A Statistical Foray into Contextual Aspects of Intertextuality

Enrique Manjavacas[a], Folgert Karsdorp[b] and Mike Kestemont[a]

[a] Department of Literature, University of Antwerp, Antwerp, Belgium
[b] Royal Netherlands Academy of Arts and Sciences, Meertens Institute, Amsterdam, The Netherlands

Abstract
Intertextuality is a highly productive concept in literary theory. The pervasiveness of intertextuality in literary texts has lead 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 intertexts: the degree of lexical similarity, and the degree to which intertexts are thematically embedded in the context. Furthermore, 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? Secondly, 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.

Keywords
Intertextuality, Text Reuse, Multi-level Modeling

1. Introduction

Intertextuality is a well-known concept from literary studies that is commonly applied to texts across various periods and languages [34, 2]. Originally proposed by post-structuralist literary theorist, Julia Kristeva [24], 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

CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands
enrique.manjavacas@gmail.com (E. Manjavacas); folgert.karsdorp@meertens.knaw.nl (F. Karsdorp); mike.kestemont@uantwerpen.be (M. Kestemont)
0000-0002-3942-7680 (E. Manjavacas); 0000-0002-5958-0551 (F. Karsdorp); 0000-0003-3590-693X (M. Kestemont)

the intertextual approach has considerable merit, as it sheds light on how texts participate in the discursive space of a culture [12]. 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 [33]. 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 (e.g., the role of the *Odyssee* in Virgil's *Aeneis* or Joyce's *Ulysses*—cases of what Genette terms "hypertextuality" [18]), to more localized phenomena such as motifs or allusions, in which the link is established from and to specific passages.

In order to situate computational approaches to intertextuality within this spectrum, Forstall and Scheirer [16] 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". These localized intertextual links—or "loci similes" in more traditional terms—have been categorized along different axes such as intentionality [15, 23, 9] –, function – parodic vs. satirical and non-satirical [18] –, or "literality" – quotation vs. mention or allusion. This taxonomic activity has lead to a considerable amount of intertext typologies, highlighting the complexity of the underlying phenomena.

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 [6, 25, 8, 43].[1] However, a variety of *contextual* factors can be easily thought of as *conditioning* the location, source and type of an intertextual link.

With no aim of exhaustiveness, it could 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 [31, 1, 11].) 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

[1]There are certainly exceptions. For example, Bamman and Crane [3] exploits syntactic information (dependency paths and word order) to extract allusions in classical Latin literature, Scheirer et al. [40] use Latent Semantic Indexing [13] to extract parallels in Latin epic, Lund et al. [27] uses local topical information extracted from anchor-based topic models to extract intra-biblical references, and Manjavacas et al. [29] examine the application of distributional semantics to help improve the detection of allusions.

“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 and considering both axes: i.e. the degree of “literality” (quotational vs. allusive) and its embedding in the thematic context.

Thus, 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—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 large number of authors and books, and known to be abounding in intertextual links. Two facts about the Patrology are worth advancing (the corpus will be thoroughly introduced in Section 2): on the one hand, the majority of authors form part of the same writing tradition sharing themes, concerns and theoretical background, which makes them commensurable from a statistical point of view. On the other hand, the main source of reference, the Bible, is shared. These two aspects will allow us to approach some of the alluded questions from a data-driven perspective.

Research Questions 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 of the paper The remaining of the present paper is structured as follows. First, Section 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 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 4 we describe the statistical models used to approach the posited questions. Finally, in Section 5, we discuss the insights that can be drawn from the models and the answers that they deliver.

2. Dataset

2.1. Sources

The main dataset used in the present paper 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 endeavour was initially undertaken by J.P. Migne between 1841 and 1855 [32]. 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 (2rd 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, mainly covering the period from late antiquity until the high medieval period. This period 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 [33], in particular, pervades the Patrology’s texts. This is partly due to the considerable number of sermons included (which departed from or even revolved around specific biblical quotations), but 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 [19].

2.2. Curation

The digital version of the Patrology was extracted from the Corpus Corporum collection [38], 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 [10]. 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.

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 shown in Listing 1 these have been kept in its original inline form, and can be extracted automatically through costumary data-wrangling techniques[2].

```
<p>Simili modo et tu, si bona
quae habes forti cautela custodire non negligis,
<pb n="0773B"/>
```

[2]In particular, we apply regular expressions to match on parenthesis formatted in the manner specified in Listing 1 and check whether the book abbreviation is in a manually curated list. In the case of a positive match, we then try to parse the chapter and verse numbers. Finally, the parsed reference is checked against the vulgate to see whether it corresponds to a real verse.

```
4  circa tabernaculum tuum, et ea quae intra illud
5  sunt tentoria suspendis. Nihil enim omnino tibi proderit
6  bona in te spiritualia congregasse, nisi diligenti
7  ea et sollicita circumspectione custodias. Hinc
8  in sacra Scriptura legimus, quia <i>posuit Deus hominem
9  in paradiso, ut operaretur, et custodiret illum (Gen. II, 15)</i>.
10 In paradiso quippe Deus hominem
11 ponit, quando delectabilem tibi spiritualium gratiarum
12 copiam gratuito largiens, in sancta et tranquilla
13 conscientia suaviter te pausare facit.</p>
```

