# Quantifying Contextual Aspects of Inter-annotator Agreement in Intertextuality Research

Enrique Manjavacas, Laurence Mellerin, Mike Kestemont

# Quantifying Contextual Aspects of Inter-annotator Agreement in Intertextuality Research

**Enrique Manjavacas**
Leiden University
Leiden, The Netherlands
enrique.manjavacas@gmail.com

**Laurence Mellerin**
Source Chrétiennes Institute
Lyon, France
laurence.mellerin@mom.fr

**Mike Kestemont**
University of Antwerp
Antwerp, Belgium
mike.kestemont@uantwerpen.be

## Abstract

We report on an inter-annotator agreement experiment involving instances of text reuse focusing on the well-known case of biblical intertextuality in medieval literature. We target the application use case of literary scholars whose aim is to document instances of biblical references in the 'apparatus fontium' of a prospective digital edition. We develop a Bayesian implementation of Cohen's $\kappa$ for multiple annotators that allows us to assess the influence of various contextual effects on the inter-annotator agreement, producing both more robust estimates of the agreement indices as well as insights into the annotation process that leads to the estimated indices. As a result, we are able to produce a novel and nuanced 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 writings of Bernard of Clairvaux. Among others, our method is able to unveil the fact that the obtained agreement depends heavily on the biblical source book of the proposed reference, as well as the underlying algorithm used to retrieve the candidate match. Finally, a discussion of the hurdles encountered by annotators supplements the results of the statistical analysis, contributing a qualitative insight into the difficulties involved in the identification of literary text reuse.

## 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 include, for instance, visualizing high-level patterns in the referential connections between collections of texts (Jänicke et al., 2015; Yousef and Jänicke, 2021), or studying the influence[1] that a given writer has had on subsequent generations (Bloom, 1973). More generally, curators of (nowadays more commonly digital) editions of literary works are concerned with making even the more subtle intertextual connections accessible to the contemporary public.

In this quest, automated retrieval algorithms play a crucial role, since they can accelerate the task of identification at different retrieval phases. At the beginning, a focus on precision can help editors dealing with the bulk of more obvious cases. Towards the end of the process, the majority of the references have been spotted, and high-recall algorithms can help suggesting potential candidates. In this context, the curation of benchmark datasets on which retrieval algorithms can be put to the test is an important milestone of applied research, since benchmark datasets are necessary not only for assessing the relative advantage of particular approaches but also in order to reliably measure the level of precision and recall that editors can expect from automated systems.

However, the curation of benchmark datasets depends on reliable annotation processes. 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. In this context, an important question—which has been however rarely approached in previous research—addresses the level of agreement that expert anno-

[1] A recent example is the HyperHamlet project, which aimed at documenting influential passages of Shakespeare on later literature in an exhaustive manner (Hohl Trillini, 2018).



*Proceedings of LaTeCH-CLfL 2021*, pages 31–42
Punta Cana, Dominican Republic (Online), November 11, 2021.

tators may reach. Still, as it has been noted before (Manjavacas et al., 2019b), inter-annotator agreement studies of intertextuality are rare.[2]

The present paper starts by addressing such research question, but moves beyond it and further aims towards an examination and understanding of the contextual factors that may affect inter-annotator agreement in intertextuality research. Thus, we not only seek to establish an estimate of the achieved inter-rater agreement, but crucially to investigate contextual aspects in the experimental design that influence the process of agreement and can therefore serve as explanatory variables for the obtained agreement scores. We approach this research by means of a re-formulation of Cohen's $\kappa$ for multiple raters (Artstein, 2017) that bases the computation of the observed agreement on a hierarchical statistical model. Thanks to the incorporation of the statistical model, we are able to infer the dependency of agreement on a number of factors of variation present in the experimental design. Finally, using modern Bayesian modeling techniques to fit the statistical model we obtain estimates of agreement that naturally incorporate uncertainty arising from the inferential process.

We target the text reuse retrieval context of ongoing efforts towards a digital edition of the sermons written by the influential medieval author Bernard of Clairvaux (1090–1153). Bernard's sermons are known for their pervasive biblical intertextuality and are, therefore, rich in potential cases of reuse. The digital edition is currently in a late-stage retrieval phase in which the goal is to exhaustively find relevant but missed cases of reuse.

In this context, we examine the effect of the following contextual factors. First, we investigate the influence of the underlying retrieval method—examining two competitor algorithms from the literature that, a priori, obtain similar performance results but target different aspects of intertextuality. Depending on the biases towards particular types of reuse, algorithms may retrieve candidate pairs that are consistently more or less prone to disagreement. Secondly, we inspect the effect of the biblical book from which the suggested source verse stems. Annotators may be more or less familiar with particular books, and may have diverging expectations on what biblical book the author is likely to borrow from. From this point of view, the source book could constitute an important factor of influence for the obtained agreement. Third, we look into the effect of the amount of lexical overlap between the suggested documents. The level of literality in the suggested matches may constitute a source for disagreement when, for example, annotators expectations on the author's style of reuse diverge. Finally, we tackle the issue of intra-biblical intertextuality, the fact that biblical verses may refer to each other and, as a result, annotators must decide which of the possible verses a Bernardine passage actually refers to.

