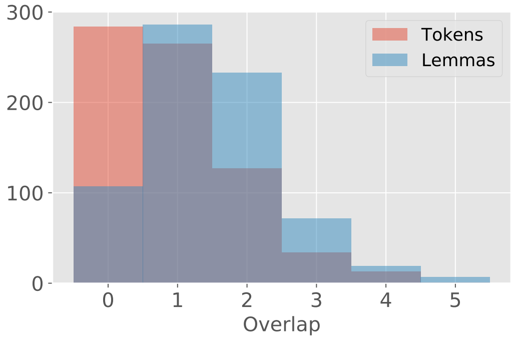


Figure 4.1: Histogram of world overlap between the annotated queries and their corresponding biblical references, distinguishing the overlap obtained by raw tokens from the overlap obtained by lemmata.

mentions and quotations have 25% or more tokens and 30% or more lemmata in common. The full distribution of token and lemma overlap for allusions shown in Figure 4.1 indicates that more than 500 (65%) instances have at most 1 token in common; about more than 400 (50%) share at most 1 lemma.

As explained in Section 2.5.1, retrieval systems in text reuse detection rely on pre-existing segmentation, or apply automatic segmentation of the original works into consecutive, equal-length chunks of texts, which are then used as queries to find cross-document matches. For semi-literal cases of reuse, this matching procedure yields good results and overlapping or adjacent matches can be easily merged into longer units of reuse. For allusive text reuse, however, such an approach seems unfeasible at the current stage, partially because the definition of the relevant query units is much harder to establish.

The annotated allusive references are mere "anchors", consisting of single words or single multi-word expressions that cannot be easily used as queries. For illustration purposes, Listing 4.1 shows an excerpt from the original dataset, in which the provided anchor is

```
     <w xml:id="lat.w.Sermo58.8.1698">Flores</w>
     <w xml:id="lat.w.Sermo58.8.1699">etiam</w>
     <w xml:id="lat.w.Sermo58.8.1700">fuerunt</w>
     <w xml:id="lat.w.Sermo58.8.1701">qui</w>
5    <w xml:id="lat.w.Sermo58.8.1702">primi</w>
     <w xml:id="lat.w.Sermo58.8.1703">crediderunt</w>
     <w xml:id="lat.w.Sermo58.8.1704">de</w>
     <w xml:id="lat.w.Sermo58.8.1705">populo</w>
     <pc xml:id="lat.pc.Sermo58.8.392">,</pc>
10   <w xml:id="lat.w.Sermo58.8.1706">primitiae</w>
     <seg xml:id="lat.sQ.Sermo58.8.j">
         <w xml:id="lat.w.Sermo58.8.1707">sanctorum</w>
         <span from="#lat.w.Sermo58.8.1707"
             to="#lat.w.Sermo58.8.1707"/>
15   </seg>
     <note xml:id="lat.sNote.Sermo58.8.j"
     n="j" place="foot" type="scripturalNote">
         <seg xml:id="lat.b.Sermo58.8.12" type="bRef">
             <bibl type="biblical">
20               <ref cRef="Vg:1_Co:15:20">1 Co 15, 20</ref>
             </bibl>
         </seg>
         <link type="allusion"/>
     </note>
25   <pc xml:id="lat.pc.Sermo58.8.393">.</pc>
     <w xml:id="lat.w.Sermo58.8.1708">Flores</w>
     <w xml:id="lat.w.Sermo58.8.1709">eorum</w>
     <w xml:id="lat.w.Sermo58.8.1710">miracula</w>
     <pc xml:id="lat.pc.Sermo58.8.394">,</pc>
30   <w xml:id="lat.w.Sermo58.8.1711">instar</w>
     <w xml:id="lat.w.Sermo58.8.1712">florum</w>
```

Listing 4.1: Excerpt from a modern digital edition of Bernard of Clairvaux's 10th sermon from the book *Sermons on the Song of Songs*, highlighting the placement of an allusive reference—lines no. 11 to no. 24—to *1 Corinthians 15:20*: "nunc autem Christus resurrexit a mortuis primitiae dormientium"

"sanctorum"—even though other words in the context, e. g. "primitiae" or "flores" also support the allusion. The usage of anchors in the original annotation is, further, reflected in the third column of Table 4.1, showing that the average number of tokens in the provided annotations is slightly over one. This is in agreement with pragmatic editorial conventions, which favor uncompromising signposting of references at anchor words over establishing particular decisions on the scope of the reference. However, from the point of view of the evaluation of IR systems, the provided editorial anchors must be turned into fully-fleshed queries. In order to accomplish this, we have conducted an annotation experiment which we shall describe next.

4.2.1 Experimental Design

The aim of the annotation was to determine the scope of a biblical reference identified by the editors in text by Bernard. From an IR perspective, the annotation task consists of delineating the appropriate input query, given the anchor word in the target text and the corresponding biblical verse. The example shown in Quote 1 illustrates the annotation process, where the anchor word provided by the editors is "scientiae" and the corresponding query annotated by one of the experts spans the sub-clause "sed dilection illa supereminet omni scientiae". Naturally, such references not always correspond to full sentences and often go over sentence boundaries.

The dataset was distributed evenly across four annotators,[2] who worked independently through a custom-built interface. All annotators were proficient readers of Medieval Latin with expertise ranging from graduate student to University professor. The annotators were familiar with the text reuse detection task and were given explicit instructions that can be summarized as follows: given a previously identified allusion between the Bernardine passage surrounding an anchor word, on the one hand, and a specific biblical verse on the other hand, annotate the *minimal textual span* in the Bernardine passage that is *maximally allusive* to the biblical verse. For the sake of simplicity, the interface only allowed continuous annotation spans and the annotated span had to include the pre-identified anchor token. Of a total of 876 initial instances, we discarded 147 cases in which annotators expressed doubts on the existence of the alleged reference or could not precisely decide the span. This decision was taken in order to ensure a high quality in the resulting benchmark data.

Determining the scope of an allusive reference is a relevant task for two reasons. Firstly, we expect this task to be reader-dependent, and thus highly subjective, given the minimal lexical overlap between the source and target passage. Measuring the agreement between annota-

[2] The annotators were, in alphabetical order, Jeroen Deploige, Jeroen De Gussem, Wim Verbaal and Dina Wouters.

tors sheds new light on the overall feasibility of the task. Secondly, the resulting annotations allow us to critically evaluate the performance of existing retrieval methods under near-perfect segmentation conditions: if the correct query is given, what is the performance of existing methods when attempting to retrieve the correct biblical verse in the source data?

4.2.2 Measuring Inter-annotator Agreement

Inter-annotator agreement coefficients such as Fleiss's κ and Krippendorff's α are typically defined in terms of labels assigned to items in a multi-class classification setup (Artstein and Poesio, 2008). In the present case, however, the annotation involves making a decision on the span of words surrounding an anchor word that better captures the allusion and it is unclear how to quantify the variation in annotation between annotators. A naïve approach defined in terms of number of overlapping words has a number of undesirable issues. For example, since the annotations are centered around the anchor word, a relatively high amount of overlap is to be expected for short annotations. Moreover, disagreements over otherwise largely agreeing long spans should weigh in less than disagreements over otherwise largely agreeing small spans. Additionally, it is unclear how to quantify the rate of agreement expected under chance-level annotation, a quantity that needs to be corrected for in order to obtain reliable and non-inflated inter-annotator agreement coefficients (Artstein, 2017). We have found that an extension of the Jaccard coefficient defined over sequences can help adapt Fleiss's κ to our case and tackle such issues.

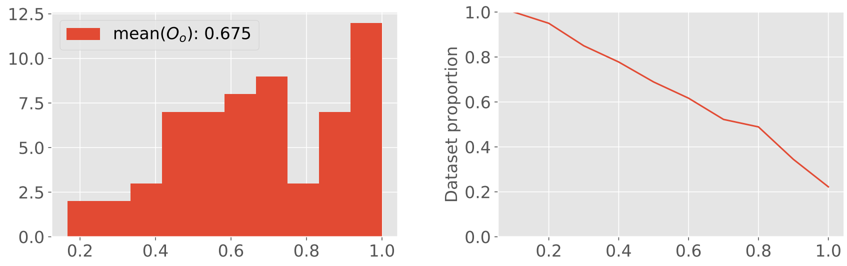
Given any pair of span annotations, s and t, we can define overlap in a similar way to the Jaccard index, as the intersection (i. e. the Longest Common Substring) over the union (i. e. the total number of selected tokens by both annotators):

$$O = \frac{LCS(s,t)}{|s| + |t| - LCS(s,t)} \tag{4.1}$$

Interestingly, this quantity can be decomposed into an agreement $A(s,t) = LCS(s,t)$ (number of tokens in common) and a disagreement score $D(s,t) = |s| + |t| - 2 \cdot LCS(s,t)$ (number of tokens not shared with the other annotator):

$$O = \frac{A}{A + D} \tag{4.2}$$

The advantage of this reformulation is that, first, it lets us see more easily how O is bounded between 0 and 1, and that, second, it gives us a way of computing the expected overlap score O_e by aggregating dataset-level A and D scores, as shown in Equation 4.3.



(a) Observed Overlap

(b) Cumulative Overlap

Figure 4.2: Observed overlap in the inter-annotator agreement experiments. On the left (a), we see the full histogram of O_o in the dataset ($N = 60$). On the right (b), we see the cumulative plot. We observe two modes in the histogram, perhaps indicating a qualitative difference in the dataset. One with high overlap scores close to 1.0 and another one at around 0.6—which is close to the overall average overlap.

$$\begin{aligned} O_e &= \frac{A_e}{(A_e + D_e)} \\ A_e &= \frac{\sum_{s,t} A(s,t)}{|s,t|} \\ D_e &= \frac{\sum_{s,t} D(s,t)}{|s,t|} \end{aligned} \tag{4.3}$$

where $|s,t|$ refers to the number of unordered annotation pairs in the dataset.[3]

O_e can be thus interpreted as the expected overlap between two arbitrary annotators. The final inter-annotator agreement score is defined following Fleiss's κ:

$$\kappa = \frac{O_o - O_e}{1 - O_e} \tag{4.4}$$

where O_o refers to the dataset average of Equation 4.2.

4.2.3 Inter-annotator Agreement Results

In order to estimate κ for our dataset, we extracted a random sample of 60 instances which were thoroughly annotated by 3 of the annotators. We obtain a $\kappa = 0.22$, which compares unfavorably with respect to commonly assumed reliability ranges. For example, values in the range $\kappa \in (0.67, 0.8)$ are considered fair agreement (Schütze, Manning, and Raghavan, 2008). While our result remains hard to assess in the absence

[3] This quantity is computed by $N \cdot k \cdot (k-1)/2$, where N is the number of annotations and k the number of annotators.

of comparable work, it is low enough to cast doubts over the feasibility of the task, which is in fact rarely explicitly questioned. The annotators informally reported that, against their expectations, the task was not straight-forward and required a considerable level of concentration and interpretation. This situation may be due to particularities of Bernard's usage of biblical language. Besides conventional, direct allusions, Bernard is also known for pointed use of single, significant allusive words, which are hard to isolate. Still it should be noted that in some instances inter-annotator agreement was high, and, as Figure 4.2 (b) shows, in 22% of all pairwise comparisons even perfect. This suggests that there exist clear differences at the level of individual allusions. We now turn to the question how well current retrieval approaches perform, given manually segmented queries.

4.3 RETRIEVAL EXPERIMENT

Given the small amounts of lexical overlap in the allusive text reuse datasets (see Table 4.1), we aim to investigate and quantify to which extent semantic information can help improving retrieval of allusive references. For this reason, we look into three types of models. We first look at purely lexical-based approaches and, second, at approaches based on distributional semantics, focusing on retrieval approaches that utilize word embeddings. Finally, we look at hybrid approaches that can accommodate relative amounts of semantic information into what is otherwise a purely lexical model. From the retrieval point of view, all approaches fall into one of two categories: retrieval methods based on similarity in vector space and retrieval methods using domain-specific similarity scoring functions. Methods based on the text alignment paradigm were left out of consideration, since by definition they rely on a high degree of lexical overlap.

4.3.1 Lexical Models

Lexical approaches only take into account the identities of the tokens or lemmata in the input documents.

4.3.1.1 Hand-crafted Scoring Function

Previous work has devised hand-crafted scoring functions that aim at retrieving intertextual relationships comparable to those found in Bernard. In particular, Forstall et al. (2015) defined a scoring function in order to retrieve allusive references and deployed it in the Tesserae online retrieval system.[4] The function is shown in Equation 4.5:

$$\mathrm{Tesserae}(s,t) = \ln \left(\frac{\sum_{w \in (S \cap T)} \frac{1}{f_{(w,s)}} + \frac{1}{f_{(w,t)}}}{d_s + d_t} \right) \tag{4.5}$$

where $f_{(w,d)}$ refers to the frequency of word w in document d—which has the goal of capturing the "relative rarity of the words in the phrases shared by the two texts"—and d_d refers to the distance in tokens between the two most infrequent words—from the set of overlapping words—in document d. The latter is aimed at capturing "phrase density" and models the fact that intertexts are generally found to "consist of compact rather than diffuse collocations".

Note that $\mathrm{Tesserae}(s,t)$ is only defined for cases in which documents share at least two words, since otherwise the denominator cannot be computed. While this presents a clear disadvantage—considering that a large number of allusions are based on less than two overlapping words—, it also lends itself to evaluation in a hybrid fashion with a complementary back-off model operating on passages with lower overlap. While originally $f_{(w,s)}$ is defined with respect to the query (or target) document, we observed such choice yielded poor performance—likely due to the small size of the documents—, and, therefore, we used frequency estimates extracted from the respective document collections instead. We refer to this model as `Tesserae`.

4.3.1.2 *Bag-of-words & Tf-Idf*

We include retrieval models based on a Bag Of Words (BOW) representation and cosine similarity for ranking. As explained in Section 2.5.2.2, a BOW space model represents a document d by a vector where the i_{th} entry represents the frequency of the i_{th} word in d. Beyond word counts, it is customary to apply the Tf-Idf transformation, which targets the fact that the importance of a word for a document is also dependent on how specific the word is for that document. Tf-Idf for the i_{th} word is computed as the product of its frequency in d, denoted $\mathrm{Tf}(w,d)$, and its inverse document frequency, $\mathrm{Idf}(w,d)$, defined by Equation 4.6:

$$\mathrm{Idf}(w,d) = \log \left(\frac{|D|}{1 + |\{d \in D : w \in d\}|} \right) \tag{4.6}$$

We refer to these retrieval models as `BOW` and `Tf-Idf`, respectively. Given document vector representations in some common space, we

[4] The retrieval system can be found available online through the following URL: {http://tesserae.caset.buffalo.edu/

can compute their similarity score based on the cosine similarity between such vectors:

$$\cos(\vec{s}, \vec{t}) = \frac{\sum_i s_i t_i}{\sqrt{\sum_i s_i^2}\sqrt{\sum_i t_i^2}} \tag{4.7}$$

4.3.2 Semantic Models

We define a number of semantic models based on distributional semantics and, in particular, word embeddings.[5] We use `FastText` word embeddings (Bojanowski et al., 2017) trained with default parameters on a large collection of Latin literature provided by Bamman and Crane (2011a), which includes 8.5GB of text of varying quality.[6]

4.3.2.1 Sentence Embeddings

We use distributional semantic models based on the idea of computing a sentence embedding through a composition function operating over the individual embeddings of words in the sentence. The most basic composition function is averaging over the single word embeddings in the sentence (Wieting et al., 2015). We can take into account the relative importance of words to a given sentence using the Tf-Idf transformation defined in Section 4.3.1 and compute a Tf-Idf weighted average word embedding. We refer to these models as $\texttt{BOW}_{\text{emb}}$ and $\texttt{Tf-Idf}_{\text{emb}}$, respectively.

4.3.2.2 Word Mover's Distance

`WMD` is a metric based on the transportation problem known as Earth Mover's Distance but defined for documents over word embeddings. `WMD` has shown excellent performance in document retrieval tasks where semantics play an important role (Kusner et al., 2015). Intuitively, `WMD` is grounded on the idea of minimizing the amount of "travel cost" incurred in moving the word histogram of a document s into the word histogram of t, where the "travel distance" between words w_i and w_j is given by their respective distance in the embedding space—we use the complement of the cosine similarity for this purpose. Formally, `WMD` is computed by finding a so-called flow matrix $T \in \mathbb{R}^{V \times V}$—where T_{ij} denotes how much of word w_i in s travels to word w_j in t—such

[5] We also experimented with an LSI retrieval model (Deerwester et al., 1990), similar to the one used by Scheirer, Forstall, and Coffee (2016), but found that it performed poorly on this dataset due to the small size of the documents, and, thus, left it out of the experiments.

[6] All the relevant materials are available online and can be downloaded through the following URL: `http://www.cs.cmu.edu/~dbamman/latin.html`.

that $\sum_{i,j} T_{i,j} c(w_i, w_j)$ is minimized. Computing WMD involves solving a linear programming problem for which specialized solvers exist.[7]

4.3.3 Hybrid Models

We look into methods that are able to encompass both lexical and semantic information.

4.3.3.1 Tesserae with a Back-off Model

Since the Tesserae scoring function is only defined for document pairs with at least two words in common, it can be deployed in combination with purely semantic models for cases where the lexical overlap is below that requirement. In that hybrid deployment the purely semantic model serves as a back-off model. In particular, we evaluate this setup using WMD as the back-off model since it proved to be the most efficient purely semantic model. We note that for this retrieval setup to be used in practice, WMD and Tesserae similarity scores must be transformed into a common scale, such, so that scores do not outweigh each other.

4.3.3.2 Soft Cosine

A more principled approach to combining lexical and semantic information is based on the soft cosine similarity function, which was discussed in Section 2.5.2.2 and we re-introduce here for clarity purposes. As already mentioned, soft cosine generalizes cosine similarity by considering not only how similar vectors s and t across feature i but more generally across any given pair of features i, j. Equation 4.8 formalizes the intuition in the notation of the present chapter:

$$\text{SoftCosine}(\vec{s}, \vec{t}) = \frac{\sum_{i,j} S_{i,j} s_i t_j}{\sqrt{\sum_{i,j} S_{i,j} s_i s_j} \sqrt{\sum_{i,j} S_{i,j} t_i t_j}} \tag{4.8}$$

In Equation 4.8, $S \in \mathbb{R}^{V \times V}$ represents a matrix where $S_{i,j}$ expresses the similarity between the i_{th} and the j_{th} word in the vocabulary. It can be seen that soft cosine reduces to cosine when S is taken to be the identity matrix.

In order to estimate the similarity between words capture by matrix S, we utilize the following two models. First, Soft-Cosine$_{\text{wn}}$ uses a similarity function based on the size of the group of synonyms extracted from the Latin WordNet (Minozzi, 2010). Specifically, the WordNet-based similarity function is defined as follows:

$$S^{WN}_{i,j} = \frac{1}{|T_i \cap T_j|} \tag{4.9}$$

[7] We use the implementation provided by the pyemd package (Laszuk, 2017)

where T_i refers to the set of synonyms of the i_{th} word.

Second, `Soft-Cosine`$_{\text{emb}}$ exploits similarities estimated on the basis of word embeddings using the following function:

$$S^{SC}_{i,j} = max(0, cos(\overrightarrow{w_i}, \overrightarrow{w_j})) \tag{4.10}$$

over embeddings $\overrightarrow{w_i}, \overrightarrow{w_j}$.

All retrieval models based on soft cosine are applied on `Tf-Idf` document representations. In agreement with previous research (Charlet and Damnati, 2017), we boost the relative difference in similarity between the upper and lower quantiles of the similarity distribution by raising S^{SC} to the n^{th} power.[8]

4.3.4 Evaluation

Given a Bernardine reference as a query formulated by the annotators and the collection of biblical candidate documents, all evaluated models produce a ranking. Using such a ranking, we evaluate retrieval performance over a set of queries using MRR[9] defined in Equation 2.6.

Additionally, we also report Precision@K (P@K)—based on how often the system is expected to retrieve the relevant document within the first k results—since it is a more interpretable measure from the point of view of the retrieval system user.

As noted in Chapter 2, ranking-based evaluation metrics like P@K and MRR are not suitable to evaluate text reuse detection systems on unrestricted data, since, in fact, most naturally occurring text is not allusive. However, the focus of the present study lies on the feasibility of allusive text detection, which we aim to elucidate on the basis of a pre-annotated dataset in which each query is guaranteed to match to a relevant document in the source collection. The results must therefore be interpreted taking into account the artificial situation, where the selected queries are already known to contain allusions and the question is how well different systems recognize the alluded verse.

4.3.5 Results

As shown in Table 4.2, the best model overall is `Soft-Cosine`$_{\text{emb}}$, achieving 21.95 MRR and 47.60 P@20, closely followed by another soft cosine-based hybrid approach: `Soft-Cosine`$_{\text{wn}}$. Interestingly, a simple `Tf-Ifd` over lemmatized input used as baseline results in strong ranking performance, surpassing all other purely lexical (including the hand-crafted function `Tesserae`) and all purely semantic models. In

[8] During development we found that raising S^{SC} to the 5th power yielded the best results across similarity functions.

[9] For clarity, we transform MRR from the original $[0 - 1]$ range into the $[0 - 100]$ range.

			Lexical	
Metric	Lemma	BOW	Tf-Idf	Tesserae
MRR		11.85	16.42	12.39
	✓	15.07	19.51	13.36
P@10		20.16	30.59	19.20
	✓	27.30	34.43	25.79
P@20		25.38	35.94	22.22
	✓	34.16	43.35	30.86
			Semantic	
		BOW$_{emb}$	Tf-Idf$_{emb}$	WMD
MRR		8.54	9.59	13.68
	✓	9.82	11.13	14.07
P@10		15.50	18.11	24.14
	✓	16.87	20.99	25.38
P@20		20.44	24.14	27.85
	✓	22.63	26.20	31.28
			Hybrid	
		SC$_{wn}$	SC$_{emb}$	T+WMD
MRR			21.41	17.01
	✓	19.75	**21.95**	16.18
P@10			37.31	29.22
	✓	35.25	**39.64**	31.14
P@20			44.31	33.61
	✓	44.44	**47.60**	38.27

Table 4.2: Retrieval results grouped by approach. All models are evaluated with tokens and lemmata as input except for SC$_{wn}$, which requires lemmatized input. Overall best numbers per metric are shown in bold letters.

agreement with general expectations, all models benefit from lemmatized input and Tf-Idf transformation (both as input representation in purely lexical models and as a weighting scheme for the sentence embeddings in purely semantic approaches). WMD outperforms any other purely semantic model, but as already pointed out, it compares negatively to the purely lexical Tf-Idf baseline. The combination of Tesserae with WMD as back-off proves useful and outperforms both approaches in isolation, highlighting that they model complementary aspects of text reuse.

		Model		
Metric	Lemma	SC_{emb}	SC_{w2v}	SC_{rnd}
MRR		21.41	19.26	18.56
	✓	21.95	20.18	20.22
P@10		37.31	33.33	31.28
	✓	39.64	36.35	35.67
P@20		44.31	39.09	36.76
	✓	47.60	43.90	43.48

Table 4.3: Comparison of soft cosine using FastText embeddings (SC_{emb}), word2vec embeddings (SC_{w2v}) and a random similarity baseline (SC_{rnd}).

In order to test the specific contribution of the similarity function used to estimate S, we compare results with soft cosine using a random similarity matrix (S_{rnd}) defined by Equation 4.11:

$$S_{i,j} = \begin{cases} 1 & i = j \\ \sim \mathcal{N}(0.5, 0.05) & otherwise \end{cases} \tag{4.11}$$

We also investigate the effect of the algorithm used to compute the word embedding matrices by comparing $\texttt{Soft-Cosine}_{emb}$ using FastText embeddings to $\texttt{Soft-Cosine}_{emb}$ using word2vec embeddings (Mikolov et al., 2013). As Table 4.3 shows, FastText embeddings, an algorithm known to capture not just semantic but also morphological relations, yields strong improvements over word2vec. Moreover, a random approach produces strong results, only under-performing the word2vec model by a small margin, which questions the usefulness of the semantic relationships induced by word2vec for the present task.

Finally, we test the relative importance of the query segmentation to the retrieval of allusive text reuse. For this purpose, we evaluate our best model ($\texttt{Soft-Cosine}_{emb}$) on a version of the dataset in which the referencing text is segmented according to a window approach, selecting n words around the anchor expression.

As Table 4.4 shows, results on manually segmented text are always significantly better than on automated segmentation. A window of 10 words around the anchor produces slightly better results than a window of 3 words—more closely matching the overall mean length of manually annotated queries. This indicates the importance of localizing the appropriate set of referential words in context, while avoiding the inclusion of confounding terms. In other words, both precision and recall matter to segmentation, an issue that has been observed previously (Bamman and Crane, 2009).

		Segmentation		
Metric	Lemma	Manual	Win-3	Win-10
MRR		21.41	13.41	13.98
	✓	21.95	14.67	14.69
P@10		37.31	25.79	25.10
	✓	39.64	25.93	26.47
P@20		44.31	31.41	31.41
	✓	47.60	32.78	34.57

Table 4.4: Effects of the segmentation approach—manual segmentation and automatic segmentation using a sliding window of 3 (Win-3) and 10 (Win-10) tokens to each side of the anchor word—on the best performing model (Soft-Cosine$_{\text{emb}}$).

‡ visibilis quaedam imago et species decoris eius
† qui est imago dei invisibilis primogenitus omnis creaturae

Figure 4.3: Example of a correctly retrieved case of allusive text reuse, where a Bernardine passage (*S. 27, 7*, text above) is matched to biblical verse *Colossians, 1:15* (text below).

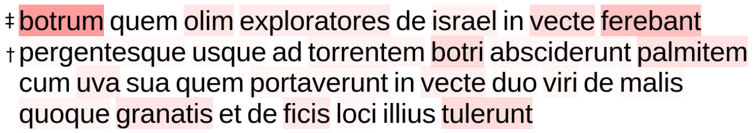
4.3.5.1 *Qualitative Inspection*

To appreciate the effect of the soft cosine similarity function using a semantic similarity matrix, it is worthwhile to inspect a hand-picked selection of items which were correctly retrieved by Soft-Cosine$_{\text{emb}}$ but not by Tf-Idf.[10]

In the first example, shown in Figure 4.3, the distributional approach adequately captures the antonymic relation between "visibilis" (‡, en. visible) and "invisibilis" (†, en. invisible), which is reinforced by the synonymy between "species" (‡, en. aspect, look) and "imago" (†) (en. image, copy). Similar mechanisms seem to be at work in Figure 4.4, where the semantic similarity between vinery-related words increases the overall similarity score ("botrus" en. grape, "palmes" en. vine-sprout, "uva" en. grape, "granatus" en. having many seeds).

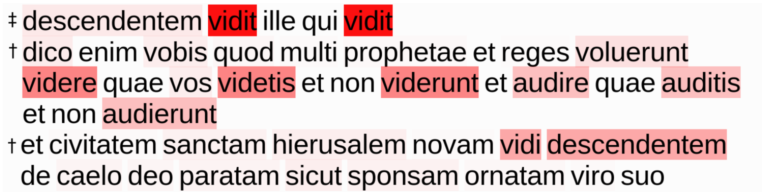
Although Soft-Cosine offers a welcomed boost in retrieval performance, many errors remain. A first and frequent category are allusions that are simply hard to detect, even for human readers, often because they are very short or cryptic such as Figure 4.6, where despite increased semantic support—"cognovissent" (en. they knew) being synonymous with "intellexerint" (en. they understood)—the match is missed.

[10] In the examples, we display the relative contribution made by each term in a sentence to the total similarity score (darker red implies higher contribution). Queries are preceded by a double dagger (‡) and biblical references by a simple dagger (†).



‡ botrum quem olim exploratores de israel in vecte ferebant
† pergentesque usque ad torrentem botri absciderunt palmitem
cum uva sua quem portaverunt in vecte duo viri de malis
quoque granatis et de ficis loci illius tulerunt

Figure 4.4: Example of a correctly retrieved case of allusive text reuse, where a Bernardine passage (*S. 154, 3*, text above) is matched to biblical verse *Numbers, 13:24* (text below).



‡ descendentem vidit ille qui vidit
† dico enim vobis quod multi prophetae et reges voluerunt
videre quae vos videtis et non viderunt et audire quae auditis
et non audierunt
† et civitatem sanctam hierusalem novam vidi descendentem
de caelo deo paratam sicut sponsam ornatam viro suo

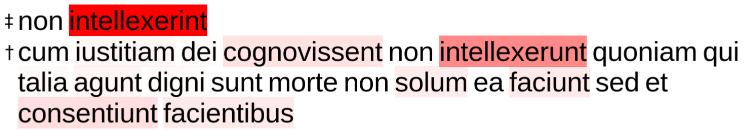
Figure 4.5: Example of an incorrectly retrieved case of allusive text reuse, where a Bernardine passage (*S. 27, 7*, text above) is matched to biblical verse *Luke, 10:24* (text in the middle), where *Apocalypse, 21:2* (text below) should have been retrieved instead.