Listing 1: Example of an XML source file snippet from Adam Scotus, corresponding to *De tripartito tabernaculo*, showcasing a passage containing an annotation of a biblical reference (Gen, II, 15.) in line number 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. While the OCR is not perfect, and the annotation cannot be deemed exhaustive, a manual inspection of a representative sample indicates that the automatic procedure manages to parse editorial annotations with high precision. More concretely, we isolated a sample of 100 instances which showed a low alignment score according to the Smith-Waterman algorithm [44], and carefully checked for the alleged reference.[3] 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.

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 [41]. For lemmatization, we use a neural network-based lemmatizer trained with PIE [28] on a corpus of medieval Latin (Capitularia), that has served as the basis to a number of Latin lemmatization studies [14, 5, 22]. As opposed to TreeTagger's lemmatizer, the neural-network based lemmatizer is able to analyze previously unseen types and is able to disambiguate between possible alternative analyses, which, as shown in the Appendix, results in more coherent topics.

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 100K tokens and from their writings we sample books with at least 100 references to the Bible, making sure that at least two books per author are held out for

[3]After a first inspection of the distribution of scores, it could be observed that scores higher than 4 consistently yielded true references, therefore these were excluded from the sample to be manually checked. Note that the exact number of this threshold is dependent on the algorithm parameters and cannot be interpreted independently.

developing purposes. From this 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.) 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 2.2.2. In total, we collected 15,195 biblical references across 24 authors.

The remaining documents of the Patrology are set apart and used for training a topic model that will be used in order to automatically capture the theme in a given passage.

2.4. Topic Modeling

An LDA topic model [4] was trained on the lemmatized version of the remaining dataset, comprising 103,687,454 tokens. The topic model was trained using the `gensim` package [36], which provides an implementation of Online LDA [20]. We fit an LDA model after removal of all words that were not strictly alphanumeric, any word that was not identified as an adjective, adverb, noun or verb, and all words that appear in a specifically designed stopword list.[4]. The hyper-parameters of the LDA algorithm were further selected on the basis of a validation study that used grid-search with the objective of maximizing topical coherence [37] on the held-out dataset. The results of the validation study are reported in the Appendix. The resulting topic model was fit on document snippets of 1,500 words, 200 topics and a vocabulary truncated to the 20,000 most frequent words.

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 a 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 common and expected direction, especially if the intertextual link is based on high lexical overlap. In order to disentangle topical from lexical similarity, the topic distributions are inferred on the original document after removal of the lexical overlap with respect to the linked document. In such case, a strong match in the respective inferred topic distributions can be interpreted as an indication of a high thematic embedding in the context on the basis that the lexical choices made in the context point at terms that co-occur in similar topics as the terms that establish the link.

3.1. Lexical Similarity

In order to extract lexical similarity, we resort to traditional methods from the text re-use literature – see, for instance [42]. We focus on the Jaccard coefficient, which is defined as the number of shared words divided by the total number of words types in the documents.

[4]This wordlist includes terms such as "dominus", "deus", etc. that were deemed irrelevant for composing topic-term distributions due to their high frequency. The wordlist together with all relevant code will be published upon publication.

In the present study, we compute the weighted version of the Jaccard coefficient, shown in Equation 3.1, which gives a more accurate value by taking into account the frequency of the words:

$$J(D_i, D_j) = \sum_{w \in D_i \cup D_j} \frac{min[c(w, D_i), c(w, D_j)]}{max[c(w, D_i), c(w, D_j)]} \tag{1}$$

In Equation 3.1 $c(w, D_i)$ refers to the number of times word w appears in document D_i. In order to weight higher the influence of 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 lemmas and apply a stopword list.[5]

3.2. Topical Similarity

Given the inferred topic distributions of given source and target documents, we resorted to information-theoretic measures relating to the distribution entropies to estimate the topic similarity of the underlying documents. In particular, we use the Jensen-Shannon divergence, shown in Equation 2:

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

which corresponds to the arithmetic mean between the Kullback-Leiber divergence 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 reverse. By taking the mean, JSD transforms the KL Divergence into a symmetric measure. Since JSD is a divergence, we transform it into a similarity by substracting 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.

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 2.4 to identify the most dominant topic in the inferred topic distribution of a given passage. Thus, a given document D_i is assigned an index pointing to topic t with highest probability in the topic distribution inferred for document D_i:

$$\underset{t}{\operatorname{argmax}}\, \theta^t_{D_i}$$

By taking such summary of the distribution we are certainly ignoring important information about the 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, it simplifies the statistical modeling considerably, while still capturing a considerable amount of topic information.