**Contributions** Our contributions are as follows. We present a novel study of the inter-annotator agreement on the task of identifying cases of biblical intertextuality in a real-world scenario. 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 as well as 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. Secondly, we statistically inspect additional factors of variation—related to both objective (style of reuse retrieved by the system) and subjective variables (knowledge of the collection from which the passages are borrowed)—that may help explain the obtained agreement scores. Specifically, we find that the biblical book from which the source of the reference stems is a significant factor of variation and overshadows the effects of other variables. Furthermore, we find that agreement is lowest with average values of lexical overlap in the candidate pair, and that the overall shape of the effect of lexical overlap on agreement diverges depending on whether the Bernardine passage was already known to contain a reference or not.

Finally, we contribute a quantitative assessment of the main hurdles to agreement that our annotators encountered during the experiment, highlighting that not only expert knowledge on the target collections can have important consequences on the assessment of intertextual links, but also that choices in the experimental design may contribute

[2] Bär et al. (2012) include 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.

| Method | Matches | Sermons | Matches/ Sermon |
|---|---|---|---|
| S-C | 296 (72) | 24 | 12.33 (± 10.66) |
| S-W | 292 (56) | 22 | 13.27 (± 18.5) |

Table 1: Summary statistics of the annotation dataset, displaying the number of matches per method (including the number of matches that had no previous annotation), the number of sermons involved, and the mean (+ standard deviation) of matches per sermon.

to inflated levels of disagreement.

## 2 Experimental Design

In the present section, we explain the underlying resources and methods used in this study.

### 2.1 Dataset

**Underlying Collections** The target collection consists in the 85 *Sermons on the Song of Songs* by Bernard of Clairvaux—made available at the Sources Chrétiennes Institute as part of the BiblIndex project (Mellerin, 2013, 2014)—which we segmented into "documents" using a sliding window of 20 tokens with an overlap of 10 tokens. The processed collection comprises 19,987 such documents. The source 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. Both collections were lemmatized using the neural Latin lemmatizer provided by the software library `pie` (Manjavacas et al., 2019a).

**Focus Dataset** The focus dataset underlying the inter-annotator agreement experiment contains candidate matches between target documents from Bernard and source documents from the Bible. The matches represent optimized guesses about potential misses by the editors in a late-stage retrieval phase during the curation of the digital edition of Bernard's sermons. The current version of this digital edition contains already a total of 6,689 manually identified biblical references. On the basis of the available annotations, we fine-tuned two text reuse retrieval algorithms: one using the local alignment algorithm `Smith-Waterman` (Smith and Waterman, 1981), with a bias towards verbatim cases, and another one based on the `Soft-Cosine` similarity measure (Sidorov et al., 2014), which takes into account lexical similarity using word embeddings and has been used in a previous study on Bernard (Manjavacas et al., 2019b).[3] These algorithms were then applied to the remaining dataset in order to find references potentially overlooked by the editors. From the candidate set of each algorithm we took 300 candidate matches for a total of 600 items. An example candidate match, retrieved with the `Smith-Waterman` algorithm is shown in Figure 1.

Some of the suggested matches involve target documents that are already annotated with a reference to the Bible, whereby the labeled source diverges from the suggested one. In principle, referencing passages in Bernard need not be restricted to a single biblical source. For this reason, we did not exclude such cases from the final target dataset. Table 1 displays dataset statistics about the focus dataset.

### 2.2 Inter-annotator Agreement

Our experiment involved a total of three expert editors of Bernard,[4] who were shown candidate matches retrieved by the algorithms. The guideline provided to the participants was limited to indicate whether the candidate match would be considered for inclusion in the 'apparatus fontium' of the prospective edition.

In order to quantify agreement, we chose a chance-corrected inter-annotator agreement index for multiple raters based on Cohen's $\kappa$. As shown in (Artstein and Poesio, 2008), a series of chance-corrected indices can be formulated as follows:

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e} \quad (1)$$

Here, $A_e$ expresses the expected agreement due to chance as the probability of agreement based on a theoretical annotator casting judgments on a random fashion. In contrast, $A_o$ expresses the observed agreement as the probability—commonly computed as the proportion of instances of agreement in the dataset—that agreement actually

[3] A careful comparison on the basis of Average Precision showed that these algorithms are likely to perform equally with a probability of 0.91 within a margin (region of practical equivalence) of 0.02 points. See Benavoli et al. (2017) for a detailed description of the Bayesian modeling approach to retrieval performance comparison taken in the present study. See Manjavacas (2021) for a thorough discussion of these retrieval methods.