A second type of error occurs when less relevant candidates are pushed higher in the rank due to semantic reinforcements in the wrong direction. For example, in Figure 4.5 we have a query together with a wrongly retrieved match ("dico enim ..." en. I say) and the true, non retrieved reference ("et civitatem ..." en. and the city). We observe that due to the high similarity of redundantly repeated perception verbs ("video" en. to see, "audio", en. to hear), the wrong match receives high similarity whereas the true reference remains at lower rank.

4.4 CONCLUSION & FUTURE WORK

Our experiments have highlighted the difficulties of automated allusion detection. Even assuming manually defined queries, the best performing model could only find the matching reference within the top 20 hits in less than half of the dataset. Moreover, the retrieval quality heavily drops when relying on windowing for query construction. This aspect calls for further research into the problem of automatic query construction for the detection of allusive reuse.

Across all our experiments, purely semantic models are consistently outperformed by a purely lexical `Tf-Idf` model. Similarly, the usage of lemmatization as a pre-processing step boosts the performance of nearly all models, which also suggests that ensuring enough lexical overlap is still a crucial aspect of allusive reuse retrieval. A similar reasoning helps explaining the superiority of `FastText` over `word2vec` embeddings, since the former is better at capturing morphological



‡ non intellexerint

† cum iustitiam dei cognovissent non intellexerunt quoniam qui talia agunt digni sunt morte non solum ea faciunt sed et consentiunt facientibus

Figure 4.6: Example of a missed case of allusive text reuse, where a Bernardine passage (*S. 8, 5*, text above) should have been matched to biblical verse *Romans, 1:32* (text below).

relationships, and lemma word embeddings suffer from data sparsity in the latter.

Overall, the hybrid models involving soft cosine show best performance, which indicates the effectiveness of such technique to incorporate semantics into BOW-based document retrieval and offers evidence that improvements in allusive reuse detection, however limited, can be gained from lexical semantics.

While the effect of adding semantic information from WordNet was shown to be less effective than leveraging word embeddings, it is still worth exploring to what extent enhanced similarity metrics defined over WordNet graphs as well as expanding the scope of semantic relationship beyond synonymy can have an impact on the retrieval of allusions (Budanitsky and Hirst, 2001).

Focusing on word embeddings, an interesting direction for future research is the application of soft cosine to text reuse detection across languages. Leveraging current advances in multilingual word embeddings (Ammar et al., 2016), multilingual word similarity matrices can be extracted on a joint embedding space. Besides joint learning of multilingual similarity spaces, a more practical approach is to align embeddings spaced across languages using word embedding alignment methods that require small amounts of translation pairs (Artetxe, Labaka, and Agirre, 2018).

5 CONTEXTUAL ASPECTS OF INTERTEXTUALITY

ABSTRACT Intertextuality is a highly productive concept in literary theory. The pervasiveness of intertextuality in literary texts has led simultaneously to a proliferation of applications with often divergent interpretations of the concept of intertextuality, as well as a recurrent interest in studying it from a computational point of view. Despite the potential of data-driven, bottom-up approaches, most computational research into intertextuality has focused on the matter of text reuse detection, exploiting surface-level properties to improve the performance of retrieval systems. In the present study, we utilize the Patrologia Latina—a substantial collection of religious texts spanning over a millennium of Latin writing (3rd to 13th centuries)—to provide a large-scale systematic study of biblical intertexts. On the basis of multi-level statistical models, we investigate two axes of intertextuality: the importance of lexical similarity, and the degree to which intertexts are thematically embedded in the context. More concretely, we investigate the extent to which the following contextual sources of variation help explain the distribution of intertexts along the aforementioned axes. First, we analyze the effect of authorship: do authors differ in the way they compose their intertexts? Second, we inspect factors related to the source collection (i. e. the Bible) to elucidate whether the authority and tradition of particular books exert an influence on the observed intertexts: do certain books trigger a more allusive or quotational intertext type? Finally, we take into account the dominant topic surrounding the intertext location and examine associations between the distribution of dominant topics and intertext types. On the one hand, our analysis indicates that both axes (lexical similarity and thematic embedding) play partially complementary roles in our computational account of intertextual types. On the other hand, we find that biblical books and, more strongly, dominant topics constitute important factors of variation, while the authorial signal remains comparatively weak.

This chapter is based on Enrique Manjavacas Arévalo, Folgert Karsdorp, and Mike Kestemont (2020). “A Statistical Foray into Contextual Aspects of Intertextuality.” In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)* (Amsterdam, The Netherlands, Nov. 18, 2020–Nov. 20, 2020). CEUR Workshop Proceedings 2723. Aachen, pp. 77–96. URL: `http://ceur-ws.org/Vol-2723/long28.pdf`

5.1 INTRODUCTION

Intertextuality is a well-known concept from literary studies that is commonly applied to texts across various periods and languages (Allen, 2000; Orr, 2010). Originally proposed by post-structuralist literary theorist, Julia Kristeva (Kristeva, 1967), intertextuality models literature as an intricate network of textual nodes that are interconnected by the "intertexts" that they share. Texts can refer to one another, for instance, through the literal integration of quotes from other works or through the inclusion of more subtle allusions to other texts. There is widespread agreement in literary studies that the intertextual approach has considerable merit, as it sheds light on how texts participate in the discursive space of a culture (Culler, 1976). In computational literary studies, intertextuality has also received ample attention, and the vast scope at which intertextuality can be studied has rendered the application of computational techniques very attractive from early on.

In spite of the considerable popularity of intertextuality in literary studies, there exists no straightforward definition of it (Moyise, 2002). Instead, a more fruitful discussion of intertextuality can be obtained by focusing on the aspects of intertextuality that scholars have exploited to generate new readings and interpretations of literary works. These aspects range from abstract structuring roles, in which an original text serves as organizational principle in the creation of another—see, for instance, the famous examples of the structural parallelisms between Homer's *Odyssee*, Virgil's *Aeneis* and Joyce's *Ulysses*—to more localized phenomena such as motifs or allusions, in which the link is established from and to specific passages.

In order to situate current computational approaches to intertextuality within this spectrum, Forstall and Scheirer (2019) introduced a useful distinction between large-scale effects and local effects of intertextuality, referring the latter to the scope of what they call "quantitative intertextuality".

Still, when considering such "loci similes", the bulk of computational studies so far have adopted a fairly narrow conception of the phenomenon, focusing on the issue of "text reuse detection", and relying on techniques that exploit string similarity—(Büchler et al., 2014b; Coffee et al., 2012a; Smith et al., 2014; Yale-DHLab, 2017). There are certainly exceptions. For example, Bamman and Crane (2008) exploits syntactic information (dependency paths and word order) to extract allusions in classical Latin literature, Scheirer, Forstall, and Coffee (2016) use LSI (Deerwester et al., 1990) to extract parallels in Latin epic, Lund et al. (2019) uses local topical information extracted from anchor-based topic models to extract intra-biblical references, and our approach to allusive reuse detection in Chapter 4 examined the application of distributional semantics to help improve retrieval performance. However, besides lexical and semantic similarity, the placement, source and type of an intertextual link can be thought of as being *conditioned* by a variety of *contextual* factors.

For example, it can be hypothesized that certain themes (e. g. "war" or "love") may be more likely than others to "trigger" references, perhaps because the author's conceptualization of that theme is indebted to a particular source. In that sense, the location of an intertext would be conditioned by its *embedding* in the triggering theme.

Moreover, writers may show preferences to borrow from particular authors, books or fragments of books. On the one hand, the influence of a particular source on a community of authors can explain the frequency of references to that particular source, due to, for instance, social biases, such as "conformist" or "anti-conformist" biases towards or against popular writers—see, for instance, recent literature from the field of Cultural Evolution (Acerbi and Bentley, 2014; Crema, Kandler, and Shennan, 2016; Mesoudi, 2011). On the other hand, the distribution of intertext types, considering, for instance, an axis of "literality" going from literal quotation to allusive reference, may be affected by mentioned influence: a particular source may exert an authoritative pressure towards a more literal style.

Furthermore, the type of reference that can be expected in a particular text may be a feature of authorial style. In this respect, we could expect to observe trends towards more or less allusive referencing as a marker of authorial preference.

Besides the degree of "literality", which is easily quantifiable in terms of lexical overlap, we need to consider a further aspect of referential style which is easily overlooked: the extent to which an intertextual unit is "prepared" by the textual context. If the textual contexts around the borrowing and the borrowed passage are handling similar themes, the intertextual link could be explained as having been facilitated by the theme similarity. A possible hypothesis in this regard is that shorter and more subtle allusions would necessitate a higher degree of contextual similarity with respect to the source passage to exist, because in the absence of such topical preparation, the audience would be more likely to miss the link. However, such a hypothesis relies on the problematic assumption that intertextual linking must be a conscious act of the writer to be perceived as such by the reader. Instead of top-down approaches to intertextuality, as the one implied in the previous hypothesis, we would like to systematically investigate factors of variation that influence the type of intertext in a bottom-up fashion, considering both axes: i. e. the degree of "literality" (quotational vs. allusive) and its embedding in the thematic context.

In the current study, we take a step back from the problem of retrieving local intertexts and present a quantitative analysis of the role of contextual factors on the placement of intertexts. Specifically, we tackle the issues of authorship, the impact of the source or referenced collection and the context theme. We make use of the Patrologia Latina (henceforth, Patrology), which is a large-scale corpus comprising numerous authors and books, known to be abounding in intertextual links. While the corpus shall be thoroughly introduced in Section 5.2, two facts about the Patrology are worth advancing now: on the one hand, the majority of authors form part of the same writing tradition, sharing themes, concerns and theoretical background. On the other hand, the main source of reference—the Bible—is shared. These two characteristics make the different authors and books commensurable from a statistical point of view and will allow us to approach the questions posed above from a data-driven perspective.

RESEARCH QUESTIONS More concretely, the research questions that we pursue in the present study are as follows:

1. Besides lexical similarity, does the thematic embedding of intertexts into their context represent an additional axis of meaningful variation?
2. As intertextual links vary along a continuum from more to less literal as well as in the degree to which they are thematically embedded in the topical context, do we observe systematic variation across authors?
3. What is the effect of tradition or authority on the referencing style of the considered authors? More specifically, do certain books of the Bible trigger particular types of reference? Does the structure of the source collection (i. e. the Bible in the present case) help explain such variation?
4. Besides authorship, do specific topics help further explain the type of reference and their topical embedding?

OUTLINE The remainder of the present chapter is structured as follows. First, Section 5.2 contains a description of the data sources underlying the study, as well as the preprocessing applied in order to produced text amenable to quantitative analysis. Next, Section 5.3 describes the computational approach used to operationalize the theoretical categories that the study targets: the type of reference along the quotation-allusion axis and the theme similarity with respect to the source passage. Next, in Section 5.4 we describe the statistical models used to approach the posited questions. Finally, in Section 5.5, we discuss the insights that can be drawn from the models and the answers that they deliver.

5.2 DATASET

5.2.1 Sources

The dataset used in the present chapter has been compiled on the basis of the Patrology, an extensive collection of editions of Latin writings, attributed to the so-called "Church Fathers" in the Christian tradition, as well as a number of other influential ecclesiastical authors. This monumental endeavor was initially undertaken by J.P. Migne between 1841 and 1855 (Migne, 1844-1855 (and 1862-1865)). The diachrony of this collection covers a reasonably balanced sample of more than a millennium of written text production, ranging from the oeuvre of Tertullian (2nd century AD) to that of Pope Innocent III (13th century AD). This resource, moreover, continues to be relevant in literary scholarship, not only because for many of the included works Migne's constitutes the most recent edition.

Despite the diverse origins of its source materials, the Patrology can be argued to represent a coherent corpus of religious Latin writings. The period that it covers coincides with the rise of Christianity, which would become the dominant religion throughout Europe by the reign of Charlemagne. The dissemination of the Bible—or rather that of its individual books, which often still circulated individually—played a major role of support in these developments. Biblical intertextuality (Moyise, 2002), in particular, pervades the texts in the Patrology. This is partly due to the considerable number of religious sermons included—which departed from or even revolved around specific biblical quotations—, and also because various aspects of medieval exegesis crucially depended on intertextual phenomena. One of the standard ways to understand the medieval Bible, for instance, was through an analogical understanding of the parallels between the Old and New Testament, also at the textual level. Therefore, it does not come as a surprise that we are not the first to use this data to study intertextuality using computational means (Ghiban and Trăuşan-Matu, 2013).

5.2.2 Curation

The digital version of the Patrology was extracted from the Corpus Corporum collection (Roelli, 2014), which offers high-quality OCR from Migne's 1853 edition in a convenient XML format. On the side of the source of the references, the Bible, we used the version of the Vulgate provided by the Perseus Digital Library (Crane, 1996). We kept the original structure of the Vulgate into *verse*, *chapter* and *book* as metadata, and added to each verse a tag indicating whether the verse is part of the Old or the New Testament.

5.2.2.1 *Gold Standard*

While the OCR'd documents from the Corpus Corporum do not include the biblical references as part of the XML markup, as Listing 5.1 shows, these have been kept in its original inline form, and can be extracted automatically through customary data-wrangling techniques. First, we applied regular expressions to match parentheses formatted in the manner specified in Listing 5.1. After extracting the book abbreviation (e. g. *Gen.*), we checked whether it appeared in a manually curated list of books. In the case of a positive match, we then try to parse the chapter and verse numbers (e. g. *II, 15*). Finally, the parsed reference (e. g. *Genesis 2:15*) was checked against the Vulgate to see whether it corresponds to a real verse.

```
   <p>Simili modo et tu, si bona
   quae habes forti cautela custodire non negligis,
   <pb n="0773B"/>
   circa tabernaculum tuum, et ea quae intra illud
 5 sunt tentoria suspendis. Nihil enim omnino tibi proderit
   bona in te spiritualia congregasse, nisi diligenti
   ea et sollicita circumspectione custodias. Hinc
   in sacra Scriptura legimus, quia <i>posuit Deus hominem
   in paradiso, ut operaretur, et custodiret illum (Gen. II, 15)</i>
       .
10 In paradiso quippe Deus hominem
   ponit, quando delectabilem tibi spiritualium gratiarum
   copiam gratuito largiens, in sancta et tranquilla
   conscientia suaviter te pausare facit.</p>
```

Listing 5.1: Example of an XML source file snippet from Adam Scotus (extracted from the *De tripartito tabernaculo*) showcasing a passage containing an annotation of a biblical reference (*Genesis 2:15*) in line no. 9.

The automatic extraction of manually coded references resulted in a dataset of 210,022 references, which facilitates large-scale computational analyses of biblical intertextuality.

VALIDATION While the OCR is not perfect, and the annotation by the editors can be expected to have missed cases of intertextuality, a manual inspection of a representative sample indicates that the automatic procedure manages to parse editorial annotations with high precision. This was tested on an isolated sample of 100 instances extracted from pairs with a particularly low lexical similarity. For this purpose, we deployed the Smith-Waterman algorithm (Smith and Waterman, 1981)—described in Section 2.5.2.3—with default parameters, and carefully checked for the alleged reference. After a first inspection of the distribution of scores over the entire dataset, it could be observed that scores higher than 4 consistently yielded true positive references,[1] and we, therefore, excluded these from the sample to be manually checked.

The set of references showing low alignment scores amounted to 35.5% of all references. From the manually checked subset, 82% of all references could be clearly found, 11% were unexpectedly located in the nearest context (due to OCR mistakes pertaining to the recognition of digits), and 7% were missed. The analysis thus reveals that about 2.45% (i. e. 7% out of 35.5% from the total) of all references are wrong, an amount that, while not fully negligible, was yet deemed to be unproblematic.

5.2.2.2 Preprocessing

We apply the same preprocessing pipeline to the Patrology and the Vulgate. First, the text is tokenized and POS-tagged using TreeTagger (Schmid, 2013). For lemmatization, we use the neural lemmatizer introduced in Chapter 3 trained on the cap corpus of Medieval Latin. As opposed to the commonly used TreeTagger lemmatizer (Schmid, 2013), the neural network based lemmatizer is able to analyze previously unseen types and to disambiguate between possible alternative analyses, which—as we shall see in Section 5.2.4—results in more coherent topics.

5.2.3 Sampling

The analysis focuses on a subset of authors that are particularly prolific and thus provide a fruitful test-bed for statistical analysis. From the entire Patrology, we sample authors who have contributed a total of at least 100,000 tokens and from their writings we sample books with at least 100 references to the Bible. We made sure that at least two books per author are held out for developing purposes. From the sampled subset, we further remove commentaries, which, due to their exegetical nature, refer to the Bible very copiously and in a less interesting manner from the point of view of intertextuality research. In total, commentaries amounted to 8 documents across all authors. The resulting subset (which amounts to 2,921,142 tokens or 2.7% of the collection) is further processed to extract passages containing references to the Bible, as described in Section 5.2.2.2. In total, we collected 15,195 biblical references across 24 authors.

Table 5.1 contains information about the authors that are contained in this subset. Besides the Latin name, we also report the number of references attributed to the authors, as well as the initials of their names, which shall serve as codes in the subsequent visualizations in this chapter.

[1] Note that the exact number of this threshold is dependent on the algorithm parameters and cannot be interpreted independently.

Name	Total Refs	Initials
Adamus Scotus	540	A-S
Alcuinus	796	A
Ambrosius Mediolanensis	661	A-M
Anselmus Cantuariensis	128	A-C
Augustinus Hipponensis	4057	A-H
Bernardus Claraevallensis	717	B-C
Fulgentius Ruspensis	356	F-R
Gerhohus Reicherspergensis	327	G-R
Gregorius I	165	G-I
Guibertus S Mariae de Novigento	624	G-S-M-d-N
Hilarius Pictaviensis	719	H-P
Hildebertus Cenomanensis	164	H-C
Hincmarus Rhemensis	454	H-R
Honorius Augustodunensis	186	H-A
Iulianus Toletanus	596	I-T
Odo Cluniacensis	1539	O-C
Paschasius Radbertus	313	P-R
Petrus Abaelardus	939	P-A
Petrus Cellensis	274	P-C
Prosper Aquitanus	203	P-A
Rabanus Maurus	597	R-M
Ratherius Veronensis	287	R-V
Tertullianus	355	T
Vigilius Tapsensis	198	V-T

Table 5.1: Information about the authors in the Patrology dataset.

The remaining documents of the Patrology were set apart and used for training a topic model, which we deploy in order to automatically capture the theme in a given passage.

5.2.4 Topic Modeling

A topic model based on the Latent Dirichlet Allocation (LDA) process (Blei, Ng, and Jordan, 2003) was trained on the lemmatized version of the remaining dataset, comprising 103,687,454 tokens. The topic model was trained using the `gensim` package (Rehurek and Sojka, 2010), which provides an efficient implementation of Online LDA (Hoffman, Bach, and Blei, 2010). We fit the model on documents processed to discard all not strictly alphanumeric words, or words not identified as adjectives, adverbs, nouns or verbs by the part-of-speech tagger shipped with TreeTagger. Moreover, we discard words that appear in a specifically designed stop-word list, including terms such as "dominus", "deus", etc. that were deemed irrelevant for composing topic-term distributions due to their high frequency.

5.2.4.1 *Optimization of LDA*

In order to ensure the quality of the inferred topics, we conducted a hyper-parameter fine-tuning experiment using grid-search with the goal of optimizing the C_V topical coherence measure (Röder, Both, and Hinneburg, 2015) on the subset of documents that were held out. Coherence measures aim at quantifying the degree to which a set of terms describes a coherent topic through the application of information-theoretic measures (i. e. how much does the appearance of a term in the topic tells us about the appearance of the other terms in the topic). Despite the limitation of topic coherence measures as proxies for topic quality in isolation, they are known to be amongst the strongest correlates of topic interpretability.

Figure 5.1 displays results from the topic evaluation experiments. Topic coherence is plotted over number of words per training document (`DocWords`), size of the vocabulary (`Top-k`), number of topics (`NumTopics`) and lemmatization model (`Model`).

As we can see, the vocabulary size (i. e. `Top-K`) has a positive influence on coherence, especially when increasing the size of the training documents and the number of topics. Overall, the neural lemmatizer yielded more highly coherent topics except for models with 1000 topics, where it lagged behind the non-disambiguating lemmatizer by a small margin in the `Top-K`=20k condition.

Based on the results of the experiment, we fitted the final model document fragments of 1,500 words, using 200 topics, a vocabulary truncated to the 20,000 most frequent words, using the lemmata extracted by the neural lemmatizer.

5.3 METHODOLOGY

In order to model the thematic embedding of intertextual references, we need to operationalize a notion of similarity of both a purely lexical and a thematic type. While lexical similarity can be easily approximated by means of set-based similarity metrics typically used in text reuse applications, the operationalization of thematic similarity in terms of similarity between the topic distributions inferred by the topic model requires certain preprocessing.

Since topic models are essentially modeling word co-occurrence patterns across documents, the presence of an intertextual link will bias the respective inferred topic distributions in a an expected way,

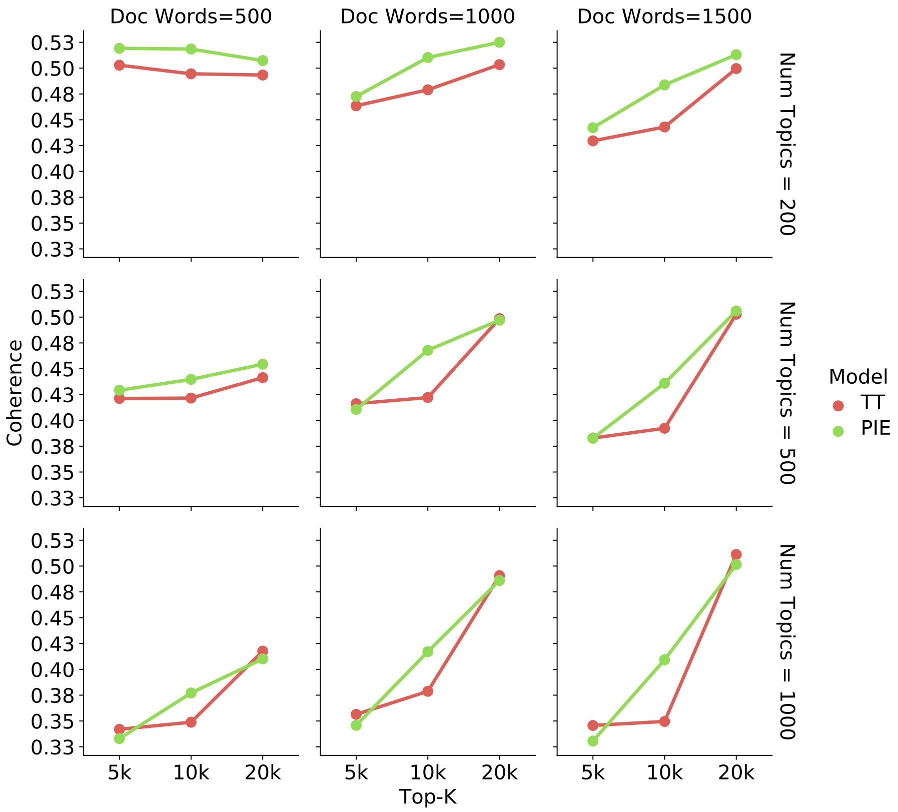


Figure 5.1: Topic Coherence evaluation over the number of topics, the size of training documents (in number of words), the size of the vocabulary and the lemmatization model.

especially if the intertextual link is based on high lexical overlap. It is, thus, important to disentangle topical from lexical similarity. For this purpose, we inferred the topic distributions after removing any lexical overlap with respect to the source document. Under such circumstances, a strong similarity between the respective inferred topic distributions must be driven by terms from the compared documents that are lexically different but display high within-topic co-occurrence. Arguably, we can interpret high topical similarity in this context as an indication of a high thematic embedding of the intertext into its textual surroundings.

5.3.1 Lexical Similarity

In order to extract lexical similarity, we resort to traditional methods from the text re-use literature – see, for instance (Seo and Croft, 2008). We focus on the Jaccard coefficient, which as shown in Section 2.5.2.1, is defined as the number of shared words divided by the total number of words types in the documents. In the present study, we compute the weighted version of the Jaccard coefficient, shown in Equation 5.3.1, which gives a more accurate value by taking into account the frequency of the words:

$$\text{Jaccard}(D_i, D_j) = \sum_{w \in D_i \cup D_j} \frac{min[c(w, D_i), c(w, D_j)]}{max[c(w, D_i), c(w, D_j)]} \qquad (5.1)$$

In Equation 5.3.1, $c(w, D_i)$ refers to the number of times word w appears in document D_i. In order to give higher weight to more literal borrowings, we represent the documents not just at the level of words but include word bi-grams and tri-grams as well. Finally, we do not consider the actual words but their lemmata and apply a stop-word list.[2]

5.3.2 Topical Similarity

Given the inferred topic distributions of a pair of source and target documents, we resort to information-theoretic measures in order to estimate the topic similarity of the underlying documents. In particular, we use the Jensen-Shannon Divergence (JSD), shown in Equation 5.2.

$$JSD(\theta_{D_i}, \theta_{D_j}) = \frac{1}{2} D_{KL}(\theta_{D_i} || \theta_{D_j}) + \frac{1}{2} D_{KL}(\theta_{D_j} || \theta_{D_i}) \qquad (5.2)$$

JSD corresponds to the arithmetic mean between the Kullback-Leiber Divergence (KLD) of the topic distribution of the i^{th} document θ_{D_i} with respect to the topic distribution of the j^{th} document $theta_{D_j}$ and the KLD of the topic distribution of the j^{th} document θ_{D_j} with respect to the distribution of the i^{th} document θ_{D_i}. By taking the mean, JSD transforms KLD into a symmetric measure. Since JSD is a divergence, we transform it into a similarity by subtracting it from one: $1 - JSD(D_i, D_j)$.

In order to guarantee rich topic representations, we consider left and right contexts of a given reference including a total of 500 words for the referencing documents, and the entire chapter-level context for the biblical text.

5.3.3 Topical Context

In order to approach RQ4, we need to capture the theme surrounding the particular intertexts. In the present study, we utilize the topic-model from Section 5.2.4 to identify the most dominant topic in the inferred topic distribution of a given passage. Thus, a given document

[2] Note that this stop-word list differs slightly from the one applied in the topic model pipeline, since the nature of the task is different.

D_i is assigned an index pointing to topic t with highest probability in the topic distribution inferred for document D_i:

$$\mathrm{Topic}(D_i) = \mathrm{argmax}_t \theta^t_{D_i} \tag{5.3}$$

By taking the single maximum probability topic as a summary of the distribution we are certainly ignoring important information about the overall composition of topics in the document—especially in high entropy topic distributions—and also limit the subsequent modeling from exploiting correlations in the distribution of topics across documents—since some topics will tend to co-occur with each other. However, this simplification makes the subsequent statistical modeling considerably easier, while still capturing a considerable amount of topic information.

5.4 DATA ANALYSIS

We approach the research questions by making use of a multivariate multi-level intercept-only model using lexical similarity (`lex`) from Section 5.3.1, and topic similarity (`topic`) from Section 5.3.2 as the outcomes. In order to analyze the effects of authorship and contextual theme as well as any source collection-level effects on the type of intertext, we specified a series of multi-level models including random intercepts for each of the levels in these grouping factors. The number of levels per grouping factor amounted to the following: author (`A`: 24 levels), biblical book (`B`: 52 levels) and dominant topic (`T`: 129 levels). The number of topics differs from the total number of topics in the topic model (200) because not all of the estimated topics are realized as dominant topic in the target dataset.

We conducted all analyses in `R` (3.6.3) (Ihaka and Gentleman, 1996) using the `brms` package (Bürkner, 2018) for model fitting. `brms` provides a user-friendly interface to specify models to be fitted with `Rstan` (Carpenter et al., 2017), and, thus, benefit from a powerful implementation of a modern Hamiltonian Monte Carlo sampler.

Despite its known strong sampling abilities, Hamiltonian Monte Carlo is not a bulletproof method, and a number of diagnostics must be checked in order to ensure the validity of the resulting posterior distributions. For instance, it is crucial that effective sample sizes are large enough, that the samples are homogeneous across chains (i. e. $\hat{R}$ values should be close to 1), and that divergent transitions are not endangering the inference. For the present experiments, we ran 4 chains with 2000 iterations per chain using the first 1000 as warmup, monitoring the mentioned diagnostics for convergence. Finally, unless otherwise specified, we chose weakly informative priors as per the defaults in the `brms` package, which, at the time of running the experiments consisted of Student T distributions with 3 degrees of freedom, location of zero, and a scale of 2,5. These weakly informative priors are agnostic to the parameter values while preventing the sampler from exploring highly unlikely regions of the parameter space.

Since we were not particularly interested in the magnitude of the effects, we did not operate on the outcome variables directly, but instead applied a normalizing transformation to center the variables around a zero-mean and a unit standard deviation. This transformation also facilitates model fitting and makes the interpretation of coefficients more accessible, especially when considering comparisons of variables in different scales.