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

Table 1

Summary statistics of the model comparison displaying leave-one-out estimates of the expected log predictive density (ELPD)—lower is better, estimates of the effective number of parameters (P), and difference 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.

Model	ELPD	ELPD (SE)	P	P (SE)	DIFF	DIFF (SE)
$M_{A \cup B \cup T}$	-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
$M_{B \cup T}$	-38430.3	264.0	257.4	6.6	-553.5	32.2
$M_{A \cup T}$	-39971.3	294.4	106.7	2.6	-2094.6	116.7

4. Data Analysis

We approach the research questions by making use of multivariate multi-level intercept-only model using lexical similarity (`lex`) from Section 3.1, and topic similarity (`topic`) from Section 3.2 as 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 each of the grouping factors. The number of levels per grouping factors amounted to the following: author (`A`: 24 levels), biblical book (`B`: 52 levels) and dominant topic (`T`: 129 levels).[6]

We conducted all analyses in `R` (3.6.3) [21] using the `brms` package [7] for model fitting. We chose weakly informative priors as per the defaults in the `brms` package, unless otherwise specified.[7] Throughout the experiment, model convergence was checked on the basis of R-hat values, number of effective samples and trace plots.

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

4.1. Model definition

The general model including all grouping factors is defined by Equation 3. The statistical model consists in a bi-variate model that includes no predictors[8], and groups observations according to three different criteria. Observations are modeled as coming from a bi-variate normal. The means are decomposed into grand means, a_l and a_t, and group-specific deviations from the mean a_l^K, a_t^K. Furthermore, the latter are modeled hierarchically as distributed themselves

[6]Note that this number corresponds to the actual number of dominant topics that appear in the dataset and therefore diverge from the total number of topics fit. This situation arises since not all of the 200 estimated topics are realized as dominant topic in the target dataset.

[7]At the time of running the experiments, these priors were Student T distributions with 3 degrees of freedom, location of 0, and a scale of 2,5.

[8]Note, however, that here is nothing inherent to the research design that prevents from including predictors. For instance, future work may want to improve the model by considering the influence of time, genre or the density of references in the surrounding passage.

according to a second multivariate normal centered around zero. Finally, following Gelman and Hill (2006, Chapter 13), covariances Σ and Σ^K are decomposed into a diagonal matrix of standard deviations that model lexical and topical variation individually and a correlation matrix that additionally targets correlations between both response variables.

$$
\begin{aligned}
\begin{bmatrix} y_l \\ y_t \end{bmatrix} &\sim \text{MVNormal}(\begin{bmatrix} \mu_l \\ \mu_t \end{bmatrix}, \Sigma) \\
\Sigma &= \begin{pmatrix} \sigma_l & 0 \\ 0 & \sigma_t \end{pmatrix} R \begin{pmatrix} \sigma_l & 0 \\ 0 & \sigma_t \end{pmatrix} \\
\begin{bmatrix} \mu_l \\ \mu_t \end{bmatrix} &= \begin{bmatrix} a_l \\ a_t \end{bmatrix} + \begin{bmatrix} a_l^A \\ a_t^A \end{bmatrix} + \begin{bmatrix} a_l^B \\ a_t^B \end{bmatrix} + \begin{bmatrix} a_l^T \\ a_t^T \end{bmatrix} \\
\begin{bmatrix} a_l^K \\ a_t^K \end{bmatrix} &\sim \text{MVNormal}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma^K) \\
\Sigma^K &= \begin{pmatrix} \sigma_l^K & 0 \\ 0 & \sigma_t^K \end{pmatrix} R^K \begin{pmatrix} \sigma_l^K & 0 \\ 0 & \sigma_t^K \end{pmatrix}
\end{aligned}
\tag{3}
$$

In Equation 3, y_l and y_t refer to the lexical and topical outcome variables, a_l^K and a_t^K refer to the varying intercepts for the lexical and topical similarity for factor K. Finally, we set the priors of all σ terms to student-t priors and the correlation term R to a flat LKJ prior [26].

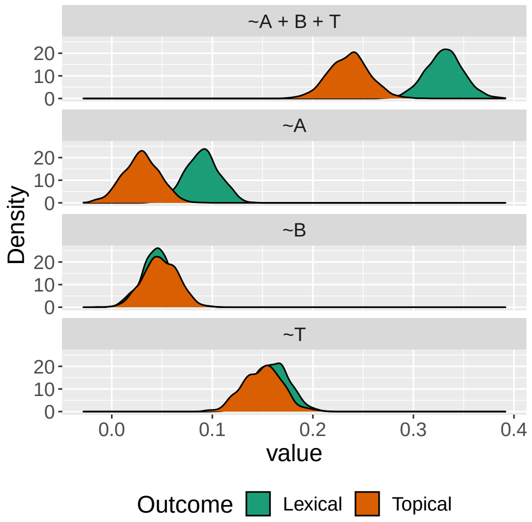
4.2. Model comparison

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 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 et al. (2017) for a description of the method and Vehtari et al. (2018) for an implementation in the **R** programming language.