[4] 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.

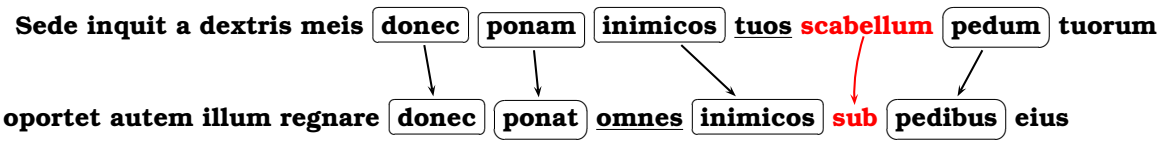


Figure 1: Example of a candidate pair retrieved by the `Smith-Waterman` algorithm, matching a passage from Bernard's $6^{th}$ Sermon-—document on top—and the biblical verse *1 Corinthians, 15:25*—document on the bottom. Boxes mark tokens that matched based on their lemmata. Underlined tokens represent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

occurs—i.e. based on the evidence provided by the collected judgments. The difference between $S$ (Bennett et al., 1954), $\pi$ (Scott, 1955) and $\kappa$ (Cohen, 1960) resides on the posited model of random annotator behaviour used to compute $A_e$, where Cohen's $\kappa$—the preferred approach—is the only one taking annotator bias towards particular labels into account.

In order to compute agreement indices for more than two raters, we follow the common approach of decomposing the total agreement into the agreement between pairs of annotators, normalizing by the total number of possible annotator pairs.

### 2.3 A Probabilistic Model for Cohen's $\kappa$

As already mentioned, the common approach to obtaining $A_o$ consists in computing the proportion of pairs of judgements that are in agreement over the total number of pairs of judgments implied by the dataset. For each candidate match $i$, the corresponding agreement $\text{agr}(i)$ is given by:

$$\text{agr}(i) = \frac{1}{\binom{c}{2}} \sum_{j=1}^{k} \binom{n_{ij}}{2} \tag{2}$$

where $c$ and $k$ refer respectively to the number of raters and labels, and $n_{ij}$ refers to the number of times label $j$ was assigned to candidate match $i$. The dataset-level $A_o$ is given by the average $agr(i)$ on the entire dataset: $A_o = \frac{1}{N} \sum_i agr(i)$, where $N$ indicates the total number of candidate matches.

Our approach differs in that we base the computation of $A_o$ on a statistical model of the possible outcomes of a pairwise judgement comparison, as we will show below. Relying on a statistical model allows us to both control for contextual factors as well as look for explanatory variables through the incorporation of statistical co-variates. Moreover, through the deployment of a statistical model we can naturally incorporate uncertainty and provide more robust estimates of the agreement indices.

We resort to a hierarchical multinomial regression model to capture the probabilities of the outcomes of the pairwise judgement comparison. For $k$ labels, this approach uses $m-1$ log-linear models, where $m = k^2$ indicates the number of possible outcomes from pairwise comparisons of casted judgements, using the remaining outcome as a "pivot". For the present case, let $\theta_{01}$ refer to the probability of the outcome in which one annotator assigns label 0 and the second assigns label 1.[5] Then, the proposed model is given by Equations 3, 4 and 5.

$$\begin{aligned} \log\left(\frac{\theta_{01}}{\theta_{00}}\right) &= \alpha_{01} + \beta_{01p} \cdot X_p + \nu_{01q} \\ \log\left(\frac{\theta_{10}}{\theta_{00}}\right) &= \alpha_{10} + \beta_{10p} \cdot X_p + \nu_{10q} \\ \log\left(\frac{\theta_{11}}{\theta_{00}}\right) &= \alpha_{11} + \beta_{11p} \cdot X_p + \nu_{11q} \end{aligned} \tag{3}$$

$$\theta_{00} + \theta_{01} + \theta_{10} + \theta_{11} = 1 \tag{4}$$

More specifically, Equation 3 shows the log-odds of the responses $\theta_{01}$, $\theta_{10}$ and $\theta_{11}$ with respect to the pivot: $\theta_{00}$. Each log-odds are computed by a multi-level linear model where $\alpha_{..}$ refers to the fixed intercepts, $\beta_{..p}$ to the fixed coefficient corresponding to the $p^{th}$ independent variable $X_p$ and $\nu_{..q}$ to the $q^{th}$-level random intercept, which captures within-group variation for the corresponding grouping factor.