5.4.1 Model Definition

The general model including all grouping factors is defined by Equation 5.4.

$$
\begin{gathered}
\begin{bmatrix} y_l \\ y_t \end{bmatrix} \sim \text{MVNormal}(\begin{bmatrix} \mu_l \\ \mu_t \end{bmatrix}, \Sigma) \\
\Sigma = \begin{pmatrix} \sigma_l & 0 \\ 0 & \sigma_t \end{pmatrix} R \begin{pmatrix} \sigma_l & 0 \\ 0 & \sigma_t \end{pmatrix} \\
\begin{bmatrix} \mu_l \\ \mu_t \end{bmatrix} = \begin{bmatrix} a_l \\ a_t \end{bmatrix} + \begin{bmatrix} a_l^A \\ a_t^A \end{bmatrix} + \begin{bmatrix} a_l^B \\ a_t^B \end{bmatrix} + \begin{bmatrix} a_l^T \\ a_t^T \end{bmatrix} \\
\begin{bmatrix} a_l^K \\ a_t^K \end{bmatrix} \sim \text{MVNormal}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma^K) \\
\Sigma^K = \begin{pmatrix} \sigma_l^K & 0 \\ 0 & \sigma_t^K \end{pmatrix} R^K \begin{pmatrix} \sigma_l^K & 0 \\ 0 & \sigma_t^K \end{pmatrix}
\end{gathered}
\tag{5.4}
$$

As we can see, the statistical model consists in a bi-variate model—with outcomes y_l and y_t—that includes no predictors, and groups observations according to three different criteria.

Certainly, there is nothing inherent to the research design that prevents from including predictors. For instance, model fit could be improved by considering predictors such as the diachrony or genre of the works to which the intertextual references belong, or the density of intertextual references in the surrounding passage. However, for simplicity we left such predictors out of the scope of the present study, and call for future work to investigate the matter.

Moreover, our statistical approach models observations of `lex` and `topic`—$[y_l, y_t]^T$—as coming from a bi-variate normal distribution. Furthermore, the means are decomposed into grand means, $[a_l, a_t]^T$, and group-specific deviations from the mean $[a_l^K, a_t^K]^T$ for group K. The latter are modeled hierarchically as coming from a yet another

Model	ELPD	ELPD (SE)	P	P (SE)	DIFF	DIFF (SE)
$\mathrm{M}_{\mathrm{A} \cup \mathrm{B} \cup \mathrm{T}}$	-37876.8	255.5	285.5	6.6	0.0	0.0
M_{T}	-39316.3	262.7	188.5	5.5	-1439.5	51.0
M_{A}	-40720.1	291.5	43.5	1.3	-2843.4	120.3
M_{B}	-40966.8	299.5	76.9	2.4	-3090.0	121.1
$\mathrm{M}_{\mathrm{B} \cup \mathrm{T}}$	-38430.3	264.0	257.4	6.6	-553.5	32.2
$\mathrm{M}_{\mathrm{A} \cup \mathrm{T}}$	-39971.3	294.4	106.7	2.6	-2094.6	116.7

Table 5.2: Summary statistics of the model comparison displaying LOO estimates of the ELPD with Standard Errors (SE)—lower is better—, estimates of the effective number of parameters (P), and differences in ELPD with respect to the best model (DIFF). All Pareto-k estimates computed in the estimation of ELPD were below 0.7, thus ascertaining the validity of the estimation procedure.

multivariate normal centered around zero and covariance matrix Σ^K. Following Gelman and Hill (2006, Chapter 13), covariances Σ and Σ^K are decomposed into a diagonal matrix of standard deviations that model lexical σ_l, σ_l^K and topical σ_t, σ_t^K variation individually and a correlation matrix R, R^K that additionally targets correlations between both response variables. Finally, we set the priors of all σ terms to student-t priors and the correlation components R to flat LKJ priors (Lewandowski, Kurowicka, and Joe, 2009).

5.4.2 Model Comparison

We first analyze the importance of the different factors on the outcome distribution through information criteria. As it is commonly done in Bayesian model comparison, we use the Expected Log (Pointwise) Predictive Density (ELPD) as test measure—which provides an estimate of the predictive accuracy of a model on new datasets (out-of-sample data). Estimates of ELPD can be efficiently obtained—i. e. without having to refit multiple models on the different data partitions—through approximate Leave One Out (LOO). In particular, we use the Pareto-smoothed Importance Sampling (PSIS-LOO) method—see Vehtari, Gelman, and Gabry (2017) for a description of the method and Vehtari, Gelman, and Gabry (2018) for an implementation in the **R** programming language.

Table 5.2 shows the results of the comparison. As we can see, the model utilizing all grouping factors ($\mathrm{M}_{\mathrm{A} \cup \mathrm{B} \cup \mathrm{T}}$) is expected to have much better predictive performance than any of the single-grouping models. For the individual factor models, we observe that theme-level grouping has stronger explanatory power than author-level or book-level grouping, with the latter two receiving ELPD scores within error of each other.

In order to better grasp the respective contribution of the book-level and author-level factors to the model's predictive performance, we fitted $M_{A \cup T}$ and $M_{B \cup T}$ and compared them to the general model $M_{A \cup B \cup T}$. The results of the comparison are shown in the last two rows of Table 5.2. As we can see, $M_{B \cup T}$ produces much better estimates than $M_{A \cup T}$, which indicates that grouping according to reference book produces a model with more explanatory power than when grouping according to author.

Finally, we can gain further insight into the relevance of the different grouping factors by inspecting the variance in the outcome distribution they explain. Leveraging the posterior predictive distribution of the outcomes from the fitted Bayesian model, we can estimate the proportion of explained variance in a general way using a similar approach to Gelman et al. (2019). First, we compute a reference variance using 1,000 samples from the posterior predictive distribution excluding variation from the grouping factors. Next, we compute the variance in 1,000 samples from the posterior predictive distribution including the grouping for which the estimate of explained variance needs to be computed. Finally, the estimate of explained variance is computed by subtracting the latter variance from the reference variance. Since, by definition, the reference variance is larger or equal than the group-specific variance this estimate of explained variance is bounded between 0 and 1. For the present case, we use the general model ($M_{A \cup B \cup T}$) and consider different combinations of grouping factors.

Figure 5.2 shows the distribution of explained variance per outcome variable considering all grouping factors ($\sim A + B + T$) and the individual grouping factors ($\sim A$, $\sim B$ and $\sim T$). As we can see, while both book-level and topic-level grouping factors have an approximately equal estimate of explained variance for the lexical and topical outcomes, the author-level grouping seems to explain a larger share than the topic-level grouping. This result seems to suggest that lexical similarity does a better job at discerning between referencing styles of authors. Still, since the author grouping—together with the book-level grouping—yielded the smallest out-of-sample predictive performance estimates, we can only postulate a mild authorship signal.

5.4.3 Inspection of grouping factors

Having inspected the relative contributions of the different grouping factors, we now consider the posterior estimates of the outcome variables at different grouping factors. As discussed in Section 5.1, our analysis of local intertextuality posits two aspects to intertextual links. Besides the degree of "literality" of an intertext, we would like to

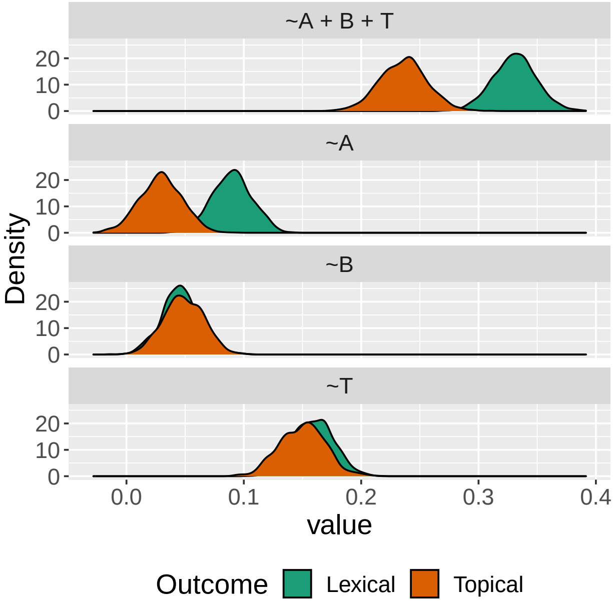


Figure 5.2: Estimates of explained variance based on the model $M_{A \cup B \cup T}$ for different grouping factors with respect to a reference model that uses no random effects. $\sim$A+B+T refers to the model including all random effects. The notation $\sim$K is used to refer to the model ignoring all random effects except K.

include its thematic embedding in the context into the analysis. By inspecting the statistical relationships between the posterior estimates of both outcome variables across grouping factors, we aim to gain insight about how these two aspects of intertextuality complement each other.

5.4.3.1 *Author grouping*

The plot in Figure 5.4 shows the mean posterior estimates for authors, averaging over books and topics.

Since we observe considerable correlation between topical and lexical similarity, we zoom in first on this aspect. Figure 5.3 shows the posterior estimates of the correlations between the dependent variables for each of the grouping factors, which, as shown in Equation 5.4 was estimated directly by the statistical model—R^K in the adopted notation.

It is important to note that the correlation estimated by the hierarchical model estimates is higher than the correlation estimated by the dataset-level maximum likelihood estimate—shown by the vertical bar in Figure 5.3. The difference is due to the effect of multilevel modeling shrinkage. In a multi-level model context, statistical estimates of target quantities are decomposed into population-level and group-level components. Following the so-called "partial-pooling" strategy, group-level information is taken into account against the

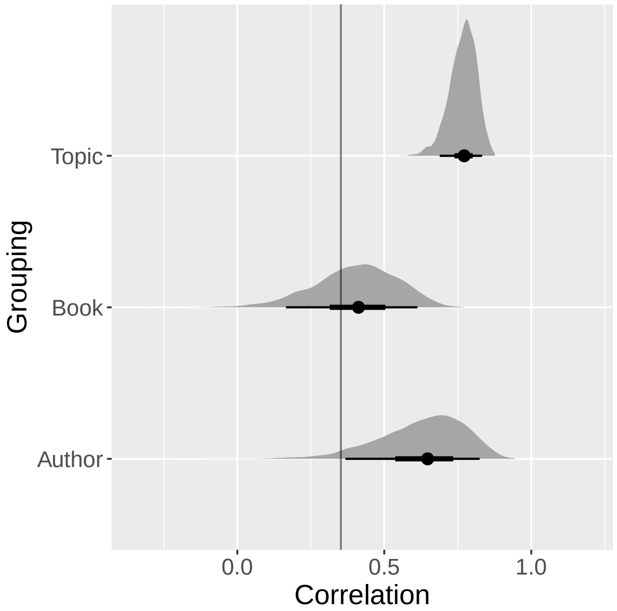


Figure 5.3: Posterior correlation estimates of the outcomes at different grouping factors including mean and 0.5 and 0.89 credible intervals. The vertical line depicts the overall empirical correlation between centered variables.

evidence contributed by the overall population. This results in flexible group-level estimates that are "shrunk" towards the population-level estimate if the number of observations within the group is small. As shown in Figure 5.5, as a result of shrinkage the author mean estimates are pushed towards the diagonal when considering book and topic grouping factors, with no author mean estimate remaining within the upper-left quadrant.

As a result of the correlation, both the upper-left and bottom-right sections of Figure 5.4 are considerably less populated. In combination with the analysis from Figure 5.2, we can interpret the high correlation between lexical similarity and topical embedding for author-level estimates in the sense that the lexical similarity axis suffices to explain the variation observed between authors.[3]

However, despite the high correlation, it is nevertheless interesting to zoom into the characteristics of extremely positioned authors. In particular, we examine authors with an outlier status according to the visualization, and the extent to which the known characteristics of their writing styles support or contravene the location assigned by the statistical model. This type of analysis has the goal of validating the explanatory power of the proposed statistical model and has, thus, a confirmatory nature. We note, however, that, upon validation, the

[3] As a reminder from Section 5.3, the estimates of topic-level similarity were computed on documents after removing the lexical overlap to avoid biases from lexical similarity, which can, thus, no longer explain the observed correlation.

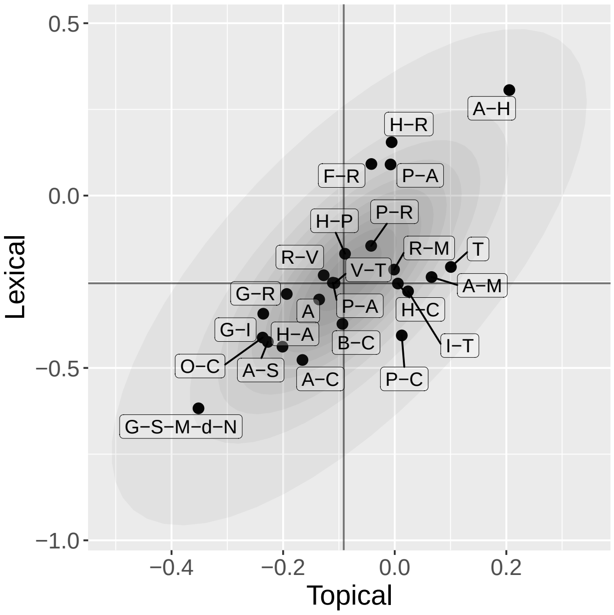


Figure 5.4: Mean posterior estimates of the bi-variate response for authors based on model $M_{A \cup B \cup T}$, averaging over books and topics. Authors are labeled using their initials (see Table 5.1 for the full table of author abbreviations).

same type of statistical model could be used for exploratory data analysis.

The first outlier is Peter Cellensis (P-C), an author who, being known for his allegorical style (Ott, 1911), appears, accordingly, in the bottom-right quadrant of the plane. A second noteworthy outlier is Bernard of Clairvaux (B-C), who, as discussed in Section 2.6.1.1, is known for his allusive referential style. As we can see, Bernard of Clairvaux appears in the vicinity of Peter Cellensis, slightly further towards the left, positioned at average topical similarity and below average lexical similarity. The statistical evidence, thus, broadly agrees with the general qualitative characterization of the author. Finally, the last two outliers that we shall highlight are Augustine of Hippo (A-H) and Guibert of Nogent (G-S-M-d-N). These authors represent extremes at the respective ends of the distribution and are, in contrast to the previous two, more difficult to characterize. Augustine of Hippo, located in the upper-right, is statistically characterized by a highly embedded and literal style. In terms of literary style, however, Augustine of Hippo is hard to characterize given the large extent and variety of his work. Augustine of Hippo's position in the plane could be, instead, explained by the prominent place of his work *De Consensu Evangelistarum*—a book an attempt at harmonizing the different accounts of Christ's life in the Gospels—in the sampled set of references (about a third of all Augustine of Hippo's references are from this book). Guibert of Nogent—a little known author from the 11th

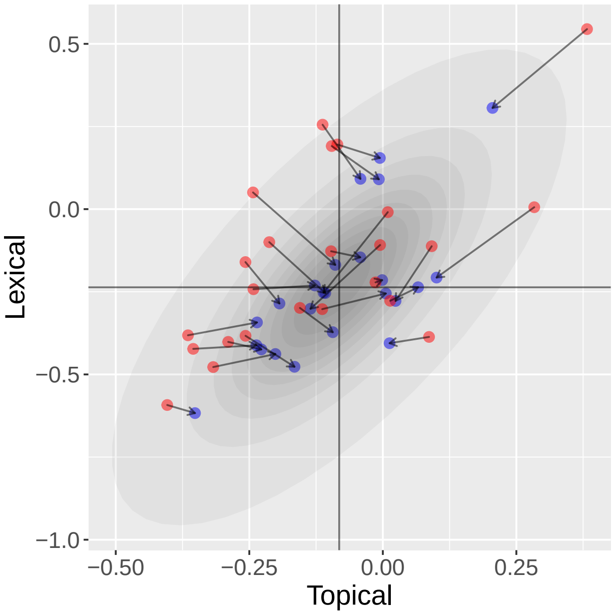


Figure 5.5: Effect of shrinkage on the mean author estimates shown as displacements from the maximum likelihood mean estimates (in red) towards the mean posterior estimates from model $\mathrm{M}_{\mathrm{A \cup B \cup T}}$ (in blue). Overall mean effects are shown in both plots, discarding variance, in order to facilitate the comparison. Note that variables were transformed to be centered around a zero mean.

century—is, in contrast, located towards the the bottom-left, leaning towards loosely embedded references with small lexical overlap.

5.4.3.2 *Book grouping*

Figure 5.6 shows the mean posterior estimates for books. We now observe a less correlated distribution, with a clear pattern emerging from the partition of the Bible into the Old and the New Testament. In general terms, biblical intertext linking to the New Testament tends towards a more quotational style. On the topical axis, the trend is less clear with a mild association of the New Testament with higher thematic embedding. After observing such a pattern, we fitted an additional model nesting the book levels into their corresponding Testament. The resulting model, however, did not yield any considerable improvements in the LOO estimates with respect to model $\mathrm{M}_{\mathrm{A \cup B \cup T}}$ and was, therefore, not further considered in the analyses.

Again, inspecting the outliers can help the interpretation of the distribution from the point of view of a confirmatory analysis. In the top of the plot we find the *Deuteronomy*, a biblical book that contains a large body of laws, blessing and courses, all of which is more likely to be quoted than alluded to. In contrast, in the more allusive quadrant of the plane—i. e. the bottom-right, we find the *Song of Songs*, a book that largely consists of love poems and a strongly allegorical style.

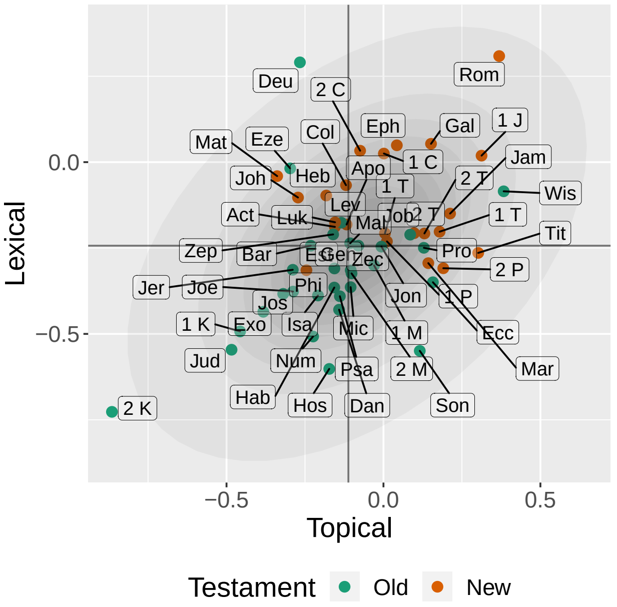


Figure 5.6: Mean posterior estimates for books from model $\mathrm{M}_{\mathrm{A}\cup\mathrm{B}\cup\mathrm{T}}$, averaging over authors and topics respectively. Dots have been highlighted according to whether the book pertains to the Old or the New Testament.

5.4.3.3 *Topic grouping*

Finally, we inspect the estimates for the topic-level grouping. Given the large number of topics and the fact that, despite our efforts to optimize the topic coherence of the topic-term distributions, topic-modeling algorithms do not provide guarantees about the interpretability of the inferred topics, care should be taken when attempting to draw conclusions from the posterior mean distribution.

Figure 5.7 displays the mean posterior estimates for topics. Similarly to the distribution of posterior means for authors, the distribution of topics shows an important degree of correlation. However, in this case there is considerable dispersion in the upper-right section. While a thorough exploration of the topics is beyond the scope of the present study, we have singled out a number of cases for commentary. For illustration, the selected topics have been highlighted in Figure 5.7 and the corresponding topic descriptors (top probability terms under the given topic) are shown below.

- **Topic 11** “propheta" (prophet), “Isaiah", “apostolus" (apostle), “Matthaeus" (Matthew), “scriptura" (Bible)
- **Topic 30** “anima" (soul), “ratio" (reason), “cogito" (to conceive), “sensus"
- **Topic 36** “fides" (faith), “veritas" (truth), “pax" (peace), “credo" (to believe)

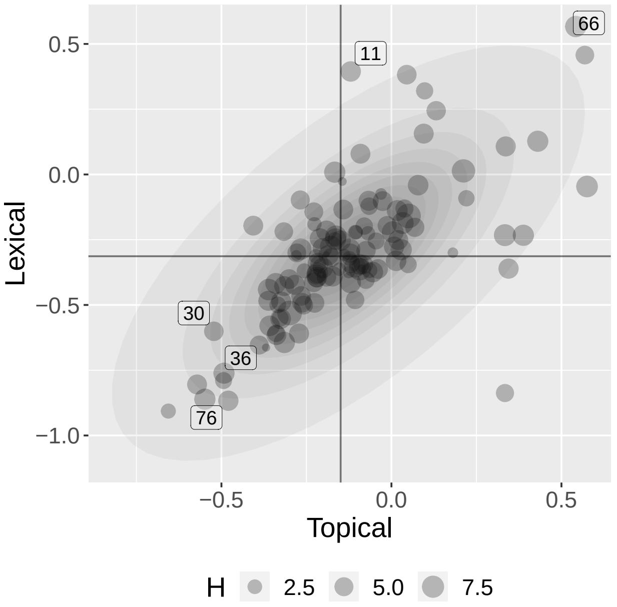


Figure 5.7: Mean posterior estimates for books and topics from model $M_{A \cup B \cup T}$, averaging over authors and topics and authors and books respectively. As a sanity check, the size of the topic is proportional to the entropy of the corresponding topic-term distribution. As we can see, no specific entropy-related patterns can be observed from the plot. Note that variables are centered around zero.

- **Topic 66** “sara" (Sarah), “ancilla" (slave), “Abraham", “angelus" (angel)
- **Topic 76** “voluntas" (will), “neccesitas" (inevitableness) , “liber" (free), “arbitrium" (judgement)

Topics 30, 36 and 76, which are located on the rather allusive quadrant of the panel, all seem to refer to moral and philosophical terms as well as to concepts relating to the human psyche. Topic 11, which points to a topic that triggers intertexts predominantly characterized by high lexical overlap, seems to relate to writings of and about prophets, apostles, etc. Such trend could indicate that references to authoritative figures are more likely to appear regardless the thematic context. Finally, Topic 66 located towards the upper-right extreme corner, thus indicating both high lexical and topical similarity, groups terms related to events that regard an important biblical figure: Abraham.

5.5 DISCUSSION

Having discussed the statistical analysis, we now proceed to address how the statistical evidence helps approaching the research questions posited in Section 5.1.

5.5.1 Thematic Embedding of Intertexts

In RQ1, we were interested in quantifying the extent to which intertext types could be characterized by two complementary axes of variation, one capturing lexical similarity and another one the thematic embedding of the borrowing passage into its context. On the basis of an operationalization of these aspects using traditional lexical similarity measures from text reuse detection and a specifically designed application of LDA topic modeling, we assessed the helpfulness of these variables for characterizing the observed variation across different grouping factors. Apriori, the intersection of both axes should produce four intertextual gradual types—depending on whether lexical and topical similarity are below or above mean—which correspond to the four quadrants shown in Figures 5.4, 5.6 and 5.7. Our analysis generally showed a correlation between both aspects, which resulted in low-density bottom-right and, especially, upper-left quadrants. The presence of high lexical similarity seems to generally trigger high topical embedding, even when controlling for lexical overlap during the estimation of topical similarity. As a result, we can conclude that the posited axes are not complementary. A hypothesis, according to which cases of allusive text reuse would rely on proportionally higher degrees of topical embedding to reinforce the intertextual link, is equally left without support. Despite the mentioned correlation, we can conclude that the inclusion of both axes can help producing a fuller picture of local intertextuality since, first, correlation varied depending on the grouping factor and, second, the position of outliers with respect to the general trend highlights the particularities of authors, books or topics and can aid future exploratory analyses.

5.5.2 Intertextual Marks of Authorial Style

With respect to RQ2, we found mild evidence of authorial signal in the type of intertext that authors place when referring to the Bible. This signal was especially pronounced on the lexical similarity axis. This result is broadly congruent with the state of the art in computational authorship identification: depending on the topical diversity of a corpus, semantic features in isolation rarely outperform more straightforwardly engineered surface features, such as word choice (Sari, Stevenson, and Vlachos, 2018). With respect to the topical embedding of intertexts, author variation was comparatively less important due to the high correlation with lexical similarity. The induced shrinkage by the hierarchical model resulted in a higher correlation than what a traditional correlation coefficient without pooling would have correspondingly under-estimated. Despite the observed correlation, the outlier status of certain authors with respect to the general trend could still be interpreted in a stylistic way—e. g. the discussed cases of Peter Cellensis and Bernard of Clairvaux.

5.5.3 Effects of Authority on Intertextual Style

With respect to RQ3, we observed a stable effect of the target collection, specifically the biblical book from which the reference originated. Model comparison showed that this effect plays a bigger role than authorial preferences in the distribution of the outcome variables. The distinction between Old and New Testament was highly relevant since it uncovered a pattern according to which New Testament books tend to elicit higher lexical similarity. Though this finding is probably not translatable to other contexts in which no single source plays such a dominant role so as to exert authoritative pressure on the type of intertext, it nevertheless highlights the importance of considering not just the borrowing and borrowed text but also structural aspects of the source collection when studying co-variates of intertextual links.

5.5.4 Effects of Topic on Intertextual Style

Finally, with respect to RQ4, the statistically most important grouping factor turned out to be the dominant topic in the borrowing passage. In this case, the correlation between lexical and topical similarity was estimated to be highest, though considerable dispersion was observed in the upper-right quadrant. Manual inspection of topics with posterior means located in significant locations illustrated that their positioning could be made sense of on the basis of the topic descriptors, even though any general theorizing on the effect of topical trends on the type of intertext must be left for future work.

5.6 FUTURE WORK

In the present chapter, we have conducted a systematic analysis of relevant factors of variation of intertextual types from a quantitative and data-driven perspective. An implicit assumption of our study, which technically underlies all computational approaches to intertextuality, is that local intertextual links depend on an explicit textual form that can be more or less rigorously identified. While in this study we exploited an already annotated collection of references, replicating our analysis on other collections depends on the automatic extraction of intertextual links. However, such analysis would require the application of text reuse detection algorithms that yield both high precision and recall for allusive cases. In order to expand the scope of quantitative intertextuality research, future efforts should, thus, aim not just at improving the task of intertextual retrieval, but also systematically evaluating the precision and recall that can be expectedly obtained. Moreover, since the effect of topic-level grouping turned out to be highly explanatory of the distribution of intertextual links, we hypothesize that such contextual interactions may turn out to be relevant for intertext retrieval applications, which can test how to incorporate them into their retrieval models.

Finally, our work relied on LDA-based topic models and therefore on topics that are not guaranteed to be interpretable. The acknowledgement of this limitation led us to refrain from an exhaustive qualitative exploration of intertext type distributional patterns at the topic-level. In the present chapter, we provided only fragmentary evidence of such topic-intertext relations: e. g. that the posterior means for lexical similarity and thematic embedding under topics related to moral and philosophical terms are low. However, we believe that future work should investigate ways in which researchers can systematically explore these topic spaces in order to elicit potentially fruitful hypotheses.

6 MATTERS OF AGREEMENT

ABSTRACT In this chapter, we report on an inter-annotator agreement experiment involving instances of text reuse retrieved by two alternative algorithms. We do so in the context of intertextuality, a concept from literary theory that emphasizes the role of references between texts. We target the application use case of textual scholars whose aim is to document intertextual links in the critical apparatus of an edition. Employing a Bayesian implementation of Cohen's κ for multiple annotators, we assess the relative utility of the algorithms, using as proxy how controversial the candidate instances of reuse are that the algorithms retrieve. Simultaneously, we produce a novel estimation of inter-annotator agreement in the context of intertextuality, exploring the challenges that arise from manually annotating a dataset of biblical references in the Medieval Latin writings of Bernard of Clairvaux. Our analysis shows that a semantically motivated algorithm retrieves candidate pairs that under circumstances are less controversial than those retrieved by an alternative text alignment algorithm with a slight bias towards literal reuse styles. Moreover, we show how our Bayesian formulation of Cohen's κ enables the inclusion of relevant factors of variation, which not only results in better estimates of agreement but also enables us to statistically explore agreement from various perspectives. Finally, a discussion of the hurdles encountered by annotators supplements the results of the statistical analysis, contributing a qualitative insight into the difficulties related to identifying instances of text reuse in literary works.

This chapter is based on Enrique Manjavacas Arévalo, Laurence Mellerin, and Mike Kestemont (2021). *Quantifying the Utility of Text Reuse Detection Algorithms through Bayesian Inter-annotator Agreement Indices.* Forthcoming

6.1 INTRODUCTION

The automatic detection of cases of text reuse in literary collections has the ultimate goal of enabling literary scholars to explore networks of intertextual references between literary works. This goal materializes in more concrete use cases for computationally-aided scholarly work, which—as discussed in Section 1.2.1—include visualizing high-level patterns in the referential connections between collections of texts (Jänicke et al., 2015; Yousef and Jänicke, 2021), studying the influence that a given writer has had on subsequent generations (Bloom, 1973), or, finally, aiding the preparation of (nowadays digital) editions of historically relevant literary works, for which editors seek to identify the sources of borrowed passages.