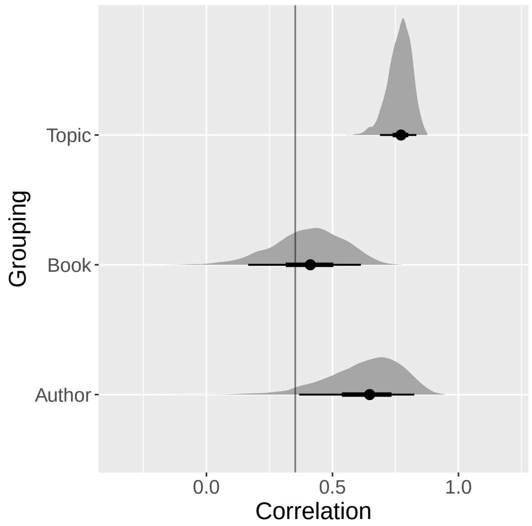
Table 1 shows the results of the comparison. As we can see, the model utilizing all grouping factors ($M_{A \cup B \cup T}$) is expected to have much better predictive performance than any of the single-grouping models. For the individual factor models, we observe that theme-level grouping has stronger explanatory power than author-level or book-level grouping, with the latter two receiving ELPDs within error of each other.

In order to better grasp the respective contribution of book-level and author-level groupings to the model's predictive performance, we fitted $M_{A \cup T}$ and $M_{B \cup T}$ and compared them to the general $M_{A \cup B \cup T}$. The results of the comparison are shown in the last two rows of Table 1. 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 modelling power of the different groupings by inspecting estimates of explained variance. For generality, our estimates are computed by subtracting a reference variance from the variance in the samples drawn from the posterior



(a) Explained variance estimates for different grouping factors with respect to the reference model with no grouping using model $M_{A \cup B \cup T}$. $\sim A+B+T$ refers to the model including all random effects. $\sim K$ refers to the model ignoring all random effects except K.



(b) Posterior correlation estimate of the outcomes at the different grouping factors including mean and 0.5 and 0.89 credible intervals. The vertical line depicts the overall empirical correlation between centered variables.

Figure 1: A comparison of explained variance estimates and posterior correlation estimates for different grouping factors in model $M_{A \cup B \cup T}$

predictive distribution of the general model ($M_{A \cup B \cup T}$) when considering different combinations of groupings. The reference variance corresponds to the variance in samples drawn from the posterior predictive when ignoring all groupings.

Figure 1a shows the results decomposed into the two different outcome variables considering all groupings ($\sim A + B + T$) and the individual groupings ($\sim A$, $\sim B$ and $\sim T$). As we can see, while both book-level and topic-level groupings have an approximately equal estimate of the explained variance for the lexical and topical outcomes, author-level grouping seems to explain a larger share than 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 yielded the smallest out-of-sample predictive performance estimates, we can only postulate a mild authorship signal.

4.3. Inspection of groupings

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 1, our analysis of local intertextuality posits two material aspects to intertextual links. Besides the degree of "literality" of an intertext, we would like to add its thematic embedding in the context, which we operationalize following the discussion in Section 3, into the analysis. By inspecting the statistical relationships between the posterior estimates of both outcome variables across groupings, we aim to gain insight about how these two aspects