$$\begin{bmatrix} \nu_{01q} \\ \vdots \\ \nu_{11q} \end{bmatrix} \sim \text{MVN}(0, \Sigma_q); \Sigma_q = \begin{bmatrix} \sigma^2_{01q} & & \\ \vdots & \ddots & \\ \sigma_{01-11q} & \cdots & \sigma^2_{11q} \end{bmatrix} \tag{5}$$

As shown in Equation 5, these group-level random intercepts are modeled jointly, coming from

[5] For illustration purposes, the formulation considers only the binary case. The extension to any number of labels is, however, straight-forward.

a multi-variate normal ($MVN$) centered around a zero-mean with a variance-covariance matrix $\Sigma_q$.[6] Letting $i$ refer to the $i^{th}$ outcome, we can turn the log-odds into actual probabilities employing the softmax function, shown in Equation 6.

$$\begin{aligned} \theta_{00} &= 1 - \sum_{i=1}^{m} \theta_{00} \cdot e^{\alpha_i + \beta_{ip} \cdot X_p + \nu_{iq}} \\ \implies \theta_{00} &= \frac{1}{1 + \sum_{i=1}^{m-1} e^{\alpha_i + \beta_{ip} \cdot X_p + \nu_{iq}}} \\ \implies \theta_i &= \frac{e^{\alpha_i + \beta_{ip} \cdot X_p + \nu_{iq}}}{1 + \sum_{i'=1}^{m-1} e^{\alpha_{i'} + \beta_{i'p} \cdot X_p + \nu_{i'q}}} \end{aligned} \tag{6}$$

As we can see, the probability of the pivot ($\theta_{00}$, in this case) can be computed as the remaining probability after subtracting the probabilities of the other $m-1$ models.

The probabilities computed by Equation 6 represent baseline probabilities of the outcomes without regard to the annotators that produced the judgements. In order to take into account the observed annotator behaviour, we fit random intercepts that capture the annotator pair underlying the observed outcome. For $c$ annotators, this approach introduces $\binom{c}{2}$ random intercepts per outcome, one for each of the pairwise combinations of the annotators in set $C$. Letting $\theta_{00}^{xy}$ be the probability of the 00 outcome for annotators x and y, $A_o$ can be computed by Equation 7.

$$A_o = \sum_{i=1}^{k} \frac{1}{\binom{c}{2}} \sum_{x,y \in C} \theta_{ii}^{xy} \tag{7}$$

Finally, we compute $\kappa$ from Equation 1 using $A_o$ from Equation 7 and the dataset-level computation of $A_e$ given by Equation 8.

$$A_e = \sum_{i=1}^{k} \frac{1}{\binom{c}{2}} \sum_{x,y \in C} P(i|x)P(i|y) \tag{8}$$

where $P(i|x)$ corresponds to the probability that annotator $x$ assigns label $i$, which we obtain as the relative proportion of $i$-judgements casted by annotator $x$.[7]

[6]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).

[7]The utilized index corresponds to a generalization of Cohen's $\kappa$ to multiple annotators. This is in contrast to the multi-rater agreement index, commonly known as Fleiss $\kappa$, which was introduced by Fleiss in (Fleiss, 1971) but that, as (Artstein and Poesio, 2007) argue, actually corresponds to a generalization of Scott's $\pi$. Following (Artstein and Poesio, 2007), we will refer to our index as multi-$\kappa$, in order to avoid confusion.

**Bayesian Modeling** 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.

## 3 Results

### 3.1 Model Validation

In order to address the effect of contextual aspects on agreement, we identify a number of relevant factors and incorporate them into the computation of the agreement index as fixed effects and random intercepts. Besides modeling the underlying annotator pair as random intercepts, we model (i) the *familiarity* (Known) of the target document (i.e. whether the Bernardine fragment was known already to contain a reference to a different biblical verse) as a binary fixed effect—this allows us to estimate agreement in cases where annotators are asked to re-assess the actual reference of a given passage. Secondly, we incorporate (ii) the *lexical overlap* between source and target documents[8] as a continuous fixed effect. Moreover, we model (iii) the underlying retrieval *method* using random intercepts, seeking to capture whether the underlying retrieval methods have a tendency to suggest more or less controversial matches. Finally, we model (iv) the biblical *source book* (Book) as random intercepts with 50 levels—one for each of the biblical books attested in the dataset—in order to test whether references to particular biblical books tend to be more or less controversial.