A matter worth studying in this context is the de facto "utility" of specific text reuse detection algorithms. Generally, studies of retrieval performance—in which a corpus is first manually or semi-automatically searched and annotated for cases of reuse and then employed as a test-bed for comparing the retrieval performance of candidate algorithms—can be viewed as studies of the usefulness of those retrieval algorithms. These studies are informative to practitioners willing to apply text reuse detection algorithms on their own corpora—although the informativeness is reduced when no out-of-sample estimates of performance are provided, and, as a result, the relative efficacy of the compared algorithms cannot be generalized to future corpora. In any case, these studies rely on existing benchmark corpora, and the production of these depends, in turn, on reliable annotation processes. In this respect, two aspects of text reuse studies in literary contexts turn the process of benchmark corpus compilation into a problematic enterprise. The first one is that the assessment of intertextual references is a highly interpretative matter. The second one is that the interpretation of these links demands a specific set of skills and expertise that is scarce and difficult to find. In this study, we approach the matter of agreement in the context of intertextuality with a double goal in mind. First, we offer a novel quantitative assessment of the difficulties of annotating reuse using inter-annotator agreement indices. Second, in the absence of benchmark corpora, we set to characterize alternative text reuse detection algorithms with respect to how controversial the links are that they retrieve.

Our study targets the application use case of editors of Medieval Latin literature, whose aim is to find intertextual links to the Bible and exhaustively document them in the accompanying "critical apparatus". The process of identifying intertextual links in the context of literary editing typically consists of several search iterations. During these searches, scholars make use of their own knowledge of both source and target collections, and are possibly assisted by text reuse detection software. In more advanced stages of the process, a large part of the bulk of reuse cases has already been collated and, thus, the role of computer-assisted identification becomes more dominant: remaining cases are likely to be elusive and computational methods can offer a vantage point. At the same time, there is a shift in the role of the automated system. While at the beginning the focus lies on maximizing intertext retrieval, with a primacy of precision, by the end higher recall and more elaborate fine-tuning are required.

Here, we focus on a late-stage iteration, where the algorithms must satisfy more specific and demanding requirements, and compare two algorithms based on different paradigms: one based on the VSM paradigm, and another one based on the text alignment paradigm. Importantly, we resort to a probabilistic formulation of inter-annotator agreement indices that allows us to deploy sophisticated statistical methods. We use the tools of multi-level statistical modeling to provide accurate and robust estimates of the expected agreement for candidate pairs retrieved by the competing algorithms, while controlling for additional factors of variation.

CONTRIBUTIONS More concretely, we make the following contributions. First, using inter-annotator agreement indices, we investigate the utility of two competing text reuse retrieval algorithms in the context of a late-stage editing phase in which the goal is to exhaustively find relevant missed cases of reuse. We implement a Bayesian variant of a popular inter-annotator agreement index that allows us to compute robust estimates of agreement in the presence of small sample sizes and control for and examine relevant factors of variation. We find that under certain circumstances a semantically motivated text reuse algorithm produces slightly higher inter-annotator agreement scores than an alternative retrieval method based on the text alignment paradigm—which has a bias towards more literal reuse styles. Second, we statistically inspect additional factors of variation—related to both objective (style of reuse retrieved by the system) and subjective (knowledge of the collection from which the passages are borrowed)—that may help explain the obtained agreement scores. Specifically, we find that the amount of lexical overlap in the candidate pair exerts a non-monotonic effect on the expected agreement, indicating that, while high agreement scores correlate with above-average lexical overlap, overly literal reuse candidates can be controversial. Furthermore, we find that the biblical book from which the source of the reference stems is a significant factor of variation. Finally, we examine the main hurdles to agreement that our annotators encountered during the experiment as perceived by the annotators themselves, highlighting not only that expert knowledge on the target collections can have important consequences in the assessment of intertextual links, but also that choices in the experimental design may contribute to inflated levels of disagreement.

OUTLINE The remainder of the chapter is structured as follows. First, in Section 6.2, we discuss the application of inter-annotator agreement indices to intertextuality research. We introduce relevant inter-annotator agreement scores in Section 6.2.1, and, in Section 6.2.2, we formulate an implementation of multi-κ using a Bayesian hierarchical model to account for statistical co-variates. Second, in Section 6.3, we detail the experimental setup, discussing the experimental design, relevant text reuse detection algorithms and the data sources underlying the study. Next, in Section 6.4, we describe the results of the experiment, providing a comparison of competing statistical model for the estimation of inter-annotator agreement (Section 6.4.1), an analysis of the estimated inter-annotator agreement indices (Section 6.4.2), and the results of the post-experimental report (Section 6.4.3). Finally, we analyze the results and derive interpretations and conclusions in Section 6.5, before finishing with pointers to future research in Section 6.6.

6.2 INTER-ANNOTATOR AGREEMENT INDICES

It is noteworthy that inter-annotator agreement studies of intertextuality are unfrequent.[1] Arguably, this may be due to the scarcity of the skill sets that are needed in order to evaluate the presence of an intertext—especially in cases of editions of non-mainstream literary works in historical languages. In any case, the lacuna in the literature not only means that no reference methodologies are available but also that the magnitude of the agreement scores resulting from new experiments is hard to interpret for lack of comparison.

For the present purposes, we follow the common practice in Computational Linguistics, a highly related field, and resort to chance-corrected inter-annotator agreement indices formulated in terms of expected and observed agreement (Artstein and Poesio, 2008). This family of indices most notoriously includes S (Bennett, Alpert, and Goldstein, 1954), π (Scott, 1955) and κ (Cohen, 1960). The application of these indices in computational studies has the goal of establishing the reliability and consistency of the annotation process with the goal of ensuring the correctness of the annotations (Artstein, 2017). In the process of bootstrapping an annotated resource as a test-bed for Machine Learning algorithms, inter-annotator agreement experiments can help determining the correctness of the resulting corpus. These experiments rely on the existence of an annotation guideline that unequivocally defines the correct annotation for a given instance,

[1] Bär, Zesch, and Gurevych (2012) includes an ad-hoc study of inter-annotator agreement of the annotation guidelines for their evaluation corpus—the Wikipedia Rewrite Corpus (Clough and Stevenson, 2011). This corpus, however, contains examples that are hardly related to literary cases of reuse.

and, thus, assumes that under the correct application of the guidelines inter-annotator agreement can be reached.

In contrast to applications in Machine Learning, where the formulation of these guidelines is certainly feasible, intertextuality research seeking to establish reference values of annotator agreement faces a significant problem: definitions of what an intertext is and instructions on how to recognize it are matters of interpretation and scholarly debate. To some extent, the objectivity of an intertextual link is even questioned by more radical literary-theoretical takes like reader-response criticism, which posits that intertexts exist inasmuch as observed by the reader (Tompkins, 1980). In light of these difficulties, we resort to inter-annotator agreement indices in order to gauge the utility of text reuse detection applications by means of a quantification of the agreement to which they lead. Thus, although our use case differs from the common setting in Computational Linguistics, and while the interpretation criteria for the indices will differ accordingly, the methodology is nonetheless transferrable.

6.2.1 Chance-corrected Indices

Following (Artstein and Poesio, 2008), the idea that lies behind of a chance-corrected agreement index is to first measure the observed agreement A_o, and then discount how much of that agreement can be expected due to chance agreement A_e. The mathematical formulation is as follows:

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e} \tag{6.1}$$

where the agreement score is normalized over the maximum amount of obtainable agreement left, once the agreement expected due to chance is discounted. For simplicity in the notation, the discussion below specifically considers the case of three annotators, but the given definitions can easily be extended to any given number of annotators.

6.2.1.1 Observed Agreement

The first question concerns the computation of the observed agreement. This can be obtained as the probability that annotators agree in their judgments conditioned on the observed behavior. In the commonly used frequentist terms, the observed agreement is quantified as the proportion of matching judgments over the total number of casted judgments. If we let $\theta_{k,k,k}$ denote the joint probability that all three annotators cast the same judgement k—where k ranges over the possible outcomes K (i. e. in our case favorable and unfavorable judgments)—the observed agreement is given by Equation 6.2:

$$A_o = \sum_k \theta_{k,k,k} \tag{6.2}$$

6.2.1.2 *Chance Agreement*

The next question concerns the computation of the agreement due to chance. This is typically computed on the basis of an estimation of the probability that a random annotator assigns a particular label—or, in our case, provides a favorable or unfavorable judgment about a given candidate pair. The approach taken to model the random annotator differentiates the coefficients mentioned above: S, π and κ.

First, S considers that an annotator operating by chance assigns labels using the maximum entropy principle, assuming a situation of perfect ignorance of the label distribution. In practice, this means that random judgments are modeled as following a uniform probability distribution, as shown in Equation 6.3:

$$P_S(k|c) = P(k) = \frac{1}{K} \tag{6.3}$$

where c refers to the c^{th} annotator.

The next index, π, uses population-level statistics and models the random annotator taking into account the overall number of positive and negative judgments. This index, however, effectively disregards individual differences between annotators, as shown in Equation 6.4:

$$P_\pi(k|c) = P(k) = \frac{\sum_{i=1}^{n} \sum_{c=1}^{|C|} \mathbb{1}(c_i = k)}{n|C|} \tag{6.4}$$

where n is the total number of candidate matches, c_i refers to the judgement made by annotator c on instance i and $\mathbb{1}(x)$ is an indicator function that evaluates to 1 if x is true.

The final index, κ, takes into account both the distribution of labels and individual annotator behavior, and models the random annotator as the overall probability that each annotator produces each judgement.

$$P_\kappa(k|c) = \frac{\sum_{i=1}^{n} \mathbb{1}(c_i = k)}{n} \tag{6.5}$$

In contrast to both S and π, κ is able to capture annotator bias, a reason why this variant is typically preferred over its alternatives.

Given the estimated probabilities modeling chance behavior, the expected agreement can be computed as the joint probability of agreement, assuming independent judgments:

$$A_e = \sum_{k} \prod_{c} P(k|c) \tag{6.6}$$

While Equations 6.3, 6.4 and 6.5 have been formulated in terms of count-based probability estimators, these definitions can be easily generalized to any given probability estimator. Here, we focus on the extension of κ, which will be used for the computations in the remaining of the chapter.

Reusing the notation introduced in Equation 6.6, the estimator for P_κ can be expressed in a more general form using Equation 6.7:

$$P_\kappa(k|c=1) = \theta_{k..} = \sum_{k' \in K} \sum_{k'' \in K} \theta_{k,k',k''} \tag{6.7}$$

This quantity corresponds to the probability of the first annotator producing k, after marginalizing over the behavior of the other two annotators. Using this notation, we can now express κ for three annotators as shown in Equation 6.8:

$$\kappa = \frac{\sum_k \theta_k - \sum_k \theta_{k..}\theta_{.k.}\theta_{..k}}{1 - \sum_k \theta_{k..}\theta_{.k.}\theta_{..k}} \tag{6.8}$$

Equation 6.8 corresponds to a multi-annotator agreement index that generalizes Cohen's κ to multiple annotators. This is in contrast to the index, commonly known as Fleiss κ, which was introduced by Fleiss in (Fleiss, 1971) but that, as Artstein and Poesio (2007) argue, actually corresponds to a generalization of Scott's π. Following Artstein and Poesio (2007), we will refer to the index defined in Equation 6.8 as multi-κ, in order to avoid confusion.

6.2.2 Statistical Modeling for the Computation of κ

When annotators examine the target dataset, a number of factors may have an influence on the outcome decision. For instance, the amount of lexical overlap between the documents paired by a text reuse algorithm may exert an effect on the annotator judgements. As will be discussed in Section 6.3, the present experimental setup includes a number of such factors. For example, even if we assume an equal retrieval performance of two text reuse detection algorithms, the style of reuse that these capture may vary, and, as a result, annotators can find it more or less difficult to agree on pairs suggested by different algorithms.

Some of these factors correspond to co-variates that split the dataset in smaller subsets. Traditional approaches to inter-annotator agreement indices face the problem that the evidence per subset sparsifies and the count-based estimates used by these methods become inefficient. In the case of the underlying retrieval algorithm, indices obtained via maximum likelihood estimation are computed on the subset corresponding to instances retrieved by each algorithm—a strategy known as "no pooling" in the jargon of multi-level statistical modeling. Moreover, other co-variates represent statistical dependencies on agreement that like, for example, the amount of lexical overlap, consist of continuous variables.

Our strategy, here, is to take advantage of the formulation of κ in terms of the joint probability given in Equation 6.8, and leverage statistical models for the estimation of these probabilities. Thus, we dispense with count-based estimates, and focus on sophisticated statistical models to estimate the inter-annotator agreement coefficients. By including any given number of relevant co-variates and grouping factors, we shall not only improve the probability estimates but also assess their influence on the output agreement scores.

6.2.2.1 *Multi-level modeling*

In order to capture the variation arising from co-variates and grouping factors, we turn to multi-level statistical models. These models are known to excel in cases where the modeled data can be clustered according to multiple, possibly hierarchical criteria and imbalanced group sample sizes. Using a strategy known as "partial pooling", multi-level models can leverage population-level information to improve the estimates for the different groups, avoiding large groups from dominating the inferential process, and naturally capturing uncertainty by directly modeling the variance associated with the parameters of each group (Gelman and Hill, 2006; McElreath, 2018).

6.2.2.2 *Bayesian Inference for* κ

Moreover, in this study we turn to Bayesian inference methods in order to fit the multi-level model. Bayesian inference has a number of advantages in this context, as it has superior modeling capacity in multi-level modeling scenarios with reduced number of cases (Gelman and Hill, 2006), and it produces a posterior distribution over model parameters, upon which further computation can be run in order to propagate parameter uncertainty to the agreement coefficients.

In order to illustrate the difference in methods, we can focus on the computation of the observed agreement from Equation 6.2. In a traditional approach to inter-annotator agreement, the computation relies on contingency tables and aims at obtaining point-wise probability estimates and confidence intervals. For three annotators, c, c' and c'', the probability of agreement is computed by counting the number of times that all three casted the same judgment and dividing by the total number of pairs in the dataset:

$$A_o = \sum_k \theta_{k,k,k} = \sum_k \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(c_i = k \land c'_i = k \land c''_i = k) \tag{6.9}$$

In contrast, a Bayesian inference approach treats $\theta_{k,k,k}$ as a random variable, and the goal is to compute a posterior distribution that places a probability on the values over which it ranges. The inference takes advantage of Bayes's theorem, which is illustrated in the following Equation:

$$P(\theta_{k,k,k}|X) = \frac{P(\theta_{k,k,k})L(X|\theta_{k,k,k})}{\sum_{\theta_{k,k,k}} P(\theta_{k,k,k})L(X|\theta_{k,k,k})} \tag{6.10}$$

In Equation 6.10, $P(\theta_{k,k,k})$ corresponds to a prior probability distribution over $\theta_{k,k,k}$—other parameters on which we wish to run inference receive similarly a prior distribution. Prior distributions can be set in order to reflect information about these parameters already available before conducting the experiment. Next, $L(X|\theta_{k,k,k})$ defines a likelihood model that computes the probability of the observed data X for a given value of the parameter. The likelihood defines the statistical model and can be used to incorporate additional variables—e. g. predictors and grouping factors—deemed relevant to the experimental design. Finally, as a result of the inference, we obtain a posterior distribution, which does not only capture a point estimate, but also the uncertainty arising from the inferential process.

6.2.2.3 Multinomial Hierarchical Model

In order to compute multi-κ for three annotators and two outcome variables (i. e. favorable and unfavorable judgments), we need to estimate 8 quantities, each of which expresses the probability of each of the 2^3 possible combinations of annotator and outcome. The approach we chose to model these quantities is to use a multinomial likelihood, with a linear model for each of the last 7 outcomes, using the remaining one as the pivot or reference value.

If we encode each possible outcome with a number from 0 to 7, as shown in Table 6.1, and take outcome 0 as the pivot value (i. e. the reference value), then the statistical model is given by Equations (6.11) to (6.13).

A1	A2	A3	code
0	0	0	0
0	0	1	1
0	1	0	2
0	1	1	3
1	0	0	4
1	0	1	5
1	1	0	6
1	1	1	7

Table 6.1: Binary mapping translating response outcomes into single numbers.

$$
\begin{aligned}
\log\left(\frac{\theta_1}{\theta_0}\right) &= \alpha_1 + \beta_{1p} \cdot X_p + \nu_{1q} \\
\log\left(\frac{\theta_2}{\theta_0}\right) &= \alpha_2 + \beta_{2p} \cdot X_p + \nu_{2q} \\
&\vdots \\
\log\left(\frac{\theta_7}{\theta_0}\right) &= \alpha_7 + \beta_{7p} \cdot X_p + \nu_{7q}
\end{aligned} \tag{6.11}
$$

$$
\theta_0 + \theta_1 + \ldots + \theta_7 = 1 \tag{6.12}
$$

$$
\begin{bmatrix} \nu_{1q} \\ \vdots \\ \nu_{7q} \end{bmatrix} \sim \text{MVNormal}(0, \Sigma_q) : \Sigma_q = \begin{bmatrix} \sigma^2_{1q} & & \\ \vdots & \ddots & \\ \sigma_{71q} & \ldots & \sigma^2_{7q} \end{bmatrix} \tag{6.13}
$$

More specifically, Equation 6.11 shows the log-odds of the responses 1 to 7 with respect to the pivot. Each log-odds are computed by a multi-level linear model where α_k refers to the fixed intercepts for the k^{th} response, β_{kp} to the fixed coefficient corresponding to the p^{th} independent variable X_p and ν_{kq} to the q^{th}-level random intercept, which captures within-group variation for the corresponding grouping factor.

As shown in Equation 6.13, these group-level random intercepts are modeled jointly, coming from a multi-variate normal centered around a zero-mean with a variance-covariance matrix Σ_q. The variance-covariance matrix is, in practice, decomposed into a diagonal variance matrix and a correlation matrix. The inferred models, thus, contain posterior distributions of the group-level correlations between random intercepts across linear models—this resembles the setup introduced by Koster and McElreath (2017).

In order to turn the log-odds into actual probabilities we employ the softmax function, shown in Equation 6.12, which is based on θ_0 being picked as the pivot:

$$\begin{aligned}\theta_0 &= 1 - \sum_{k=1}^{7} \theta_0 \cdot e^{\alpha_k + \beta_{kp} \cdot X_p + \nu_{kq}} \\ \implies \theta_0 &= \frac{1}{1 + \sum_{k=1}^{7} e^{\alpha_k + \beta_{kp} \cdot X_p + \nu_{kq}}} \\ \implies \theta_k &= \frac{e^{\alpha_k + \beta_{kp} \cdot X_p + \nu_{kq}}}{1 + \sum_{k'=1}^{K} e^{\alpha_{k'} + \beta_{k'p} \cdot X_p + \nu_{k'q}}} \end{aligned} \tag{6.14}$$

As we can see, the probability of the reference model (θ_0, in this case), can be computed as the remaining probability after subtracting the probabilities of the other 7 models.

6.3 EXPERIMENTAL SETUP

6.3.1 Experimental Design

As mention in Section 6.1, we focused on inter-annotator agreement in a late-stage phase of the editing process. Our experiment involved a total of three expert editors of Bernard,[2] who were shown candidate matches retrieved by the algorithms. The guideline provided to the participants was limited to whether the annotator would consider a candidate match worth being included in a prospective edition. The matches were extracted from the digital edition of Bernard of Clairvaux's sermons (Bernard of Clairvaux, 1998, 2006), introduced in Section 2.6.1.

At this stage, the written sources have already been analyzed for biblical intertexts—in the case of Bernard's sermons, previous iterations used fully manual annotation—and the algorithms must be specifically fine-tuned to identify any possible remaining cases of reuse.[3] More concretely, we exploited the existing resources in the following manner. First, the existing annotated instances were treated as gold annotations, on which we fine-tuned two competing text reuse detection algorithms—these will be described in Section 6.3.2. We, then, filtered out the positive cases already annotated and applied the fine-tuned algorithms to extract more relevant candidates. From the resulting sets of each algorithm, the 300 most likely pairs were

[2] The annotators were Jacqueline Picard, Yasmine Ech Chael and Laurence Mellerin, from the BiblIndex project. The biblical analysis was prepared by Jean Figuet, Marie-Imelda Huille and Laurence Mellerin.

[3] We note that this setting not only offers a realistic scenario, but also increases the informativeness of the experiment, since it forces annotators to focus on more involved passages.

sampled for the inter-annotator experiment according to the similarity scores.

Annotators were asked to produce binary relevance judgments, indicating whether the considered candidate pair constitutes a reference that should be included in the edition of Bernard. During the design of the experiment, we discarded continuous or gradual relevance scales, such as those common in semantic evaluation tasks Semantic Textual Similarity (Agirre et al., 2012), or classifications of reuse into different types, ranging from literal quotations to allusions—see Büchler (2013, p. 77), Hohl Trillini and Quassdorf (2010) and Bamman and Crane (2008) for examples of such categorizations. First, gradual scales complexify the computation and analysis of agreement scores, requiring the modeler to take into account weights in the disagreements between annotators. Moreover, they can be misused by annotators by allocating difficult instances to a mid-level range in the scale. Similarly, non-binary judgments require extensive annotator training that in our case can be considered tangential to the goals of the study. Instead, a binary setting forces annotators to make clear-cut decisions.

6.3.2 Text Reuse Algorithms

Two algorithms were considered for the retrieval of intertextual references: one based on the GVSM paradigm that has recently re-emerged under the name of soft cosine, and the Smith-Waterman algorithm, which is based on the text alignment paradigm. We refer to Sections 2.5.2.2 and 2.5.2.3 for a description of the algorithms, and only note here that the word embedding space discussed in Appendix A.1 was employed for estimating the word-similarity matrix needed by the soft cosine algorithm.

6.3.3 Candidate Selection

The input collection, i. e. Bernard's sermons, was segmented into documents using a sliding window of 20 tokens with an overlap of 10 tokens, which resulted in a total of 19,987 documents. The target collection is the Vulgate Bible available in digital form from the Perseus repository (Crane, 1996). For the Vulgate, we follow the traditional segmentation into verses, which amounts to 36,663 documents. As mentioned in Section 2.6.1, both collections were lemmatized using the neural lemmatizer introduced in Chapter 3, and the retrieval algorithms were fed lemmata instead of the corresponding inflected forms. We fine-tuned both algorithms—i. e. hyper-parameters and similarity threshold values—to optimize retrieval according to the F-measure, equally balancing precision and recall. Similarly to other text alignment algorithms, the complexity of `Smith-Waterman` is quadratic, and in order to be applied to large collections, this algorithm commonly

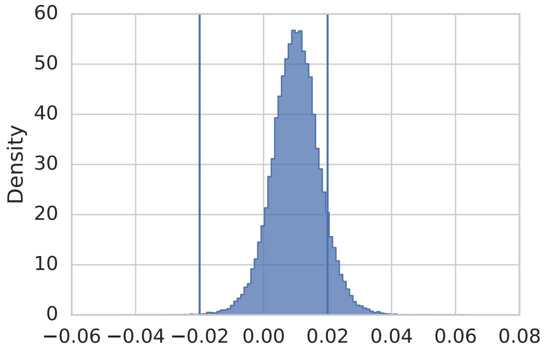


Figure 6.1: Cross-validated performance comparison between competing algorithms `Smith-Waterman` and `Soft-Cosine` using the Bayesian correlated t-test with a ROPE of 0.02 over AP scores. The distribution of posterior draws of performance differences over folds indicates a probability of 0.91 that the algorithms perform equally well for the chosen ROPE.

requires a pre-filtering step by a more efficient algorithm optimized for recall. For this purpose, we deployed the VSM-based `Soft-Cosine` using a threshold of 0.19 similarity. Candidate pairs below this threshold were discarded from comparison by `Smith-Waterman`, which implies a reduction of 99.74% in the number of required comparisons (totaling above 700 million).

6.3.3.1 *Retrieval Model Comparison*

It is noteworthy that these algorithms tend to be sensitive to different aspects of intertextuality. Specifically, the `Soft-Cosine` algorithm tends to give more weight to semantic relations, while the `Smith-Waterman` scores pairs more highly when the similarity derives from large sequences of shared word tokens—favoring, thus, cases of literal reuse. While this difference suggest interpretations of results in terms of the relevance of reuse styles in the target texts, a further reason to select these algorithms for comparison is that they perform strongly, and their precision-recall trade-off on this dataset is comparable.

This latter point was assessed empirically in a cross-validated comparison using the correlated t-test introduced by Corani and Benavoli (2015). As discussed in Section 2.6.3.1, this t-test infers a posterior distribution of performance differences over the CV folds—we use 10 folds—that can be used to answer the question of performance superiority in easy-to-interpret probabilistic terms. The comparison procedure involves defining a Region Of Practical Equivalence (ROPE), that specifies a performance difference lower-bound below which the competing algorithms can be considered equal (Benavoli et al., 2017).

We used the `baycomp` software package to estimate the posterior distribution of performance differences (Benavoli et al., 2017), selecting a

	Matches		Sermons	Matches/Sermons
	Known	Unknown		
Soft-Cosine	296	72 = 24.3%	24	12.33 (± 10.66)
Smith-Waterman	292	56 = 19.2%	22	13.27 (± 18.5)
Shared	74		22	

Table 6.2: Summary statistics of the annotation dataset, displaying the number of matches per method—including the number of matches unknown to have previous annotations—, the number of sermons involved, and the mean—and standard deviation—of matches per sermon. Finally, the row column shows the total number of matches that were retrieved by both methods.

ROPE of 0.02 points in AP. Figure 6.1 displays the posterior distribution of differences in AP obtained by the `Smith-Waterman` and `Soft-Cosine` algorithms.[4] Vertical lines highlight the chosen ROPE of 0.02. As we can see, while the mode is slightly shifted towards the right—indicating a slight preference towards `Smith-Waterman`—most of the distribution falls within the ROPE, and the estimated probability of the hypothesis that both methods are equivalent at this ROPE corresponds to 0.91.

6.3.4 Dataset

As mentioned in Section 6.3.1, our dataset consists of 300 document pairs per method. Table 6.2 provides statistics on the dataset. As we can see, the instances selected per method display a mild overlap. Moreover, the total number of instances per method can be further differentiated with respect to whether the matches underlie no previously known match or a previously known match to a different biblical reference. Cases of known matches pointing to a different biblical verse are typically due to the known fact of intra-biblical intertextuality, such as textual echoes of the Old Testament within different books of the New Testament, and parallel accounts in the synoptical Gospels.

In the case of intra-biblical intertextuality, a passage from Bernard referring to a biblical verse can simultaneously point at other biblical verses which are referred to by the first biblical verse. Deciding which biblical verse is the actual target of the reference poses additional interpretational problems, which—as we shall see in Section 6.4.3—complicates the annotation of resources.

[4] We take `Smith-Waterman` as the reference algorithm.

6.4 RESULTS

We now describe the results of the experiment. First, in Section 6.4.1 the statistical models are detailed, specifying the fitting procedure and model evaluation. Then, in Section 6.4.2, we report the inter-annotator agreement scores obtained by manipulating the posterior distributions of the fitted models, and explore the effects of the dependent variables and grouping factors on the obtained agreement estimates. Finally, in Section 6.4.3, we provide the results of the qualitative post-experimental report.

6.4.1 Statistical Inference

6.4.1.1 Model Fitting

The specification and estimation of multinomial multi-level models is a challenging procedure that is currently not well supported in commonly used statistical packages such as `lme4`. Recently, packages such as `Rstan` (Carpenter et al., 2017) have made available the option to estimate the Bayesian variants of these models relying on efficient Hamiltonian Monte Carlo sampling methods. In the present study, we utilize the `brms` library (Bürkner, 2018). For the present experiments, we ran 4 chains with 2000 iterations per chain using the first 1000 as warmup, and monitored the convergence diagnostics mentioned in Section 5.4. Finally, we choose weakly informative priors—as specified by the defaults of the `brms` package—to avoid exploring highly unlikely regions of the parameter space.

6.4.1.2 Statistical Model Comparison

We present several models, considering various combinations of dependent variables and grouping factors for the varying intercepts. For each model, we compute the Widely Applicable Information Criterion (WAIC). This measure allows for a comparison of model plausibility in terms of both predictive performance and model complexity or overfitting, and let us weigh relative model capacity according to the multiple-model framework—see Symonds and Moussalli (2011) and McElreath (2018). We consider a total of 5 models of increasing complexity, and seek to establish the relevance of the information taken into account by the different models through model comparison.

- The first model, `m`, is a baseline model that has a single intercept and adds neither predictors nor grouping factors.
- The second model, `m.m`, seeks to capture the influence of the retrieval method and, thus, uses the underlying method type as binary predictor.