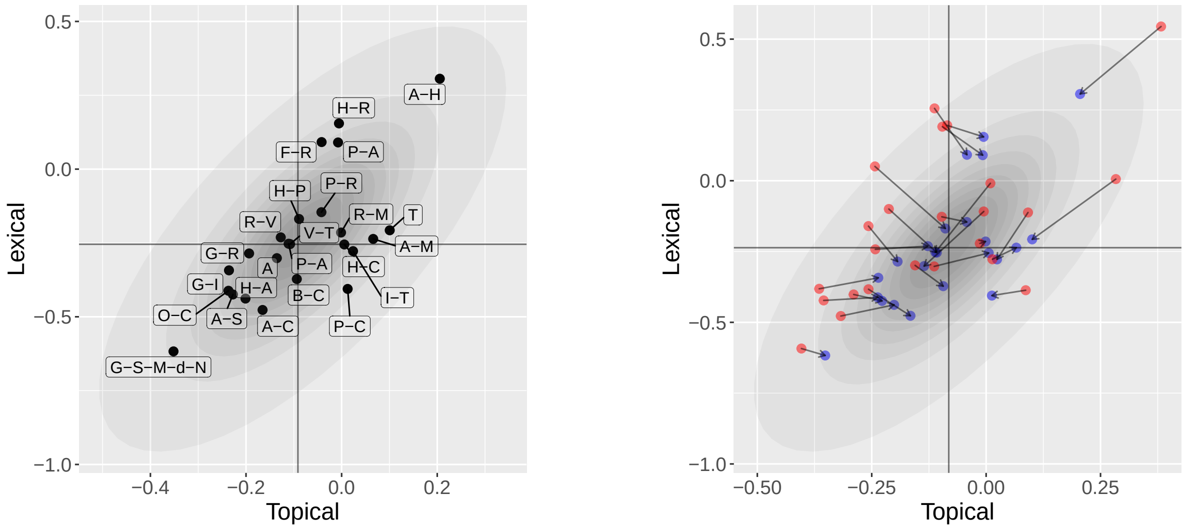


Figure 2: Left: mean posterior estimates for authors from model $M_{A \cup B \cup T}$, averaging over books and topics. Authors are labeled using their initials (see Appendix for the full table of author abbreviations.) Right: effect of shrinkage on the estimates shown as displacements from the maximum likelihood estimates (in red) towards the mean posterior estimates from model $M_{A \cup B \cup T}$ (in blue). Overall mean effects are shown in both plots to facilitate the comparison. Note that variables were transformed to be centered around a zero mean.

of intertextuality complement each other.

Author grouping The left plot in Figure 2 shows the mean posterior estimates for authors, averaging over books and topics. Overall, we observe considerable correlation between topical and lexical similarity. For reference, Figure 1b shows the posterior estimates of the correlation across outcomes for each of the groupings.

It is important to note that the observed correlation is exacerbated by the effect of multi-level modeling shrinkage. As shown in the right plot in Figure 2, author estimates are pushed towards the diagonal when considering book and topic groupings, 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 the plot are considerably less populated. In combination with the analysis from Figure 1a, we can interpret the high correlation in the sense that the lexical similarity axis suffices to explain the variation observed between authors.[9] However, it is nevertheless interesting to investigate the relative position of outliers. For instance, Petrus Cellensis (P-C), an author known for his allegorical style [35], appears in the bottom-right section indicating a more allusive style in which references are more than average thematically embedded. Bernardus Claraevallensis (B-C), known for his constant biblical allusions [30], similarly appears to the right of Petrus Cellensis. Finally, Augustinus Hipponensis (A-H) and Guibertus Mariae de Novigento (G-S-M-d-N) represent the extremes at the sections respectively to the upper-right, characterized by a highly embedded style, and to the bottom-left, leaning towards loosely connected references.

[9] As a reminder from Section 3, the estimates of topic-level similarity were computed on documents after removing the lexical overlap to avoid biases from lexical similarity.

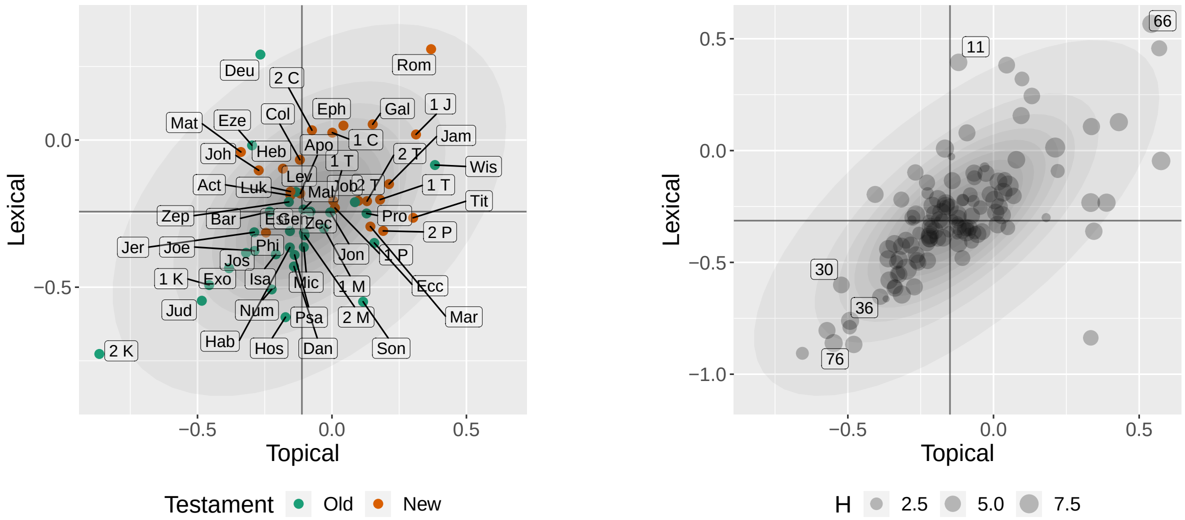


Figure 3: Mean posterior estimates for books and topics from model $M_{A \cup B \cup T}$, averaging over authors and topics and authors and books respectively. Left: The points have been highlighted according to whether the book pertains to the Old or the New Testament. Right: as sanity check, the size 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.