First, we test the explanatory power of these

[8]We compute lexical overlap using the weighted Jaccard similarity shown in the following Equation:

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

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.

| Model | ELPD (SE) | P | ELPD$\Delta$(SE) |
|---|---|---|---|
| m.klMB | -1570.05 (35.62) | 99.08 | 0.00 |
| m.kMB | -1733.29 (33.45) | 96.15 | -163.24 (16.78) |
| m.MB | -1793.92 (31.55) | 92.70 | -223.87 (19.10) |
| m.M | -2136.21 (21.63) | 12.12 | -566.15 (28.64) |
| m | -2150.97 (20.84) | 8.98 | -580.92 (29.13) |

Table 2: Evaluation of the statistical models in terms of ELPD. First column shows the absolute ELPD (higher ELPD indicates a better model fit). The second columns shows an estimate of the effective number of parameters. The third column displays the absolute difference in ELPD with respect to the best model.

effects using a statistical model comparison approach based on the expected log pointwise predictive density (ELPD) as measure. This quantity provides an estimate of the predictive accuracy of a model on out-of-sample datasets. These estimates can be efficiently obtained—i.e. without having to refit the model on the different splits—using an approximatation to leave-one-out (LOO).[9]

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.[10] Specifically, m is a baseline model that has a single intercept and adds neither predictors nor grouping factors. The second model—m.M—adds the underlying retrieval method as binary predictor. The third model—m.MB—adds varying intercepts corresponding to the biblical source book of the candidate reference. The fourth model—m.kMB—includes an additional binary predictor for the "familiarity" with the Bernardine passage. Finally, model m.klMB adds a continuous predictor accounting for lexical overlap.

Table 2 shows the results of the model comparison. As we can see the full model m.klMB can be identified as the model with superior predictive performance. We take this as a justification for the inclusion of all considered predictors and base any further inference on this model.

[9]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.

[10]Details about the model fitting process as well as the software used for inference are shown in Appendix A.

| | -Book | | | +Book | | |
|---|---|---|---|---|---|---|
| Known | -95% | $\kappa$ | +95% | -95% | $\kappa$ | +95% |
| | | | Soft-Cosine | | | |
| False | 0.66 | 0.76 | 0.84 | 0.22 | 0.72 | 0.97 |
| True | 0.66 | 0.74 | 0.82 | 0.29 | 0.73 | 0.97 |
| | | | Smith-Waterman | | | |
| False | 0.45 | 0.58 | 0.29 | 0.08 | 0.54 | 0.93 |
| True | 0.52 | 0.62 | 0.71 | 0.16 | 0.62 | 0.95 |

Table 3: Median, lower and upper 95% quantiles for posterior agreement scores obtained with the full model (m.mlBK), while keeping similarity at the zero-centered mean value. (-/+ Book refers to whether variation stemming from the source book is taken into account or not.)

### 3.2 Effects of Contextual Factors

We now inspect the effects of the different contextual factors on the output inter-annotator agreement index. Table 3 shows the resulting inter-annotator agreement scores by a number of contextual factors. As we can see, the average agreement is fairly high across conditions ranging from 0.54 to an eventual 0.76. In order to gain a better picture of the underlying phenomena, we first zoom in on the effect of the retrieval method.

**Effects of the retrieval method** Figure 2 shows the posterior distributions obtained for the $\kappa$ scores, computed using Equation 1. The plot on the left-hand side of Figure 2 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 $\nu_{..q}$ from the inferred multi-variate normal distribution from which the random intercepts are modeled to arise—see Equation 5. For each MCMC draw of parameters, we add the sampled $\nu_{..q}$ 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.

As we can see, the agreement is decisively higher for candidates suggested by the Soft-Cosine retrieval algorithm. However,