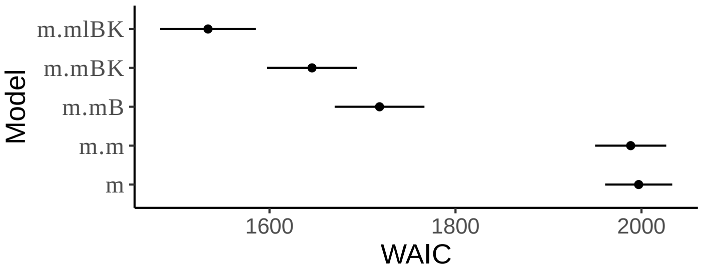


Figure 6.2: Visualization of the resulting WAIC scores and standard error bars.

- The third model, `m.mB` adds varying intercepts corresponding to the biblical book to which the candidate reference belongs. These random intercepts help modeling whether references to particular biblical books tend to be more or less controversial.

- The fourth model, `m.mBK`, includes additional varying intercepts on the "familiarity" with the Bernardine passage. This factor corresponds to whether the Bernardine fragment was known to have a reference—albeit to a different biblical verse—and allows us to estimate agreement in cases where annotator re-assess the actual reference of a given passage.

- Finally, model `m.mlBK` adds a fixed predictor accounting for lexical overlap. We chose to compute lexical overlap using the weighted Jaccard similarity, which is shown in Equation 6.15.

$$J(D_i, D_j) = \sum_{w \in D_i \cup D_j} \frac{min[c(w, D_i), c(w, D_j)]}{max[c(w, D_i), c(w, D_j)]} \tag{6.15}$$

where D_i refers to the i^{th} document, and $c(w, D_i)$ refers to the count of word w in document D_i.

As it is customary, we rescale the predictor variable to be centered around a zero-mean and unit standard deviation. With the inclusion of lexical overlap as a predictor, we seek to explore whether more literal quotations tend to be more or less controversial.

Table 6.3 displays the results of the WAIC-based model comparison. For ease of visualization, Figure 6.2 plots the WAIC estimates together with standard error bars. Based on the Akaike weight—interpretable as the probability that a model will produce superior predictions on unseen data, we can conclude that the model including all predictors and random effects can be identified as the model with superior predictions. As we can see in Figure 6.2, error bars do not overlap with the second best model. We can thus base any further inference on this model.

	Model	WAIC (SE)	P	WAICΔ(SE)	Weight
1	m.mlBK	1533.88 (51.44)	120.72		1.00
2	m.mBK	1645.71 (48.28)	116.35	111.83 (22.07)	0.00
3	m.mB	1718.35 (48.28)	110.84	184.47 (26.20)	0.00
4	m.m	1988.38 (38.25)	14.74	454.50 (39.65)	0.00
5	m	1996.99 (36.09)	7.21	463.12 (40.86)	0.00

Table 6.3: Evaluation of the statistical models in terms of WAIC. First column shows the absolute WAIC with Standard Errors (SE)—lower WAIC indicates a better model fit. The second column shows an estimate of the effective number of parameters. The third column displays the absolute difference in WAIC with respect to the best model. The last column shows the Akaike weight.

6.4.2 Inter-annotator Agreement

Through conditioning on different values of the dependent variables or different levels of the grouping factors, we can account for and tease apart the effects of those components on the obtained agreement scores. We, thus, present the results discriminating between combinations of the underlying method (independent variable) and familiarity with the borrowing passage—i. e. whether the passage is known to contain references to other biblical verses—(grouping factor).

6.4.2.1 Effects of Retrieval Method on Agreement

Figure 6.3 shows the posterior distributions obtained for the κ scores, computed using the definition of Equation 6.8. The plot on the left-hand side of Figure 6.3 shows the resulting scores obtained for references to an "average" biblical book. These estimates, thus, ignore the variability arising from the fact that references to particular books may result in more or less inter-annotator agreement. The plot on the right-hand, however, includes this variability through marginalization. Technically, the marginalization procedure is accomplished by sampling ν_{kq} from the inferred multi-variate normal distribution of the target random effect. For each Markov Chain Monte Carlo (MCMC) draw of parameters, we add the sampled ν_{kq} value, before computing the output softmax. For the case of books, this marginalization results in a posterior that corresponds to the agreement that we could expect for a reference to any (possibly unobserved) given book. For documentation purposes, detailed point estimates of the agreement scores, including the median as well as the 95% quantiles, are shown in Table 6.4.

As we can see, the agreement is decisively higher for pairs with a known reference. Using the posterior distributions, the probability of agreement being higher for pairs with a known reference can be

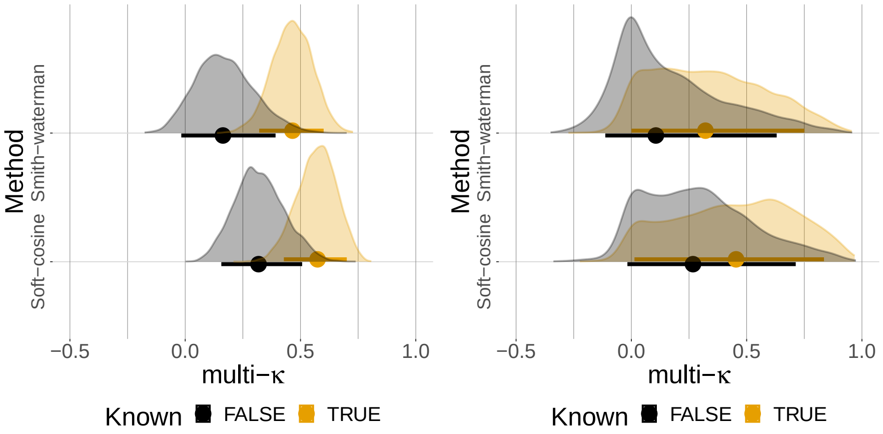


Figure 6.3: Posterior multi-κ scores inferred from the full model (`m.mlBK`), displayed according to underlying method—on the y-axis—and whether the candidate borrowing passage is known by the editors to contain a reference to a different biblical verse—on the x-axis. Word overlap is kept to the zero-centered mean value. Left plot and right plot differ on whether the variation arising from the source biblical book is excluded or not. The mean estimate is shown by a point and the 0.89 credible interval is shown by the surrounding horizontal bar.

easily computed using a bootstrap. Specifically, we sample 10,000 draws from the posterior and count the proportion of times that the agreement is higher under one condition than under the other. In this case, the bootstrapped probability amounts to 0.96 in favor of the agreement being higher for known references. Analogously, agreement scores are likely to be higher for `Soft-Cosine`, with a probability of 0.82 for unseen candidate pairs and 0.81 for pairs containing a known reference.

However, when the variability stemming from books is taken into account through marginalization, we obtain very wide posterior distributions, as shown by the right plot, and the probabilities of the differences between agreement scores decrease.[5] This is a strong indication of the importance of the target reference book for annotator behavior and supplements the evidence from the WAIC comparisons in Section 6.4.1.2, where including book-level varying intercepts resulted in a large WAIC reduction of 270.03 points—model `m.mB` vs. model `m.m`—, corresponding to a 59.4% WAIC reduction with respect to the total WAIC reduction of the best model—model `m.mlBK` vs. model `m.m`.

[5] The probability for the agreement in known cases being higher is now 0.65 for `Soft-Cosine` and 0.69 for `Smith-Waterman`, while the probability for the agreement in `Soft-Cosine` being higher is now 0.44 for known cases and 0.61 for unknown.

		-Book			+Book		
Method	Known	Lower	κ	Upper	Lower	κ	Upper
`Soft-Cosine`	FALSE	0.13	0.32	0.54	-0.07	0.24	0.81
	TRUE	0.39	0.57	0.72	-0.01	0.41	0.88
`Smith-Waterman`	FALSE	-0.05	0.16	0.44	-0.18	0.08	0.74
	TRUE	0.28	0.47	0.63	-0.04	0.29	0.82

Table 6.4: Median, lower and upper 95% quantiles for posterior agreement scores obtained with the full model (`m.mlBK`), while keeping similarity at the mean value.

6.4.2.2 *Effects of Similarity on Agreement*

While in Section 6.4.2.1 we controlled for lexical overlap by keeping it at the mean, we now focus on the dependency relation between lexical overlap and agreement scores. Typically, effects of predictors in Bayesian linear models can be assessed by directly inspecting the posterior distribution of the coefficient. In the present case, however, the dependency of the agreement score on the predictor is not directly modeled. Instead, the agreement score is obtained on the basis of posterior estimates of different linear models and the dependency is directly modeled on the individual outcomes of the linear models.

In order to visualize these dependencies we resort to counterfactual plots. Counterfactual plots allow for the inspection of direct and indirect statistical dependencies through the visualization of model predictions obtained when modifying an independent variable across its entire range—i. e. including values for which no observation is attested in the original dataset (McElreath, 2018).

Figure 6.4 visualizes the direct dependency of lexical overlap on the individual outcomes of the multinomial model. As we can see, the effect of lexical overlap on the two outcomes that signify three-way agreement—i. e. `000` and `111`—may suggest a linear positive relationship in which an increased lexical overlap is associated with a higher probability of three favorable judgments and lower probability of three unfavorable judgments.

However, this analysis does not take into account agreement due to chance, and, thus, we continue by inspecting the indirect dependency of lexical overlap on agreement. Figure 6.5 shows counterfactual plots of the posterior κ scores over the entire range of observed lexical overlap. Similar to Figure 6.3, results are split depending on whether we take book-level variability into account (bottom plots) or not (top plots).

The plots on top, ignoring book-level variability, indicate that an increase in lexical overlap is linked with an increase in agreement, but only up to a cutoff point after which the relationship flips over.

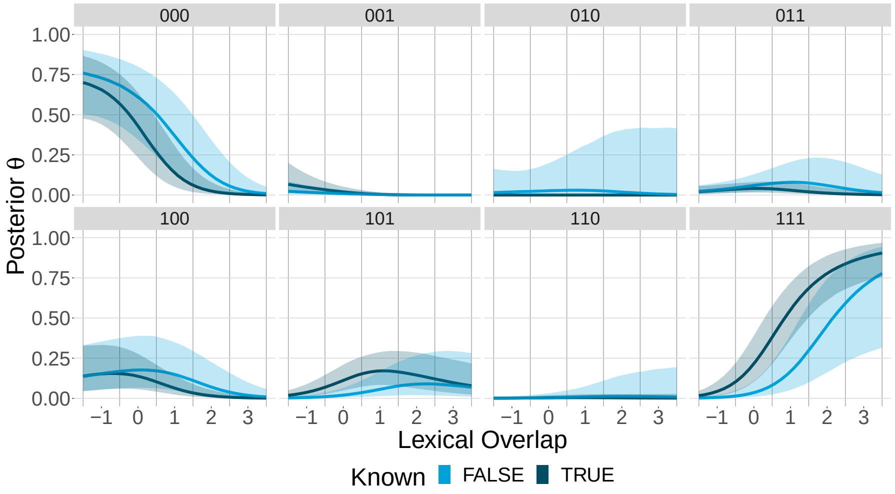


Figure 6.4: Posterior estimates of the individual multinomial outcomes—on the y-axis—over the predictor on the entire range of observed lexical overlap—on the x-axis. The coding for the outcomes corresponds to the one mapped in Table 6.1. The estimates are separated by color conditioned on whether the candidate borrowing passage is known to contain a reference to a different biblical verse. The mean effect is represented by a line and a shaded area around the line visualizes the 0.89 credible interval.

This effect is strongest for known cases, for which the existence of an effect—i. e. the slope of the trend being non-zero—is attested with a high probability. For unknown cases, however, the effect is more uncertain, since 95% credible intervals possibly include straight lines. This is expected, since the number of unknown cases is lower. The cutoff point seems to be just above mean similarity for known cases and a bit higher for unknown cases—although the latter needs to be nuanced by the large credible intervals mentioned before.

Similarly to the results discussed in Section 6.4.2.1, incorporating uncertainty about the books through marginalization results in much wider posteriors, indicating that straight horizontal lines are included within the 50% credible intervals for the agreement on instances with unknown biblical references. For instances with known references, however, high overlap is very certainly associated with low agreement.

6.4.3 Post-experimental Report

In order to kickstart a qualitative discussion of disagreement-related issues, we extracted a set of document pairs in which one of the annotators systematically disagreed with the other two, and asked her to elucidate the reasons for the disagreement. The annotator in charge of the discussion was the one with the highest level of familiarity with Bernard. The subset consists of instances retrieved with the `Smith-Waterman` algorithm and has, thus, a slight bias towards literal

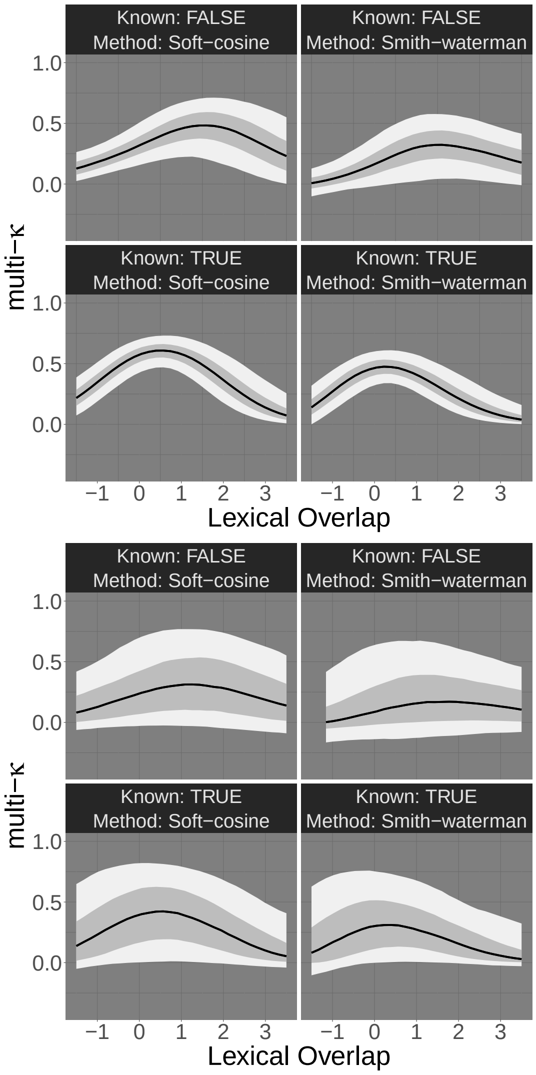


Figure 6.5: Posterior multi-κ scores over lexical overlap. Lexical overlap is scaled such that a unit on the x-axis indicates a standard deviation away from the zero-mean. Top and bottom plots differ respectively on whether the variation coming from the books is excluded or not. Black lines indicate median κ scores with 0.5 and 0.89 credible intervals shown by shaded grey areas.

quotations. The main raised points respond to the following three aspects, illustrated in the examples shown in Table 6.5.

Table 6.5: Examples from the inter-annotator agreement dataset, showcasing different types of agreement problems. The Bernardine chunk on the left is accompanied by the retrieved candidate in the center and an alternative verse proposed by the annotator during the post-experimental report in the right. Words in bold correspond to lexical overlap with the biblical references, while words in italics indicate a relevant fragment left out by the applied segmentation. Biblical references contain hyper-links re-directing to Perseus online version that includes English translations.

Bernardine Chunk	Proposed Verse	Alternative Verse
vestra, et in exitu vestro de lacu miseriae et de luto faecis, **cantastis et ipsi Domino canticum novum** ***quia mirabilia facit*** *S. 1, 9 (SC 414, p. 72)*	quando domus aedificabatur post captivitatem canticum huic David **cantate Domino canticum novum** cantate Domino omnis terra *Psalms, 95:1*	psalmus David **cantate Domino canticum novum quoniam mirabilia fecit** salvavit sibi dextera eius et brachium sanctum eius *Psalms, 97:1*
carnale matrimonium constituit **duos in carne una**, cur non magis spiritualis copula duos coniunget in uno spiritu? Denique *S. 8, 9 (SC 414, p. 192)*	et erunt **duo in carne una** itaque iam non sunt duo sed una caro *Mark, 10:8*	quam ob rem relinquet homo patrem suum et matrem et adherebit uxori suae et erunt **duo in carne una** *Genesis, 2:24*
blanditiis, seduci fallaciis, nec iniuriis frangi, **toto corde, tota anima, tota virtute diligere** est. *S. 20, 5 (SC 431, p. 136)*	ille respondens dixit diliges Dominum Deum tuum ex **toto corde tuo et ex tota anima** tua et ex omnibus viribus tuis et ex omni mente tua et proximum tuum sicut te ipsum *Luke, 10:27*	et diliges Dominum Deum tuum ex **toto corde** tuo et ex **tota anima** tua et ex tota mente tua et ex **tota virtute** tua hoc est primum mandatum *Mark, 12:30*

Continued on next page

Table 6.5 – continued from previous page

Bernardine Chunk	Proposed Verse	Alternative Verse
cognoscentur. Hinc rursus Pater ad Filium: **Sede**, inquit, **a dextris meis, donec ponam inimicos tuos** ***scabellum pedum tuorum*** *S. 6, 5 (SC 414, p. 144)*	ad quem autem angelorum dixit aliquando **sede a dextris meis quoadusque ponam inimicos tuos scabillum pedum tuorum** *Hebrews, 1:13*	david canticum dixit Dominus Domino meo **sede a dextris meis donec ponam inimicos tuos scabillum pedum tuorum** *Psalms, 109:1*

6.4.3.1 Segmentation Related Problems

The first encountered issue relates to ambiguity problems arising from the segmentation approach employed in order to break down Bernard's sermons into passages. As discussed in Section 6.3.3, we applied a rather arbitrary segmentation approach to Bernard's sermons, using a sliding window of 20 words with an overlap of 10 words. This strategy resulted in a number of difficult candidate pairs in which the annotators have to decide subjectively whether to validate a candidate pair in the presence of fuzzy segmentation. These problems have a significant incidence on the annotator disagreements and generate a lack of consistency, even by the same annotator.

For example, the first instance in Table 6.5 corresponds to an example of such a problem. The words "quia mirabilia facit" (en. becasue [he] made miracles) have been left out by the applied segmentation. Without these words, two annotators were inclined to accept *Psalms, 95:1*, a verse with which the overlap is high. The dissident annotator, however, rejected it under the assumption that the fitting reference was instead *Psalms, 97:1*, for which the missing words provide stronger evidence. As we can see, these segmentation-related issues already point towards a second difficulty, which consists in the biblical knowledge required for the interpretation of the intertextual references.

6.4.3.2 Knowledge of the Bible

In addition to the problem mentioned above, when dealing with biblical texts it is important to take intra-biblical intertextuality into account. As mentioned in Section 6.3.4, texts from the Old Testament are quoted in the New Testament, and the synoptical Gospels are known to contain parallel accounts of the same events. Disagreements can appear when annotators diverge with respect to which of the parallel variants they consider to be the actual source of the biblical reference.

The second example in Table 6.5 refers to a general idea that first appears in *Genesis, 2:24*, which is the unity of man and woman becoming one flesh through marriage. Two annotators, however, validated the suggested reference to *Mark, 10:8*, even though in the typical Bernardine style, the reference is most likely to allude to the original passage, rather than a direct quote of the passage in the Gospel. This example already suggests a third source of disagreement, which corresponds to the familiarity with the referential practices of the borrowing author.

6.4.3.3 Knowledge of Bernard

Finally, annotators must combine their knowledge of the Bible with other abilities regarding the borrowing author. In the case of patristic literature, authors can be observed to hold a general preference towards specific biblical passages. For example, a biblical passage has a higher probability of being quoted by an author if he uses it in daily prayers. Moreover, an author of exegetical commentaries of a biblical book may quote this book more often than others. The last two examples highlight this source of disagreement.

In the third example in Table 6.5, the Bernardine chunk lies in a context at the end of a paragraph in which the main points of a previous argumentation are being summarized. In that argumentation, *Mark, 12:30* has been referenced explicitly and in the current location it is being referred to implicitly. *Luke, 10:27*, however, is a more closely related match in terms of lexical overlap, which may lead annotators with more superficial knowledge of Bernard's oeuvre to select it.

In the last example in Table 6.5, Bernard refers to a passage that appears both in a Psalm and in the *Letter to the Hebrews*, in which the Psalm is, in turn, referenced. An expert annotator of Bernard can identify that the introduction formula contains a decisive clue: Bernard puts these words in the Father's mouth addressing to Son ("Pater ad Filium", en. the father to the son). Moreover, in the context surrounding this passage, Psalms are being repeatedly referenced, as evidenced by the usage of the word "psalmist" (not shown in the example).

6.5 DISCUSSION

Our study has shown how to apply Bayesian statistical methods to the computation of inter-annotator agreement indices. On the basis of a multi-level model, we were able to isolate the influence of co-variates—such as the underlying retrieval method or the effect of lexical overlap—on agreement. We showed that candidates retrieved by `Soft-Cosine` produce slightly higher agreement scores with a probability of at least 0.8 than instances retrieved by `Smith-Waterman`. Two aspects, however, nuanced that result.

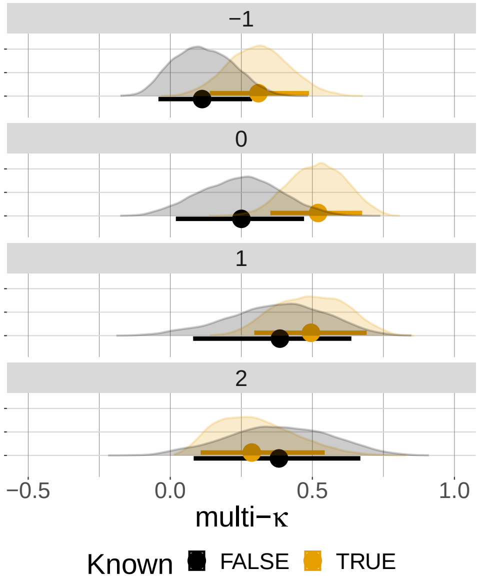


Figure 6.6: Posterior multi-κ densities with respect to familiarity of the Bernardine passage. Densities are computed at different lexical overlap values ranging from -1 to 2 standard deviations.

First, while consistent, the effect of method on the output agreement is smaller than the effect of a Bernardine passage being already known to contain a reference to a different biblical verse. In particular, unknown candidates turned out to be more controversial. This is congruent with the experimental setup—a late-stage retrieval phase—in which few unknown intertextual links are to be expected. However, the statistical dependency is slightly more subtle. As shown in Figure 6.6, an interaction between lexical overlap and familiarity can be observed. While at lower levels of lexical overlap agreement is higher for cases with known references, the difference is reduced at higher levels, and it flips sign at more than 2 standard deviations from the average lexical overlap.

Second, there was a large variability stemming from the biblical book that is targeted by the reference. In order to inspect this matter, Figure 6.7 shows the multi-κ densities per book. While no overall pattern can be observed, a number of illustrative points are worth highlighting. For instance, a rather counter-intuitive result is that the book of the *Song of Songs* appears as the most controversial, even though this book is the most frequently referenced book by Bernard—in fact Bernard's Sermons revolve around it, a fact known to all annotators. This can be interpreted as another source of annotation bias corresponding to diverging editorial purposes. Strictly speaking, many of the suggested matches involving the book of *Song of Songs* can be considered true intertextual references. However, annotators may disagree as to whether an exhaustive underlining of these references

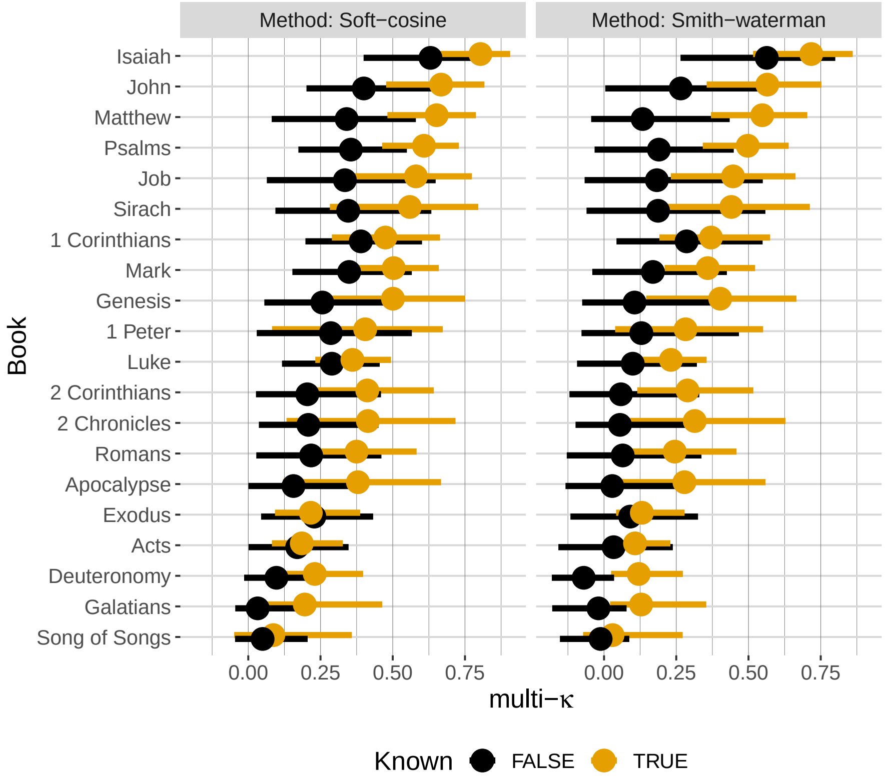


Figure 6.7: Posterior multi-κ densities per book. Lexical overlap is kept to the centered mean value of zero.

may be beneficial to readers, since the referencing can be repetitive and obvious.

Third, our analysis of the effect of lexical overlap on agreement unveiled an interesting non-monotonic effect by which higher lexical overlap is correlated with higher agreement up until an above-average value. The decrease in agreement for lexical overlap higher than mentioned tipping point may indicate that overly literal quotations can be controversial.

Finally, even though the primary goal of our inter-annotator experiment was not to assess the reliability of the annotation process, our results show that the agreement can reach as high a score as 0.5 with high probabilities—see Figure 6.3. This contributes a first assessment of inter-annotator agreement in the field of intertextuality more generally, and, more specifically in computational approaches to Literary Text Reuse Detection.

6.5.1 Limitations

In terms of modeling choices, it must be noted that the current approach is certainly limited to a small number of annotators. Since the multinomial likelihood computes $2^n - 1$ linear models for n annotators, a large number of annotators may render our approach unfeasible. Future work should investigate alternative likelihood formulations that scale better with the number of annotators. Moreover, our current formulation gives equal prior treatment to all annotators, even though, as shown in our experiment, varying levels of background knowledge can influence the resulting agreement. In the future, statistical formulations of agreement on the basis of individual annotator judgments may be able to incorporate this imbalance using differentiated priors per annotator.

Furthermore, our approach to segmentation introduced a set of additional hurdles that may be partially responsible for the observed disagreement. Even though the chance-corrected agreement index that we have implemented takes into account individual annotator biases, overlapping segmentation may result in candidate pairs with high degrees of ambiguity that prevents annotators from a consequential treatment. In the post-experiment report, we were able to confirm this with a series of examples. In the future, two strategies can be used to improve segmentation. The first one consists in applying linguistically informed segmentation targeting naturally occurring punctuation or deploying statistical sentence boundary detection algorithms. The second one consists in deploying a post-retrieval merging step that combines contiguous passages if doing so results in increased evidence for the source of the reference. In any case, the example of faulty segmentation indicates that experimental choices can have an impact on the obtained inter-annotator agreement.

6.6 CONCLUSION & FUTURE WORK

Our work highlights the importance of Bayesian modeling for inter-annotator agreement studies. On the one hand, Bayesian inference allows us to incorporate a number of sources of variation in the computation of inter-annotator agreement, with the goal of improving the estimates and enabling statistical analysis. On the other hand, partial pooling strategies from multi-level modeling enable us to obtain estimates of the agreement scores, even in the presence of few data points. This was the case when computing agreement for previously unknown references. A frequentist approach to agreement would have not obtained robust estimates considering that the number of data points amounted to a few dozens. Finally, sampling-based Bayesian inference offers posterior distributions that can be easily manipulated to obtain distributions of desired inter-annotator agreement, and to answer statistical questions in easy-to-interpret probabilistic terms.

Our study focused on a specific corpus of an author with particular reuse patterns and styles. Conclusions drawn on topics such as the relative utility of different algorithms must be interpreted against the background that the type of reuse found in other corpora would surely challenge the outcome. In this respect, future work can be expected to profit from retrieval approaches that can model statistical patterns of reuse in different corpora. In particular, Machine Learning approaches may be able to overcome the limitations of current approaches, which operationalize reuse in terms of direct similarity and diverge only in the amount of semantic information that is taken into account or the weight given to lexical overlap. From that point of view, it would be interesting to quantify to what extent an end-to-end Machine Learning system achieving comparable performance is able to produce higher agreement in the annotators, and ultimately improve the utility of computer-assisted editing.

7 CONCLUSION

In this last chapter, our goal is double. First, in Section 7.1, we zoom out from the close-up perspective of the previous chapters and synthesize key take-away lessons that can be extracted from a cross-sectional analysis of the obtained results. Second, in Section 6.6, we contemplate what lines of research future efforts in computational intertextuality may explore and what kind of retrieval systems it may deliver.