Book grouping The left plot in Figure 3 shows the mean posterior estimates for books.

We now observed 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.[10] Again, inspecting the outliers can help the interpretation of the distribution. 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.

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.

The right plot in Figure 3 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 paper, we have singled out a number of topics for commentary. For illustration, the selected topics have been highlighted in Figure 3 and the corresponding topic descriptors (top probality terms

[10] 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 $M_{A \cup B \cup T}$ and was therefore not further considered in the analyses.

under the given topic) are shown in List 1.

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

List 1.: Topic descriptors for a selection of illustrative topics

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. Discussion

After having carried out the analyses, we now proceed to address how the statistical evidence helps approaching the research questions advanced in Section 1.

With respect to RQ1, we explore to what extent the decomposition of the intertext type into the aspects of lexical similarity and thematic embedding proved helpful for characterizing the observed variation across the different grouping factors. Apriori, the intersection of both axes should produce four intertextual trends depending on whether lexical and topical similarity are below or above mean. These trends correspond to the four quadrants shown in Figure 2 and Figure 3. However, our analyses generally showed a correlation between both aspects, which resulted in low-density bottom-right and, especially, upper-left quadrants. As a result, we can conclude that overall allusive cases of intertextuality do not rely on proportionally higher degrees of topical embedding to reinforce the intertextual link. Complementarily, 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. However, despite the mentioned correlation, we can conclude that the inclusion of both axes provides a fuller picture of local intertextuality since (i) correlation varied depending on the grouping factor and (ii) the position of outliers with respect to the general trend highlights the particularities of particular authors, books or topics that would be otherwise missed.

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 [39]. With respect to the topical embedding of intertexts, author variation was less important due to high correlation with lexical similarity that was unveiled by the shrinkage induced through partial pooling. However, the outlier status of some authors with respect to the general trend could still be interpreted in a stylistic way (e.g. the discussed cases of Petrus Cellensis and Bernardus Claraevallensis).

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.

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 to 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.

6. Future Work

In the present paper, 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 paper, 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 systematic ways in which researchers can systematically explore such topic spaces in order to elicit potentially fruitful hypotheses.

References

[1] A. Acerbi and R. A. Bentley. “Biases in cultural transmission shape the turnover of popular traits”. In: *Evolution and Human Behavior* 35.3 (2014), pp. 228–236.

[2] G. Allen. *Intertextuality*. Routledge, Mar. 2000. ISBN: 9780203131039. DOI: 10.4324/9780203131039. URL: https://www.taylorfrancis.com/books/9780203131039.

[3] D. Bamman and G. Crane. “The logic and discovery of textual allusion”. In: *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data*. 2008.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent Dirichlet allocation”. In: *Journal of Machine Learning Research* (2003). ISSN: 15324435. DOI: 10.1016/b978-0-12-411519-4.00006-9.

[5] T. vor der Brück, S. Eger, and A. Mehler. “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, 2015, pp. 105–113. DOI: 10.18653/v1/W15-3716. URL: http://aclweb.org/anthology/W15-3716.

[6] M. Büchler et al. “Towards a Historical Text Re-use Detection”. In: *Text Mining: From Ontology Learning to Automated Text Processing Applications*. Ed. by C. Biemann and A. Mehler. Cham: Springer International Publishing, 2014, pp. 221–238.

[7] P. C. Bürkner. “Advanced Bayesian multilevel modeling with the R package brms”. In: *R Journal* (2018). ISSN: 20734859. DOI: 10.32614/rj-2018-017. eprint: 1705.11123.

[8] N. Coffee et al. “The Tesserae Project: intertextual analysis of Latin poetry”. In: *Literary and Linguistic Computing* 28.2 (July 2012), pp. 221–228.

[9] G. B. Conte. “The Rhetoric of Imitation: Genre and Poetic Memory in Virgil and Other Latin Poets”. In: *The Classical World* (1988). ISSN: 00098418. DOI: 10.2307/4350270.

[10] G. Crane. “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*. 1996, pp. 3–10.

[11] E. R. Crema, A. Kandler, and S. J. Shennan. “Revealing patterns of cultural transmission from frequency data: equilibrium and non-equilibrium assumptions”. In: *Nature Publishing Group* (Dec. 2016), pp. 1–10.

[12] J. Culler. “Presupposition and Intertextuality”. In: *MLN* 91.6 (1976), pp. 1380–1396. ISSN: 00267910, 10806598. URL: http://www.jstor.org/stable/2907142.