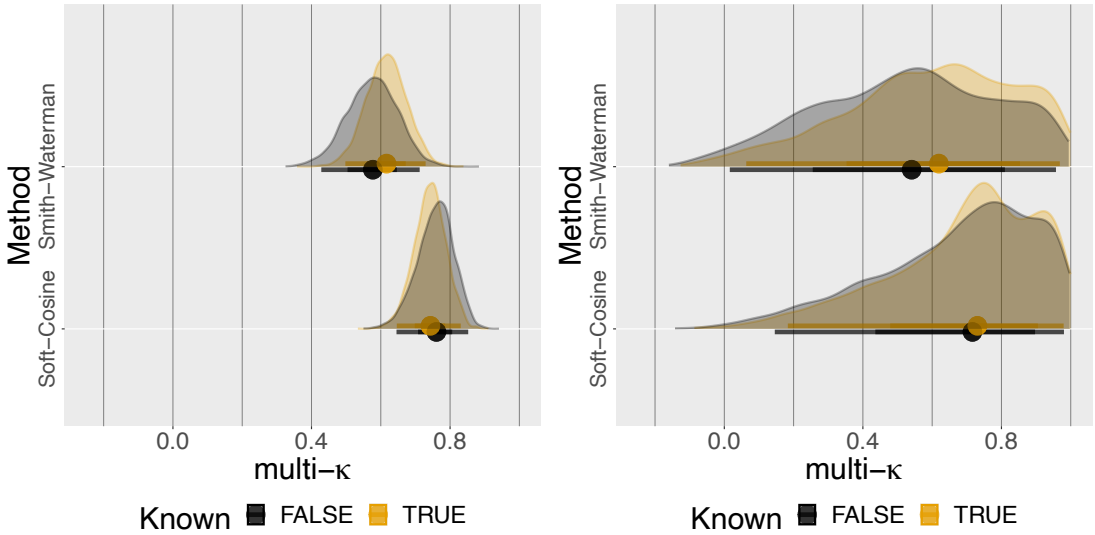


Figure 2: Posterior multi-$\kappa$ scores inferred from the full model (`m.mlBK`), displayed according to retrieval method, on the y-axis, and whether the candidate borrowing passage is known to contain a reference to a different biblical verse, on the x-axis. Word overlap is kept to the zero-centered mean value. Left plots and right plots differ on whether the variation coming from the books is excluded or not. The mean estimate is shown by a point with a 0.89 probability interval (shown by the surrounding horizontal bar).

when the variability stemming from books is taken into account through marginalization, we obtain very wide posterior distributions, as shown by the right plot. This is a strong indication of the importance of the target reference book for annotator behavior and supplements the evidence from the ELPD comparisons in Section 3.1, where including book-level varying intercepts resulted in a large ELPD increase of 342.29 points—model `m.MB` vs. model `m.M`—, corresponding to a 58.92% ELPD increase with respect to the total increase between worst and best models—i.e. model `m` vs model `m.klMB`.

**Effects of lexical overlap** Figure 3 shows the effect of lexical overlap on agreement under different combinations of underlying retrieval method and familiarity of the target passage, using counter-factual plots. These plots visualize the statistical dependency relying on the posterior predictions for the entire range of the lexical overlap variable—i.e. including values for which no observation is attested in the original dataset (McElreath, 2018). The plots on the left hand-side do not take into account variability arising from the source book, while those in the right hand-side do. As we can see, the effect of overlap on agreement is primarily decreasing and, at least for unknown passages, monotonic—i.e. an increase in overlap is associated with a decrease in agreement. Judging by the small credible intervals, this seems to suggest that annotators find it easy to agree on candidates with very low overlap—probably because the match can be discarded. At average and above-average overlap values the effect of overlap is neutralized and, in the case of target passages with known reference, the effect is even inverted—i.e. starting at an above-average overlap value agreement increases with an increase in overlap.

Similarly to the other considered predictors—i.e. the effect of underlying retrieval method—we find that the shape and magnitude of the effect overlap on agreement is much more uncertain when the biblical source book is taken into account.

### 3.3 Post-experimental Report

In order to gain insight on the sources of disagreement, 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, based on self-report capacity and experience. Four illustrative examples along explanations are shown in Appendix B.

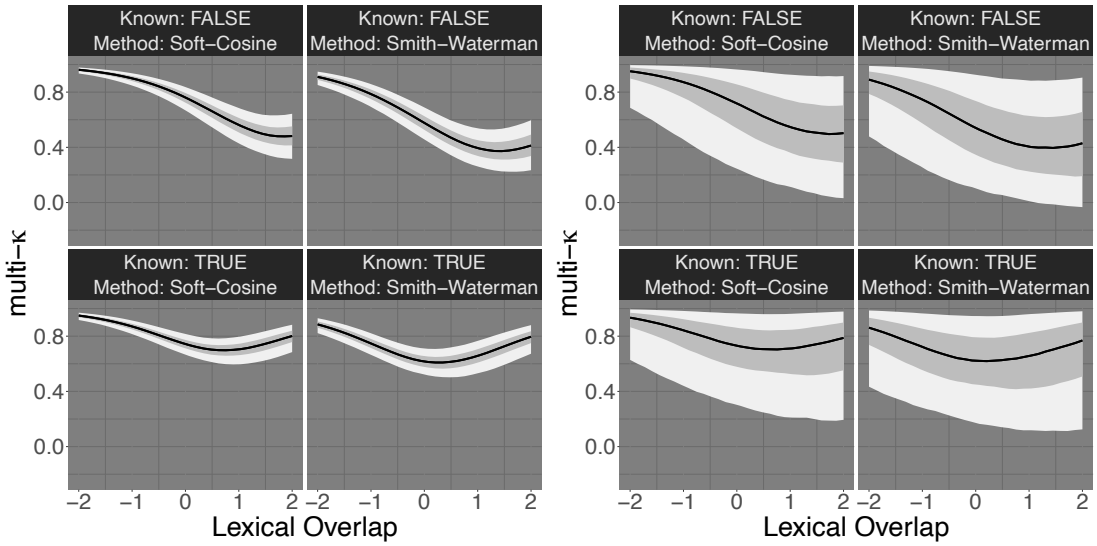


Figure 3: Posterior multi-$\kappa$ scores over lexical overlap. Note that the lexical overlap scale is centered such that a unit on the x-axis indicates a standard deviation away from the zero-mean. Left plot and right plot differ respectively on whether the variation coming from the books is excluded (left) or not (right). Black lines indicate median $\kappa$ scores with credible intervals at 0.5 and 0.89 probabilities shown in shaded grey areas.