With respect to the latter goal, our considerations have a double nature. On the one hand, we try to *guess* what type of improvements future research can aim to achieve in relation to current developments and promising approaches in the fields of IR and NLP. On the other hand, our considerations constitute a *suggestion* that looks beyond the immediately apparent future research and points at less evident paths. We take into account not only the point of view of the application developer but also that of the literary scholar. In doing so, we draw from the experience behind the present research and discuss retrieval system design choices that can significantly contribute to the enhancement of the experience of literary scholars.

7.1 SYNOPSIS

For presentation purposes, we have structured the conclusions drawn from the present research along a series of key points, which we proceed to discuss in the following sections.

7.1.1 No Model Fits All

Our work contributed to the understanding of the current state of retrieval methods in Literary Text Reuse Detection by conducting an exhaustive set of model evaluation and comparison experiments, involving several corpora and evaluation measures. As discussed, this is in contrast to previous research, which has arguably overlooked the value of baseline comparison and has, instead, focused on deploying personalized ad-hoc retrieval systems, which are typically not evaluated on benchmark corpora but on the specific corpora of interest. As we argued, baselines are not only necessary for assessing progress in the field—since they serve as a test-bed against which new approaches can be compared—but, being typically transparent and highly interpretable methods, they also represent a valuable tool in order to understand the difficulties that a particular task entails. Moreover, our evaluation procedure involving Cross Validation (CV) not only allowed us to estimate the expected performance of calibrated systems on future datasets, but also showed that, even in the absence of statistical learning, hyper-parameter fine-tuning can infer biases towards the intertextual style in the training split and result in overfitting.

The results of our experiments highlighted two strong contenders—a semantically inspired GVSM and the Smith-Waterman text alignment algorithm—, yet yielded no clear overall "winner". Instead, the conclusion to be drawn is that—as explored in Chapter 5—patterns of text reuse vary across datasets in multiple ways and, as a result, no model fits all text reuse retrieval needs.

An illustrative example is Bernard of Clairvaux. As shown in Chapter 4, Bernard often displays a highly allusive intertextual style, in which connections among words in the same semantic field offer cues that a semantically inspired algorithm like the soft cosine can exploit. However, allusive intertexts constitute only 15% of references in the BiblIndex digital edition of Bernard's *Sermons on the Song of Songs*. The remaining bulk of references consists of literal (31%) and semi-literal quotations (i. e. mentions: 54%)—see Table 4.1 in Chapter 4 for the exact numbers—which are better served by a text alignment algorithm with a specifically fine-tuned gap penalty. Thus, if, as shown in Chapter 4, soft cosine was able to offer an important boost for the retrieval of allusions over a purely lexical alternative model, Smith-Waterman performed slightly better when considering the entire dataset, ignoring allusive intertexts and exploiting references based on lexical matching only—see Figure A.3 for a visualization of the CV results. In contrast, when considering the entire dataset, soft cosine was shown to lag behind the text alignment algorithm slightly—as highlighted in Figure 6.1 through a focused cross-validated comparison—, owing to the fact that hyper-parameters controlling the weight given to semantic relations must be calibrated to strike a balance between allusive and literal reuse styles.

Besides the experiments in Chapter 2.4, we approached the question of the relative merit of retrieval algorithms from a different point of view. In Chapter 6, we re-purposed inter-annotator agreement indices in order to assess the relative utility of retrieval algorithms in the context of a late-stage phase of the editorial process. This is a novel approach that seeks to exploit the vantage point of a real-world use case for the comparison of model performance—an aspect that is commonly absent from evaluation comparisons, even in studies in the IR and NLP communities. Although, by experimental design, the comparison results cannot be extrapolated to other datasets, they showed a slight tendency of soft cosine towards retrieving less controversial cases of reuse than those retrieved by Smith-Waterman. Arguably, literal quotations are less likely to have been overseen—both by algorithms and annotators—and their appearance in a late-stage retrieval phase is more likely to be controversial.

7.1.2 The Role of Semantics

As soon as one considers types of intertextual references other than literal quotations, questions arise as to whether models of lexical and sentential semantics can help and how to incorporate them into retrieval systems. In Chapter 4, we saw that distributional models of sentential semantics based on word embeddings failed to retrieve allusive references at acceptable performance levels in the context of Bernard's Sermons. At least in this case, the reason for the ineffectiveness of sentential semantics seems to lie in the fact that, as highlighted above, Bernard's allusive style relies heavily on lexical semantics, often revolving around a semantic field introduced by a reduced number of terms in isolation from the context.

> **Source** *1 Corinthians 3:6* "ego plantavi Apollo rigavit sed Deus incrementum dedit"
>
> "I planted [the seed] (plantavi), Apollos watered [it] (rigavit) but God made it grow"
>
> **Target** *S. 65, 1* "Illam loquor, quae implevit terram, cuius et nos portio sumus: vineam grandem nimis, Domini plantatam manu, emptam sanguine, rigatam verbo, propagatam gratia, fecundatam Spiritu."
>
> "That grapevine, I mean, which fills up the earth, that we too are part of, a vast vineyard, planted by the hand of God, obtained with blood, watered by Christ (verbo), bred by grace, made fruitful by the Spirit."

Quote 2

In Quote 2, an example of an allusive reference is shown in which the semantic field of "planting"—based on the terms "planto" (en. I plant) and "rigo" (en. I water plants) in *1 Corinthians, 3:6*—is used and expanded by Bernard into the more specific field of "cultivating a vineyard". Since distributional models of sentential semantics based on word embeddings tend to conflate the meaning representations of multiple words into a single representation of the sentence—for example, by means of weighted addition—, the signal contributed by the words introducing the semantic field gets lost in the output representation, and the reference cannot be retrieved.

To the extent that the distribution of intertexts in a corpus can be explained by corpora-level reuse patterns—for example, in the case of Bernard, a tendency for recurrent semantic fields to introduce references—Machine Learning models could be deployed in order to mine these patterns and incorporate them into a semantically inspired retrieval system.

In this PhD thesis, however, we relied on the GVSM in order to equip the retrieval system with a lexico-semantic component. Two limitations of our implementation hint at possibilities for future research.

First, our semantic component only captured semantic similarities at the word level, despite the fact that the underlying Tf-Idf lexical model benefited from the inclusion of n-grams at higher levels. Retrieval systems could benefit from considering semantic similarity between n-grams (and skip-grams) that capture sub-sentential re-phrasings and can, thus, boost the overall similarity of relevant passages. Preliminary experiments inducing distributional meaning representations of n-grams based on the aggregation of word-level representations (Zhao et al., 2017) failed to produce any further boosts on the dataset of Bernard's allusions. However, its application to corpora with different allusive styles may prove beneficial.

The second point relates to the fact that in the current implementation word similarity weights are applied equally on all word combinations, regardless of the relevance of the word pair for the intertextuality of the target corpus. In our case, the soft cosine similarity was applied on top of Tf-Idf weighted bag-of-words representations, which are sensitive to frequency-based word-level relevances. However, here again, automatic weighting of word similarities on the basis of corpus patterns extracted through (un)-supervised means could enhance retrieval performance.

The statistical exploration of the Patrology presented in Chapter 5 utilized topic models in order to obtain a richer quantitative characterization of intertextual styles. On top of an axis of variation depicting lexical similarity, we assessed the extent to which a second axis of variation depicting the thematic embedding of the link helps explaining the observed variation. Even though the two axes showed explanatory power, the estimates for lexical similarity and topical embedding showed high correlation across a number of factors, which indicates that topic-level semantics may have a moderate effect on retrieval.

7.1.3 Relevance of Lemmatization

Owing to the nature of the corpora used in the present PhD thesis, lemmatization featured prominently. As argued in Chapter 3, the lemmatization of Western historical languages poses challenges of their own, and requires specific modeling choices in order to provide strong performance. The reasons were related to unstable orthography and the fact that Western historical languages have more complex morphological systems than their modern standardized variants.

Overall, lemmatization can be expected to produce a boost in retrieval performance—especially if a robust open-set disambiguating lemmatizer is used—, since the amount of captured lexical overlap is likely to increase when the input text is lemmatized. This was the case for the dataset of Bernard's *Sermons on the Song of Song*, where—as shown in Table 4.1—the lexical overlap increased most strongly for the text reuse type classified as "mentions"—going from 0.26 to 0.31 in terms of Jaccard similarity. Moreover, in the case of literal quotations lemmatization did not improve lexical overlap, since no re-phrasing is present in these instances. Finally, in the case of allusive text reuse lemmatization brought Jaccard up by a very small amount (an increase of 0.02 starting from 0.02), which is to be expected considering that, as discussed in the previous section, Bernardine allusions are more predominantly based on semantic rather than lexical cues.

A second way in which lemmatization helps is by making it possible to obtain higher quality word embeddings. By applying lemmatization to the corpus used to train the word embedding matrix, the resulting embeddings capture word-level semantic relations more faithfully. This conclusion is derived from the experiments reported in Appendix A.1, which underlie the construction of the word embedding space for Latin used throughout the present case studies. The evidence indicates that while both `FastText`—a word embedding algorithm that exploits sub-word information—and `Word2Vec` benefit from lemmatization, the improvements in the latter case are more striking. Not only does lemmatization boosts performance in a word similarity task by 4 points on average in the case of `Word2Vec`, but it also produces a strong robustness against the influence of other hyper-parameters. Finally, our experiments showed that, in the absence of an adequate lemmatizer for a target corpus, the retrieval performance of re-phrased quotations, mentions, and even allusions could still be enhanced by applying the soft cosine similarity on embeddings trained with the `FastText` algorithm—although in this case, careful fine-tuning of `FastText` hyper-parameters is necessary.

7.1.4 Expectations on Inter-Annotator Agreement

In Chapters 4 and 6 we dealt with questions related to inter-annotator agreement of intertextual links. Our research contributed a first assessment of the expected inter-annotator agreement in two different settings. In Section 4.2, we focused on the more specific matter of identifying the span of words in the target passage that is responsible for the allusion. Our experiments highlighted that—even though the lack of comparable research hinders a contextualized interpretation of the results—the obtained Fleiss's κ is situated within a low range of agreement ($\kappa = 0.22$).

In contrast, Chapter 6 looked into agreement on candidate pairs retrieved by two algorithms—soft cosine and `Smith-Waterman`. In those experiments, the focus was on statistically controlling for contextual factors, leveraging a probabilistic definition of Cohen's κ and multilevel statistical models to obtain the probability estimates. The results indicate that agreement, in this case, is likely to be low when the referring passage is not known by the editors. This result must be interpreted in light of the experimental setting, which was set in a late-stage phase where editors have already scanned the collection in search for intertextual references. More importantly, our experiments highlighted that agreement depends highly on context. In particular, the biblical book from which the borrowed passages stem—and, to a lesser extent, the lexical overlap between the linked documents—contributed a large proportion of variance to the estimates of agreement.

Our experiments provide useful pointers to future intertextuality-focused research aiming at characterizing agreement either from a theoretical point of view—e. g. a epistemological consideration of intertextual links—or from a practical perspective—with the goal of establishing a reliable annotation process for the curation of labeled resources. First, contextual factors should be controlled for, since they may not only act as statistical confounders, but also yield interesting insights about what aspects may influence the agreement among experts. Second, the degree of variability contributed by contextual factors suggests that these influence annotators towards divergent types of behavior, which can compromise the quality of the resulting resources. In view of this consideration, a viable strategy for constructing labeled resources may be to focus on a single annotator, assuming that the resulting dataset captures not just the reuse style in the texts of the considered corpus but also the patterns of interpretation that characterize the annotator. To the extent that the annotator is consistent in her interpretational patterns, this can be a promising approach.

7.2 BACK TO THE FUTURE

A question currently impending over the field of Literary Text Reuse Detection concerns the application of contemporary NLP approaches based on Machine Learning methods and, in particular, on Deep Neural Networks. More specifically, we shall discuss the application of two types of approaches. First, unsupervised large-scale pre-training using Language Models and, second, neural architectures for text matching trained in an end-to-end fashion.

7.2.1 Contextualized Word Embeddings

Current research in NLP has been transformed by the introduction of contextualized word embeddings—i. e. vectorial representations of the meaning of words that take into account the usage of the words within their textual context. Beyond the appeal of the built-in word sense disambiguating capabilities that these representations entail, their success can be arguably traced back to a less theoretical reason: the fact that the architectures used for learning these representations can leverage much larger datasets than traditional word embedding architectures.

Contextualized word embeddings go back to the introduction of probabilistic Language Models trained on the basis of Neural Networks (Bengio et al., 2003). The idea behind a Language Model is to learn a joint probability function of sentences, decomposing this probability into the product of the probabilities of the words conditioned on previous words. Originally, the application of Neural Networks to this task was motivated by the goal of overcoming the limitations of alternative count-based approaches. These count-based approaches struggled to cope with compositionality and syntactic productivity in language—which warrants poor probability estimates for sequences unseen during training—and the dense representations learned by neural architectures were seen as a promising alternative approach.

More generally, the appeal of dense representations soon transcended the task of LM once their utility for downstream tasks was, first, discovered (Collobert and Weston, 2008; Collobert et al., 2011), they were, second, shown to capture intriguing linguistic regularities (Mikolov, Yih, and Zweig, 2013), and, finally, data-efficient methods were devised that delivered higher quality embeddings by leveraging larger datasets (Mikolov et al., 2013). Subsequent research quickly reaped the benefits of applying word embeddings as word-level feature representations in order to improve the state of the art of tasks across the board.

Interestingly, the quest for methods for extracting improved dense representations led back again to Language Models based on Neural Networks, which were shown to posses much stronger model capacity than the shallow log-linear models used for computing word embeddings. First, architectures based on bi-directional RNNs were shown to produce reasonable improvements, which were further enhanced through the application of a dedicated fine-tuning phase on the target datasets (Howard and Ruder, 2018; Peters et al., 2018). While these architectures still targeted a LM loss, the second generation—best represented by the BERT architecture (Devlin et al., 2019)—dropped this formal requirement and achieved further improvements by including a set of combined objectives that are, indeed, inspired by LM but that do not formally comply with the original mechanism of decomposing the probability of sentences into that of the individual words.

Considering the progress that these architectures have brought to the field, their application to the retrieval of intertextual references is, thus, an inevitable venue for future research. In this respect, the first challenge that we faced was the limited size of the linguistic resources for our target languages. Besides the 8.5GB of highly noisy Latin text reported by Bamman and Crane (2011a)—consisting, mostly, of OCR'd scans of varying quality—, the largest dataset of clean Latin that we could collect added up to circa 165 million tokens, which hardly compares to the scale of datasets commonly used to pre-train contextualized word embeddings—for comparison, the original BERT model for English was trained on 3 billion tokens. Preliminary experiments applying a BERT variant model (Liu et al., 2019)—trained on these 165 million tokens with the `transformers` library (Wolf et al., 2020)—to the string transduction lemmatizer described in Chapter 3 yielded no improvements over previous results. Moreover, the application of purely semantic models to the retrieval of allusive text reuse in Bernard was already shown to have stalled, and, accordingly, the application of contextualized word embeddings obtained with the mentioned BERT model failed to yield satisfactory results.

Recently, (Bamman and Burns, 2020) introduced a BERT model for Latin that was trained on 642.7 million tokens—extracted, among other sources, from the larger 8.5GB noisy dataset through a set of data selection techniques. The resulting model showed performance improvements in tasks such as part-of-speech-tagging, text infilling and word sense disambiguation, and could yield improvements for the detection of intertextual references.

Provided the availability of a large-scale modern Language Model, we envision three possibilities for the deployment of these contextualized word embeddings. First, despite our failed attempts, contextualized word embeddings may still be useful in places where traditional word embeddings are currently deployed. In the context of the present thesis, this corresponds to applications of hybrid retrieval models like the soft cosine, which, as indicated above, can be enhanced with similarities computed on top of dense representations at different n-gram levels. Second, contextualized word embeddings can be deployed in order to extract sub-phrasal paraphrases, which can be used to identify less literal cases of reuse. A final domain of application is as additional input features in end-to-end models—which we shall discuss below.

7.2.2 End-to-End Models

At least partially, the rise of Neural Networks in current NLP and IR research owes to the fact that they can automate the process of feature extraction. Given a corpus annotated with labels of interest, an architecture that produces the desired output and an appropriate training objective, Neural Networks can fully automate the annotation process from the raw input, exploiting correlations between patterns in the input and the target labels. In the case of intertextual references, this end-to-end training paradigm bears the promise to faithfully capture the type of intertext that annotators—or, as argued above, a single annotator—have identified in the source collection.

From the point of view of neural architectures, deep matching networks appear as the most promising candidates. Deep matching networks target a general problem in which two fragments of text—e. g. a query and an indexed document, or a source and a target passage—must be matched with respect to a notion of relevance. Multiple incarnations of this problem exist across the fields of IR and NLP—some of which were covered in Section 2.2, and others yet include Question Answering, Dialogue Systems or Web Search.

Deep matching networks can be categorized into two major classes—representation-based and interaction-based models—depending on what informational aspects are modeled and what tasks are targeted. While the first type is concerned with capturing semantic similarity and relatedness, exploiting the compositional structure of the input text and matching the two texts from a global perspective—i. e. modeling the overall meaning of the input—, the second type focuses on weighing the relative importance of individual terms and phrases, targeting literal or slightly re-phrased local expressions that may constitute just a small proportion of the whole input texts.

Without delving into the concrete details of particular architectures, the general form of deep matching networks can be decomposed into a representation function ϕ of the input texts D_i and D_j, and a matching function $F(\phi(D_i), \phi(D_j))$ of the computed representations. As argued by Guo et al. (2016), the focus on global semantics that characterizes representation-based approaches leads to the development of relatively complex networks for implementing ϕ. These networks are tasked with obtaining abstract meaning representations of entire passages—for instance, DSSM (Huang et al., 2013) uses a multi-later perceptron, C-DSSM (Gao et al., 2014; Shen et al., 2014) extends the DSSM through the addition of a Convolutional Neural Network (CNN) (LeCun and Bengio, 1995), ARC-I (Hu et al., 2014) similarly uses a CNN and MaLSTM (Mueller and Thyagarajan, 2016) uses an LSTM (Hochreiter and Schmidhuber, 1997). By contrast, the matching function F of representation-based approaches is a comparatively simple, non parametric function like the cosine or the Manhattan similarity.

Interaction-based approaches, however, typically employ a relatively simple representation function ϕ—e. g. a map of the input into a sequence of sparse or dense representations of the input words—and F is implemented by more involved modules operating over term interaction matrices. Given a pair of documents represented as sequence of (possibly dense) embeddings of the input words, an interaction matrix contains word-level relevance scores between the words in the first document and the words in the second document. After computing these matrices, interaction-based models are tasked with producing a single relevance score through a process that involves a series of steps in which the interaction matrix is hierarchically compressed into a fixed-length vector. For example, DeepMatch (Lu and Li, 2013) uses Topic Models to compute a topic-level interaction matrix and ARC-II (Hu et al., 2014) and MatchPyramid (Pang et al., 2016) use multiple layers of convolutional networks to compute interactions at different levels—words, phrases and sentences. Finally, models have also been proposed that aim to bridge between purely representational and interaction-based matching models—e. g. MV-LSTM (Wan et al., 2016) or DUET (Mitra, Diaz, and Craswell, 2017).

On first sight, interaction-based matching models represent the most promising approach towards end-to-end training for Literary Text Reuse Detection, since modeling local interactions represents a more flexible approach to accommodate the diverse set of patterns that intertextual links entail. This is more so since deep matching models are not limited to matching on purely lexical levels, but can compute interactions at higher-levels through the stacking of multiple layers or—like the example of the DeepMatching architecture shows—through the incorporation of external components like Topic Models or contextualized word embeddings. In contrast, matching on globally obtained semantic representations is likely to encounter similar issues to those faced by purely semantic models from Chapter 4.

However, the application of interaction-based models in this field faces two problems. The first one is related to speed concerns during deployment. In contrast to representation-based approaches, which strongly decouple a computationally inexpensive matching function F from the computation of the individual document representations $\phi(D_i)$ and $\phi(D_j)$, interaction-based approaches must compute a comparatively more costly F for each new pair of documents. Thus, representation-based approaches can delegate the computation of more costly ϕ modules to off-line phases and quickly compute F for new documents against cached representations of large collections. In contrast, the application of interaction-based models for the detection of intertextual links on large corpora demands further attention.

The second one refers to the problem of false positives in literary text reuse corpora. As we could experience on preliminary experiments with interaction-based models—using the `MatchZoo` library (Guo et al., 2019) on the Latin Patrology and the dataset of Bernard's references—, the construction of strongly discriminative negative examples requires dedicated research. The training objective of deep matching networks commonly involves training instances of two types: positive pairs extracted directly from the corpus annotations, and automatically constructed negative examples that do not constitute true intertextual links but resemble positive pairs enough in order to allow networks to learn—a setup that resembles Siamese Network training (Bromley et al., 1994; Chopra, Hadsell, and LeCun, 2005). In our experience, finding discriminative negative examples was faced with the problem that the obtained instances lacked discriminative power if not enough filtering was applied or, when efficient set-similarity algorithms—like those from Section 2.5.2.1—were used for filtering, negative examples were retrieved that resembled positive examples to a worrying degree. Finding negative examples that strike the right level of discriminativity turned out to be overly challenging.

7.2.3 Beyond End-to-End Models

Finally, two considerations from the point of view of the user of Literary Text Reuse Detection engines are due.

The first one questions whether end-to-end systems are ultimately the desideratum for future research. Even if successful, end-to-end systems are bound to capture specific patterns of reuse, since, in the absence of general purpose intertextuality corpora, these systems will have to be trained on corpora of specific authors. However, a prospective user will often seek to find patterns different from those present in the training corpus. In contrast, an alternative approach to intertextual engines built in a modular fashion may be better tailored to the needs of literary scholars. Such a modular engine, in which specialized components target different aspects of reuse—e. g. lexical overlap, semantic relatedness between phrases, topic similarity and even narrative structures or literary motifs—as well as interactions between these aspects, has the potential to empower the user to choose what particular style of reuse she wants to focus on at any given point.

Finally, prospective users not only demand strong performance from a retrieval system but are also often observed to seek understanding of the mechanisms that lead the underlying algorithms to propose candidate pairs of reuse. In the case of the algorithms tested in this PhD thesis and described in Section 2.5.2, the underpinnings of the algorithms can easily be repurposed to obtain cues about which aspects of the documents are responsible for the relevance score. For example, the alignment induced by text alignment algorithms can easily be visualized and proposed as the explanation. Moreover, in Section 4.3.5.1, we showed how to compute semantic relevance of terms with the soft cosine algorithm. In the case of opaque approaches—known as black-box approaches—that employ Machine Learning techniques, interpretations and explanations about why of a pair of documents is being proposed are certainly less straightforward to compute, and future research will profit from current efforts regarding the interpretability of Neural Networks in NLP (Alishahi et al., 2020; Linzen, Chrupała, and Alishahi, 2018; Linzen et al., 2019).

A APPENDIX

A.1 FINE-TUNING OF WORD EMBEDDINGS

Word embeddings feature prominently in the experiments reported in Chapters 2, 4 and 6. In the context of the application of soft cosine, word similarity weights are computed on the basis of cosine similarity over word embeddings. In contrast to modern languages, no standard embedding spaces are available for Latin. Moreover, training corpora are restricted to a few hundred million words, and, furthermore, obtaining high-quality word embeddings is complicated by diachronic shifts in language usage for texts originating in a span of over a millennium. For these reasons, we suspect that exhaustive hyper-parameter fine-tuning can be crucial in order to achieve strong performance in word similarity tasks. Therefore, we conducted an intrinsic evaluation of two alternative methods to obtain word embeddings—`FastText` (Bojanowski et al., 2017) and `Word2vec` (Mikolov et al., 2013)—and tested the influence of several hyper-parameters on the quality of the representations.

A.1.1 Training Corpus

For training, we used the *Corpus Corporum* (Roelli, 2014), which comprises about 162 million tokens of a diachronically representative sample of Latin, a large part of which corresponds to the Latin Patrology. The corpus was lemmatized with the neural lemmatizer described in Section 2.6.1, trained on the `cap` corpus.

A.1.2 Word Similarity Benchmark Corpus

In order to control for the quality of the embeddings, we used the word similarity benchmark for Latin by Sprugnoli, Passarotti, and Moretti (2019) in order to evaluate embeddings over a number of hyper-parameter combinations.[1] The benchmark dataset consists of 2,758 tests involving a focus word, a synonym and three distractors, which are words that are seemingly unrelated to the focus word and the synonym. Each test is considered to be successfully solved when

[1] For the present experiments, we used the implementation of `Word2vec` included in the `gensim` package (Rehurek and Sojka, 2010) and the official implementation of `FastText` available through the developers repository on the following URL: `https://github.com/facebookresearch/fastText`.

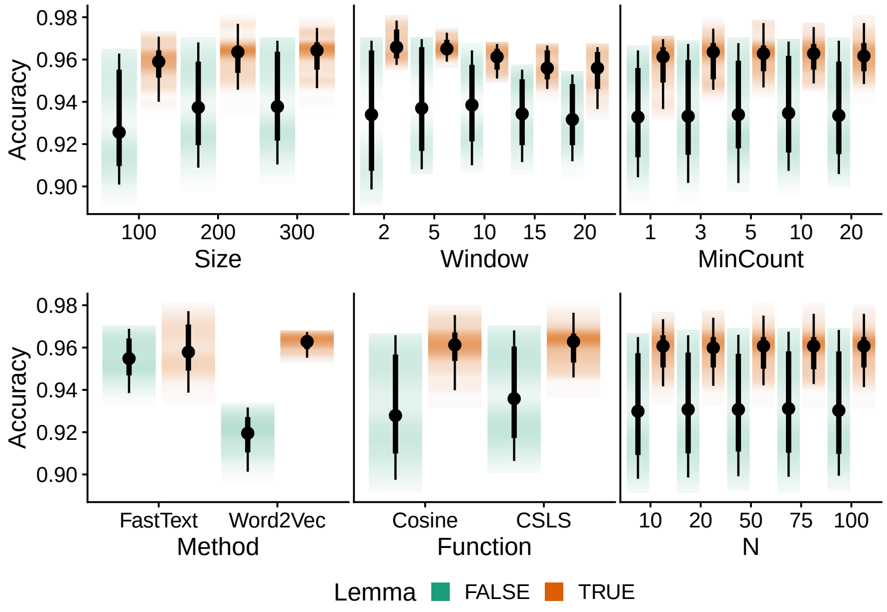


Figure A.1: Comparison of `Word2vec` and `FastText` on the Latin word similarity benchmark by Sprugnoli, Passarotti, and Moretti (2019). From left to right and top to bottom, "Size" refers to the embedding dimensionality, "Window" to the size of the context window, "MinCount" to the word frequency threshold, "Method" to the training algorithm, "Function" to the function used to estimate the semantic similarity and "N" to the number of neighbors used in the CSLS computation.

the similarity between the focus word and the synonym is higher than the similarity between the focus word and any of the distractors. We used cosine similarity as well as the Cross-Domain Similarity Local Scaling (CSLS), which is a measure of similarity that aims at mitigating hubness-related issues in multi-dimensional spaces and that has been shown to improve word translation retrieval tasks (Lample et al., 2018).

Since `Word2vec` models cannot generate an embedding for words that have not been encountered during training, test cases including out of vocabulary words must be dropped to ensure fair comparison. Therefore, after removing out of vocabulary words only test cases were kept that contained the focus word, the synonym and at least one distractor.

A.1.3 Results

Figure A.1 shows the results of a grid-search over different combinations of parameters, including whether the training data was lemmatized or not.

Overall, the inclusion of a lemmatization step produced the best results. For non-lemmatized training data, the distribution of scores is generally much broader, which means that hyper-parameter tuning becomes more important in the absence of lemmatization. In terms of hyper-parameters—embedding dimension (Size), context window (Window) and frequency threshold (MinCount)—a larger dimensionality seems to improve results (although at 300 performance can be seen to plateau), context window peaks at 10 before it starts to drop and the frequency threshold does not seem to have a noticeable impact on performance.

Overall, FastText was the superior method and, more importantly, it seems unaffected by lemmatization. This is an interesting result for practitioners who wish to obtain embeddings for historical languages without available lemmatizers. However, `Word2Vec` on lemmatized input showed the most consistent results, regularly achieving high performance regardless of other hyper-parameter configurations. Finally, CSLS produced consistently better results than cosine and the number of neighbors used by CSLS does not seem to have a big influence on the output scores.

A.2 VISUALIZATION OF MODEL COMPARISONS

This section provides visualizations to the model comparison experiments from Chapter 2.

A.2.1 Cross Validation Results

Figures A.2 and A.3 display the distribution of AP scores obtained through CV for the `blueletterbible` and `Sermons` datasets respectively. Similarly, Figure A.4 shows the distribution of NDCG scores on the `openbible` dataset. These visualizations complement the text reuse retrieval results discussed in Section 2.6.3 for the hybrid and the ranking-based evaluation settings.