[13] S. Deerwester et al. “Indexing by latent semantic analysis”. In: *Journal of the American Society for Information Science* (1990). ISSN: 10974571. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[14] S. Eger, R. Gleim, and A. Mehler. “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)*. 2016, pp. 1507–1513.

[15] J. Farrell. “Intention and intertext”. In: *Phoenix* (2005). ISSN: 00318299.

[16] C. W. Forstall and W. J. Scheirer. *Quantitative Intertextuality. Analyzing the Markers of Information Reuse*. Springer, 2019.

[17] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press, 2006. ISBN: 9780511790942. DOI: 10.1017/CBO9780511790942. URL: http://ebooks.cambridge.org/ref/id/CBO9780511790942.

[18] G. Genette. *Palimpsestes: La littérature au second degré*. Seuil, 1982.

[19] I. C. Ghiban and Ş. Trăuşan-Matu. “Network Based Analysis of Intertextual Relations”. In: *Advances in Information Systems and Technologies*. Ed. by Á. Rocha et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 753–762. ISBN: 978-3-642-36981-0.

[20] M. Hoffman, F. R. Bach, and D. M. Blei. “Online learning for latent dirichlet allocation”. In: *advances in neural information processing systems*. 2010, pp. 856–864.

[21] R. Ihaka and R. Gentleman. “R: A Language for Data Analysis and Graphics”. In: *Journal of Computational and Graphical Statistics* 5.3 (Sept. 1996), pp. 299–314. ISSN: 1061-8600. DOI: 10.1080/10618600.1996.10474713. URL: http://www.tandfonline.com/doi/abs/10.1080/10618600.1996.10474713.

[22] M. Kestemont and J. D. Gussem. “Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning”. In: *J. Data Min. Digit. Humanit.* 2017 (2017). URL: https://jdmdh.episciences.org/3835.

[23] G. N. Knauer. “Die Aeneis und Homer. Studien zur poetischen Technik Vergils mit Listen der Homerzitate in der Aeneis”. In: *The Classical World* (1965). ISSN: 00098418. DOI: 10.2307/4345826.

[24] J. Kristeva. “Bakhtine, le mot, le dialogue et le roman”. In: *Critique.* (1967).

[25] Y. D. H. Lab. *Intertext*. https://github.com/YaleDHLab/intertext. 2017.

[26] D. Lewandowski, D. Kurowicka, and H. Joe. “Generating random correlation matrices based on vines and extended onion method”. In: *Journal of Multivariate Analysis* (2009). ISSN: 0047259X. DOI: 10.1016/j.jmva.2009.04.008.

[27] J. Lund et al. “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, June 2019, pp. 3978–3987. DOI: 10.18653/v1/N19-1399. URL: https://www.aclweb.org/anthology/N19-1399.

[28] E. Manjavacas, Á. Kádár, and M. Kestemont. “Improving Lemmatization of Non-Standard Languages with Joint Learning”. 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, June 2019, pp. 1493–1503. URL: https://www.aclweb.org/anthology/N19-1153.

[29] E. Manjavacas, B. Long, and M. Kestemont. “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*. Minneapolis, USA: Association for Computational Linguistics, June 2019, pp. 104–114. URL: https://www.aclweb.org/anthology/W19-2514.

[30] B. P. Mcguire. “Bernard of Clairvaux”. In: *A Companion to Philosophy in the Middle Ages*. Oxford, UK: Blackwell Publishing Ltd, Nov. 2007, pp. 209–214. ISBN: 9780470996669. DOI: 10.1002/9780470996669.ch28. URL: http://doi.wiley.com/10.1002/9780470996669.ch28.

[31] A. Mesoudi. *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. University of Chicago Press, 2011.

[32] J. P. Migne. *Patrologiae Cursus Completus. Series Latina (217 + 4 vols.)* Garnier frères, 1844-1855 (and 1862-1865).

[33] S. Moyise. “Intertextuality and Biblical Studies: A Review”. In: *Verbum et ecclesia* 23.2 (2002), pp. 418–431.

[34] M. Orr. “Intertextuality”. In: *The Encyclopedia of Literary and Cultural Theory*. Oxford, UK: John Wiley & Sons, Ltd, Dec. 2010. DOI: 10.1002/9781444337839.wbelctv2i002. URL: http://doi.wiley.com/10.1002/9781444337839.wbelctv2i002.

[35] M. Ott. “Peter Cellensis”. In: *The Catholic Encyclopedia*. Robert Appleton Company, 1911, Vol. 11. URL: http://www.newadvent.org/cathen/11762b.htm.

[36] R. Rehurek and P. Sojka. “Software Framework for Topic Modelling with Large Corpora”. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (2010). ISSN: 2951740867.

[37] M. Röder, A. Both, and A. Hinneburg. “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, 2015, pp. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: http://dl.acm.org/citation.cfm?doid=2684822.2685324.