**Segmentation Related Problems** The first issue relates to ambiguity problems arising from the approach employed in order to segment Bernard's sermons into documents. Bernard's Sermons were segmented using a sliding window of 20 words with an overlap of 10 words. This strategy resulted in a number of difficult cases in which the annotators have to decide subjectively whether to validate a candidate pair in the presence of fuzzy segmentation. For example, eventually segmentation left crucial words out of the target document, which lead to an artificial increase in the pool of candidate verses that can be interpreted as source. These problems have a significant incidence on the annotator disagreements and generate a lack of consistency, even by the same annotator.

**Knowledge of the Bible** When dealing with biblical texts, it is important to take intra-biblical intertextuality into account, a known phenomenon that consists in internal borrowing within the Bible. As an example, verses from the Old Testament are frequently referenced in the New Testament, and the so-called synoptical Gospels are known to contain parallel accounts of the same events. As a result, 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.

**Knowledge of Bernard** Finally, annotators must combine their knowledge of the Bible with other abilities regarding the borrowing author. In the case of the current case study, dedicated scholarship can show authors 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. Depending on the level of familiarity with the author's preferences, annotators choices will be in disagreement.

## 4 Conclusion & Future Work

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 on agreement and show how the inter-annotator agreement scores vary depending on the value of the independent variables.

While the overall average inter-annotator agreement reached is fairly high, the amount of uncertainty arising from the source book resulted in very wide posterior distributions which drastically nuance the reported coefficients.

Our approach fits random intercepts in order to capture individual annotator behaviour and should, therefore, scale reasonably well to higher number of annotators. However, the multinomial approach presented in the current research

requires fitting a number of linear models that is quadratic on the number of labels, and, thus, more complex tasks may become unfeasible to model in the same manner. Future work should address this shortcoming. Moreover, our approach is limited to annotation tasks defined in terms of categorical outcomes, and, thus, ordinal or continuous outcomes would require further research in order to be accommodated.

As the post-experimental report highlighted, some of the experimental design choices introduced artificial hurdles to agreement, and future research should take this into account in order to produce a more robust experimental settings.

Finally, while our study constitutes one of the first dedicated to the problem of inter-annotator agreement in intertextual studies, the specific experimental setting and design should be complemented in the future by other case studies in order to offer a more general picture of the matter.

## References


Ron Artstein. 2017. Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer.

Ron Artstein 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.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Daniel Bär, 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*.

Alessio Benavoli, Giorgio Corani, Janez Dem ˇsar, and Marco Zaffalon. 2017. Time for a Change: A Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis. *The Journal of Machine Learning Research*, page 36.

E. M. Bennett, R. Alpert, and A. C. Goldstein. 1954. Communications Through Limited Response Questioning. *Public Opinion Quarterly*, 18(3):303.

Harold Bloom. 1973. *The Anxiety of Influence: A Theory of Poetry*. Oxford University Press.

Paul Christian Bürkner. 2018. Advanced Bayesian multilevel modeling with the R package brms. *R Journal*.

Paul Clough and Mark Stevenson. 2011. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1):5–24.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Gregory Crane. 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*, pages 3–10.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Regula Hohl Trillini. 2018. *Casual Shakespeare : Three Centuries of Verbal Echoes*. Routledge.

Ross Ihaka and Robert Gentleman. 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Stefan Jänicke, Thomas Efer, Marco Büchler, and Gerik Scheuermann. 2015. Designing Close and Distant Reading Visualizations for Text Re-use. In *Computer Vision, Imaging and Computer Graphics - Theory and Applications*, Communications in Computer and Information Science, pages 153–171, Cham. Springer International Publishing.

Jeremy Koster and Richard McElreath. 2017. Multinomial analysis of behavior: statistical methods. *Behavioral Ecology and Sociobiology*, 71(9):138.

Enrique Manjavacas. 2021. *Computational approaches to intertextuality. from retrieval engines to statistical analysis : thesis.* Proefschriften UA-LW : letterkunde: 2021: 2.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019a. 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)*, pages 1493–1503, Minneapolis, Minnesota. Association for Computational Linguistics.

Enrique Manjavacas, Brian Long, and Mike Kestemont. 2019b. 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*, pages 104–114, Minneapolis, USA. Association for Computational Linguistics.