In the visualizations, "S-W" refers to `Smith-Waterman`, "Set" to `Set-based` and "S-C" to `Soft-Cosine`. Moreover, we used "CV" to refer to results obtained via Cross Validation (CV)—in which the best hyper-parameters of each training fold are applied to the remaining data. In contrast, "Oracle" refers to the distribution of scores obtained by the best possible hyper-parameters on each of the test split of each fold. These distribution, thus, corresponds to an estimation of the empirical maximum performance per fold.

A.2.2 Bayesian Model Comparison

Figures A.5 and A.6 visualize the results of the pairwise performance comparisons of the retrieval systems across the `blueletterbible` and

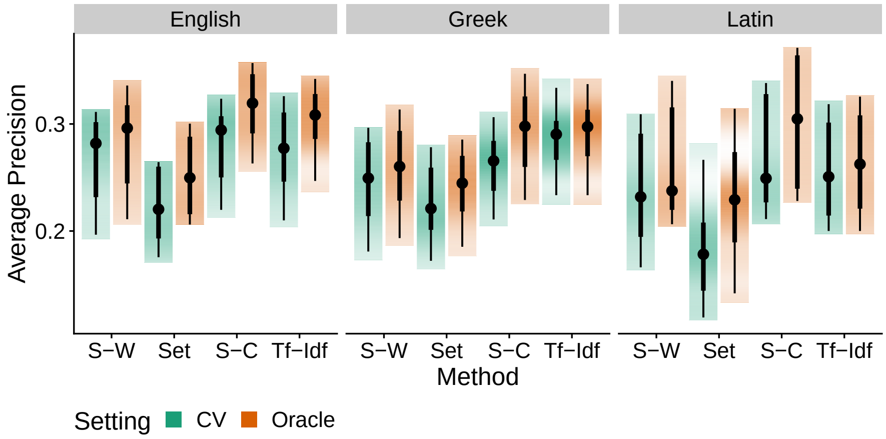


Figure A.2: Cross-validated distribution of AP scores over 10 folds for Bible versions in three different languages, using the the `blueletterbible` dataset.

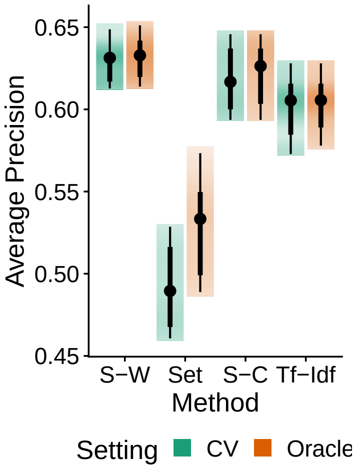


Figure A.3: Cross-validated distribution of AP scores over 10 folds for the `Sermons` dataset.

`Sermons` datasets, using the Bayesian hierarchical model described in Section 2.6.3.

Each triangle plot summarizes the posterior distribution of expected differences on unseen datasets using a dot-plot over the probability simplex. Each dot in the plot corresponds to one draw of the posterior, and the position in the simplex indicates the contributed evidence to each of the three hypothesis: one in support of the practical equivalence of both systems as specified by the ROPE, and two more in support of the relative improvement of one of the systems over the other. Furthermore, the probability of each hypothesis—i. e. the proportion of posterior samples in favor of each of the hypotheses—is provided on the vertices of the simplex. For all model comparisons, the ROPE was specified at 1% difference in AP or NDCG

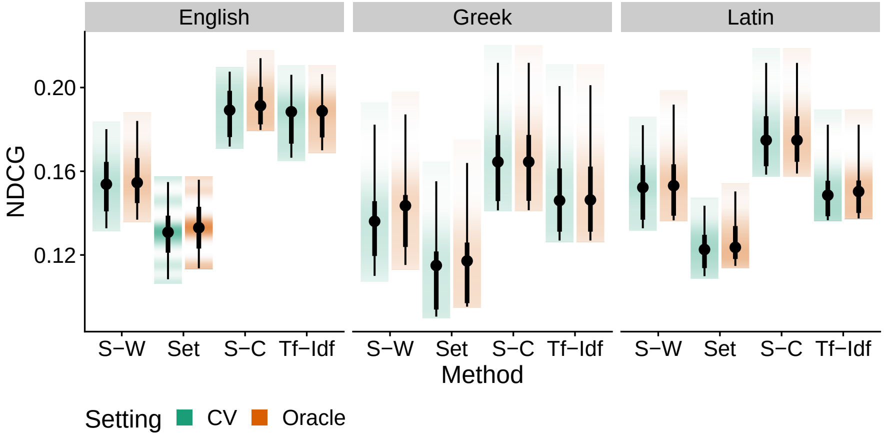


Figure A.4: Cross-validated distribution of NDCG scores over 10 folds for the `openbible` dataset using Bible versions in different languages.

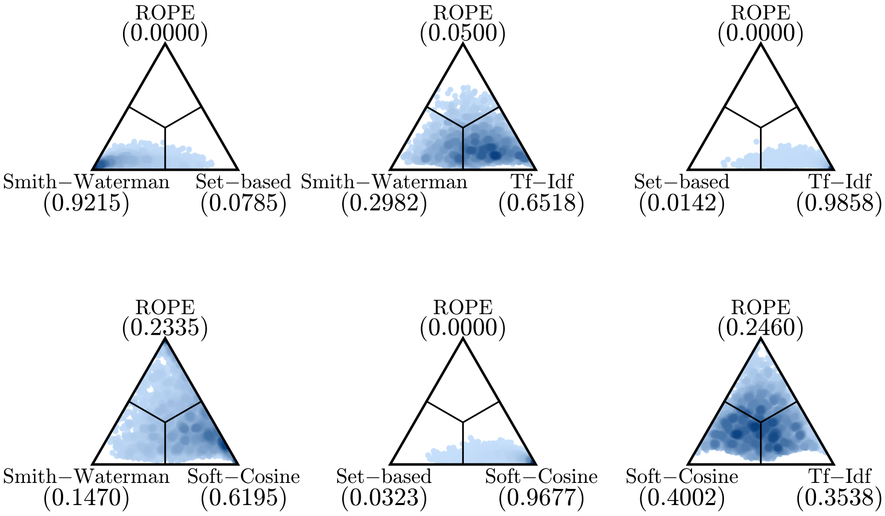


Figure A.5: Visualization of the Bayesian model comparison using cross-validated AP scores across the `blueletterbible` datasets on three Bible versions and the `Sermons` dataset.

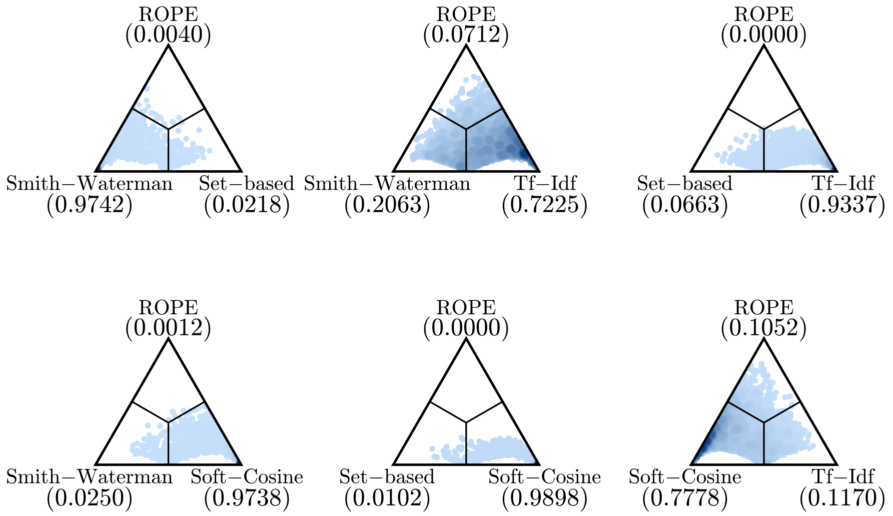


Figure A.6: Visualization of the Bayesian model comparison using cross-validated NDCG scores across three Bible versions on the `blueletterbible` dataset.

BIBLIOGRAPHY

Acerbi, Alberto and R. Alexander Bentley (2014). "Biases in Cultural Transmission Shape the Turnover of Popular Traits." In: *Evolution and Human Behavior* 35.3, pp. 228–236.

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg (2017). "Fine-Grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks." In: *ICLR '17*. ISSN: 00392499. DOI: 10.1161/STR.0b013e318284056a.

Agirre, Eneko, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre (2012). "SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity." In: **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, pp. 385–393.

Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo (June 2013). "*SEM 2013 shared task: Semantic Textual Similarity." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 32–43. URL: https://www.aclweb.org/anthology/S13-1004.

Akbik, Alan, Duncan Blythe, and Roland Vollgraf (2018). "Contextual String Embeddings for Sequence Labeling." In: *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 1638–1649.

Alishahi, Afra, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, eds. (Nov. 2020). *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.blackboxnlp-1.0.

Allen, Graham (Mar. 2000). "Intertextuality." In: *Intertextuality*. Routledge. ISBN: 978-0-203-13103-9. DOI: 10.4324/9780203131039.

Alshomary, Milad, Michael Völske, Tristan Licht, Henning Wachsmuth, Benno Stein, Matthias Hagen, and Martin Potthast (2019). "Wikipedia Text Reuse: Within and Without." In: *Advances in Information Retrieval*. Ed. by Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra. Cham: Springer International Publishing, pp. 747–754. ISBN: 978-3-030-15712-8.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (Oct. 1990). "Basic local alignment search tool." In: *Journal of Molecular Biology* 215.3, pp. 403–410. ISSN: 00222836.

DOI: 10.1016/S0022-2836(05)80360-2. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022283605803602.

Alzahrani, Salha, N. Salim, and A. Abraham (2012). "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, pp. 133–149.

Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith (2016). *Massively Multilingual Word Embeddings*. arXiv: 1602.01925.

Arasu, Arvind, Venkatesh Ganti, and Raghav Kaushik (2006). "Efficient Exact Set-Similarity Joins." In: *Proceedings of the 32nd International Conference on Very Large Data Bases*, pp. 918–929. ISBN: 1595933859.

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu (2013). "A Practical Algorithm for Topic Modeling with Provable Guarantees." In: *International Conference on Machine Learning*. PMLR, pp. 280–288.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (July 2018). "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 789–798. DOI: 10.18653/v1/P18-1073. URL: https://www.aclweb.org/anthology/P18-1073.

Artstein, Ron (2017). "Inter-Annotator Agreement." In: *Handbook of Linguistic Annotation*. Springer, pp. 297–313.

Artstein, Ron and Massimo Poesio (2007). *Inter-Coder Agreement for Computational Linguistics*. This reference corresponds to an extended version of the survey article appearing in the Computational Linguistics journal under the same name. URL: https://dces.essex.ac.uk/Research/nle/arrau/icagr.pdf.

Artstein, Ron and Massimo Poesio (Dec. 2008). "Inter-Coder Agreement for Computational Linguistics." en. In: *Computational Linguistics* 34.4, pp. 555–596. ISSN: 0891-2017, 1530-9312. DOI: 10.1162/coli.07-034-R2.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate." In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1409.0473.

Bamman, David and Patrick J Burns (2020). "Latin BERT: A Contextual Language Model for Classical Philology." In: *arXiv preprint arXiv:2009.10053*. arXiv: 2009.10053.

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

Bamman, David and Gregory Crane (2009). "Discovering Multilingual Text Reuse in Literary Texts." In: *Perseus Digital Library*.

Bamman, David and Gregory Crane (2011a). "Measuring Historical Word Sense Variation." In: *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. ACM, pp. 1–10.

Bamman, David and Gregory Crane (2011b). "The ancient Greek and Latin dependency treebanks." In: *Language technology for cultural heritage*. Springer, pp. 79–98.

Bär, Daniel, Torsten Zesch, and Iryna Gurevych (2012). "Text Reuse Detection Using a Composition of Text Similarity Measures." In: *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*.

Baron, Alistair and Paul Rayson (May 2008). "VARD2:A Tool for Dealing with Spelling Variation in Historical Corpora." In: *Postgraduate Conference in Corpus Linguistics*. Postgraduate Conference in Corpus Linguistics. Aston University, Birmingham. URL: https://eprints.lancs.ac.uk/id/eprint/41666/ (visited on 03/16/2021).

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (June 2014). "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 238–247. DOI: 10.3115/v1/P14-1023. URL: https://www.aclweb.org/anthology/P14-1023.

Barteld, Fabian, Katharina Dreessen, Sarah Ihden, and Ingrid Schröder (2017). "Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200-1650) Korpusdesign, Korpuserstellung Und Korpusnutzung." In: *Mitteilungen des Deutschen Germanistenverbandes* 64.3, pp. 226–241. ISSN: 0418-9426. DOI: 10.14220/mdge.2017.64.3.226.

Bayardo, Roberto J., Yiming Ma, and Ramakrishnan Srikant (2007). "Scaling up all pairs similarity search." In: *16th International World Wide Web Conference, WWW2007*, pp. 131–140. ISBN: 1595936548. DOI: 10.1145/1242572.1242591.

Benavoli, Alessio, Giorgio Corani, Janez Dem˘sar, and Marco Zaffalon (2017). "Time for a Change: A Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis." en. In: *The Journal of Machine Learning Research*, p. 36.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin (2003). "A neural probabilistic language model." In: *The journal of machine learning research* 3, pp. 1137–1155.

Bennett, E. M., R. Alpert, and A. C. Goldstein (1954). "Communications Through Limited Response Questioning." In: *Public Opinion Quarterly* 18.3, p. 303. ISSN: 0033362X. DOI: 10.1086/266520. URL: https://academic.oup.com/poq/article-lookup/doi/10.1086/266520.

Bergmanis, Toms and Sharon Goldwater (2018). "Context Sensitive Neural Lemmatization with Lematus." In: *North American Chapter* *of the Association for Computational Linguistics: Human Language Technologies Volume 1*. ISBN: 0-02-590111-7. DOI: 10.1007/s00259-011-1787-z.

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization." In: *Journal of Machine Learning Research* 13.10, pp. 281–305.

Bernard of Clairvaux (1998). *Sermons Sur Le Cantique 16-32*. Ed. by Raffaele Fasseta and Paul Verdeyen. Vol. 2. Sources Chrétiennes. Paris.

Bernard of Clairvaux (2000). *Sermons Sur Le Cantique 33-50*. Ed. by Raffaele Fasseta and Paul Verdeyen. Vol. 3. Sources Chrétiennes. Paris.

Bernard of Clairvaux (2003). *Sermons Sur Le Cantique 51-68*. Ed. by Raffaele Fasseta and Paul Verdeyen. Vol. 4. Sources Chrétiennes. Paris.

Bernard of Clairvaux (2006). *Sermons Sur Le Cantique 1-15*. Ed. by Raffaele Fasseta and Paul Verdeyen. Second. Vol. 1. Sources Chrétiennes. Paris.

Bernard of Clairvaux (2007). *Sermons Sur Le Cantique 69-86*. Ed. by Raffaele Fasseta and Paul Verdeyen. Vol. 5. Sources Chrétiennes. Paris.

Bernhard, Delphine and Iryna Gurevych (2008). "Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites." In: *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. EANL '08. Columbus, Ohio: Association for Computational Linguistics, pp. 44–52. ISBN: 9781932432084.

Bhagat, Rahul and Eduard Hovy (May 2013). "What Is a Paraphrase?" In: *Computational Linguistics* 39.3, pp. 463–472. ISSN: 0891-2017. DOI: 10.1162/COLI_a_00166.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation." In: *Journal of Machine Learning Research*. ISSN: 15324435. DOI: 10.1016/b978-0-12-411519-4.00006-9.

Bloom, Harold (1973). *The Anxiety of Influence: A Theory of Poetry*. Oxford University Press.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information." In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Bollmann, Marcel and Anders Søgaard (2016). "Improving Historical Spelling Normalization with Bi-Directional LSTMs and Multi-Task Learning." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, pp. 131–139.

Breiman, Leo (2001). "Statistical Modeling: The Two Cultures." In: *Statistical science* 16.3, pp. 199–231.

Broder, Andrei Z. (1997). "On the resemblance and containment of documents." In: *Proceedings of the International Conference on Compression and Complexity of Sequences*. IEEE, pp. 21–29. doi: 10.1109/sequen.1997.666900.

Bromley, Jane, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah (1994). "Signature verification using a" siamese" time delay neural network." In: *Advances in neural information processing systems*, pp. 737–737.

Büchler, M., G. Franzini, E. Franzini, and Maria Moritz (2014a). "Scaling historical text re-use." In: *2014 IEEE International Conference on Big Data (Big Data)*, pp. 23–31.

Büchler, Marco (2013). "Informationstechnische Aspekte Des Historical Text Re-Use." PhD thesis. Universität Leipzig.

Büchler, Marco, Philip R Burns, Martin Müller, Emily Franzini, and Greta Franzini (2014b). "Towards a Historical Text Re-Use Detection." In: *Text Mining: From Ontology Learning to Automated Text Processing Applications*. Springer, pp. 221–238. isbn: 978-3-319-12654-8 978-3-319-12655-5. doi: 10.1007/978-3-319-12655-5\{\\_\}11.

Büchler, Marco, Gregory Crane, Maria Moritz, and Alison Babeu (2012). "Increasing Recall for Text Re-Use in Historical Documents to Support Research in the Humanities." In: *Theory and Practice of Digital Libraries*. Ed. by Panayiotis Zaphiris, George Buchanan, Edie Rasmussen, and Fernando Loizides. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 95–100. isbn: 978-3-642-33290-6.

Budanitsky, Alexander and Graeme Hirst (2001). "Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures." In: *Workshop on WordNet and Other Lexical Resources*. Vol. 2, p. 2.

Bürkner, Paul Christian (2018). "Advanced Bayesian Multilevel Modeling with the R Package Brms." In: *R Journal*. issn: 20734859. doi: 10.32614/rj-2018-017.

Burrows, Steven, Martin Potthast, and Benno Stein (June 2013). "Paraphrase Acquisition via Crowdsourcing and Machine Learning." In: *ACM Transactions on Intelligent Systems and Technology* 4.3, pp. 1–21. issn: 2157-6904. doi: 10.1145/2483669.2483676.

Camps, Jean-Baptiste, Elena Albarran, Alice Cochet, and Lucence Ing (Apr. 2019). *Geste: un corpus de chansons de geste*. Version v02. doi: 10.5281/zenodo.2630574. url: https://doi.org/10.5281/zenodo.2630574.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017). "Stan : A Probabilistic Programming Language." In: *Journal of Statistical Software* 76.1. issn: 1548-7660. doi: 10.18637/jss.v076.i01. url: http://www.jstatsoft.org/v76/i01/.

Caruana, Rich (1997). "Multitask Learning." In: *Machine learning* 28.1, pp. 41–75.

Chakrabarty, Abhisek, Onkar Arun Pandit, and Utpal Garain (2017). "Context Sensitive Lemmatization Using Two Successive Bidirectional Gated Recurrent Networks." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1481–1491. DOI: 10.18653/v1/P17-1136.

Charlet, Delphine and Geraldine Damnati (2017). "SimBow at SemEval-2017 Task 3: Soft-Cosine Semantic Similarity between Questions for Community Question Answering." In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 315–319. DOI: 10.18653/v1/S17-2051. URL: http://aclweb.org/anthology/S17-2051.

Chaudhuri, Surajit, Venkatesh Ganti, and Raghav Kaushik (2006). "A primitive operator for similarity joins in data cleaning." In: *Proceedings - International Conference on Data Engineering*. Vol. 2006. IEEE, p. 5. ISBN: 0769525709. DOI: 10.1109/ICDE.2006.9.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio (2014). "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches." In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, pp. 103–111. DOI: 10.3115/v1/W14-4012.

Chopra, Sumit, Raia Hadsell, and Yann LeCun (2005). "Learning a similarity metric discriminatively, with application to face verification." In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE, pp. 539–546.

Chrupala, Grzegorz, Georgiana Dinu, and Josef van Genabith (May 2008). "Learning Morphology with Morfette." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair) Khalid Choukri. Marrakech, Morocco: European Language Resources Association (ELRA). ISBN: 2-9517408-4-0.

Clough, Paul and Mark Stevenson (Mar. 2011). "Developing a Corpus of Plagiarised Short Answers." In: *Language Resources and Evaluation* 45.1, pp. 5–24. ISSN: 1574-020X. DOI: 10.1007/s10579-009-9112-1.

Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Christopher W Forstall, Roelant Ossewaarde, and Sarah L Jacobson (2012a). "The Tesserae Project: Intertextual Analysis of Latin Poetry." In: *Literary and Linguistic Computing* 28.2, pp. 221–228.

Coffee, Neil, Jean-Pierre Koenig, Poornima Shakti, Roelant Ossewaarde, Chris Forstall, and Sarah Jacobson (Sept. 2012b). "Intertextuality in the Digital Age." In: *Transactions of the American Philological Association* 142. DOI: 10.2307/23324457.

Cohen, Jacob (Apr. 1960). "A Coefficient of Agreement for Nominal Scales." In: *Educational and Psychological Measurement* 20.1, pp. 37–46. ISSN: 0013-1644. DOI: 10.1177/001316446002000104. URL: http://journals.sagepub.com/doi/10.1177/001316446002000104.

Collobert, Ronan and Jason Weston (2008). "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning." In: *Proceedings of the 25th International Conference on Machine Learning*. Helsinki, Finland: Association for Computing Machinery, pp. 160–167. ISBN: 9781605582054. DOI: 10.1145/1390156.1390177. URL: https://doi.org/10.1145/1390156.1390177.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). "Natural Language Processing (Almost) from Scratch." In: *Journal of Machine Learning Research* 12, pp. 2461–2505.

Conte, Gian Biagio (1988). "The Rhetoric of Imitation: Genre and Poetic Memory in Virgil and Other Latin Poets." In: *The Classical World*. ISSN: 00098418. DOI: 10.2307/4350270.

Corani, Giorgio and Alessio Benavoli (2015). "A Bayesian Approach for Comparing Cross-Validated Algorithms on Multiple Data Sets." In: *Machine Learning* 100.2, pp. 285–304.

Corani, Giorgio, Alessio Benavoli, Janez Demšar, Francesca Mangili, and Marco Zaffalon (Nov. 2017). "Statistical Comparison of Classifiers through Bayesian Hierarchical Modelling." en. In: *Machine Learning* 106.11, pp. 1817–1837. ISSN: 1573-0565. DOI: 10.1007/s10994-017-5641-9.

Cotterell, Ryan, Alexander Fraser, and Hinrich Schütze (2015). "Joint Lemmatization and Morphological Tagging with LEMMING." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ISBN: 0921-8009. DOI: 10.1016/j.ecolecon.2009.11.007.

Crane, Gregory (1996). "Building a Digital Library: The Perseus Project as a Case Study in the Humanities." In: *Proceedings of the First ACM International Conference on Digital Libraries*, pp. 3–10.

Crema, Enrico R., Anne Kandler, and Stephen J. Shennan (Dec. 2016). "Revealing Patterns of Cultural Transmission from Frequency Data: Equilibrium and Non-Equilibrium Assumptions." In: *Nature Publishing Group*, pp. 1–10.

Creutz, Mathias Johan Philip et al. (2018). "Open Subtitles Paraphrase Corpus for Six Languages." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Crystal, D. (2001). *Language and the Internet*. Cambridge University Press.

Culler, Jonathan (1976). "Presupposition and Intertextuality." In: *MLN* 91.6, pp. 1380–1396. ISSN: 00267910, 10806598.

Davis, Jesse and Mark Goadrich (2006). "The Relationship between Precision-Recall and ROC Curves." In: *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*. New York, New York, USA: ACM Press, pp. 233–240. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143874.

De Saussure, Ferdinand (2011). *Course in General Linguistics*. Columbia University Press [1916].

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (1990). "Indexing by Latent Semantic Analysis." In: *Journal of the American Society for Information Science* 41.6, pp. 391–407. ISSN: 10974571. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Dereza, Oksana (2018). "Lemmatization for Ancient Languages: Rules or Neural Networks?" In: *Conference on Artificial Intelligence and Natural Language*. Springer, pp. 35–47.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Dietterich, Thomas G (1998). "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms." In: *Neural computation* 10.7, pp. 1895–1923.

Doan, AnHai, Alon Halevy, and Zachary Ives (2012). *Principles of Data Integration*. Elsevier. ISBN: 9780124160446. DOI: 10.1016/C2011-0-06130-6.

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

Dubin, David (Mar. 2004). "The Most Influential Paper Gerard Salton Never Wrote." In: *Library Trends* 52.

Eger, Steffen, Rüdiger Gleim, and Alexander Mehler (2016). "Lemmatization and Morphological Tagging in German and Latin: A Comparison and a Survey of the State-of-the-Art." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Paris, France: European Language Resources Association (ELRA), pp. 1507–1513. ISBN: 978-2-9517408-9-1.

Ehrman, Bart D (1997). *The New Testament: A historical introduction to the early Christian writings*. Oxford University Press Oxford.

Erjavec, Tomaž (2015). "Reference Corpus of Historical Slovene Goo300k 1.2." In:

Erwin, Harry and Michael Oakes (2012). "Correspondence Analysis of the New Testament." en. In: *Proceedings of the Language Resources and Evaluation Conference LREC 2012. Workshop on Language Resources and Evaluation for Religious Texts*. Istanbul, Turkey.

Farrell, Joseph (2005). "Intention and Intertext." In: *Phoenix-usa*. ISSN: 00318299.

Fellbaum, Christiane (Nov. 2012). "WordNet." In: *The Encyclopedia of Applied Linguistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc. ISBN: 9781405198431. DOI: 10.1002/9781405198431.wbeal1285. URL: http://doi.wiley.com/10.1002/9781405198431.wbeal1285.

Fleiss, Joseph L. (1971). "Measuring Nominal Scale Agreement among Many Raters." In: *Psychological Bulletin* 76.5, pp. 378–382. ISSN: 0033-2909. DOI: 10.1037/h0031619.

Forstall, Christopher W and Walter J Scheirer (2019). *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*. Springer.

Forstall, Christopher, Neil Coffee, Thomas Buck, Katherine Roache, and Sarah Jacobson (2015). "Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level n-Gram Matching." In: *Digital Scholarship in the Humanities* 30.4, pp. 503–515. ISSN: 2055768X. DOI: 10.1093/llc/fqu014.

Franzini, Greta, Marco Passarotti, Maria Moritz, and 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 CLiC-It 2018*. Torino: Accademia University Press.

Gaizauskas, Robert, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough, and Scott Piao (2001). "The METER Corpus: A Corpus for Analysing Journalistic Text Reuse." In: *Proceedings of the Corpus Linguistics 2001 Conference*. Vol. 1.

Gal, Yarin and Zoubin Ghahramani (2016). "A Theoretically Grounded Application of Dropout in Recurrent Neural Networks." In: *Advances in Neural Information Processing Systems*, pp. 1019–1027.

Ganascia, Jean-Gabriel, Peirre Glaudes, and Andrea Del Lungo (June 2014). "Automatic detection of reuses and citations in literary texts." In: *Literary and Linguistic Computing* 29.3, pp. 412–421. ISSN: 0268-1145. DOI: 10.1093/llc/fqu020. eprint: https://academic.oup.com/dsh/article-pdf/29/3/412/9951344/fqu020.pdf. URL: https://doi.org/10.1093/llc/fqu020.

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch (June 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*. Atlanta, Georgia: Association for Computational Linguistics, pp. 758–764.

Gao, Jianfeng, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng (Oct. 2014). "Modeling Interestingness with Deep Neural Networks." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 2–13. DOI: `10.3115/v1/D14-1002`. URL: `https://www.aclweb.org/anthology/D14-1002`.

Gelman, Andrew, Ben Goodrich, Jonah Gabry, and Aki Vehtari (2019). "R-squared for Bayesian Regression Models." In: *The American Statistician* 73.3, pp. 307–309. DOI: `10.1080/00031305.2018.1549100`. eprint: `https://doi.org/10.1080/00031305.2018.1549100`. URL: `https://doi.org/10.1080/00031305.2018.1549100`.

Gelman, Andrew and Jennifer Hill (2006). "Data Analysis Using Regression and Multilevel/Hierarchical Models." In: *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press. ISBN: 978-0-511-79094-2. DOI: `10.1017/CBO9780511790942`.

Genette, Gérard (1982). *Palimpsestes: La Littérature Au Second Degré*. Seuil.

Gesmundo, Andrea and Tanja Samardži (2012). "Lemmatisation as a Tagging Task." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 368–372.

Ghiban, Ioan Cristian and Ştefan Trăuşan-Matu (2013). "Network Based Analysis of Intertextual Relations." In: *Advances in Information Systems and Technologies*. Ed. by Álvaro Rocha, Ana Maria Correia, Tom Wilson, and Karl A. Stroetmann. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 753–762. ISBN: 978-3-642-36981-0.

Gorman, Vanessa (Dec. 2019). *Vgorman1/Greek-Dependency-Trees: Ancient Greek Prose Dependency Treebanks*. `https://github.com/vgorman1/Greek-Dependency-Trees/tree/1.0.1`. DOI: `10.5281/zenodo.3596076`.