[38] P. Roelli. “The corpus corporum, a new open Latin text repository and tool”. In: *Bulletin du Cange - Archivum Latinitatis Medii Aevi* (2014). ISSN: 09948090. DOI: 10.5167/uzh-171105.

[39] Y. Sari, M. Stevenson, and A. Vlachos. “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 E. M. Bender, L. Derczynski, and P. Isabelle. Association for Computational Linguistics, 2018, pp. 343–353. URL: https://www.aclweb.org/anthology/C18-1029/.

[40] W. Scheirer, C. Forstall, and N. Coffee. “The sense of a connection: Automatic tracing of intertextuality by meaning”. In: *Digital Scholarship in the Humanities* (2016). ISSN: 2055768X. DOI: 10.1093/llc/fqu058.

[41] H. Schmid. “Probabilistic part-ofispeech tagging using decision trees”. In: *New methods in language processing*. 2013, p. 154.

[42] J. Seo and W. B. Croft. “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, 2008, p. 571. ISBN: 9781605581644. DOI: 10.1145/1390334.1390432. URL: http://portal.acm.org/citation.cfm?doid=1390334.1390432.

[43] D. A. Smith et al. “Detecting and modeling local text reuse”. In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. 2014, pp. 183–192. ISBN: 9781479955695. DOI: 10.1109/JCDL.2014.6970166.

[44] T. F. Smith and M. S. Waterman. “Identification of common molecular subsequences”. In: *Journal of Molecular Biology* (1981). ISSN: 00222836. DOI: 10.1016/0022-2836(81)90087-5.

[45] A. Vehtari, A. Gelman, and J. Gabry. “loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models”. In: *R package version* 2.0 (2018), p. 1003.

[46] A. Vehtari, A. Gelman, and J. Gabry. “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. In: *Statistics and Computing* (2017). ISSN: 15731375. DOI: 10.1007/s11222-016-9696-4. eprint: 1507.04544.

Table 2
Information about authors in the dataset.

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

A. Author Information

Table 2 displays a list of the authors included in the present study together with the total number of references in the dataset and the initials used in Figure 2 to identify the author.

B. Topic Modeling

Figure 4 displays results from the topic evaluation experiments. We grid-searched the optimal number of words per training document (`DocWords`), vocabulary size (`Top-k`), number of topics (`NumTopics`) and lemmatization model (`Model`) with respect to the C_V coherence measure [37]. 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. 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 and in the best combinations (`Top-K`= 20k).

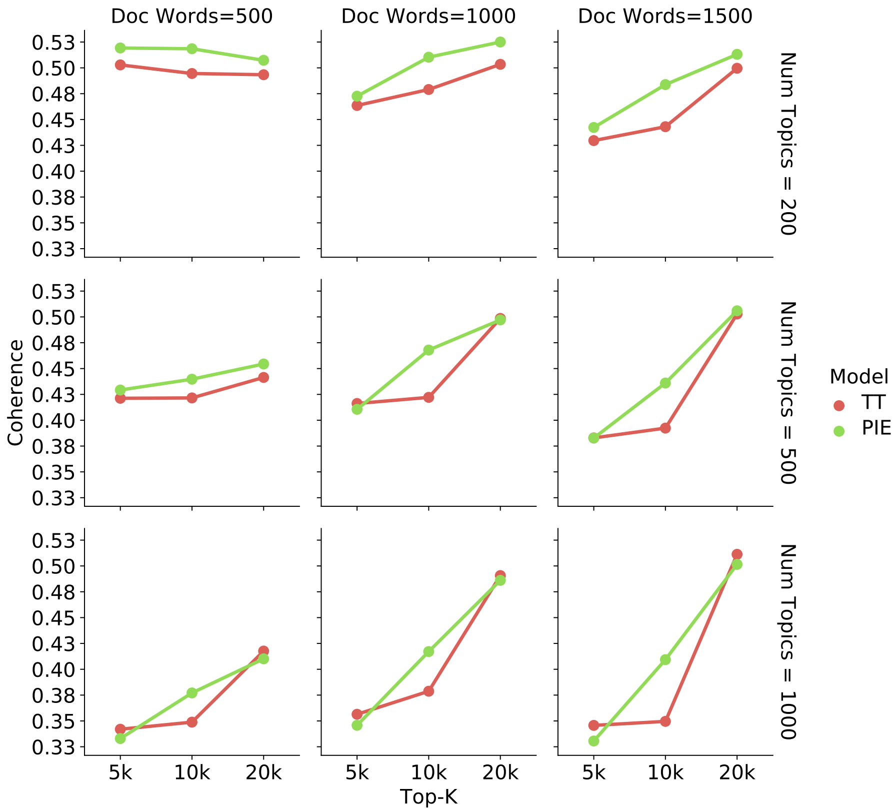


Figure 4: Topic Coherence evaluation over number of topics, training document size (in number of words) and vocabulary size.