Richard McElreath. 2018. *Statistical Rethinking*. Chapman and Hall/CRC.

Laurence Mellerin. 2013. Methodological issues in biblindex, an online index of biblical quotations in early christian literature. In Laurence Mellerin, Markus Vinzent, and Hugh Houghton, editors, *Biblical Quotations in Patristic Texts (Papers Presented at the Sixteenth International Conference on Patristic Studies Held in Oxford 2011)*, volume 2 of *Studia Patristica*, pages 11–32. Peeters.

Laurence Mellerin. 2014. New ways of searching with biblindex, the online index of biblical quotations in early christian literature. In Claire Clivaz, Gregory Andrew, and Hamidovic David, editors, *Digital Humanities in Biblical, Early Jewish and Early Christian Studies*, volume 2 of *Digital Humanities in Biblical, Early Jewish and Early Christian Studies*, pages 175–192. Brill.

William A. Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3):321.

Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504.

T.F. Smith and M.S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Aki Vehtari, Jonah Gabry, Yuling Yao, and Andrew Gelman. 2018. loo: Efficient leave-one-out cross-validation and waic for bayesian models. *R package version*, 2(0):1003.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Tariq Yousef and Stefan Janicke. 2021. A Survey of Text Alignment Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1149–1159.

## A Model Fitting

In the present study, we deploy Bayesian linear models as implemented by the `brms` library (Bürkner, 2018), an **R** (3.6.3) package (Ihaka and Gentleman, 1996) providing a user-friendly interface to `Rstan`, which utilizes a powerful Hamiltonian Monte Carlo sampler. In order to ensure the validity of the resulting posterior distributions, we make sure that the following diagnostics check: first, effective sample sizes are large enough; secondly, the samples are homogeneous across chains (i.e. $\hat{R}$ values should be close to 1), and, finally, divergent transitions are kept to a minimum. This was accomplished using 4 chains for 2000 iterations with the first 1000 used as warmup iterations. Some models required fine-tuning the "adapt delta" and the "maximum tree depth" parameters.

Moreover, we choose weakly informative priors to avoid exploring highly unlikely regions of the parameter space.

## B Post-experimental Report Examples

Table 4 shows a number of candidate pairs that illustrate common sources of disagreement. The subset consists of instances retrieved with the `Smith-Waterman` algorithm and has, thus, a slight bias towards literal quotations.

The first instance in Table 4 corresponds to an example of a segmentation problem. The words "quia mirabilia facit" 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 these intertextual references.

The second example in Table 4 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 to the Gospel. This example already suggests a further source of disagreement, which corresponds to the familiarity with the referential practices of the borrowing author.

In the third example in Table 4, Bernard's 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 explicitely and in the current location it is being referred to implicetely. *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 4, 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 adressing to Son ("Pater ad Filium"). 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).

| Bernardian 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***<br>*S. 1, 9 (SC 414, p. 72)* | quando domus aedificabatur post captivitatem canticum huic David **cantate Domino canticum novum** cantate Domino omnis terra<br>*Psalms, 95:1* | psalmus David **cantate Domino canticum novum quoniam mirabilia fecit** salvavit sibi dextera eius et brachium sanctum eius<br>*Psalms, 97:1* |
| carnale matrimonium constituit **duos in carne una**, cur non magis spiritualis copula duos coniunget in uno spiritu? Denique<br>*S. 8, 9 (SC 414, p. 192)* | et erunt **duo in carne una** itaque iam non sunt duo sed una caro<br>*Mark, 10:8* | quam ob rem relinquet homo patrem suum et matrem et adherebit uxori suae et erunt **duo in carne una**<br>*Genesis, 2:24* |
| blanditiis, seduci fallaciis, nec iniuriis frangi, **toto corde, tota anima, tota virtute diligere** est.<br>*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<br>*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<br>*Mark, 12:30* |
| cognoscentur. Hinc rursus Pater ad Filium: **Sede**, inquit, **a dextris meis, donec ponam inimicos tuos *scabellum pedum tuorum***<br>*S. 6, 5 (SC 414, p. 144)* | ad quem autem angelorum dixit aliquando **sede a dextris meis quoadusque ponam inimicos tuos scabillum pedum tuorum**<br>*Hebrews, 1:13* | david canticum dixit Dominus Domino meo **sede a dextris meis donec ponam inimicos tuos scabillum pedum tuorum**<br>*Psalms, 109:1* |

Table 4: Examples from the dataset, showcasing different types of agreement problems. The first one highlights **segmentation** issues, and the second one and the last two relate respectively to *diverging degrees of familiarity* with the **Bible** and **Bernard**. The Bernardine chunk on the left is accompanied by the retrieved candidate in the center and a better 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 the BiblIndex online version that includes English translations.