Guo, Jiafeng, Yixing Fan, Qingyao Ai, and W. Bruce Croft (Oct. 2016). "A Deep Relevance Matching Model for Ad-Hoc Retrieval." In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM. DOI: `10.1145/2983323.2983769`.

Guo, Jiafeng, Yixing Fan, Xiang Ji, and Xueqi Cheng (2019). "Matchzoo: A learning, practicing, and developing system for neural text matching." In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1297–1300.

Harrington, J. Matthew et al. (Jan. 2021). *Perseids-Publications/Harrington-Trees: Release v2.0.1*. `https://github.com/perseids-publications/harrington-trees`. DOI: `10.5281/zenodo.4465074`.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). "The Elements of Statistical Learning." In: *The Elements of Statistical Learn-*

ing. Springer Series in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-84857-0. DOI: 10.1007/b94608.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long Short-Term Memory.” In: *Neural Computation* 9.8, pp. 1735–1780.

Hoffman, Matthew, Francis R Bach, and David M Blei (2010). “Online Learning for Latent Dirichlet Allocation.” In: *Advances in Neural Information Processing Systems*, pp. 856–864.

Hohl Trillini, Regula (Jan. 2018). *Casual Shakespeare : Three Centuries of Verbal Echoes*. en. Routledge. ISBN: 978-1-351-12094-4. DOI: 10.4324/9781351120944.

Hohl Trillini, Regula and Sixta Quassdorf (May 2010). “A ’Key to All Quotations’? A Corpus-Based Parameter Model of Intertextuality.” In: *Literary and Linguistic Computing* 25.3, pp. 269–286. ISSN: 0268-1145. DOI: 10.1093/llc/fqq003. eprint: https://academic.oup.com/dsh/article-pdf/25/3/269/2919622/fqq003.pdf.

Howard, Jeremy and Sebastian Ruder (July 2018). “Universal Language Model Fine-tuning for Text Classification.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 328–339. DOI: 10.18653/v1/P18-1031. URL: https://www.aclweb.org/anthology/P18-1031.

Hu, Baotian, Zhengdong Lu, Hang Li, and Qingcai Chen (2014). “Convolutional Neural Network Architectures for Matching Natural Language Sentences.” In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, pp. 2042–2050. URL: https://proceedings.neurips.cc/paper/2014/hash/b9d487a30398d42ecff55c228ed5652b-Abstract.html.

Huang, Po-Sen, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck (2013). “Learning deep structured semantic models for web search using clickthrough data.” In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 2333–2338.

Ihaka, Ross and Robert Gentleman (Sept. 1996). “R: A Language for Data Analysis and Graphics.” In: *Journal of Computational and Graphical Statistics* 5.3, pp. 299–314. ISSN: 1061-8600. DOI: 10.1080/10618600.1996.10474713.

Inan, Hakan, Khashayar Khosravi, and Richard Socher (2017). “Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling.” In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=r1aPbsFle.

Jaccard, Paul (1901). "Etude de la distribution florale dans une portion des Alpes et du Jura." In: *Bulletin de la Societe Vaudoise des Sciences Naturelles* 37, pp. 547–579. DOI: 10.5169/seals-266450.

Jänicke, Stefan, Thomas Efer, Marco Büchler, and Gerik Scheuermann (2015). "Designing Close and Distant Reading Visualizations for Text Re-Use." en. In: *Computer Vision, Imaging and Computer Graphics - Theory and Applications*. Ed. by Sebastiano Battiato, Sabine Coquillart, Julien Pettré, Robert S. Laramee, Andreas Kerren, and José Braz. Communications in Computer and Information Science. Cham: Springer International Publishing, pp. 153–171. ISBN: 978-3-319-25117-2. DOI: 10.1007/978-3-319-25117-2_10.

Juršic, Matjaz, Igor Mozetic, Tomaz Erjavec, and Nada Lavrac (2010). "Lemmagen: Multilingual lemmatisation with induced ripple-down rules." In: *Journal of Universal Computer Science* 16.9, pp. 1190–1214.

Keersmaekers, Alek, Wouter Mercelis, Colin Swaelens, and Toon Van Hal (Aug. 2019). "Creating, Enriching and Valorizing Treebanks of Ancient Greek." In: *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*. Paris, France: Association for Computational Linguistics, pp. 109–117. DOI: 10.18653/v1/W19-7812.

Kestemont, Mike, Guy De Pauw, Renske van Nie, and Walter Daelemans (2016). "Lemmatization for Variation-Rich Languages Using Deep Learning." In: *Digital Scholarship in the Humanities* 32.4, pp. 797–815.

Kestemont, Mike and Jeroen De Gussem (2017). "Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning." In: *J. Data Min. Digit. Humanit.* 2017.

Kingma, Diederik P. and Jimmy Lei Ba (2015). "Adam: A Method for Stochastic Optimization." In: *International Conference on Learning Representations 2015*, pp. 1–15. ISSN: 09252312. DOI: http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503.

Knauer, Georg Nikolaus (1965). "Die Aeneis Und Homer. Studien Zur Poetischen Technik Vergils Mit Listen Der Homerzitate in Der Aeneis." In: *The Classical World*. ISSN: 00098418. DOI: 10.2307/4345826.

Knowles, G. and Z. Mohd Don (2004). "The Notion of a "Lemma". Headwords, Roots and Lexical Sets." In: *International Journal of Corpus Linguistics* 9.1, pp. 69–81.

Kondratyuk, Daniel, Tomáš Gavenčiak, and Milan Straka (2018). "LemmaTag: Jointly Tagging and Lemmatizing for Morphologically-Rich Languages with BRNNs." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 4921–4928.

Korkiakangas, Timo and Matti Lassila (2013). "Abbreviations, Fragmentary Words, Formulaic Language: Treebanking Mediaeval Charter Material." In: *Proceedings of the Third Workshop on Annotation* *of Corpora for Research in the Humanities*. Ed. by F. Mambrini, M. Passarotti, and C. Sporleder, pp. 61–72.

Koster, Jeremy and Richard McElreath (Sept. 2017). "Multinomial analysis of behavior: statistical methods." In: *Behavioral Ecology and Sociobiology* 71.9, p. 138. ISSN: 0340-5443. DOI: 10.1007/s00265-017-2363-8. URL: http://link.springer.com/10.1007/s00265-017-2363-8.

Kristeva, Julia (1967). "Bakhtine, Le Mot, Le Dialogue et Le Roman." In: *Critique*.

Kruschke, John K. (May 2013). "Bayesian Estimation Supersedes the t Test." eng. In: *Journal of Experimental Psychology. General* 142.2, pp. 573–603. ISSN: 1939-2222. DOI: 10.1037/a0029146.

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger (2015). "From Word Embeddings to Document Distances." In: *International Conference on Machine Learning*, pp. 957–966.

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2018). "Word translation without parallel data." In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Laszuk, Dawid (2017). "Python Implementation of Empirical Mode Decomposition Algorithm." In:

LeCun, Yann and Yoshua Bengio (1995). "Convolutional networks for images, speech, and time series." In: *The handbook of brain theory and neural networks* 3361.10, p. 1995.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning." In: *nature* 521.7553, pp. 436–444.

Lee, John (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. ISBN: 0736-587X.

Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman (2014). "Mining of Massive Datasets." In: *Mining of Massive Datasets*. Cambridge: Cambridge University Press. ISBN: 978-1-139-92480-1. DOI: 10.1017/CBO9781139924801.

Levenshtein, Vladimir (1966). "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." In: *Soviet Physics Doklady*.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe (2009). "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." In: *Journal of Multivariate Analysis*. ISSN: 0047259X. DOI: 10.1016/j.jmva.2009.04.008.

Liebl, Bernhard and Manuel Burghardt (2020). "Shakespeare in the Vectorian Age: An Evaluation of Different Word Embeddings and NLP Parameters for the Detection of Shakespeare Quotes." In: *Proceedings of the the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*.

Linzen, Tal, Grzegorz Chrupała, and Afra Alishahi, eds. (Nov. 2018). *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and*

Interpreting Neural Networks for NLP. Brussels, Belgium: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W18-5400.

Linzen, Tal, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, eds. (Aug. 2019). *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W19-4800.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg (2016). "Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies." In: *Transactions of the Association for Computational Linguistics* 4, pp. 521–535.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Roberta: A robustly optimized bert pretraining approach." In: *arXiv preprint arXiv:1907.11692*. arXiv: 1907.11692.

Lu, Zhengdong and Hang Li (2013). "A Deep Architecture for Matching Short Texts." In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, pp. 1367–1375. URL: https://proceedings.neurips.cc/paper/2013/hash/8a0e1141fd37fa5b98d5bb769ba1a7cc-Abstract.html.

Lulu, Leena, Boumediene Belkhouche, and Saad Harous (Nov. 2016). "Overview of Fingerprinting Methods for Local Text Reuse Detection." In: *2016 12th International Conference on Innovations in Information Technology (IIT)*. IEEE, pp. 1–6. ISBN: 978-1-5090-5341-4. DOI: 10.1109/INNOVATIONS.2016.7880050.

Lund, Jeffrey, Piper Armstrong, Wilson Fearn, Stephen Cowley, Emily Hales, and Kevin Seppi (June 2019). "Cross-referencing Using Fine-grained Topic Modeling." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3978–3987. DOI: 10.18653/v1/N19-1399. URL: https://www.aclweb.org/anthology/N19-1399.

Mambrini, Francesco (Jan. 2020). *The Perseids Project: Daphne Trees*. https://perse-ids-publications.github.io/daphne-trees/.

Manjavacas Arévalo, Enrique, Ákos Kádár, and Mike Kestemont (June 2019). "Improving Lemmatization of Non-Standard Languages with Joint Learning." In: *Proceedings of the 2019 Conference of the North*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1493–1503. DOI: 10.18653/v1/N19-1153. URL: https://www.aclweb.org/anthology/N19-1153.

Manjavacas Arévalo, Enrique, Folgert Karsdorp, and Mike Kestemont (2020). "A Statistical Foray into Contextual Aspects of Intertextuality." In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)* (Amsterdam, The Netherlands, Nov. 18, 2020–Nov. 20, 2020). CEUR Workshop Proceedings 2723. Aachen, pp. 77–96. URL: http://ceur-ws.org/Vol-2723/long28.pdf.

Manjavacas Arévalo, Enrique and Mike Kestemont (2021). *Evaluation in Text Reuse Detection for Literary Texts*. Forthcoming.

Manjavacas Arévalo, Enrique, Brian Long, and Mike Kestemont (June 2019). "On the Feasibility of Automated Detection of Allusive Text Reuse." In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 104–114. DOI: 10.18653/v1/W19-2514. URL: https://www.aclweb.org/anthology/W19-2514.

Manjavacas Arévalo, Enrique, Laurence Mellerin, and Mike Kestemont (2021). *Quantifying the Utility of Text Reuse Detection Algorithms through Bayesian Inter-annotator Agreement Indices*. Forthcoming.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze (2008). "Introduction to Information Retrieval." In: *Introduction to Information Retrieval*. Cambridge: Cambridge University Press. ISBN: 978-0-511-80907-1. DOI: 10.1017/CBO9780511809071.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky (June 2014). "The Stanford CoreNLP Natural Language Processing Toolkit." In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, pp. 55–60. DOI: 10.3115/v1/P14-5010. URL: https://www.aclweb.org/anthology/P14-5010.

Matthiessen, Francis Otto (1968). *American renaissance: Art and expression in the age of Emerson and Whitman*. Vol. 230. Oxford University Press.

McCarthy, Philip M and Danielle S McNamara (2012). "The user-language paraphrase corpus." In: *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches*. IGI Global, pp. 73–89.

McElreath, Richard (Jan. 2018). *Statistical Rethinking*. Chapman and Hall/CRC. ISBN: 9781315372495. DOI: 10.1201/9781315372495. URL: https://www.taylorfrancis.com/books/9781315362618.

Mcguire, Brian Patrick (Nov. 2007). "Bernard of Clairvaux." In: *A Companion to Philosophy in the Middle Ages*. Oxford, UK: Blackwell Publishing Ltd, pp. 209–214. ISBN: 978-0-470-99666-9. DOI: 10.1002/9780470996669.ch28.

Mellerin, Laurence (2013). "Methodological Issues in Biblindex, an Online Index of Biblical Quotations in Early Christian Literature." In: *Biblical Quotations in Patristic Texts (Papers Presented at the Sixteenth*

International Conference on Patristic Studies Held in Oxford 2011). Ed. by Laurence Mellerin, Markus Vinzent, and Hugh Houghton. Vol. 2. Studia Patristica. Peeters, pp. 11–32.

Mellerin, Laurence (2014). "New Ways of Searching with Biblindex, the Online Index of Biblical Quotations in Early Christian Literature." In: *Digital Humanities in Biblical, Early Jewish and Early Christian Studies*. Ed. by Claire Clivaz, Gregory Andrew, and Hamidovic David. Vol. 2. Digital Humanities in Biblical, Early Jewish and Early Christian Studies. Brill, pp. 175–192. ISBN: 978-90-04-26432-8. DOI: 10.1163/9789004264434_012.

Mesoudi, Alex (2011). *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. University of Chicago Press.

Metzler, Donald, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel (2005). "Similarity Measures for Tracking Information Flow." en. In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management - CIKM '05*. Bremen, Germany: ACM Press, p. 517. ISBN: 978-1-59593-140-5. DOI: 10.1145/1099554.1099695.

Migne, Jacques Paul (1844-1855 (and 1862-1865)). *Patrologiae Cursus Completus. Series Latina (217 + 4 Vols.)* Garnier frères.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space." In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (June 2013). "Linguistic Regularities in Continuous Space Word Representations." In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751. URL: https://www.aclweb.org/anthology/N13-1090.

Minozzi, Stefano (2010). "The Latin WordNet Project." In: *Akten Des 15. Internationalen Kolloquiums Zur Lateinischen Linguisti*. Ed. by Peter Anreiter and Manfred Kienpointner. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck Bereich Sprachwissenschaft, pp. 707–716.

Mitra, Bhaskar, Fernando Diaz, and Nick Craswell (2017). "Learning to Match using Local and Distributed Representations of Text for Web Search." In: *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. Ed. by Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich. ACM, pp. 1291–1299. DOI: 10.1145/3038912.3052579. URL: https://doi.org/10.1145/3038912.3052579.

Moretti, Franco (2000). "Conjectures on World Literature." In: *New left review* 1. URL: `https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature`.

Moritz, Maria, Johannes Hellrich, and Sven Büchel (2018). "A Method for Human-Interpretable Paraphrasticality Prediction." In: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 113–118.

Moritz, Maria, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler (2016). "Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and Its Application to Bible Reuse." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, pp. 1849–1859. ISBN: 978-1-945626-25-8. DOI: `10.18653/v1/d16-1190`.

Moyise, Steve (2002). "Intertextuality and Biblical Studies: A Review." In: *Verbum et ecclesia* 23.2, pp. 418–431.

Mueller, Jonas and Aditya Thyagarajan (2016). "Siamese Recurrent Architectures for Learning Sentence Similarity." In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. Ed. by Dale Schuurmans and Michael P. Wellman. AAAI Press, pp. 2786–2792. URL: `http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12195`.

Nadeau, Claude and Yoshua Bengio (Sept. 2003). "Inference for the Generalization Error." en. In: *Machine Learning* 52.3, pp. 239–281. ISSN: 1573-0565. DOI: `10.1023/A:1024068626366`.

Needleman, Saul B. and 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, pp. 443–453. ISSN: 00222836. DOI: `10.1016/0022-2836(70)90057-4`.

Ng, Andrew (2016). "What artificial intelligence can and can't do right now." In: *Harvard Business Review* 9.11.

Nivre, Joakim et al. (2016). "Universal Dependencies v1: A Multilingual Treebank Collection." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA).

Olson, Charles (1947). *Call Me Ishmael*. Reynall and Hitchcock.

Orr, Mary (2003). *Intertextuality: Debates and Contexts*. Polity Press, p. 246. ISBN: 0-7456-0621-0.

Orr, Mary (Dec. 2010). "Intertextuality." In: *The Encyclopedia of Literary and Cultural Theory*. Oxford, UK: John Wiley & Sons, Ltd. DOI: `10.1002/9781444337839.wbelctv2i002`.

Ott, Michael (1911). "Peter Cellensis." In: *The Catholic Encyclopedia*. Robert Appleton Company, Vol. 11.

Pang, Liang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng (2016). "Text Matching as Image Recognition." In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence,* *February 12-17, 2016, Phoenix, Arizona, USA*. Ed. by Dale Schuurmans and Michael P. Wellman. AAAI Press, pp. 2793–2799. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11895.

Passarotti, Marco, Marco Budassi, Eleonora Litta, and Paolo Ruffolo (May 2017). "The Lemlat 3.0 Package for Morphological Analysis of Latin." In: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Gothenburg: Linköping University Electronic Press, pp. 24–31.

Pearl, Judea and Dana Mackenzie (2018). *The Book of Why. The New Science of Cause and Effect*. New York: Basic Books. ISBN: 978-0-465-09760-9.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations." In: *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1. Association for Computational Linguistics, pp. 2227–2237. ISBN: 978-1-948087-27-8. DOI: 10.18653/v1/n18-1202. arXiv: 1802.05365. URL: http://aclweb.org/anthology/N18-1202.

Pettersson, Eva, Beáta Megyesi, and Joakim Nivre (2014). "A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text." In: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Association for Computational Linguistics, pp. 32–41. DOI: 10.3115/v1/W14-0605.

Piotrowski, Michael (2012). "Natural Language Processing for Historical Texts." In: *Synthesis Lectures on Human Language Technologies* 5.2, pp. 1–159. ISSN: 19474040. DOI: 10.2200/S00436ED1V01Y201207HLT017.

Potthast, Martin, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein (2013a). "Overview of the 5th International Competition on Plagiarism Detection." In: *CEUR Workshop Proceedings*.

Potthast, Martin, Matthias Hagen, Michael Völske, Jakob Gomoll, and Benno Stein (Sept. 2012). *Webis Text Reuse Corpus 2012*. DOI: 10.5281/zenodo.1341602. URL: https://doi.org/10.5281/zenodo.1341602.

Potthast, Martin, Matthias Hagen, Michael Völske, and Benno Stein (Aug. 2013b). "Crowdsourcing Interaction Logs to Understand Text Reuse from the Web." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 1212–1221.

Potthast, Martin, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso (Aug. 2010). "An Evaluation Framework for Plagiarism Detection." In: *Coling 2010: Posters*. Beijing, China: Coling 2010 Organizing Committee, pp. 997–1005.

Potthast, Martin, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso (2009). "Overview of the 1st International Competition on Plagiarism Detection." In: *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*, p. 1.

Potthast, Martin, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso (June 2011). *PAN Plagiarism Corpus 2011 (PAN-PC-11)*. DOI: 10.5281/zenodo.3250095. URL: https://doi.org/10.5281/zenodo.3250095.

Press, Ofir and Lior Wolf (Apr. 2017). "Using the Output Embedding to Improve Language Models." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 157–163. URL: https://www.aclweb.org/anthology/E17-2025.

Regneri, Michaela and Rui Wang (July 2012). "Using Discourse Information for Paraphrase Extraction." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 916–927.

Rehurek, Radim and Petr Sojka (2010). "Software Framework for Topic Modelling with Large Corpora." In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50. ISSN: 2951740867.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the Space of Topic Coherence Measures." In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*. New York, New York, USA: ACM Press, pp. 399–408. ISBN: 978-1-4503-3317-7. DOI: 10.1145/2684822.2685324.

Roelli, Philipp (2014). "The Corpus Corporum, a New Open Latin Text Repository and Tool." In: *Bulletin du Cange - Archivum Latinitatis Medii Aevi*. ISSN: 09948090. DOI: 10.5167/uzh-171105.

Rus, V., Rajendra Banjade, and Mihai C. Lintean (2014). "On Paraphrase Identification Corpora." In: *LREC*.

Salmi, Hannu, Petri Paju, Heli Rantala, Asko Nivala, Aleksi Vesanto, and Filip Ginter (Sept. 2020). "The Reuse of Texts in Finnish Newspapers and Journals, 1771–1920: A Digital Humanities Perspective." In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pp. 1–15. ISSN: 0161-5440. DOI: 10.1080/01615440.2020.1803166.

Sari, Yunita, Mark Stevenson, and Andreas Vlachos (2018). "Topic or Style? Exploring the Most Useful Features for Authorship Attribution." In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Association for Computational Linguistics, pp. 343–353.

Scheirer, Walter, Christopher Forstall, and Neil Coffee (Apr. 2016). "The sense of a connection: Automatic tracing of intertextuality by meaning." In: *Digital Scholarship in the Humanities* 31.1, pp. 204–217.

ISSN: 2055-7671. DOI: 10.1093/llc/fqu058. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqu058.

Schmid, Helmut (2013). "Probabilistic Part-of-Speech Tagging Using Decision Trees." In: *New Methods in Language Processing*, p. 154.

Schulz, Sarah, Guy De Pauw, Orphée De Clercq, Bart Desmet, Veronique Hoste, Walter Daelemans, and Lieve Macken (2016). "Multimodular Text Normalization of Dutch User-Generated Content." In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.4, p. 61.

Schütze, Hinrich, Christopher D Manning, and Prabhakar Raghavan (2008). *Introduction to Information Retrieval*. Vol. 39. Cambridge University Press.

Scott, William A. (1955). "Reliability of Content Analysis: The Case of Nominal Scale Coding." In: *Public Opinion Quarterly* 19.3, p. 321. ISSN: 0033362X. DOI: 10.1086/266577. URL: https://academic.oup.com/poq/article-lookup/doi/10.1086/266577.

Seo, Jangwon and W. Bruce Croft (2008). "Local Text Reuse Detection." In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '08*. New York, New York, USA: ACM Press, p. 571. ISBN: 978-1-60558-164-4. DOI: 10.1145/1390334.1390432.

Shen, Yelong, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil (2014). "Learning semantic representations using convolutional neural networks for web search." In: *Proceedings of the 23rd international conference on world wide web*, pp. 373–374.

Sidorov, Grigori, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto (2014). "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model." In: *Computacion y Sistemas* 18.3, pp. 491–504. ISSN: 14055546. DOI: 10.13053/CyS-18-3-2043.

Smith, David A., Ryan Cordel, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson (2014). "Detecting and Modeling Local Text Reuse." In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pp. 183–192. ISBN: 978-1-4799-5569-5. DOI: 10.1109/JCDL.2014.6970166.

Smith, David A., Ryan Cordell, and Elizabeth Maddock Dillon (Oct. 2013). "Infectious texts: Modeling text reuse in nineteenth-century newspapers." In: *2013 IEEE International Conference on Big Data*. IEEE, pp. 86–94. ISBN: 978-1-4799-1293-3. DOI: 10.1109/BigData.2013.6691675. URL: http://ieeexplore.ieee.org/document/6691675/.

Smith, T.F. and M.S. Waterman (Mar. 1981). "Identification of Common Molecular Subsequences." In: *Journal of Molecular Biology* 147.1, pp. 195–197. ISSN: 00222836. DOI: 10.1016/0022-2836(81)90087-5.

Søgaard, Anders, Anders Johannsen, Barbara Plank, Dirk Hovy, and Héctor Martínez Alonso (2014). "What's in a p-Value in NLP?" In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 1–10. DOI: 10.3115/v1/W14-1601.

Sprugnoli, Rachele, Marco Passarotti, and Giovanni Moretti (2019). "Vir is to Moderatus as Mulier is to Intemperans-Lemma Embeddings for Latin." In: *CLiC-it*.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." In: *Journal of Machine Learning Research* 15, pp. 1929–1958. ISSN: 15337928. DOI: 10.1214/12-AOS1000.

Stahlberg, Lesleigh Cushing (2008). *Sustaining Fictions: Intertextuality, Midrash, Translation, and the Literary Afterlife of the Bible*. Vol. 486. A&C Black.

Symonds, Matthew R. E. and Adnan Moussalli (Jan. 2011). "A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion." In: *Behavioral Ecology and Sociobiology* 65.1, pp. 13–21. ISSN: 0340-5443. DOI: 10.1007/s00265-010-1037-6. URL: http://link.springer.com/10.1007/s00265-010-1037-6.

Szymański, Piotr and Kyle Gorman (2020). "Is the Best Better? Bayesian Statistical Model Comparison for Natural Language Processing." en. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2203–2212. DOI: 10.18653/v1/2020.emnlp-main.172.

TEI Consortium, eds (2020). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. URL: http://www.tei-c.org/Guidelines/P5/ (visited on 02/16/2021).

Tang, Gongbo, Fabienne Cap, Eva Pettersson, and Joakim Nivre (2018). "An Evaluation of Neural Machine Translation Models on Historical Spelling Normalization." In: *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 1320–1331.

Tenney, Ian et al. (May 15, 2019). "What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations." In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. arXiv: 1905.06316. URL: http://arxiv.org/abs/1905.06316 (visited on 03/16/2021).

Tompkins, J.P. (1980). "An Introduction to Reader-Response Criticism." In: *Reader-Response Criticism. From Formalism to Post-Structuralism*. Johns Hopkins University Press, pp. ix–xxvi.

Van Reenen, P. and M. Mulder (1993). "Een Gegevensbank van 14de-Eeuwse Middelnederlandse Dialecten Op Computer." In: *Lexikos* 3, pp. 259–279.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). "Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC." In: *Statistics and Computing*. ISSN: 15731375. DOI: 10.1007/s11222-016-9696-4.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2018). "Loo: Efficient Leave-One-out Cross-Validation and WAIC for Bayesian Models." In: *R package version* 2.0, p. 1003.

Verbart, André (Jan. 1995). *Fellowship in* Paradise Lost*: Vergil, Milton, Wordsworth*. en. Brill Rodopi. ISBN: 978-90-5183-882-4.

Vesanto, Aleksi, Asko Nivala, Heli Rantala, Tapio Salakoski, Hannu Salmi, and Filip Ginter (May 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*. Gothenburg: Linköping University Electronic Press, pp. 54–58.

Voloshinov, Valentin Nikolaevich and Michail M Bakhtin (1986). *Marxism and the Philosophy of Language*. Harvard University Press.

Voorhees, Ellen M (1999). "The TREC-8 Question Answering Track Report." In: *TREC 8*. DOI: 10.1017/S1351324901002789.

Wan, Shengxian, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng (2016). "A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations." In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. Ed. by Dale Schuurmans and Michael P. Wellman. AAAI Press, pp. 2835–2841. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11897.

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu (2015). "Towards Universal Paraphrastic Sentence Embeddings." In: *CoRR* abs/1511.0.

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Wong, S. K. M., Wojciech Ziarko, and Patrick C. N. Wong (1985). "Generalized vector spaces model in information retrieval." In: *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '85*. New York, New York, USA: ACM Press, pp. 18–25. ISBN: 0897911598. DOI: 10.1145/253495.253506. URL: http://portal.acm.org/citation.cfm?doid=253495.253506.

Wong, S. K.M., W. Ziarko, V. V. Raghavan, and P. C.N. Wong (June 1987). "On modeling of information retrieval concepts in vector spaces." In: *ACM Transactions on Database Systems* 12.2, pp. 299–321. ISSN: 0362-5915. DOI: 10.1145/22952.22957. URL: https://dl.acm.org/doi/10.1145/22952.22957.

Xiao, Chuan, Wei Wang, Xuemin Lin, Jeffrey Xu Yu, and Guoren Wang (2011). "Efficient similarity joins for near-duplicate detection." In: *ACM Transactions on Database Systems* 36.3, pp. 1–41. ISSN: 03625915. DOI: 10.1145/2000824.2000825.

Yale-DHLab (2017). *Intertext.* `https://github.com/YaleDHLab/intertext`.

Yousef, Tariq and Stefan Jänicke (Feb. 2021). "A Survey of Text Alignment Visualization." In: *IEEE Transactions on Visualization and Computer Graphics* 27.2, pp. 1149–1159. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: `10.1109/TVCG.2020.3028975`.

Zhao, Zhe, Tao Liu, Shen Li, Bofang Li, and Xiaoyong Du (Sept. 2017). "Ngram2vec: Learning Improved Word Representations from Ngram Co-occurrence Statistics." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 244–253. DOI: `10.18653/v1/D17-1023`. URL: `https://www.aclweb.org/anthology/D17-1023`.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.

van Halteren, Hans and Margit Rem (Dec. 2013). "Dealing with Orthographic Variation in a Tagger-Lemmatizer for Fourteenth Century Dutch Charters." In: *Language Resources and Evaluation* 47.4, pp. 1233–1259. ISSN: 1574-0218. DOI: `10.1007/s10579-013-9236-1`.

vor der Brück, Tim, Steffen Eger, and Alexander Mehler (2015). "Lexicon-Assisted Tagging and Lemmatization in Latin: A Comparison of Six Taggers and Two Lemmatization Models." In: *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 105–113. DOI: `10.18653/v1/W15-3716`.